Usage-Based Linguistic Informatics 5

Spoken Language Corpus and Linguistic Informatics

Edited by
Yuji Kawaguchi, Susumu Zaima
and Toshihiro Takagaki

John Benjamins Publishing Company

Spoken Language Corpus and Linguistic Informatics

Usage-Based Linguistic Informatics

Volume 5

Spoken Language Corpus and Linguistic Informatics
Edited by Yuji Kawaguchi, Susumu Zaima and Toshihiro Takagaki

Spoken Language Corpus and Linguistic Informatics

Edited by

Yuji Kawaguchi
Susumu Zaima
Toshihiro Takagaki
Tokyo University of Foreign Studies

John Benjamins Publishing Company
Amsterdam / Philadelphia

∞™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Control Number: 2006047915

ISBN 90 272 3317 9 (Hb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Contents

1. The Second International Conference on Linguistic Informatics

2. Workshop on Spoken Language Corpora —C-ORAL-ROM and UBLI—

2.1. C-ORAL-ROM

2.2. UBLI

3. Linguistic Informatics

Message from the President

Setsuho IKEHATA *(President, Tokyo University of Foreign Studies)*

On this occasion of hosting the Second International Conference on Linguistic Informatics, as part of the 21st Century COE Program, Tokyo University of Foreign Studies, I would first like to extend my sincere appreciation to our five lecturers from France, Great Britain, the United States, and Italy. I also express my gratitude to the members of C-ORAL-ROM. Outstanding scholars from every area of the world—Italy, France, Spain, Portugal, and Turkey—participated in the collaborative workshop on spoken language corpora that was held yesterday. These leading authorities have been kind enough to take time out of their extremely busy schedules and present lectures at our workshop and conference.

I remember that it was late autumn in 2002 when the COE Program was officially launched. In the several months leading up to the program applications, the Center leader, Professor Yuji Kawaguchi, led a series of heated debates on the research programs. The proposals were drawn up, subjected to an inquiry by the review committee, and finally the program got underway.

The aim of this COE Program is to organically integrate linguistics and language education in order to develop cutting-edge online teaching materials, “TUFS Language Modules,” in numerous languages by using computer science as a foundation. The graduate courses of our university meet all the necessary conditions for the sufficient achievement of collaboration between linguistics and language education. Rather than merely pursuing linguistic theory in foreign language research, we have advocated the importance of reconsidering linguistic theory through language education programs. It can be said that this type of bidirectional feedback between linguistic research and language education is one of the unique scholastic characteristics of our university.

Almost four years have quickly flown by since the selection of this COE Program. The program is fast approaching its most crucial stage of compiling the final results. In my executive role at the University, which is a little distant from the actual research and education scene, I observed the developments that were being made throughout the program. I am unwavering in my conviction regarding the findings of the program and regarding the maintenance and development of the Center at our University following the conclusion of the program.

This conviction can be attributed to my unfailing confidence in my colleagues, who have promoted research as core members of the program, as well as in the graduate students and other young researchers, who have literally worked day and night for the completion of this program. It is also an indication of my sincere gratitude and the respect that I hold for the support and cooperation received in a variety of ways from outside the University, mainly from the five professors who will be presenting lectures today.

The theme of this conference is “What is Linguistic Informatics—Contributions of Linguistics, Applied Linguistics, and Computer Sciences,” and it has been continuously examined since the conception of this program. I understand that the aim of the conference is to reexamine this theme on the basis of the more than three years of research activities and achievements.

The development of all disciplines and the exploration of new frontiers of knowledge have taken place through the repeated questioning of the reason for the existence of a learning and the inherent nature of that learning.

My specialization is in the field of history. History is one of the oldest humanities disciplines; yet, even in this field, I think it would be fair to say that ever since the latter half of the nineteenth century, which is said to be the era of modern history, there has never been a period of time when the question “What is history?” has not been asked.

If we are to advocate and attempt to establish this new discipline of linguistic informatics, it is inevitable that this questioning will become increasingly more urgent. One answer to this question, which I believe is perhaps the most effective answer, is to introduce to the wider world the achievements generated from this new discipline and to seek out appraisals of this work.

I am hopeful that the array of dissertations, literary works, and the course material in the 17 language modules, which opens a new horizon of research and education, embodies the essence of this new discipline.

I would like to conclude by expressing the hope that today’s conference will be of assistance in developing this still relatively new discipline of “linguistic informatics” and in unlocking new possibilities for learning.

December 10, 2005

1.
The Second International Conference on Linguistic Informatics

Introduction

Yuji KAWAGUCHI *(Center of Excellence (COE) Program Leader)*

"Linguistic Informatics" is a research field named by the Center of Excellence (COE) program of the Graduate School of Tokyo University of Foreign Studies (TUFS). It implies the systematic integration of computer science, linguistics, and language education. The "Usage-Based Linguistic Informatics (UBLI)" was designed for the purpose of improving language education in Japan. The initial aim was to realize more efficient multilingual education, while bringing language education to the fore via the utilization of computer technology, and to elaborate advanced educational materials via the utilization of linguistic theory. After the launch of the program, the first international conference was held at TUFS for two days—December 13 and 14, 2003. It might not be surprising that the core research domain of "linguistic informatics" needed more precision during this early stage of the program. Despite this, contributions in the first conference were mainly related to two key research domains.

The first domain was represented by computer-assisted linguistics and corpus linguistics. It is evident that, when researching linguistic structures in detail, the computer-assisted corpus analysis is essential. A look at the papers from the first conference reveals that there exists a wide range of genres and individualities in language corpora, such as medieval literature, bilingual databases, linguistic atlas data, workplace language, and natural dialogue.

The second domain included the studies concerning the relevance between linguistic theory and second language acquisition. That the analysis of natural dialogue is necessary for developing conversation textbooks is a prime example of this. The development of the "TUFS Language Modules," web-based language teaching materials covering 17 different languages, would also have been impossible without the foundations of linguistic theory and second language acquisition. In fact, the Pronunciation Modules have been designed based on phonetic and phonological theory; the Dialogue Modules are built on a notional and functional syllabus; the Cross-linguistic Grammar Modules are an attempt to synthesize more than ten different grammatical systems from a typological viewpoint, and the Vocabulary Modules are classified with regard to their semantic categories based on the lexical taxonomy of Japanese elaborated by the National Institute for Japanese Language, which gave a cross-linguistic perspective to the

Vocabulary Modules.

A large number of researchers belonging to research organizations in Japan and other countries were invited to attend the first conference. The gross attendance was 300, and there was a lively exchange of opinions. Several reports were also presented by the postgraduate students of TUFS. The collection of papers from the first international conference was published by John Benjamins in the spring of 2005 as the first volume in the "Usage-Based Linguistic Informatics" series[1]. As a result of this first conference, the essence of linguistic informatics, the construction of which is our aim, became progressively more precise[2].

An ongoing analysis of linguistic usage, particularly at the grammatical level, based on language corpora has been conducted in the UBLI. However, a large portion of that analysis concerns written language, and spoken language has been in the minority[3]. It is only in recent years that research on spoken language corpora has been conducted in earnest. The UBLI has conducted field surveys since the very beginning of the COE program and has built spoken language corpora for French, Spanish, Italian (Salentino dialect), Russian, Malaysian, Turkish, and Japanese. Since building corpora involves persistent work that requires many long hours, such as for transcribing, it is only recently that we have been able to analyze the spoken language corpora. There is an ongoing research on learner corpora in Japanese and English as well; however, this is also a very recent field of research. These steady fundamental studies have shown us a fruitful but challenging direction toward the analysis of spoken language corpora and its application to our teaching materials. The construction of spoken language corpora and its linguistic analysis are crucial for foreign language teaching and learning because they would provide authentic usages in various communicative aspects of verbal interactions. In this manner, the main theme of the second international conference became more substantial.

On December 9 of 2005, a workshop entitled "Spoken Language Corpora — its Significance and Application —" was held in conjunction with C-ORAL-ROM, a consortium researching the spoken Romance

[1] Kawaguchi, Yuji, Susumu Zaima, Toshihiro Takagaki, Kohji Shibano, and Mayumi Usami (2005) *Linguistic Informatics State of the Art and the Future,* Usage-Based Linguistic Informatics 1, Philadelphia/Amsterdam, John Benjamins.

[2] For a more detailed description, see Yuji Kawaguchi, "Foundations of Center of Usage-Based Linguistic Informatics (UBLI)" in this volume, pp.9-33.

[3] Takagaki, Toshihiro, Susumu Zaima, Yoichiro Tsuruga, Francisco Moreno-Fernández, and Yuji Kawaguchi (2005) *Corpus-Based Approaches to Sentence Structures,* Usage-Based Linguistic Informatics 2, Amsterdam/Philadelphia: John Benjamins.

languages[4]. On the following day, December 10, the second international conference on Linguistic Informatics was held. Three different lectures were given on the state of grammatical research into spoken language, the pragmatic analysis of spoken language, and the application of spoken language corpora to education. A general discussion was also opened between the lecturers and the C-ORAL-ROM members. There was an audience in excess of 300 people over the two days, and the meeting was a success. In this way, further precisions were given to the subject of research for the center for linguistic informatics. This chapter presents the following three contributions, all of which were presented at the second international conference.

Claire Blanche-Benveniste, in her “Linguistic Analysis of Spoken Language —The Case of French Language—,” first introduces as bibliographical information important studies conducted on four Romance languages. Explaining multi-dimensional approaches to spoken texts, she highlights several syntactic devices of spoken French: focalization, dislocation, and parenthesis. She demonstrates many examples of lexical restrictions on French grammar. Finally, considering the methodological issues in describing spoken French grammar, she proposes to distinguish between two grammatical levels. Further, she points out the short-comings in handling statistical tools and the important phenomena of interruptions or repairs to grasp better spoken French grammar.

In “Challenges for English Corpus Linguistics in Second Language Acquisition Research,” Susan Conrad discusses some of the most important issues in second language acquisition research for educators and policy makers in the United States that the current corpus linguistics work does not address. She then proposes a new type of corpus, which would be connected to the National Adult ESOL Labsite project and includes speech from low-level adult immigrant students, a classroom component, and links to the video database.

Massimo Moneglia and Emanuela Cresti, in their “C-ORAL-ROM —Prosodic Boundaries for Spontaneous Speech Analysis—,” present in a concise way the C-ORAL-ROM corpus (Integrated Reference Corpora for Spoken Romance Languages) and explain, in particular, the annotation of the reference units for both syntactic analysis and multimedia representation of speech data. They discuss the value of the annotation of prosodic breaks to determine utterance and boundaries in comparison with concurrent syntactic, pragmatic, and acoustic methods.

[4] See Emanuela Cresti and Massimo Moneglia (2005) *C-ORAL-ROM : integrated reference corpora for spoken Romance languages*, Amsterdam/Philadelphia: John Benjamins.

Foundations of Center of Usage-Based Linguistic Informatics (UBLI)

Yuji KAWAGUCHI *(Center of Excellence (COE) Program Leader)*

1. Linguistic Informatics

The Center of Usage-Based Linguistic Informatics (UBLI) — a 21st Century COE Program — was adopted by the Ministry of Education, Culture, Sports, Science and Technology in September 2002 as a five-year plan. The goal of the 21st Century COE Program was to raise the current level of various disciplines in Japan to a point where they could be globally competitive. Subsequent to that, in addition to achieving international research levels, the emphasis shifted to the development of young researchers. As the name suggests, under the 21st Century COE, Centers of Excellence are to be formed in their respective disciplines through five years of research. Following our 2004 interim assessment, this year will be the final year for the UBLI.

The aim of the UBLI is for the systematic integration of computer science with linguistics and language education. The name given to this area of study is “linguistic informatics,” and over the course of five years, we will ultimately create a research area called “linguistic informatics.” Underlying the design of this COE program was an intent to improve and reform language education in Japan’s higher education, especially at universities. The original starting point was to aim for the realization of more efficient multilingual education, while bringing language education to the fore via the utilization of computer technology, and striving to advance educational content via the utilization of linguistic theory. It would appear that setting this kind of goal is important for the monolingual country that is Japan. Furthermore, since there are very few situations in which languages other than Japanese are used on a daily basis, it seems that, even for cross-cultural understanding, there is a need for foreign language education to be positioned as something beyond simply a means of communication, and to promote cross-cultural understanding through multilingual education from an early stage.

“Linguistic informatics,” as according to the UBLI, is an academic field, which is distinctly and strongly colored by application, and which ultimately leads to improvements and upgrades in language education. Some researchers regard linguistic informatics as a division of applied linguistics in a broad sense of the word. If this is the case, then why not call it applied linguistics? Why confine it to the term “linguistic informatics”? I would like to begin my explanation from this point. However, prior to this explanation, there is something which should be brought to the readers’ attention. The term “linguistic informatics” referred to here, does not denote natural language processing, machine translation, or computer linguistics. Even supposing for the moment that it is associated with these fields, this would be “linguistic informatics” in the very narrowest of meanings. As I have remarked previously, the goal of the UBLI is not related to language processing using computer technology or to linguistic analysis in itself. We are aiming to determine what should be done to linguistic theory and educational practice, with the aid of computer technology, so that they can further meet the needs of society. This point must first be stressed in order to dispel all the misunderstandings related to “linguistic informatics.”

Now then, if we take an overview of the research methods used at the UBLI, we can see that we have corpus linguistics and computer linguistics research. We also have research based on discourse analysis, second language acquisition, and descriptions of language proficiency. Furthermore, we can see that educational technology practices also feature prominently in our research. This shows that “linguistic informatics” is an academic field that has been put together by combining a wide variety of methods and concepts from other disciplines. It also shows that, more often than not, the distinction between this and other disciplines is vague and difficult, and they share various theories and methodologies. The fact that “linguistic informatics” is an applied discipline is indeed confirmed in the significance of this point.

Since the inception of applied linguistics, scientific research in foreign language education has occupied a central position. Even today, these circumstances remain unchanged. However, in recent years, we have reached a point where, particularly in English and other languages, the results of corpus linguistics, which use computers to analyze large volumes of linguistic data, are now being applied to language education. At the same time, technology in language education which utilizes the Internet has also been developing at a rapid pace. “Linguistic informatics” is the term which tried to capture this new trend in research. We were still at a stage where no name has been given to the field of research which, based on computer technology, uses corpus linguistics and computer linguistics techniques to analyze data on actual language use and attempts to reflect the results in language education. Although this is “the application of corpus analysis and language-usage analysis to educational practice,” to call this applied linguistics would probably result in the scope of applied linguistics being restricted to an all too narrow domain. Therefore, at the UBLI, we ventured to call this field of research, which uses computers to analyze language usage and which attempts to link this analysis to more efficient and advanced educational practice, “linguistic informatics.”

2. Fields of Basic Research in Linguistic Informatics

In order to bring the abovementioned field of research to fruition, the UBLI is organized into four research groups (the Linguistic Informatics Group, the Linguistics Group, the Language Education Group, and the Computer Science Group), and each group proceeds with their research while maintaining close coordination. The Linguistics Group uses computers to analyze corpora and it conducts research on language use data. The Language Education Group conducts analyses on learner corpora and natural discourse based on second language acquisition theory, and it also conducts research on a language proficiency descriptive model. The Computer Science Group carries out the design and development of computer-assisted language processing and the e-learning system. Finally, the Linguistic Informatics Group pulls together the basic research conducted by the other three groups, and attempts to make language education more sophisticated and efficient. In this section I will explain how the basic research from each of the Linguistics, Language Education, and Computer Science groups comes together.

2.1. Analysis of Linguistic Usage and Phonetic Analysis

Until now, research in theoretical linguistics has concentrated on the structures of language systems and their functions. On the other hand, as a result of the rapid progress of computer technology, it is now possible to process massive amounts of language corpora, at levels which language researchers could not possibly have processed previously.[1] As research of vast spoken language corpora has advanced, it has become progressively evident that there exists vast numbers of linguistic variations within language communities which are presumed to be homogeneous or else have been researched while eliminating heterogeneous parts to a certain degree in analysis. By using data on actual language usage to verify phenomena which had been problematic in language research, we have found that natural biases or tendencies can be found for many linguistic phenomena. It seems that linguistic interest is also in the process of shifting toward clarification of the diversity, and associated mechanisms, in the actual realization of systems and functions from the language systems and the functions themselves. In recent years, the importance of research on linguistic usage has been increasing.

[1] McNery and Wilson (2003).

Previously, UBLI published in Japanese, *Analyses in Sentence Structures in Corpus Linguistics*, Working Papers in Linguistic Informatics 3 (September 2004), *Lexicon and Grammar in Corpus Linguistics*, Working Papers in Linguistic Informatics 7 (October 2005), and in English, *Corpus-Based Analyses on Sentence Structures*, Linguistic Informatics II (April 2004). Several research papers were added to the final collection of papers, and they were published in the spring of 2005 by the Dutch publishing company, John Benjamins, as *Corpus-Based Approaches to Sentence Structures*, the second volume in the "Usage-Based Linguistic Informatics" series. Each of studies analyzes the various forms and syntactic structures which appear in the large-scale language corpora.

The results from our research on language use will be applied to the Web-based teaching materials developed by the UBLI. Possible examples include the difference in frequencies of verbs in the written language and the spoken language, frequently occurring collocations, the frequency of specific sentence structures, adjectives and their positioning, and the usage of cases. Incidentally, research on the very nature of linguistic data is also vital in corpus analysis. From such research, the fundamental question of "What is usage in linguistics?" is asked, and at this time, language research comes face to face with linguistic variation. Recognition of the importance of usage-based linguistic analysis is reignited, and we should appreciate the significance of using computer science.[2]

Linguistic symbols are units in which sound and meaning are inextricably linked. Research on the phonetic aspect of linguistic symbols has occupied an important place ever since linguistics came into being. At the UBLI as well, phonetic or phonological analysis on language use has been carried out in parallel with corpus linguistics. *Cross-Linguistic Perspectives in Phonetics — Phonetic Description and Prosodic Analysis*, Working Papers in Linguistic Informatics 4 was released in October 2004. Then, in December 2005, *Prosody and Sentence Structures,* Linguistic Informatics IV was released, and later, papers by overseas collaborators were added to this title, and this was released in the spring of 2006 by John Benjamins, as *Prosody and Syntax Cross-Linguistic Perspectives*. At the COE, in addition to research on phonetic and phonological variations of single sounds, research is also being focused on prosodic structure — earlier research outcomes of which are believed to have been poorly reflected in educational practice. Within this research, with regard to Asian languages, the conversation teaching materials developed by the UBLI were regarded as

[2] Yuji Kawaguchi, "Usage-Based Approach to Linguistic Variation — Evidence from French and Turkish —", 247-267, in this series.

phonetic corpora, and analysis of the accents and intonations was conducted.[3] These results will also be ultimately used to improve the teaching materials of phonetics.

2.2. Construction and Analysis of Language Corpora

At the UBLI, we have planned the construction of multilingual corpora that meet our research objectives. In particular, in order to analyze actual language usage, it becomes increasingly important to construct natural spoken language corpora. In addition to Japanese, for which construction of a corpus began during the initial stages of the program being adopted, from 2004, recordings of conversations were taken on location for French, Spanish, Russian, Malaysian, Turkish, and Italian (Salentino dialect). Spoken language corpora were constructed from between five and 20 hours. Of these, since comparable spoken language corpora did not exist for Russian, Malaysian, or Turkish, great significance can be found in the very construction of the corpora. Linguistic analyses using these corpora are also underway.[4]

In the Language Education Group, analysis of the spoken Japanese corpus is also being conducted from a perspective of dialogue analysis, and in particular a perspective of social psychology, targeting the same spoken language corpus. One of the outcomes has been the April 2005 release of *Natural Dialogue Analysis and Conversation Training. Pursuing the Creation of an Integrated Module*, Working Papers in Linguistic Informatics 6. By comparing the previously developed conversation teaching materials with the spoken Japanese corpus, detailed analyses were carried out from such perspectives as discourse function and politeness. These analyses are the basic research for implementing natural conversation in conversation teaching materials.[5]

If the abovementioned corpora are grammar function, discourse function and so-called language function-specific corpora in a broad sense of the term, then at the UBLI, we have also constructed a research objective-specific corpora. Among the more important corpora are learner language corpora, which in recent years have been gaining attention in research

[3] Yuji Kawaguchi et al. (2006), intonation analysis of Indonesian, Filipino, Turkish, and Japanese.

[4] For instance, Selim Yılmaz, “Viewpoint and Postrheme in Spoken Turkish”, 269-286, and Isamu Shoho, “Nonreferential Use of Demonstrative Pronouns in Colloquial Malay”, 287-301 in this volume.

[5] See Usami (2004) for the necessity of analyzing natural conversation.

on second language acquisition.[6] Recently, the importance of combining native corpora and learner language corpora has been recognized. While native corpora have led to an increased accuracy in teaching material descriptions, at the same time, the development of teaching materials and curricula are being sought to bring to fruition more effective language education using learner corpora. In this way, the construction and analysis of learner corpora can be said to be an important field of research for applying the results of the analysis of language corpora to educational practice, and for striving for an efficient language education;[7] and it can be thought of as one of the major areas of study in linguistic informatics.

Previously, the Language Education Group announced its basic research on learner corpora for Japanese learners of English in the December 2004 publication *Second Language Pedagogy, Acquisition, Evaluation*, Working Papers in Linguistic Informatics 5. During 2006, the group will also publish its findings on its research into learning corpora for the Japanese and English languages. Furthermore, in March 2006, the Language Education Group also published Linguistic Informatics V, *Studies in Second Language Teaching and Second Language Acquisition*. Papers by overseas collaborators were added to this title, and this was released in the summer of 2006 by John Benjamins, as *Readings in Second Language Pedagogy and Second Language Acquisition in Japanese Context*.

3. Computer science and TUFS Language Modules

As was remarked at the outset, "linguistic informatics" is the academic field that attempts to integrate linguistic theory and educational practice on a computer science base. Following on from the trend of recent years for the development of language teaching materials using CALL and networks, the UBLI has been developing "TUFS Language Modules" — Web-based language teaching materials covering 17 different languages. Not only do these teaching materials use the latest techniques available in educational technology,[8] but they are structured with content to which linguistic theory has been applied, and it could be argued that they are the most visible

[6] For instance, Part I by Sylviane Granger, Granger et al. (2002) for the necessity of learner language corpora. Part II and Part III contain arguments related to an analysis of the interlanguage and research on foreign language education using learner language corpora.

[7] Hunston (2002) 206-212.

[8] For example, in conversation teaching materials, XML technology which supports UTF-8/16 is implemented, and then using a program, the XML data and the audio and video data are synchronized through an MXSML server with JavaScript. See Lin et al. (2004). The same text also refers to the development of TUFS Language Modules in general.

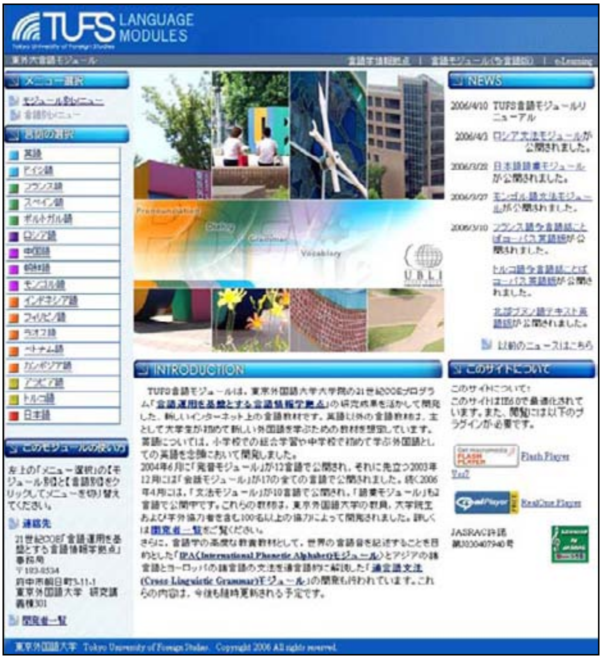
academic results of “linguistic informatics.”

One of the TUFS language module types, the “Pronunciation Module” was released in 2003 in 12 languages, including English, German, and French. In 2004, “Dialogue Modules” were published in all 17 languages: English, German, French, Spanish, Portuguese, Russian, Chinese, Korean, Mongolian, Indonesian, Filipino, Laotian, Cambodian, Vietnamese, Arabic, Turkish, and Japanese; and one of these was implemented as some undergraduate courses of the Faculty of Foreign Studies at TUFS. Then in the spring of 2006, “Grammar Modules” were released in 11 languages: German, French, Spanish, Russian, Chinese, Mongolian, Filipino, Cambodian, Vietnamese, Turkish, and Japanese; and “Vocabulary Modules” were released in 11 languages. Amongst these, the development of Web-based teaching materials for Mongolian, Laotian, and Cambodian were world firsts.

As was remarked at the outset, the UBLI program is a project aimed at innovations for foreign language education at higher education in Japan. Therefore, it is envisaged that the language teaching materials, other than for English, are teaching materials mainly for university students learning a new foreign language for the first time. As their name suggests, TUFS Language Modules were designed based on a “module-type notion.” Specifically, the idea is that they are divided into four types of modules: pronunciation, dialogue, grammar, and vocabulary; and while each module is mutually independent to a degree, they come together to form a cohesive set of teaching materials. In this sense, it could probably be argued that TUFS Language Modules take a perspective which focuses on structure. Naturally, since they are Web-based teaching materials, they can be modularized, and consequently, they can be more efficiently corrected and revised. Furthermore, by utilizing hyperlinks, they are capable of providing a sense of unity. These benefits are the reason why module-type teaching materials were adopted.

A structuralistic linguistic view is reflected in the module structure comprising four parts: pronunciation, dialogue, grammar, and vocabulary. One linguistic universality is the “double articulation of language.” Let us suppose there is an expression which appears in a dialogue. The French greeting “Salut, ça va?” (*Hello. How are you?*), for example. First, the words are divided by primary articulation into the smallest linguistic signs having meaning. In this case there are three: SALUT, ÇA, and VA. Next, for example, SALUT is further divided by the second articulation into /saly/ which is a combination of four phonemes. According to this hypothesis, both levels of articulation form parts of the linguistic structure while functioning independently from each other; or if expressed in linguistic terms, the monemes and phonemes form a linguistic structure while remaining mutually independent. Some correlations can be drawn to the module-type linguistic view. The following section briefly describes each of the pronunciation, dialogue, grammar, and vocabulary modules.

TUFS Language Modules



2006/4/10 TUFS言語モジュールリニューアル

2006/4/3 ロシア文法モジュールが公開されました。

2006/3/28 日本語語彙モジュールが公開されました。

2006/3/27 モンゴル語文法モジュールが公開されました。

2006/3/10 フランス語今言語版ことばコーパス英語版が公開されました。

トルコ語今言語版ことばコーパス英語版が公開されました。

北部ブヌン語テキスト英語版が公開されました。

以前のニュースはこちら

TUFS言語モジュールは、東京外国語大学大学院の21世紀COEプログラム「言語運用を基盤とする言語情報学拠点」の研究成果を活かして開発した、新しいインターネット上の言語教材です。英語以外の言語教材は、主として大学生が初めて新しい外国語を学ぶための教材を想定しています。英語については、小学校での総合学習や中学校で初めて学ぶ外国語としての英語を念頭において開発しました。

2004年6月に「発音モジュール」が12言語で公開され、それに先立つ2003年12月には「会話モジュール」が17の全ての言語で公開されました。続く2006年4月には、「文法モジュール」が10言語で公開され、「語彙モジュール」も2言語で公開中です。これらの教材は、東京外国語大学の教員、大学院生および学外協力者を含む100名以上の協力によって開発されました。詳しくは開発者一覧をご覧ください。

さらに、言語学の高度な教育教材として、世界の言語音を記述することを目的とした「IPA(International Phonetic Alphabet)モジュール」とアジアの諸言語とヨーロッパの諸言語の文法を通言語的に解説した「通言語文法(Cross Linguistic Grammar)モジュール」の開発も行われています。これらの内容は、今後も随時更新される予定です。

左上の「メニュー選択」の【モジュール別】と【言語別】をクリックしてメニューを切り替えてください。

連絡先

21世紀COE「言語運用を基盤とする言語情報学拠点」事務局
〒183-8534
府中市朝日町3-11-1
東京外国語大学 研究講義棟301

開発者一覧

このサイトについて↑
このサイトはIE6.0で最適化されています。また、閲覧には以下のプラグインが必要です。

Yes?

JASRAC許諾
第J030407940号

東京外国語大学 Tokyo University of Foreign Studies. Copyright 2006 All rights reserved.

http://www.coelang.tufs.ac.jp/modules/index.html

3.1. *Pronunciation Modules*

The Pronunciation Modules consist of a "practical course" and a "theoretical course." Each course was developed based on a different design concept.

In the "practical course," the design of the teaching materials aims to be as user-friendly as possible. For this reason, ordinary day-to-day vocabulary and expressions are used as examples, and a phonetic view which contrasts Japanese is introduced. The course is devised so that learners pick up pronunciation through practice and training. Furthermore, three stages are envisaged in the procedures for acquiring speaking and listening skills, and learning pronunciation proceeds according to those levels. The first stage emphasizes the correctness of individual sounds, the second stage places emphasis on smoothness, and the third stage pursues peculiarities of the individual language and fluency. Naturally, it would not do to have the same acquisition procedures for all 17 languages. While there are some languages in which effort is placed on acquiring segmental sounds, there are other languages for which time is spent on practicing suprasegmental sounds. Nevertheless, by having experts estimate the acquisition procedure they believe to be ideal for each language; this will lead to teaching resources in which the phonetic qualities of each individual language will be highlighted.

The "theoretical course" pursues self-study material so that people who have already learnt the "practical course," or who already have knowledge in the language, can increase their skills. It is supposed that this course will be used as supplementary teaching materials in a university course, for example, and technical terms and IPA (International Phonetic Alphabet) are used so that the phonetic and phonological basics of the language can be learnt. It is generally said that instruction in pronunciation needs to be adjusted to the age of the learner. However, since these modules target university students, who have passed the critical stage of language acquisition, the instruction in pronunciation has not been limited to only practical aural comprehensions of phonetics. It also includes the acquisition of phonetic and phonological knowledge, such as the use of minimal pairs, links with IPA Module, the fluctuation of phonemes, and neutralization and the functional load of phonemes. Furthermore, it refers to demarcative functions, contrastive functions, and enunciative functions, which possess suprasegmentals such as rhythm, accent and intonation; and by linking to the Dialogue Module, learners are able to practice example sentences which are uttered in more natural spoken environments. The theoretical course has been implemented into undergraduate coursework of French since 2005, and evaluations of both teaching materials and learning were conducted.

3.2. Dialogue Modules

Whereas the Pronunciation Modules have been developed based on a focus on form, the Dialogue Modules are teaching materials, which, rather than being based on the form of a language, emphasize language proficiency, and more particularly, they emphasize communicative competence. The context in which dialogues are formed and discourse strategies, etcetera, take more of a focus-on-meaning stance. The syllabus employed for this reason is a notional and functional syllabus which is used broadly in language education for communication. Of these, the focus in Dialogue Modules is particularly on the communicative function. In developing the Dialogue Modules, CALL teaching materials for five languages (German, French, Spanish, Portuguese, and Chinese) were surveyed. Based on the results, the material was organized into functional classifications as referred by Wilkins

(1976) and van Ek (1990), and ultimately 40 fundamental language functions, such as “greeting someone” and “thanking someone,” were selected.[9] There are two courses in the Dialogue Modules currently released: “for lesson use” and “for student use.” They have both been developed using the same content, but the design of the respective teaching materials differs greatly.

The “for lesson use” pages suppose that they are being used in a university course, etcetera. The screens are designed to be all-encompassing. In the classroom, the instructor analyzes the needs of the learners, and learning commences from the corresponding language functions. First, the dialogue is played, and the students are made to imagine the context of the speech. Next, they learn by role-playing the entire dialogue. At this point it is important that students are made aware that, rather than sentences of the target language, the aim of the learning is discourse. Also, it is important that the students are provided with language resources which are as real as possible. At the UBLI, basic research is proceeding to incorporate natural discourse in the Dialogue Modules.[10] It could be said that the “for lesson use” pages are teaching materials for classroom practice based on a communicative approach.

The “for student use” pages have been designed so that learners can acquire the four language skills of listening, speaking, reading, writing on their own. Four learning models were established so that learning could take place from various aspects, according to the goals of the learner. Model 1 is “Listening and Speaking (Role Play)” where a student assumes one of the roles and practices the dialogue. If a learner can respond immediately and without looking at the text, then this will lead to an improved speaking ability. Model 2 is “Reading and Speaking (Reading Aloud)” where students practice reading in time with the sounds that they hear. Model 3 is “Listening and Writing (Dictation)” where students practice writing down as they listen to speech. This model is effective in improving a learner’s listening comprehension. Finally, Model 4 is “Reading and Writing (Copying)” where students practice transcribing the text that they read. With the “for student use” pages, by clarifying the acquisition procedures for the target language, learners can proceed with their study, while being conscious of their target language skills. In addition to this, in the Dialogue Module for English, a Teacher’s Manual has been prepared which proposes detailed commentary

[9] Yuki, Abe and Lin (2005) 339-342.

[10] Usami, Mayumi (ed.) (2005) *Natural Dialogue Analysis and Conversation Training ? Pursuing the Creation of an Integrated Module* (in Japanese), Working Papers in Linguistic Informatics 6, the 21st Century COE “Usage-Based Linguistic Informatics”, Graduate School of Area and Culture Studies, see also Suzuki, Matsumoto and Usami (2005).

for use in the classroom as well as task-based instruction examples.[11] In the future, by devising similar manuals for other languages, it is believed the learning environment of the Dialogue Modules will be improved.

3.3. Grammar Modules

In the past, it was common for grammar instruction syllabi to contrast the native language and target language, and to be based on experiential intuition. Even today, this situation has barely changed. It is extremely difficult to find a grammar syllabus which is believed to clearly enhance learning effects. Nevertheless, in English, during the 1980s it had already been demonstrated that acquiring grammar according to an acquisition order was faster than learning naturally.[12] However, in order to design syllabi based on similar experimental studies for the 17 languages covered by the UBLI, much more long-term and steady research is still required. Consequently, the design of the current Grammar Module concentrates on morphological and syntactic commentary with a focus on the form of the language. Naturally, grammatical items are arranged with consideration given to difficulty and practicality. Furthermore, example sentences which are only for explanation have been excluded, and examples which are close to actual language use are included. In some languages, the module has been designed so that learners can be aware of misuses and correct usages, and there is even commentary on frequency of actual usage and biases in sentence structures.[13] Although the course titles vary depending on the language, by establishing several study courses, such as “Ability Development Course,” “Standard Course,” the teaching materials become cognizant to a degree of the procedures for grammar acquisition. Also, by arranging the same speech on the cards, commentary and example sentences, the modules have also been designed so that the input to the learner is as great as possible. Although these kinds of devices are included, it could be said that the Grammar Module is by and large based on the traditional syllabi.

3.4. Vocabulary Modules

Vocabulary Modules record between approximately 500 and 900 basic vocabulary, although the numbers vary depending on the language. The

[11] Yoshitomi, Asako (ed.) (2004) *Eigo for KIDS: Eigo de Hanaso! Teacher’s Manual*. A similar manual is currently being developed for a Japanese conversation module.

[12] Ellis (1989).

[13] Biber, Conrad and Leech (2002) *Longman Student Grammar of Spoken and Written English* presents a single-language scale, but to realize this for 17 languages would not be easy.

vocabulary from Level 4 of the Japanese Language Proficiency Test forms a common foundation, and vocabulary specific to each language has been added. Vocabulary search and synonym search interfaces have been built in, meaning that it can also be used as a rudimentary dictionary of basic vocabulary. The semantic categories for the vocabulary have been based on *Bunrui Goihyo*, Lexical Taxonomy of Japanese, elaborated by the National Institute for Japanese Language.[14]

According to the theory of universal grammar, it is believed that all languages possess the exact same grammatical structure; and in language acquisition, only lexical learning plays a part, so there is no need for structural learning.[15] If we ascribe to this viewpoint, then it means that lexical learning is purely the accumulation of elements used in language as stock in memory. Leaving this radical hypothesis aside, in reality, language acquisition involves structural learning as well as lexical learning. Furthermore, in recent years, it has been determined that, rather than learning individual and isolated vocabulary, comparatively large amounts of vocabulary can be learnt in a short time by repeatedly and intentionally learning meaningful groups of words, called “chunks.”

In addition to the elementary dictionary-type functions mentioned above, the exercises have been set up so that learners can learn about 200 basic words through exercises. In lexical learning, it seems that rather than learning the individual words by rote, it is important to learn by supposing a systematic network between the individual words. To this end, two types of study courses have been established in the exercises: “learning by situation” such as overseas travel, sports, and roads; and “learning by semantic category” such as adjectives, and things worn or carried and associated actions. This should enable words to be grouped and for them to be learnt as vocabulary networks.[16] Furthermore, the Vocabulary Modules are closely linked to the Pronunciation, Dialogue, and Grammar modules, so that students can learn the vocabulary, by checking the pronunciation of the vocabulary they have learnt, learning grammatical characteristics, and by matching the vocabulary to dialogue situations.

4. Multilingual Learning in a Ubiquitous Environment

The TUFS Language Modules assume learners can understand Japanese. However, conversely, what if the Japanese teaching materials in the TUFS Language Modules could be studied in various other languages? The TUFS

[14] *Bunrui Goihyo* (2003) Revised Edition, The National Institute for Japanese Language.
[15] See the critique in 1. Grammar, Radford (1997).
[16] See for instance LNT (Lexical Network Theory) in Norvig and Lakoff (1987).

Language Modules (multilingual version) attempt to achieve this. At the UBLI, we have been developing modules on a trial basis for non native Japanese speakers to learn Japanese. In spring 2006, Pronunciation Modules and Dialogue Modules were set up so that persons, who understand English, French, Chinese, Korean, Mongolian, or Turkish, could learn Japanese, http://www.coelang.tufs.ac.jp/english/module/ for details.

4.1. From Linguistic Theory to Education

What exactly does it mean to apply linguistic theory to educational practice using computer technology? Two instances that represent this in tangible forms are the “IPA (International Phonetic Alphabet) Module” and the “Cross-Linguistic Grammar Module.”

During development of the Pronunciation Module, while aiming for the acquisition of speech and phonetic knowledge in the various languages, the use of IPA as cross-linguistic phonetic notation was considered. At the same time, development of the “IPA Module” began. The IPA Module contains specialized knowledge which is essential, not only in the learning of foreign languages, but also in learning phonetics. Furthermore, as a result of developing the theoretical course in the Pronunciation Module, a link with the IPA Module became possible, and phonetic and phonological theory was able to be applied to the educational practice of individual languages.

The current IPA Module was developed, based on the 1996 revised version. An image and explanation of the vocal organ has been developed, and a list of vowels, consonants (pulmonic), and consonants (non-pulmonic) have been presented. By clicking on any of the phonetic symbols, users can hear the sound of that symbol. Other phonetic symbols, diacritics, and other symbols for suprasegmental sounds are also listed. Both the Japanese and English versions of the IPA Module have been released, http://www.coelang.tufs.ac.jp/ipa/ (Japanese) or http://www.coelang. tufs.ac.jp /ipa/english/index.htm (English).

The “Cross-Linguistic Grammar Module”[17] is a little removed from the practical interest of acquiring the four skills necessary for communication in an individual language. It deals with studying general linguistic issues. This module is for learners to acquire expertise related to grammar research, one of the components of linguistics. Study courses are set up in the Cross-Linguistic Grammar Module, similar to the Grammar Module. The learning objective is to take a broad cross-linguistic view of the “general grammatical characteristics” which are shared across various languages, by

[17] Makoto Minegishi, “Developing Grammatical Modules Based on Linguistic Typology”, in this volume for the concept and development of “Cross-Linguistic Grammar Module.”

taking grammatical items as examples from the Grammar Modules for those languages. What is called “grammar” in the teaching materials for the various languages varies widely, from the use or agreement of nouns and verbs, to word order. By taking a cross-sectional view of a certain grammar entry across more than ten languages, in addition to a contrastive linguistic interest, we can also give consideration to what “grammar” really is, and even what human language is. Two study courses are available: the “Step by Step Course” in which the arrangement of components in a sentence is studied in an orderly sequence; and a “Functional Course” in which characteristic expressions of each language are compared by predicate function, conative function, and presentational function, etcetera.

4.2. Second Language Acquisition Research

A difference between second language acquisition by university students, who have passed the critical stage, and the acquisition of a first language by young children over a long period of time, is that, in general, the period of acquisition for university students is short. For this reason, the social context during acquisition may have a significant impact. Previously, second language acquisition by university students had been centered around instructors in lessons at university. However, as a result of Internet technology and other innovations, recently, second language acquisition has no longer been bound to just the classroom. Now it has become possible for such acquisition to be centered around the learners, and for it to occur at places other than the classroom. In response to these kinds of demands of the times, the TUFS Language Modules were registered as e-learning materials at the Tokyo University of Foreign Studies from 2005. All students at the university can now login to their own account, and can study anywhere at anytime. They can also check their study history. With some languages, the modules are being used as teaching materials or supplementary teaching materials in an undergraduate course, but most students use them as self-study materials within the e-learning system.

University students, who have a mature character and who are well-educated, are able to take responsibility for their own learning. In other words, we should be able to regard second language acquisition for these students as autonomous learning. Within each of the modules of the TUFS Language Modules, the stages of learning are shown using sections or steps, and the path for learning guidance designed by the developers is clearly expressed. However, in network-mediated autonomous learning, there is no guarantee that learning will proceed according to that path. For example, in implementing autonomous learning in a classroom, the instructor gives consideration to the beliefs of the learners and to learning strategies, and because they know the characteristics of the individual learners, motivation should be enhanced and learning efficient. However, making network-based language learning autonomous is not all that easy. While there are some learners who can study autonomously, there are others who cannot. There is also a myriad of strategies, and it appears that it is difficult for a learner to select a strategy that suits him/herself and then to monitor him/herself. Although it may seem true that Web-based teaching materials, such as the TUFS Language Modules, provide learners with an unrestricted learning environment and enhance the possibilities of language learning, in order that effective autonomous learning can proceed in such an environment, there needs to be a system that will assist with setting goals and managing the progress of learning, with selecting teaching materials, and with pointing learners in the right direction. In the case of learning based on the e-learning system which is currently being run at the university, since the Language Modules are positioned as more supplementary teaching materials, instructors are able to point students (learners) in the right direction, and it is possible to manage the progress of their learning. On the other hand, in the case of autonomous learning, learners need to be able to evaluate their own learning achievements to take the place of classroom tests. In 2004, the Language Education Group conducted a questionnaire survey to evaluate the teaching materials in the Pronunciation Modules.[18] Then, in 2005, this was extended, and a detailed evaluation of teaching materials was conducted for the Dialogue Modules. Furthermore, a large-scale questionnaire was conducted on the degree to which the descriptions of language proficiency listed in the “Common European Framework of Reference for Languages (CEFR)” apply to Japanese learners. The results and observations from the questionnaire were listed in the *Development of Teaching Materials, Evaluation, Second Language Acquisition,* Working Papers in Linguistic Informatics 10, published in March 2006. Eventually, the plan is to present the descriptions of language proficiency for Japanese learners in the form of a Can-Do list. Based on these descriptions of language proficiency, in the future, it will be possible to set language proficiency levels using the TUFS Language Modules to a certain degree for the 17 languages.

5. Lectures, Workshops and International Conferences

As was remarked at the outset, the primary goal of the 21st Century COE Program was to select outstanding research projects in various academic fields, and by appropriating an ample budget to each project over

[18] For further details, *Second Language Pedagogy, Acquisition, Evaluation*, the 5th volume of the “Working Papers in Linguistic Informatics”, Chapter 2: Evaluation, pp.35-102.

five years, form world-class research centers, and raise the standard of Japan's research organizations. To this end, from the initial stages, the UBLI had planned two international conferences. Also, since 2002, we have invited numerous researchers from the fields of linguistics and language education, and have held lectures. Copies of the lectures have been published in *Linguistics, Applied Linguistics, Information Technology*, Working Papers in Linguistic Informatics 2 in March 2004, and in the 9th volume of the series, *Symposium, Lecture, Research Report* in February 2006.

5.1. The First International Conference on Linguistic Informatics

The outline of the first international conference was decided at the end of 2002 when the 21st Century COE was adopted. Subsequently, the First International Conference on Linguistic Informatics was held at the Tokyo University of Foreign Studies for the two days of December 13 and 14, 2003. At the first conference, research papers were presented on two key research domains.

The first key domain was research typified by computer-assisted linguistics and corpus linguistics. Needless to say, when researching linguistic structures in detail, computer-assisted corpus analysis is essential. On the premise of computer-assisted language processing, a schema is produced based on linguistic analysis, and it is important to remember that this is applied to the language processing platform.[19] This can be argued to be a field of research that is built on a collaboration between linguistic theory and computer science. Furthermore, as indicated by Francisco Moreno-Fernández, when building a language corpus, we must consider what kind of language data will be represented by the corpus.[20] From just looking at the reports from the first conference, there are a wide range of genre and individualities in language corpora, like medieval literature, bilingual databases, linguistic atlas data, workplace language, and natural dialogue. It can be said that the first conference presented the concept of linguistic analysis using various corpora.

As a result of the deepening of corpus analyses in recent years, it is now known that numerous linguistic variations can be seen in language communities. It could probably be argued that we can no longer understand the reality of language usage without directing our attention to linguistic variation. In general, when researching linguistic variation, two different

[19] For instance, Christian Leclère, "The Lexicon-Grammar of French Verbs: a syntactic database" 29-45, in Yuji Kawaguchi et al. (2005).

[20] McEnery and Wilson (2003) 77-81 and Francisco Moreno-Fernández, "*Corpora* of Spoken Spanish Language. The Representativeness Issue", in Kawaguchi et al. (2005) 120-144.

approaches are imagined. The differences between the two approaches lie in how the register, style, gender, age, interpersonal relationships and other social contexts are perceived. First, there is the perspective where social context is thought to be an independent variable. If this standpoint is adopted, the social variables are connected to linguistic variation, and the relationship between the two becomes something for quantitative analysis. Let us call this the quantitative approach. In contrast to this, there is the viewpoint where it is regarded that social context is not an independent variable, and social context and linguistic variation always form a set. In this viewpoint, describing the linguistic variation and context in as much detail as possible becomes the problem. This is called the qualitative approach. For example, the research by Kanetaka Yarimizu et al. on standardization in the linguistic atlas of regional French is a typical example of the quantitative approach.[21] In this analysis, the social context of the geographical expanse of standardization is believed to be a variable that is independent of standard form. On the other hand, the socio-pragmatic analysis of workplace language by Janet Holmes could be said to be a qualitative approach.[22] This is because, in a certain workplace, learning by experience the verbal behavior and techniques appropriate for that workplace suggests that it is integrated into the social context of the workplace, and so the social context and verbal behavior form one set.

The other key domain of the first conference was the studies surrounding the relevance between linguistic theory and second language acquisition. Mayumi Usami's assertion that the analysis of natural dialogue is necessary for developing conversation textbooks is a prime example of this. Development of the TUFS Language Modules would also have been impossible without the foundations of linguistic theory and second language acquisition. As mentioned earlier, the Pronunciation Modules have been designed, based on phonetic and phonological theory, and the Dialogue Modules are built on a notional and functional syllabus.

A large number of researchers were invited from research organizations in Japan and abroad to attend the first conference. Over the two days, the gross attendance was 300, and there was a lively exchange of opinions. Several reports were also presented by postgraduate students of the university. The collection of reports from the First International Conference on Linguistic Informatics was published by John Benjamins in the spring of 2005 as the first volume in the "Usage-Based Linguistic Informatics" series.

[21] Yarimizu Kanetaka, Yuji Kawaguchi, Masanori Ichikawa, "Multivariate Analysis in Dialectology A Case Study of the Standardization in the Environs of Paris", in *op. cit.*, 99-119

[22] Janet Holmes, "Socio-pragmatic Aspects of Workplace", in *op. cit.*, 196-220.

As a result of this first conference, the concept of linguistic informatics, the construction of which is our aim, became progressively clearer. First, I would like to define what is meant by language usage that is ultimately applied to educational practice.

There are four types of Language Modules: pronunciation, grammar, vocabulary and dialogue. In the Pronunciation Modules, it would appear that data on social phonetic variations should be applied.[23] In the Grammar Modules and the Vocabulary Modules, it is important that reference be able to be made to the question of how the grammatical rules and vocabulary are related to the actual usage. For example, Biber, Conrad and Leech (2002) provide us with one model. When explaining grammatical items, comment is given on how frequently that item appears in language use of specific genre, and on what kind of characteristic collocations there are. However, to achieve this for 17 languages would not be an easy feat. Far from it, if we take a language like French which is taught throughout the world, the fact that there are no syllabi based on linguistic usage for the negative "...pas" or the personal pronoun "on," demonstrates that the adoption of such a point of view is imperative. If usage data is applied to the Dialogue Modules, then that will inevitably be natural dialogue data.

The Linguistics Group at the UBLI has conducted ongoing analysis of linguistic usage based on corpora. It has conducted much basic research on grammatical phenomena. However, most of that analysis concerns corpora of written language, and spoken language has been in the minority. It has only been in recent years that research on spoken language corpora has been conducted in earnest.[24] Even if we consider the shift from listening to reading and writing, both written language and spoken language are important in second language acquisition. It is not as if listening skills and

[23] In the French Pronunciation Module (theoretical course), as phonemes are introduced, reference has been made to the fluctuation of phonemes, and basic explanation has been given for minimal phonetic variation. Although it may be ideal for learners to be able to learn about geographic variations, doing so would make the content complex and wide-ranging. There is a possibility that the content would become inappropriate as teaching material for beginners. At present, with regard to languages for which multiple protocol are envisaged, such as Portuguese-Portuguese and Brazilian-Portuguese for example, two separate Pronunciation Modules are developed, and they each have their own separate pronunciation description.

[24] In the case of French, it was 1987 when transcribing guidelines were proposed for the construction of spoken language corpora; and it was in 1990 and beyond when the studies of the spoken language analysis first begun to be published. Claire Blanche-Benveniste et C. Jeanjean (1987) *Le français parlé: transcription et édition*, Paris: Didier Érudition and Claire Blanche-Benveniste (ed.) (1990) *Le français parlé: études grammaticales*, Paris: Éditions du C.N.R.S..

reading skills are in some kind of subordinate-superior relationship. There are instances of learners, who have only undertaken training in listening, who develop reading skills naturally. It goes as far as hypotheses which assert that both skills are interrelated language processing mechanisms.

Since the early stages of the adoption of the 21st Century COE program, the UBLI has conducted field surveys, and has built spoken language corpora for French, Spanish, Italian (Salentino dialect), Russian, Malaysian, Turkish, and Japanese. Since building corpora involves persistent work that requires many long hours, such as for transcribing, it has really only been recently that we have been able to analyze the spoken language corpora. Research on learner corpora in Japanese and English has also been ongoing, but this is also a field of research that has only begun quite recently.[25]

In January and October of 2005, a symposium and national conference were held at the Tokyo University of Foreign Studies. Round-table discussions with COE program promoters were held, and there was open debate on how linguistic theory and educational practice should be integrated.[26] As a result of these two discussions, the direction for forming a center for linguistic informatics became clear. In other words, by further promoting the corpus linguistic analysis centered around written languages, and by inviting overseas researchers conducting cutting-edge research on the analysis of spoken language, and discussing with them the significance of research on spoken language corpora, discourse analysis of corpora, the construction of learner language corpora and the application of research findings to language education, the belief was that the subject of linguistic informatics research could be more stringently defined.

5.2. Workshops and The Second International Conference on Linguistic Informatics

On December 9, 2005, a workshop entitled “Spoken Language Corpora — its Significance and Application —” was held in conjunction with C-ORAL-ROM, a consortium researching the spoken Romance languages. In addition to a report on the results of the C-ORAL-ROM project,[27]

[25] Interest in actual research into classroom discourse had been expressed previously as part of lesson analyses, etcetera, but it was not until the 1990s before learner language corpora were built as part of second language acquisition research and foreign language pedagogy, empirical research using these corpora began, and the importance was first recognized.

[26] For details on the symposium, “Is the Integration of Linguistic Theory and Language Education Possible?”, Working Papers in Linguistic Informatics 9, 124-139. For details on the national conference, Susumu Zaima, “German Language Research Methodology Based on Language Use — Language Use, Application and Evaluation —”, 309-329 in this volume.

[27] For further details, see Cresti and Moneglia (2005).

Emanuela Cresti compared language processing at the UBLI and at C-ORAL-ROM.[28] From the UBLI, reports were presented on research into spoken language corpora for Malaysian, Turkish, and Japanese. On the following day, December 10, the Second International Conference on Linguistic Informatics was held. Following lectures on the state of grammatical research into spoken language, the pragmatic analysis of spoken language, and the application of spoken language corpora to education, a general discussion was held between the lecturers and the C-ORAL-ROM members.[29]

There was an audience in excess of 300 people over the two days, and the meeting was a success. In this way, even further clarification was given to the subject of research for the center for linguistic informatics to attempt to use computer technology and apply linguistic theory to education. In other words, the most important academic contribution of linguistic informatics is the analysis of various linguistic variations and discourse functions that appear in actual usage, by using written language and spoken language corpora based on computer science and corpus linguistics. At the same time, learner language corpora would be constructed and analyzed. Then, by incorporating the findings of these analyses into educational practice, a more efficient and advanced language education could be achieved. This is the goal of linguistic informatics. In this instance, the term “educational practice” refers to the class-mediated and network-mediated autonomous learning in the TUFS Language Modules, which were developed by UBLI.

The UBLI has received research grants over the five years from 2002. This year is the final year. We invited experts in the fields of corpus linguistics, analysis of linguistic usage, and language education from overseas and the Second Workshop on Corpus Linguistics — Research Domain — was held on September 14, 2006. And on September 15, in order to have these five years of research objectively evaluated, round-table discussions were held with the COE program promoters, to examine the formation and academic results of the Center for Usage-Based Linguistic Informatics (UBLI). By being objectively evaluated by overseas experts while acknowledging critical comments and recommendations, it is my hope that, together with the Tokyo University of Foreign Studies, the formation of the UBLI was the first step toward international recognition as a research center of global standards.

[28] Emanuela Cresti, “Some Comparisons between UBLI and C-ORAL-ROM”, 125-152 in this volume.

[29] For the report on The Second International Conference on Linguistic Informatics, see Chapter 1 of this paper.

References

Biber, Douglas, Susan Conrad and Geoffrey Leech (2002) *Longman Student Grammar of Spoken and Written English*, Essex: Longman/Pearson Education Limited.

Brugman, Claudia and George Lakoff (1988) “Cognitive topology and lexical networks”, Steven Small et al. (eds.) *Lexical Ambiguity Resolution*, Morgan Kaufman, 477-508.

Cresti, Emanuela and Massimo Moneglia (2005) *C-ORAL-ROM : integrated reference corpora for spoken Romance languages*, Amsterdam/Philadelphia: John Benjamins.

van Ek, Jan (1980) *Threshold Level English*, Oxford: Permgamon Press.

Ellis, Rod (1989) “Are classroom and naturalistic acquisition the same ?: A study of the classroom acquisition of German word order rules”, *Studies in Second Language Acquisition*, 11, 305-328.

Granger, Sylviane, Joseph Hund and Stephanie Petch-Tyson (Eds.) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching,* Amsterdam/Philadelphia: John Benjamins.

Hunston, Susan (2002) *Corpora in Applied Linguistics,* Cambridge Applied Linguistics, Cambridge: Cambridge University Press.

Kawaguchi, Yuji, Susumu Zaima, Toshihiro Takagaki, Kohji Shibano and Mayumi Usami (2005) *Linguistic Informatics State of the Art and the Future*, Usage-Based Linguistic Informatics 1, Philadelphia/Amsterdam, John Benjamins.

Kawaguchi, Yuji, Ivan Fónagy and Tsunekazu Moriguchi (2006) *Prosody and Syntax : Cross-Linguistic Perspectives*, Usage-Based Linguistic Informatics 4, Philadelphia/ Amsterdam, John Benjamins.

Koike, Ikuo (ed.) (2003) *Oyo Gengogaku Jiten* (A Dictionary of Applied Linguistics), Kenkyusha.

Koike, Ikuo (ed.) (2004) *Dai-ni Gengo Shutoku Kenkyu no Genzai* (Current Status of Research into Second Language Acquisition), Taishukan Publishing.

Lin, ChunChen, Kentaro Yuki, Kazuya Abe, Naoyuki Naganuma (2004) *TUFS Tagengo e-learning system Kaiwa Kyozai Kaihatsu* (Development of Conversation Teaching Materials for TUFS Multilingual e-learning System), Working Papers in Linguistic Informatics 1: TUFS Language Modules, Tokyo University of Foreign Studies — Graduate School, 21st Century COE “Center of Usage-Based Linguistic Informatics (UBLI)”, 115-121.

McEnery, Tony and Andrew Wilson (2003) *Corpus Linguistics,* 2nd Edition, Edinburgh Textbooks in Empirical Linguistics, Edinburgh: Edinburgh University Press.

Norvig, P. and G. Lakoff (1987) “Taking: A study in lexical network theory”, J. Asket et al. (eds.) *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 13, 195-206.

Radford, Andrew (1997) *Syntax : a minimalist introduction*, Cambridge: Cambridge University Press.

Suzuki, Takashi, Koji Matsumoto and Mayumi Usami (2005) “An Analysis of Teaching Materials Based on New Zealand English Conversation in Natural Settings — Implications for the Development of Conversation Teaching Materials —”, Yuji Kawaguchi et al. (eds.), *Linguistic Informatics State of the Art and the Future*, Amsterdam/Philadelphia: John Benjamins, 295-315.

Takagaki, Toshihiro, Susumu Zaima, Yoichiro Tsuruga, Francisco Moreno-Fernández and Yuji Kawaguchi (2005) *Corpus-Based Approaches to Sentence Structures*, Usage-Based Linguistic Informatics 2, Amsterdam/Philadelphia: John Benjamins.

Usami, Mayumi (2005) “Why Do We Need to Analyze Natural Conversation Data in Developing Conversation Teaching Materials ? — Some Implications for Developing TUFS Language Modules —”, Yuji Kawaguchi et al. (eds.), *Linguistic Informatics State of the Art and the Future*, Amsterdam/Philadelphia: John Benjamins, 279-294.

Wilkins, D.A. (1976) *Notional syllabuses*, Oxford : Oxford University Press.

Yoshitomi, Asako, Tae Umino and Masashi Negishi (2006) *Readings in Second Language Pedagogy and Second Language Acquisition In Japanese Context*, Usage-Based Linguistic Informatics 3, Philadelphia/Amsterdam, John Benjamins.

Yuki, Kentaro, Kazuya Abe, Chunchen Lin (2005) “Development and Assessment of TUFS Dialogue Module — Multilingual and Functional Syllabus —”, *Linguistic Informatics State of the Art and the Future*, Yuji Kawaguchi et al. (ed.), Amsterdam/Philadelphia: John Benjamins, 333-357.

Appendix

COE Program Promoters

Yuji KAWAGUCHI (French and Turkish Linguistics), Susumu ZAIMA (German Linguistics), Nobuo TOMIMORI (Romance Linguistics), Toshihiro TAKAGAKI (Spanish Linguistics), Yoichiro TSURUGA (French Linguistics), Ikuo KAMEYAMA (Russian Literature), Hideki NOMA (Korean Linguistics), Kohji SHIBANO (Information Technology), Makoto MINEGISHI (Theoretical Linguistics), Mayumi USAMI (Social Psychology of Language)

Research Projects in 2002-2006

Linguistic Informatics:

Developments of TUFS Language Modules, Linguistic Culture Portal Site, Multilingual Corpora, Teaching Materials for Advanced Liberal Arts Courses, Discourse Analysis, Publications of Linguistic Informatics and Working Papers in Linguistic Informatics

Yuji KAWAGUCHI, Makoto MINEGISHI, Kohji SHIBANO : TUFS Language Modules

Yuji KAWAGUCHI, Yoichiro TSURUGA, Toshihiro TAKAGAKI, Susumu ZAIMA : Linguistic Culture Portal Site, Multilingual Corpora

Makoto MINEGISHI, Ikuo KAMEYAMA, Yuji KAWAGUCHI : Teaching Materials for Advanced Liberal Arts Courses

Mayumi USAMI : Discourse Analysis

Yuji KAWAGUCHI, Toshihiro TAKAGAKI, Susumu ZAIMA, Yoichiro TSURUGA, Mayumi USAMI, Makoto MINEGISHI, Ikuo KAMEYAMA, Nobuo TOMIMORI, Kohji SHIBANO : Publications of Linguistic Informatics and Working Papers in Linguistic Informatics

Linguistics:

Corpus-Based Analysis of Linguistic Usages, Prosody and Syntax in Cross-Linguistic Perspectives

Yoichiro TSURUGA : Verbal class in French — Frequency analysis and construction —, Impersonal constructions in French

Yuji KAWAGUCHI : Diachronic research on negative constructions in French, Corpus-based analysis of French conditional

Naotoshi KUROSAWA : Word order of modifier and modified constituent in Latin and Portuguese

Kiyoko SOHMIYA : Aspects of marked constructions as seen in corpora

Kazuyuki URATA : Diachronic research on the subjunctive in English

Susumu ZAIMA, Takashi NARITA : Corpus-based research on verb construction in German

Toshihiro TAKAGAKI : Construction of a Spanish corpus and the development of relevant tools to advance Spanish language research

Hidehiko NAKAZAWA : Corpus-based analysis of Russian aspect, Utilization of a corpus for research on Russian verbs

Takayuki MIYAKE : Research on the syntactic characteristics of Chinese verbs based on corpus analysis

Keiko MOCHIZUKI : Comparative study of compound verbs in Japanese and Chinese that express "causal phenomena" and "resultant phenomena" and their corresponding English sentence structures

Shinjiro KAZAMA : Descriptive study of grammar using spoken and literary corpora

Isamu SHOHO : The causes and results of marked word order in the Malaysian language

Satoko YOSHIE : Construction of a Wakhi vocabulary corpus

Shinji YAMAMOTO : Italian language in the 21st century

Futoshi KAWAMURA : Database of case-marking for Old Japanese adjectives

Yuji KAWAGUCHI, Tsunekazu MORIGUCHI, Nobuo TOMIMORI, Hiroko SAITO, Masashi FURIHATA, Yoshio SAITO : Prosody and syntax in ambiguous sentences, Prosodic analysis of speech through the TUFS Dialogue Module

Applied Linguistics :
Discourse Analysis, Second Language Acquisition, Assessment of TUFS Language Modules

Mayumi USAMI : Construction and analysis of a multilingual corpus of spoken language, Basic research on methodology for natural conversation analysis, Development of a basic transcription system for Japanese, Korean, Chinese and English.

Tae UMINO : Construction and analysis of Japanese learner-language corpus, Basic research aimed at the development of learner's manual for 'Japanese Dialogue Module'

Asako YOSHITOMI : Construction of an English learner language corpus, Revision of the English Dialogue Module teacher's manual

Masashi NEGISHI, Hideyuki TAKASHIMA, Masanori ICHIKAWA, Koyo YAMAMORI : Development of a Language Proficiency Scale, Assessment of TUFS Language Modules

Computer Sciences :
e-learning, Natural Language Processing

Hiroshi SANO : Construction of an educational material corpus for Japanese language education

Chun Chen LIN : Construction of e-learning system of TUFS Language Modules

TUFS Language Modules (Supervisors)

IPA Module, Cross-Linguistic Grammar Module

IPA	Yoshio SAITO, Hiroshi NAKAGAWA, Yukie MASUKO
Cross-Linguistic Grammar	Makoto MINEGISHI, Shinjiro KAZAMA

Language Modules
Pronunciation (P), Dialogue (D), Grammar (G) and Vocabulary (V) Modules

English	Keizo NOMURA (G), Hiroko SAITO (P), Kazuyuki URATA (G,V), Asako YOSHITOMI (D)
German	Takashi NARITA (P,D,G), Akiko MASAKI (P), Susumu ZAIMA (V)
French	Yuji KAWAGUCHI (P,D,G,V), Akira MIZUBAYASHI (D)
Spanish	Shigenobu KAWAKAMI (P,D,G), Toshihiro TAKAGAKI (G,V)
Portuguese	Naotoshi KUROSAWA (P,D,G,V), Chika TAKEDA (D)
Russian	Hidehiko NAKAZAWA (P,D,V,G)
Chinese	Kazuyuki HIRAI (P,D), Takayuki MIYAKE (G,V)
Korean	Eui-sung CHO (P), Koichi IKARASHI (D), Hideki NOMA (G,V)
Mongolian	Renzo NUKUSHINA (D), Hideyuki OKADA (G,V), Yoshio SAITO (P)
Indonesian	Masashi FURIHATA (P,D,G,V)
Filipino	Tsunekazu MORIGUCHI (P), Michiko YAMASHITA (D,G,V)
Lao	Reiko SUZUKI (P,D,G,V)
Cambodian	Hiromi UEDA (P,D,G,V), Tomoko OKADA (P,D,G,V)
Vietnamese	Yoshio UNE (P,D,G,V), Hiroki TAHARA (P)
Arabic	Robert RATCLIFFE (P,D,G,V)
Turkish	Takahiro FUKUMORI (P), Mutsumi SUGAHARA ((D,G,V)
Japanese	Yohei ARAKAWA (V), Futoshi KAWAMURA (G), Yumiko SATO (P), Tae UMINO (D,G)

Linguistic Analysis of Spoken Language — The Case of French Language —

Claire BLANCHE-BENVENISTE

0. Introduction

The first inquiries about Spoken languages began in Europe in the early sixties for educational purposes, especially for second language teaching. Electronic corpora followed later on, mostly beginning in the nineties, with different objectives: speech recognition, prosodic and conversational analysis, socio-linguistics, etc.. The research I was involved in during a long period was mainly oriented toward corpus-based grammar for French language. French electronic corpora did not reach the size of other European languages such as English, Spanish or Portuguese, but it has nevertheless created a new field in grammatical, semantic and prosodic issues in French linguistics. I also found great interest in comparative studies between French and other Romance languages. Spoken language research brought in new and important insights on the way of considering lexical and grammatical units, on how different types of speakers relate to their language or on the way they access to meaning during their talk.

I will focus on some topics:

(1) a brief bibliographical survey of main existing spoken corpora

(2) a multi-dimensional approach to spoken texts, using prosodic, pragmatic, grammatical and semantic aspects; prosodic groupings and focalisations (using C-Oral-Rom as a basis)

(3) Some syntactic devices: focalizations, dislocations, parenthesis

(4) Lexical restrictions upon grammar

(5) Methodological issues about double-layered grammar, statistics usages, dysfluencies

1. Bibliographical Survey

1.1. C-ORAL-ROM corpus

C-ORAL-ROM, edited by Cresti and Moneglia (2005), provides a sample corpus of 300.000 words for each of four Romance Languages: Italian, French, Spanish and Portuguese. Total recording amounts to 121 hours, 1427 different speakers and 1.200.000 words transcriptions. Each corpus is given morpho-syntactic tagging, text-to-speech and speech-to-text correspondence by text to speech alignment and prosodic analysis using

WinPitch software (Philippe Martin).

1.2. Recent history

I will summarize here the short history of recent bibliographical sources given by C-ORAL-ROM within four national frameworks, extending it for French and providing some other sources I was involved in. A general survey shows how Spoken Language research was first established during the 1970s by some pioneering researchers, how it enlarged in the 1990s with electronic tools and how it blew up in the 2000s with English leadership. It is worth notice that such institutions as National Academies for language supported the enterprises for spoken Italian, spoken Spanish and spoken Portuguese. A noticeable extension must be reminded for spoken Spanish in Latin America, for spoken Portuguese in Brazil, in Africa and in Asia and for Spoken French in Canada. In such cases, speakers are very eager to compare usual practices inside and outside "ancient metropolis".

1.2.1. Italian

LABLITA (Florence) is an open diachronic corpus, begun in the 1970s in the perspective of first language acquisition and progressively developing until now for prosodic and grammatical analysis. A sample called *Corpus di Italiano Parlato* (60.000 words) was published by Cresti (2000) with the support of Accademia della Crusca.

Other main speech corpora appeared in the Nineties, the first and most famous one being LIP, created in Rome in 1993 specially for lexical research (57 hours, 500.000 words). Then followed aligned text-to-speech corpora such as AVIP (1997), API (1999), CLIPS (2001). Specific corpora were created for radio and television, for adult second language acquisition and for normal and impaired first language acquisition.

1.2.2. Spanish

Main oral corpora also begin in the Nineties, the first and most famous being CORLEC (Corpus Oral de Referencia de la Lengua Espanola, 1991), enlarged in 1997 into CREA, 1.100.000 words (Corpus de Referencia del Espanol Actual), within a cooperation with the Real Academia Espanola. Then followed several specialized corpora, for instance VALESCO for conversations, VUM for dialects, CLUVI for bilingual dialogue (Spanish-Galician). Important American Spanish corpora developed in cooperation with Spanish researchers, as reported by Moreno (2005).

1.2.3. Portuguese

The first and most famous spoken corpus, Corpus de Português

Fundamental, created in Lisbon by CLUL (Centro de Linguistica das Universidades de Lisboa) started in the 1970s, with 700.000 words. Later on, it took place in a larger one, CRPC (Corpus de Referência do Português Contemporâneo), in constant development, which presently amounts to 2.500.000 words for spoken language. Spoken materials were collected in Portugal, in European Portuguese speaking islands, in Brazil, in Africa and in Asia.

An important collection, conducted on the basis of dialect Atlases, developed from the 1970s on (ALEPG, 3.500 hours) with a special extension for lexicon related to fish activity (210 hours) and for activities in Algarve (100 hours).

Specific corpora were created for first language acquisition, normal and impaired, for broadcasting, for speech recognition and synthesis.

1.2.4. French

Two spoken French corpora issued in the early sixties, in the perspective of second language acquisition.: Le *Français Fondamental* (75 hours), and *Corpus d'Orléans* (350 hours), being presently transcribed in electronic form.

In the seventies, linguistic research on spontaneous spoken data began in University de Provence, under the leadership of GARS (Groupe Aixois de Recherche en Syntaxe). These data were progressively converted into an electronic corpus, completed in the late 90s by a computational research group, DELIC. The corpus presently amounts to 2.500.000 words. A sample (around 60.000 words) was published in 1999 by Blanche-Benveniste, Rouget and Sabio[1]. Several publications dealing with grammatical analysis of spoken language phenomena were issued (Blanche-Benveniste's editings, 1984, 1987, 1997, 2000).

Specific corpora were created for broadcast (Lindqvist, publication in 2001), for conversational analysis (Kebrat-Orecchioni), for prosodic analysis (Mertens, Morel), for dialectology, for children acquisition, for language pathology, for technical purposes (LIMSI, Habert).

Belgian and Swiss corpora were launched in the 1980s: Willems (1983) for Francophones de Gand (38 hours, transcription), Francard in Louvain (373 hours, transcription and electronic tools). Very large spoken corpora were collected and studied in French speaking places in Canada, specially in Quebec and Ontario. It would deserve a specific bibliography.

[1] Blanche-Benveniste, Rouget et Sabio, 1999, *Choix de textes de français parlé : trente-six extraits*. Paris : Champion (collection : les français parlés, textes et études).

1.2.5. Extensions

French resources can be enlarged by referring to Blanche-Benveniste and Jeanjean (1987) for the period anterior to 1990 an to Cappeau and Seijido's recent inventory (2005) for the following one.

The oldest collection of spoken French recordings was created in 1911 by F. Brunot. It contains several types of productions. One was made to enhance French normative pronunciation, by giving famous literary or political models, such as Appolinaire reciting his poems or Dreyfus delivering a lecture. Another one was devoted to second language acquisition; one dealt with dialects (Ardennes, Limousin), and another one with language pathology. These recordings are housed by the National Library (BnF, Bibliothèque nationale de France) as "Archives de la parole".

The most recent one in the due to phoneticians and prosodists, who are presently building in Toulouse a corpus called PFC (Phonologie du Français contemporain) intended to reach 500 hours. A specific one, CLAPI, is being built in Lyon for studying interactions (200 hours).

Several non-French linguists collected – an are still collecting - their own corpora in order to teach French as a second language: in Sweden, in Danemark, in Finland or in Germany (Schmale). In 1984, Schmale published a sample of the collected corpora. In 2001, C. Lindqvist published a sample of radio and television recordings.

We should also mention the existence of numerous spoken corpora built by non-linguists scientists, whose recordings and transcriptions could be useful for linguists, in case we could favour some inter-disciplinary analysis. The most impressive are made by historians collecting oral sources (F. Descamps 2001)[2]: Service historique de l'Armée de l'Air, Direction des Archives du Ministère des Affaires Etrangère, Comité pour l'Histoire Economique et Financière de la France, Service Historique de l'Education, Institut National de la Recherche Agronomique, etc.

Present-day largest corpora are those made for English, for instance the British National Corpus (10.000.000 words). Recent Corpus for Netherlands language is reaching the same dimension. Such large corpora enable linguists to describe the whole grammatical and lexical usages evidenced by the speakers, as in Biber and alii's Grammar of Spoken and Written English (1999). Smaller dimensions often present some phenomena with so few occurrences that it is impossible to draw any valuable generalization.

I got involved personally in some non-French corpora, by directing

[2] Florence DESCAMPS, 2001, *L'historien, l'archiviste et le magnétophone. De la constitution de la source orale à son exploitation.* Paris : Comité pour l'Histoire Economique et Financière de la France (804 p.).

PhDs in EPHE in Paris – Outi Duvallon (2003) for Finish, Il-il Yatsiv (2004) for Hebrew - or by collaborating with colleagues on Spanish, Catalan or Italian data.

In every case, it is worth mentioning how expansive are spoken corpora. Transcriptions are extremely expansive because they require cautious and minute attention. Recent financial evaluations[3] stated that transcribing one word, from the first record-hearing to the last computerized transcript, amounts to around one euro. Large spoken corpora can only be created by large financial consortiums.

1.2.6. What use are spoken corpora ?

In present-day research, computational and communication sciences are interested by spontaneous spoken data: automatic analysis of speech, training of speech recognition systems, speech synthesis, prosodic components, fine-tuning of automatic segmentation tools and labelling, etc.

Linguists get more and more utterly involved in oral researches, partly because they stopped relying only on intuition and partly because quantitative links between grammar, prosody and lexicon seem to very important for descriptive purposes. New perspectives have evolved on grammaticality, on grammar and prosody on the relations between frequency and linguistic structures, (J. Bybee and P. Hopper 2001). Spoken corpora play a role in several linguistic applications:

- first and second language acquisition
- Comparison between standard and regional pronunciation
- Comparison between different types of prosody
- Comparison between children and adult speech
- Comparison between speeches in different social parts of a population
- The role of styles in oral speech
- Data base for comparing impaired and normal speech

I would like to express some personal insights gathered during my experiences with different spoken data.

- Spoken language has strong inner variation, according to genres and situations. It cannot be reduced to familiar conversations.
- It is very important to study at the same time low-level speech in spontaneous situations and high-level speech in public ceremonious situations (with the same speakers when possible).
- Children up to 12 years are able to parody high-level speech from adults, which give useful insights on their sometimes hidden competence

[3] S. Gillis for Netherlands Corpus, personal communication.

- Speakers build their utterances. They don’t rely on lexicon and intonation alone, as it was often said
- In several languages, embedded subordinate syntax is related to situations types: it never occurs in familiar conversations and it does occur in argumentative speeches and in technical explanations.
- When people talk about their profession., professional speech extends upon ordinary speech. Utterances supposedly restrained to written language thus enter in day-life speech, in several languages.
- Formulaic expressions play an important role.
- Ordinary speech has many lexical repetitions, so that lexical density is lower in spoken utterances than in their written equivalents (Halliday 1985).
- Quantitative difference between very frequent items and less frequent ones is higher than in written counterparts.
- In several languages, low-level speech has few noun-phrases acting as subjects, while ceremonious speech has more.
- In several languages, oral discourse structures have a very important rhetorical component: symmetric and anti-symmetric devices, grammatical configurations and regular lexical repetitions.
- So called “dysfluencies” (repairs and retractions) are more than performance accidents. They reveal important cognitive constructions, for instance in the development of grammatical phrases: syntactic frames coming before lexical filling; lapses and reformulations showing a large working memory, etc.
- When transcription conceals such strategies (for example when corpora are “re-written” for publications), many linguistic phenomena disappear: turn taking, topic management, negotiations of meaning , fillers, feed-backs, hedges, discourse markers, politeness strategies, repairs, markers for stating an opinion (I think) and markers for giving a reason (because).

2. Multi-Dimensional Approach

My report will be limited to French Spoken corpora. Spoken French entails new grammatical descriptions, usually discarded because written data had prominence. Descriptions take into account new dimensions such as prosody, frequency in item occurrences, and several non-canonical devices. Units cannot be delimitated in the same way for spoken and written data. Written delimitations between words and sentences have no strict correspondence with oral units.

2.1. Describing utterances

According to usual descriptions of French language, spoken corpora contain a high amount of non-canonical patterns, which would mean disorder, non-grammatical linkages and a high percentage of incomplete utterances

(Le Goffic 1994). However, it may be the case that some traditional descriptive arrays fail to describe linguistic realities, and can be replaced by more convenient tools. For instance, it can be shown that finite verbal sentence cannot figure as the sole syntactic unit. Comparing spoken and written English, Halliday (1985) could state that, contrary to usual prejudice, spoken language possesses high subordinate imbrications, more than in the written counterparts.

2.2. New tools for prosodic units

Technological tools enable us to take into account prosodic cues without being a specialist. Aligned corpora provide a fist improvement. Visible prosodic graphs, denoting pitch, rhythm, intensity, speed and pauses give useful controls (Ph. Martin 2005).

According to Cresti and Moneglia's hypothesis [Cresti and Moneglia (2005)], prosody can supply the searchers with adequate units for intra and inter-language comparisons. For instance, in all four Romance languages utterance length is correlated to speed and to dialogic turns (lowest length in telephone speech, in the four languages). But, according to prosodic delimitations, coordinative markers do not appear at the same place in the utterances. In all four languages, pseudo-subordinations, by *que* or *che,* play an important role (see, further, parenthetic verbs)..

2.3. Some prosodic groupings and de-groupings in French

Semantic, pragmatic and prosodic factors indicate groupings beyond syntactic units. Some seem to be partly iconic, such as presupposition coming before central assertion and consequence coming after. Other ones rely on symmetrical devices, such as particular comparisons types or positive and negative equilibrium. Other ones show specific de-groupings, known in rhetorical tradition as epexegesis.

2.3.1. Iconic non-symmetrical groupings

Pre-suppositions

Presupposed elements tend to come before central assertion. Specific items such as causal *comme* (meaning roughly "as it has been well established") never occur in a second position:

> Comme je pouvais plus lever le bras, je me suis fait opérer la clavicule
>
> (as I could not lift my arm, I had my collarbone operated)
>
> Comme je suis assez romantique, enfin, ou sentimental, je ne m'en cache pas du tout
>
> (As I am rather romantic, well, or sentimental, I don't conceal it)

Conditional presupposition entailing negative consequence are generally uttered before the consequence. Such is the case for a frequently used expression *avoir beau* (whatever one does…, although), which could never come after the uttered consequence :

> J'aurais beau faire le tour de France, il n'y a rien à faire
>
> (although I could touring around France, it would not do)
>
> J'ai beau avoir plus de soixante ans euh moi j'en ai soixante-sept ben je peux vous dire que – j'ai pas le temps de dire ouf (Although I am more than sixty me I am sixty-seven well I can tell you that I have no time to say ouf)
>
> Ils auront beau essayer de parler comme un Français leur accent il est là
>
> (Although they try to speak like French people their accent it is here)

Consecutions

A reverse disposition holds for consecutive elements. Consecutive clauses usually come after the nucleus part, in a sort of iconic disposition. Consecutive marker (*si bien que, so that*) never occurs before the assertive one.

> La peur du gendarme fait peut-être lever le pied, si bien que il y a peut-être un peu moins d'accidents
>
> (fearing policemen makes perhaps lift the foot, so that there are less accidents)
>
> On s'est trouvé bloqué à Melun, si bien que, à Melun on a recherché encore de nouveau une maison
>
> (We got stuck in Melun, so that in Melun we try again to find a house)
>
> je lis énormément, énormément, énormément, si bien que, comme j'ai beaucoup d'amis, on m'apporte des livres
>
> (I read a lot, a lot, a lot, so that, as I have many friends, they bring me books)

Such examples are very frequent. As a consequence, the grouping together of two prosodic units, one acting as a dependent-melodic sequence and the other as an assertive one can receive the general meaning of presupposed cause and consecutive consequence, even with no explicit morphological marker

> Il avance – je recule (he walks on, I walk back)

This very simple scheme gives way to numerous frozen sequences, such as

> Plus beau que ça, tu meurs (more beautiful than that, you die)

2.3.2. Symmetrical groupings

A very frequent symmetrical grouping is one pairing a positive verbal expression with a negative one, as if there was a kind of balance between both, which allows several meanings, according to contexts.

Alternative in interrogative contextes :

Est-ce que j'ai dormi ? Est-ce que j'ai pas dormi ? Je n'en sais rien
(Did I sleep ? Did I did not sleep ? I don(t know)

Alternative in temporal measure

un coup ça marche un coup ça marche
(once it works, once it does not work)
Je n'en sais rien parce que, une fois il passe, une fois il ne passe pas, ça dépend
(I don't know because once he comes and once he does not. It depends)
Parfois ils comprennent, parfois ils comprennent rien du tout
(sometimes they understand, sometimes they don't understand anything)

Alternative meaning "whatever"

Tenir les corbières , qu'il fasse froid, qu'il fasse pas froid…
(To hold the baskets, being cold or not cold)

Alternative in selecting an item among a set :

Vous avez des têtes qui plaisent, des têtes qui plaisent pas
(you've got faces that please, faces that don't please)
Il y a des éléments qu'on peut donner et d'autres qu'on peut pas donner
(There are elements you can give and others you cannot give)

Opposite sides of a regular causal/ consequence sequence :

Si je leur dis "non", ils commencent à pleurer ; si je leur dis "oui" ils commencent à s'amuser
(If I tell them "no", they start crying; if I tell them "yes", they start playing around)
tu appuies sur le rouge, ça marche; tu appuies sur le vert, ça s'arrête
(you push the red one, it works ; you push the green one, it stops).

Such processes using opposite polarities result in semantic equivalences of indefinite pronouns and indefinite adverbs: *whenever, whatever, in some cases, some, sometimes*. It shows an example of how syntactic means are complementary with morphological and lexical units.

Conditionals

Two conditionals verbs group to form an whole utterance expressing hypothesis first and consequence next, without any morphological link being expressed:

On habiterait la Grèce, ce serait un atrium
(should we live in Greece, it would be called atrium)
Il le dirait, il faudrait pas le croire
(he would say it, you should not believe him)

on (n')aurait pas assumé cette position correctement – je crois qu'on se serait fait montrer du doigt par les petits copains

(ha we not assumed this position correctly, I think we should have bee finger pointed by our little friends)

2.3.3. De-grouping

Prosodic schemes may act as "de-grouping" processes, cutting syntactic patterns into smaller pieces. Here are two examples, one which can scatter a syntactic sentence into many sub-constituents; another one in which a syntactic sentence breaks into another one, using it as a "nest".

Epexegesis

Prosodic units sometimes cut a syntactic construction into two prosodic parts. For instance, a complement can be divided from its governing verb by a final prosodic break (signalled here by a period):

Alors il a écrit. Au directeur. A l'inspecteur. Au préfet.

(then he wrote. To the director. To the inspector. To the prefect)

Il m'a marché sur le pied. Exprès (Béguelin (2002))[4]

(he trod on my foot. On purpose)

In such cases, prosodic units do not correspond with syntactic units (Deulofeu (2002)).

Parenthesis

Parenthetical sequences, having their own prosodic and syntactic coherence, cut the flow of another prosodic and syntactic sequence, which act as its "host". The dividing device is marked by several prosodic means (lower register, higher speed, lesser accentuation). Parenthetical sequences usually mean a sudden change in linguistic attitude. They may express several kinds of "discourse breaks": change in temporal referring,, comment on what was just being said, interactive markers, etc. Parenthetical organizations are much more numerous in spoken language than they are in written counterparts. In some speakers discourse, parenthetical sequences sometimes amount to one third of the total utterances. Here are some examples (I use graphic parenthesis to facilitate the reading):

J'aimais mieux l'autre route parce que avant (maintenant ça a change) c'était une petite route de montagne

(I preferred the other road because before (now it changed) it was a small mountain road)

[4] Example quoted by Béguelin, *Cahiers de Linguistique Française* n° 20, apud F. Zay. The period here corresponds to a final break in prosodic pattern.

Ils ont (je m’en souviens bien) changé d’avis dès le lendemain
(they had (I remember) changed their mind the following day)

Recognition of such patterns can be difficult, because they cannot be identified only by prosodic cues. The difference between hosting and hosted elements are very often more sensible in meaning and syntactic coherence than in prosodic indications, although prosody is indispensable. A multi-dimensional approach then proves very useful.

3. Some Syntactic Devices

There is no specific syntax that could be attributed to spoken use and that could not be found in written language. Nevertheless, some syntactic pattern seem to occur more frequently in spoken usages than in written equivalents. Three examples will be proposed here: one about focalisations, involving the notion of “independent utterance”; one about right and left dislocations, involving the notion of strict or loose co-reference within syntactic patterns; and one about “parenthetical verbs”, inviting us to separate verbal morphological appearance and verbal syntactic function.

3.1. Focalisations

One focalization type based on *c’est….qu-*. received much attention in French linguistics (Lambrecht (2001)). There are many other types, not so well described, among which one which is frequently used with focalizing modifiers meaning “at least, only”, mainly marked by position and prosody, appearing more often in spoken discourses than in written ones.

Position marking uses a pre-verbal space: focalized elements come before the subject + verb sequence... Contrary to a well-known belief, French can place a complement before the verb it is attached to, in such examples as (Sabio (1995)):

des cerises il voulait (cherries, he wanted)
à ceux-là je voulais parler (to those guys, I wanted to talk)

The complement phrase *des cerises* relates to verb *il voulait* ; complement phrase *à ceux-là* relates to the verbal phrase *je voulais parler.* This marked world-order OSV can apply to all verbal constructions, without any lexical restrictions[5]. In such cases, the pre-posed complement gets an utterance-final prosody, as if it were an answer to a question; *what did he want ? Cherries*

[5] The pre-posed complement is a general scheme, contrary to more limited ones like in *les cerises, j’adore j’aime, je déteste,* (cherries, I love, I hate), valid only for some verbs (opinion, sentiment) and having a different tonal pattern.

he wanted. The verb, not being in the scope of the focalization, only receives a post-final prosody, without any modulation:

des cerises	*il voulait*
Pre-posed focalized complement	verb it is attached to
Final prosody	post-final prosody

Such a focalized complement can figure alone, without the verb it is syntactically and semantically attached to. It may then realize as an independent utterance:

Que voulait-il ? Des cerises.

(what did he want ? Cherries)

Such a focalization type does not restrict to complement. It may also affect subjects. But in the case of subjects, it would be impossible to mention any pre-posing device, because subjects are usually pre-posed to the verb they are attached to, with or without any focalization. Intonation is the only markers for subjects made of nominal phrases, or pronominal forms as elle, elles, nous, vous. This specific intonation is very sensitive when a focalized subjects are affected by a restrictive modality *rien que, au moins* (no one but, at least):

Non-focalized subject	focalized subject
Les amis viendront	rien que les amis viendront
(friends will come	only friends will come,)
elle viendra	rien qu'elle viendra, elle au moins viendra
(she will come	only she will come, she at least will come)
vous viendrez	rien que vous viendrez
(you will come	only you will come,)

We find in the corpus:

Mais enfin au moins deux pompiers s'étaient joints aussi au groupe

(but then at least two fire-men joined also the group)

Here the nominal phrase au *moins deux pompiers* is strongly accentuated and marked with final intonation contour.

Weak clitic pronouns cannot bear such tonal status and accentuation force.

Il viendra	*rien qu'il viendra,

Il and *ils* have accentuated correspondents[6] *lui, eux*:

[6] Pronoun *on* has no accentuated correspondent; it cannot be focalized.

Non-focalized subject	focalized subject
il viendra	lui viendra, rien que lui viendra
ils viendront	eux viendront, rien qu’eux viendront

We find many instances of *eux* used as a “strong” subject in the corpus :

Eux le parlent (them speak it)
Moi je suis blanc et eux sont noirs (me I am white and them are black)
Eux aussi atterrissent en Algérie (them too land into Algeria)
Eux ne disaient rien (them said nothing)
Eux devraient manifester sous les fenêtres du gouvernement
(them should manifest under the government windows)
eux portent toujours en congé
(them always go on holidays)

Such “strong” focalized pronouns can even occur in non-verbal utterances, as in the following example where *eux* third occurrence is associated to a modal assertion, *peut-être* (perhaps):

- Je pense que c’est plutôt eux qui sont racistes, c’est pas nous, c’est eux
- *Eux* peut-être, oui

First and second person singular have no available form, because “strong” forms *moi, toi,* cannot function as subjects:

Je viendrai	*rien que je, *rien que moi viendrai
(I’ll come	only I will come)
tu viendras	*rien que tu , *rien que toi viendras,
(you’ll come	*only you’ll come)

In such a case, focalization can only be expressed by syntactic devices. A frequently used one is the verbal clause, *il n’y a que… qui*:

Il n’y a que moi qui te pose des questions
(there is only me who ask you questions)
Il y a que moi qui peux les avoir
(there is only me who can get them)

This recourse extending beyond strictly necessary situations can be used with other items which could do without, nominal phrases or strong pronouns *lui, eux*:

Il n’y a que le docteur qui ne le sait pas
(there is only the doctor who does not know it)
il n’y a que lui qui est né là-bas

(there is only him who was born there)

Another solution is the use of “clitic doubling”: a weak clitic pronoun acting as a subject near the verb and its strong version being used with the restrictive modality:

Toi au moins tu es intelligent
(you at least you are intelligent)

As it was seen for the precedent device, clitic doubling often extends beyond its strict necessity domain. It applies to pronouns and nominal phrases that could do without:

Lui au moins il sait
(he at least he knows)
eux au moins ils ont poussé
(them at least they have grown up)
les enfants au moins ils se sentent entourés
(children at least felt (people) took care of them)

This rapid sketch among focalization devices shows how heterogeneous the domain is. It also shows that we cannot describe it by using only a strict functional perspective. We can describe some grammatical recourses by stating how necessary they are in given situations; but we must admit that, very often, they are used beyond strictly informative needs. Economic schemes and redundancies are deeply entangled.

3.2. Left dislocations

Syntactic dislocations is a famous topic in French linguistics (Blasco-Dubelcco (1999)). It seemed specially interesting when dislocation applies to subjects, because other Romance languages don’t use it the same way as French does. Some linguists thought it could be treated as a typological feature and some even argued that French was evolving toward a state in which all grammatical subjects would undergo dislocation the way some young people do it (Ashby (1988), Lambrecht (1994)).

I will focus here on one dislocation type: the one containing a “strong” pronoun *moi, toi, lui, elle, nous, vous, eux, elles,* that comes before the subject +verb sequence. When it seems to be semantically related to a “weak” clitic pronoun acting as a subject, it could be analysed as a case of “clitic doubling”, as it happens so often with *moi je, toi tu*:

moi je trouve que c’est bien
moi je pense que ça va réellement se développer
moi j’avais des sous un peu là-bas

Moi les garagistes je me méfie
(me, the garagists, I am cautious, indeed)

But in many other cases, the strong fronted pronoun bears no anaphoric relation to any weak pronoun inside the verbal phrase. It acts as a thematic element which gives the whole utterance a kind of personal frame: "as far as I am concerned, as far as you are concerned, things work this way":

nous, il y a Coralie qui a commence la flûte
(us, there is Coralie who began (playing) flute)
nous les brebis vont dans la colline
(us, sheep go out in the hills)
moi il y a tous les égouts

A noticeable property is that such fronted pronoun phrases do not bear any prepositional mark that could induce a relational link with the following verb. For instance, verb *falloir* ("be necessary") normally requires a sort of dative complement marked by preposition *à*. But fronted thematic pronouns have no preposition at all, as shown by *eux*, instead of *à eux*:

mais eux il leur fallait du lait
(but them we had to give milk)
moi il m'est arrivé de nombreuses histories
moi il m'a fichu une paire de baffes
moi il me reste après à constituer de magnifiques dossiers

They are sometimes related to a possessive determiner inside the verb phrase: *Moi mon…, toi ta….*

Moi ma soeur elle était venue à la maison
Moi il a remplacé mon père
Moi ma mère me gardait mes enfants
Moi ma commission est la même

It could be treated as a sort of indirect anaphoric relation; but it seems more convenient to analyse it, as in previous cases, as thematic elements. The same analysis could also prevail when such thematic pronouns have an anaphoric correspondent inside the verb phrase.

In spite of the appearance, they can be analysed here also as thematic elements, the anaphoric relation being a supplementary semantic information, but not a basic syntactic one.

3.3. Parenthetical verbs phrases

A parenthetical status was recognized for long for “inquit” verbs such as *il dit, dit-il* (he says), which take place either at the beginning of an utterance:

Il dit qu’il fait beau aujourd’hui (he says that weather is fine to-day)

or inside the utterance, or at the end, (sometimes with subject inversion):

Il fait beau, dit-il/ il dit, aujourd’hui,

il fait beau aujourd’hui, dit-il / il dit

Many other verbs can classify in similar ways, for instance opinion, belief, appearance verbs, *je crois, je trouve, je pense, il paraît, il semble:*

Je crois qu’il fait beau, il fait beau, je crois

(I believe that the weather is fine, the weather is fine, I believe)

on prenait surtout les femmes, paraît-il

(they took specially women, it seems)

Such verbs were labelled “weak verbs” because they show strong restrictions on usual verbal properties: almost exclusive use of first person singular; no interrogative nor negative markers; no temporal, locative nor manner adjuncts. Even when they apparently command a that-construction, they don’t really subordinate anything. They act as modifying elements, in the same way as some adverbial components do.

Tomasello (2003) once showed the importance of such facts in English for evaluating how children acquire subordinating devices. According to his findings, they start learning parenthetical verbs like *I think*, in a formulaic way, when they are only three year old. People unfamiliar with syntactic analysis could understand the phenomenon as being the acquisition of subordination as a general device. But they don’t handle real subordinate phrases like *I explain that* until they are four or five. What people misunderstand as subordination is only the frequent handling of parenthetical verbs.

4. Lexical Restrictions upon Grammar

Lexical restrictions can be observed in many fields (Sinclair (1991)). They entails some correctives on grammatical schemes.

4.1. Long dependencies constructions

They are limited to a narrow set of verbs: *dire, falloir, vouloir,* (C-Oral-Rom 1, 154)

C’est là que je dis que le communisme a existé

(it is the place where I say communism existed)

C’est moi qu’il faut qui parle maintenant
(it is me that have to speak now)
Qu’est-ce qu’il faut qu’on fasse pour lutter contre l’invasion
(what do we need to do to fight against invasion)
Qu’est-ce que tu veux que je fasse avec ça
(what do you want me to do with it)

A sound description would relate such constructions to the small list of verbs they are attach to and would avoid generalizations presenting them as plausible with all subordinating verbs.

4.2. Post-posed subjects of “*Dit-il*” *type*

Post-posed subjects of *dit-il* type look very specific of written discourses. They are clitic pronouns linked to the governing verb in the way a suffix could be. They mostly occur in two particular contexts: interrogative and quotative devices. Contrary to what could be expected, we find such post-posed subjects in the spoken French corpus, but they undergo strong lexical restrictions. Here are some results for three most frequent verbs *être, avoir, dire.*

Dire has only 2 occurrences, one being said by a lawyer:
Et pourtant, dit-il, il n’a pas été invité au mariage.

With *être* and *avoir,* the post-position is slightly more frequent, but mainly when the verbs are used as auxiliaries: 42 occurrences for *est-il,* 30 for *a-t-il*. In both cases, lexical restrictions are dramatically restrained: 85% of the *est-il* examples (36/40) are used for one and the same impersonal verb *se passer* (to happen)

Que s’est-il passé ?

60% of the *a-t-il* examples (18/30) are used for one and the same impersonal verb *il y a* (there is)

Combien y a-t-il de nichées environ par an ?

This little grammatical domain gives a striking illustration of how spoken language grammar can be limited by lexical items (Tognini-Bonelli 2005).

4.3. Relative pro-nouns

French relatives could be presented as a declension system:

	without semantic nor morphological marking					
nominative	*qu-i*					
accusative	*que*					
genitive	*dont*					
	with semantic marking		with morphological marking			
prepositional phrases	- Human	+ Human	m sg	f sg	m pl	f pl
	quoi	*qui*	*lequel*	*laquelle*	*lesquels*	*lesquelles*

But these relative forms are not used on the same level. Only *qu-i* and *que* are frequently used. Oblique forms *dont, quoi, qui lequel,* appear with strong restrictions. Here is a survey of spoken usages for *dont* and *lequel.*

Dont

Relative pronoun *dont* is taught in elementary school, but teachers find it difficult and many speakers (young children, low-scholarship people) very rarely use it. Some speakers tend to use use it in frozen phrases such as *la façon dont* (the way in), which amounts to about 15% of all instances in the corpus:

On fait trop attention à la façon dont on parle
(people pay too much attention to the way they speak)

Pragmatic situations in which *dont* can be frequently heard are professional speech (radio, TV, scientific lectures, political addresses, technical explanations, literary comments). But even in such situations, speakers often hesitate and switch from *dont* to *que* or to other items. On a total of 500 uses of *dont*, 11% show explicit hesitations or repairs.

il y a toute une série de manipulations *qu'on dont on* ne peut pas entrer voir en détail (brulure 15, 9)
(there are a lot of manipulations which of which one cannot get into)

Dont is used in the dependence of verbs, nouns or quantitative expressions. When governed by a verb, it comes with *parler de* (to speak of), *se rappeler, se souvenir de (*to remember), *se servir de* (to use), *être fier de* (to be proud of), *se passer de* (to do without):

Il y a des éléments dont on ne se souvient pas du tout

(there are elements of which you don't remember at all)

Il leur fallait certaines choses dont nous on pouvait se passer (Barb 1,8)
(they wanted some things which we could do without)

The most frequent association is the one between *dont* and *parler de,* (to speak of, to tell), which represents more than half of the total uses with verbs:

Il y avait ce lavoir dont je vous parle (Pr 54, 12,4)
(there was this washing place of-which I tell you)

We have to state that the most probable use of pronoun *dont* in modern spoken French is in the dependence of verb *parler.*

When governed by a noun, *dont* is far less frequent and different speakers have different ways of using it, some ignoring it totally. The following example was said by an administrator speaking of his profession:

Voici le document dont nous présentons un extrait
(here is the document of-which we present an extract)

Here lies the main difference between skilled and non-skilled users. Some people may have a large display of such constructions and other may have none. When people use it seldom, the main use happens to be related with nouns designing human beings, *dont la famille, dont la mère, dont le père..* (of which family, of-which mother, of which father..):

c'est une personne dont la famille est venue spontanément en Algérie (Alger 6,1)
(it is a person of-which family came spontaneously to Algeria)

ce sont des gens qui n'ont jamais travaillé – dont les parents n'ont jamais travaillé (voyages 5,5)
(they are people who never worked, of which parents never worked)

The third main use of relative *dont* is with numerals. It then gets more or less a meaning which could be explained by "among which":

J'ai été arrêté près de deux ans dont un an couché (Po 4Ap 172,12)
(I had to stop for nearly two years, among which one year lying)

Les trois autres frères – dont mon grand-père, ont fait leur vie ailleurs (Alger 10,3)
(the three other borthers among which my grand-father, made they life in anotther place)

A much rarer pattern is one where *dont* means "about which, about whom", with declarative verbs *dont je dis que, dont je pense que :*

cette société dont je t'ai dit qu'on l'appelait la holding (Cast 1,5)
(this society about which I told you it was named a holding)

Here is a brief survey of how 500 instances of relative pronoun *dont* do occur in the Spoken French Corpus. If we take off hesitations and repairs (around 11%), the distribution can be stated for 443 examples as follow:

dont + verb	49 %	(between 30 and 50 % with *parler*)
dont + noun	25 %	
la façon dont	15 %	
dont plusieurs	10 %	
dont je dis que	1 %	

Dont is related to some forty different verbs, the most frequent being *parler* the other being, in regressive order *:*

dont on a besoin	(of which you need)
dont il fait partie	(of which it is a part)
dont je me souviens/ rappelle	(of which I remember)
dont je m'occupe	(of which I take care)
dont on se sert	(of which we use)

Many speakers tend to use *que* instead of *dont* :

Je voulais faire un stage de formation que j'avais besoin (Convoc 7,16)
si tu as acheté quelque chose qui correspond à ce que tu as besoin (plais 5,9)

A same speaker (a shopkeeper in next example) may use both for the same verb, within 3 seconds distance:

Avant même d'écouter ce que le client avait besoin […] il fallait savoir ce dont le client avait besoin
(Prévoy 1,10-1,15).

Obviously, the distribution cannot be stated only with grammatical surroundings. In a flectional perspective, we could imagine *de + lequel* taking the place. But such is not the case.

Lequel

I take off interrogative pronouns (*son but est lequel ?, je me rappelle plus lequel*) and a specific non-prepositionnal type, which only occurs in a lawyer’s speech[7]:

Il y a un échange de correspondance, lequel est fait par notre service

Among 250 remaining instances of masculine *lequel, lesquels,* nearly 40% are attached to preposition *dans,* with almost always the same general meaning,” the place in which I live, or work”:

Le groupe dans lequel on est, le milieu dans lequel on vit

Le milieu professionnel dans lequel j’évolue, le secteur dans lequel je travaille

Feminine *laquelle, lesquelles* show another tendency : nearly 15% are fixed on one frozen expression expressing causality :

C’est la raison pour laquelle j’ai décidé d’arrêter

(it is the reason why I decided to stop)

les raisons pour lesquelles il est arrivé dans le sang […]

(the reasons why it came into the blood)

In both cases, formulaic usages are important. Moreover, instances of *duquel, desquels, de laquelle, desquelles* are so scarce (some 20 examples in the whole corpus) that they cannot be interpreted as complementary forms for *dont.* The whole scheme about relative declension is inadequate.

4.4. Subjunctive

Subjunctive mood is limited to one and only tense, the present (simple and compound form). Imperfect subjunctive disappeared, except in a few frozen expression, mainly used by politicians, or in formal situations (*qui l’eût cru, dussè-je, encore êut-il fallu…*). But, as far as we can rely on our spoken French corpus, subjunctive is widely used in present tense, with a rather large lexical dispersion[8], although there is an important lexical fixation on one verb governing subjunctive, the verb *falloir* (must).

I examined three very frequent verbs, *faire, dire, aller,* which have distinct morphological forms for first and third person singular subjunctive: *fasse* (185 occurrences), *dise* (83 occurrences) *aille* (78 occurrences). These three subjunctives mainly occur under thee government of verb *il faut*:

Il faudrait que ça se fasse vite

(it must that it be done quickly)

Il faut que tout le monde aille le voir

(it must that every one go to see it)

[7] Such uses are supposed not to exist in modern French (Grevisse-Goosse (1987)). But they still show in professional discourse (lawyers, administrative staff, etc.).

[8] The situation is totally different in Canadian French, as Shana Poplack (2001) could show.

Il faut que je vous dise la vérité

(it must that I tell you the truth)

Il faut que + (je, il) fasse amounts to more than half of total occurrences
Il faut que + (je, il) aille amounts to a little less than half of total occurrences

Il faut que + (je, il) dise amounts to one third or one fourth of total occurrences, according to different situations ; less in formal ones, more in informal ones.

Other frequent elements governing subjunctive for these three verbs are:

- Verbs: *vouloir, attendre* and more than 30 other verbs like il vaut mieux, j'aimerais bien, j'ai bien peur,

- Prepositional phrases + *que*, the most frequent one being *pour que,* and other ones like *avant que, plutôt que, jusqu'à ce que, le temps que* (this one is not usually identified as a prepositional phrase, but it does function as such) :

le temps que j'aille prévenir Alexandre que qu'on fasse tout ça et on est retourné dans la chambre

(time for me tell Alexandre…we returned into the room)

le temps que j'aille à la cuisine, elle, elle a fait la catastrophe

(time for me go to the kitchen, she did the wrong thing)

le projet a un peu germé, le temps que tout le monde se décide

(the project grew up, time for everyone to decide))

-relative clauses,

- qu'est-ce que tu imagines comme restaurant ? – Ben, un restaurant qui soit quand même assez grand

(what do you imagine for a restaurant - well, a restaurant that would be large enough)

- injonctive meaning

ben qu'elle fasse un prêt, ce sera bien plus simple

(well, let her ask for a loan, it will be much easier)

The number of verbal and prepositional items governing the subjunctive amounts to more than 50, so that lexical dispersion can be considered as rather large. That is why we can say that, in French language as it is spoken in France nowadays, present subjunctive is a freely used grammatical device, without lexical restrictions, except the one figured by *il faut*.

4.5. Quantitative importance of frozen expressions

Let us take an example for the difference between intuitive and corpus-based data, concerning one type of verbal complementation. LADL methods (Leclère 2002), using intuition-based data, presents French verb *planter* (to plant) as a "cross construction" verb, displaying three arguments, agent, theme and locative, into two main realizations (table 37M). Both theme and locative are presented as belonging to the same botanic semantic field:

(1) Je plante le jardin (locative) de roses (I plant the garden with roses)

(2) je plante des roses dans le jardin (locative) (I plant roses in the garden)

The locative argument is supposed to receive two grammatical status, one as a direct object, planter le jardin, and one as a prepositional phrase, dans le jardin. The problem is that corpus-based approaches, using written corpora as well as oral ones, do not confirm the existence of construction (1)

Observing 1820 occurrences of *planter* out of a written corpus, Tsuruga finds that "the object N1 realized in the active voice is very rare": only 5 instances of construction (1) within the defined semantic field (Tsuruga 2005: 213-234). In the oral corpus, I only found one for this semantic field, but I found 12 instances with non-vegetable lexical items, such as:

Planter un clou quelque part (to knock a nail somewhere)

and many occurrences of a familiar expression *se planter*, meaning "to fail" (more than 30 % of total utterances):

Il faut pas qu'on se plante là-dessus, il faut pas qu'on dise n'importe quoi

(We must not fail here, we must say anything wrong)

We find expressions that could be considered as a sort of "passive voice" of construction (1):

le domaine […] est planté de cépages

les deux villes les plus euh plantées en – en vignoble

Une tranchée plantée d'arbres, (trench planted with trees)

un terrain planté de (en) vignobles. (field planted with vineyards)

but they are not really passives, as long as they never have active counterparts. Past participles *planté, plantées* function here in the way a modifying adjective would do. On the other hand, (2) is well attested (Tsuruga found 1401 examples in the written corpus), but without any relation with construction (1).

Conclusion: according to corpus-based data, the alleged "cross construction", a sub-type of transformational relations, simply does not exist. Tsuruga insists on what makes the difference between intuitive and corpus-based methods:. "The operation based on intuition consists in making in mind simple and typical examples and in asking oneself if they are acceptable or not […] it seems impossible in this operation to imagine more than typical and extremely simplified examples which are more or less idealized" (p.226-7). He notes that all occurrences do not have the same weight and that no system can be define as completely monolithic and coherent. So the "cross construction" is very rare and almost non-existent" (p. 227).

5. Methodological Issues

5.1. Double-layered grammar

When describing spoken French grammar from the point of view of the speakers, we cannot put all grammatical phenomena on the same level, nor treat them as driven by one and only one type of linguistic competence. Mixing up free-occurring devices with lexically restricted ones, or socially free expressions with professionally conditioned ones would not be a sound method.. I proposed to distinguish two grammatical levels (Blanche-Benveniste (1990)). One could be called "grammaire première" (primary grammar), the one everyone knows at the age of six or seven, before going to school, without explicit teaching. The other one could be called "grammaire seconde" (secondary grammar); it is the one we are learning all our life long, by going to school, looking at old states of our language, reading books, getting professional habits, having particular cultural behaviours, etc. French language has a very strong division between both types of competence. Three examples can be briefly given here: nominal *en*, nominal *dont* and post-posed subject of the *dit-il* type.

5.1.1. Nominal en

Pronoun *en* has many different meanings and constructions. Some, like quantitative *en* , freely used by everyone, even by young children, can be found easily in the corpus:

> J'en ai mis tout un tas devant la porte d'entrée
> (I put a whole pile in front of the entrance door)
> des témoignages, j'en ai des tonnes
> (testimonies, I have by tons)
> j'en ai des piles entières (I have got huge heaps of them)

Grammar books present another nominal *en* construction, in which *en* is a kind of possessive pronoun referring to the relation with a non-human entity. Using a possessive determiner would be un-grammatical here:

> Ce monument, le souvenir en a disparu, *son souvenir a disparu –
> (the remembrance of it disappeared, *its remembrance disappeared)

Ce livre, j'en ai oublié le titre , *j'ai oublié son titre
(this book, I forgot the title of it, *I forgot its title)

We almost never find this construction in ordinary spoken French, except for one lexical item, *en avoir, en garder le souveni*r (to have or to keep remembrance of it):

Les conjugaisons, moi j'en ai pas spécialement des souvenirs merveilleux
J'en ai un très bon souvenir
J'en ai gardé un bon souvenir.

Few people use this construction with other lexical items, except in public or media speech:

J'en ai quelques détails aussi
J'en ai toujours eu une mauvaise opinion
On en a des relevés et des commentaires

But everyone is exposed to the *en* construction when reading or when listening to some radios and TV. I would say that nominal en is not part of a prime grammar and that it appears only in second grammar usage.

5.1.2. Relative pronouns dont and lequel, dit-il type

As we saw earlier, relative pronouns *dont* and *lequel* have few occurrences in what I call "primary grammar". *Dont* appears frequently only with verb *parler*,

la chose *dont* on parle,

with some five or six other verbs and in semi-frozen expression

la façon *dont* on le fait

Lequel mainly occurs in two expressions,

le milieu dans lequel on vit
les raisons pour lesquelles on fait cela

But in public speech or in written texts, which trigger the use of secondary grammar, *dont* and *lequel* do occur freely, without such lexical restrictions. The same could be said for post-posed subjects of the *dit-il* type. The difference does not allow us to oppose a grammar of written French to a grammar of spoken French. Spoken French involves so much inner variation to be treated as an homogeneous class of expression. Primary and secondary grammar gives a more convenient opposition.

5.2. Some short-comings in handling statistical tools

As is well-known for written as well as for spoken language, very frequent items tend to split into several units having different syntactic and semantic properties. For instance, the most frequent verb in French, *être* (to be), can be a full status lexical verb expressing existence, but it very often functions as an auxiliary, *être venu* (to have come), or as part of a cleft sentence: *c'est l'hiver qu'il fait froid* (it is in winter that it is cold). Many apparent verbal occurrences do not function as syntactical verbs at all:

Auxiliaries and modals

Elle va être fâchée, elle peut avoir raté son train
(she goes to be angry, she may have issed the train)

or discourse markers

N'est-ce pas, je veux dire, voyons, tu vois, allez, tiens
(isn't it, I mean, let us see, you see, go, hold)

It is still worse with nouns (Flaux et Van de Velde (2000)). Many items only have a morphological appearance of nouns but they are not syntactical nouns, such as quantifiers and hedges:

Un tas d'amis, une espèce de sable, un putain de métier, elle a l'air heureuse
(a heap of friends, a kind of sand, un whore of work, she has a happy look)

Many frequent adjectives mainly act as discourse markers

bon, vrai, sûr, possible
(good, true, sure, possible)

Most frequent verbs, nouns and adjectives tend to grammaticalize in many other languages, as shown by Italian in C-Oral-Rom edition for

bello, grande, certo, vero

That is why statistical data must be handled with care (C-Oral-Rom 1, 159-61). Without a thoroughly established categorization, they cannot give sound results.

5.3. Dysfluencies ?

"Disfluencies" is the term used for interruptions, retractions, repairs, etc. According to Cresti and Moneglia's measures, dysfluencies would occur in 30% of the total utterances. Here is an example of an article being uttered three times in Italian:

non vedevo bene la la la strada.
(I did not see the the the road)

Such phenomena are recognized some utility for pragmatic means (interactions, self-control, etc.), but they are generally discarded from syntactic and lexical analysis as pure rubbish.

Recent studies (Fox and Jasperson (1995)) have however given more grammatical and semantic importance to dysfluencies: lexical strategies in approximates and hedgings; interesting insights into meaning "on the making"; syntactic planning coming before lexical planning (Blanche-Benveniste 2003). Here are some remarks about syntax and lexical heads.

Nominal determiners are very often repeated ahead of the nominal phrase (*ce, ce*), as well as prepositions in front of prepositional phrases, *dans la, dans la:*

Ils ont installé ce, ce système

Des places de parking creusées dans la, dans la falaise

A similar scheme occurs with verbal phrases : repetitions are very frequent on weak clitic subjects, coming at the first place in verbal phrases, *je, je,* or on clitic subjects and auxiliaries*, ils ont, ils ont*:

Je, je, je vais le faire bientôt

Ils ont, ils ont réduit un maximum

In both cases, determiners and subjects signal the nature of the phrase-to-come, without any lexical inside. I suggest an explanation: speakers would give first the syntactic frame, with no lexical fillers, and they would only give the whole phrase, syntax and lexicon together, in a second time. We can figure it with unfinished graphic lines corresponding to lexical vacuum:

```
Ils ont installé ce ……..
                 ce système

Des places de parking creusées dans la……
                               dans la falaise

Je………………….
Je… ………………
je vais le faire bientôt

Ils ont…………………
  ils ont réduit un maximum
```

Another type of repetition is the enrichment one. In a first stage, the speaker utters a nominal phrase with a lexical head, and in a second stage, the same lexical head, enriched with approximates, quantitfiers or other modifiers: *une guerre, un genre de guerre civile*:

> Il lui mettait une couronne, voilà, un genre de couronne
> Il y a une guerre, un genre de guerre civile
> Pas de récompense, pas l'ombre d'une récompense

The same occurs with verbal phrases, a lexical verb item being given first, *est commune*, and then enriched with a modal verb, *devrait être commune*:

> Celle qui est commune à chacun, qui devrait être commune à chacun
> Cette complicité qu'on a, enfin qu'on devrait avoir entre frère et sœur
> Qui gardait son secret, qui avait su, qui avait su garder ce secret

Speakers never act in a reverse way, giving for instance first the enriched version and then the poorer one. In such apparent "dysfluencies", speakers give precious insights on semantic and syntactic developments. That is why getting rid of such phenomena is a linguistic mutilation.

6. Conclusion

Because of the importance of grammar and lexicon inter-relations, intuition is helpless for many domains in spoken language. I would like to quote a conclusion by Tsuruga:

> "Intuitionally the construction with *avec* is well possible, but in fact we found no occurrence […] Everyone is in agreement, since Saussure, in that the "langue" can only be observed through the "parole", and we must add that the "parole" cannot be sufficiently represented by intuitionally forged simple examples […] We are now in an era when we have various means of examining an impressioningly large-scale corpus" (Y. Tsuruga 2005: 227-8).

Though spoken language corpora are not as large-scaled as we wish, they nevertheless allow us to describe some important aspects of "parole" and to draw a picture of what could be the relations between "parole" and "langue". Grammatical schemes that could figure out the "langue level" do not extend steadily upon lexical items; they shift according to different degrees in grammaticalization and according to speakers' behaviours. By studying spoken language corpora, we get convinced that grammatical competence is moving and flexible and that it cannot be viewed as a monolithic capacity.

Bibliography

ANDERSEN, Hanne-Leth et NØLKE, Henning, (éds.), *Macro-syntaxe et macro-sémantique, Actes du colloque international d'Aarhus, 17-19 mai 2001*. Berne : Peter Lang (Sciences pour la communication).

ASHBY, W.J., 1988, “The Syntax, Pragmatics and Socio-linguistics of left-and right-dislocation in French”, *Lingua* 75, 203-239.

BEGUELIN, Marie-José, 2002, “Routines macro-syntaxiques et grammaticalisation : l’évolution des clauses en *n’importe*”, in H.L.ANDERSEN et H. NØLKE, 43-70.

BERRENDONNER, Alain, 1990, “Pour une macro-syntaxe”, *Travaux de Linguistique* n° 112, 31-49.

BERRENDONNER, Alain, 2002, “Morpho-syntaxe, pragma-syntaxe, et ambivalences sémantiques”, in H-L ANDERSEN et H.NØLKE, pp.23-42.

BIBER, D., JOHANSSON, LEECH, CONRAD & FINEGAN, 1999, *Longman Grammar of Spoken and Written English*. London: Longman.

BILGER, Mireille, 1998, “Le statut micro et macrosyntaxique de ET”, in M. BILGER, F.GADET, et K. van den EYNDE (éds.), *Analyse linguistique et approches de l’oral. Recueil d’étude offerts en hommage à Claire Blanche-Benveniste*. Louvain/Paris : Peeters, 91-102.

BLANCHE-BENVENISTE, Claire, 1990, “Grammaire première et grammaire seconde : l’exemple de EN”, *Recherches Sur le Français Parlé* 10, 51-73.

BLANCHE-BENVENISTE, Claire, 1999, *Approches de la langue parlée en français.* Paris : Ophrys.

BLANCHE-BENVENISTE, Claire, 2002 “Macro-syntaxe et micro-syntaxe : les dispositifs de la rection verbale”, in A.L. ANDERSEN et H. NØLKE, 95-118.

BLANCHE-BENVENISTE, Claire, 2003, “Phrase et construction verbale”, in M. Charolles, Le Goffic et M.A. Morel (éds.), *Y a-t-il une syntaxe au-delà de la phrase ?* Colloque de Paris-3.

BLANCHE-BENVENISTE, Claire, 2005, “L’étude grammaticale des corpus de langue parlée en français”, in G. WILLIAMS (éd.), *La Linguistique de corpus.* Rennes : Presses Universitaires de Rennes, 47-67.

BLANCHE-BENVENISTE, Claire et JEANJEAN, Colette, 1987, *Le français parlé : édition et transcription.* Paris : Didier-Erudition.

BLANCHE-BENVENISTE, Claire, BILGER, Mireille, ROUGET, Christine et Van den EYNDE, Karel, 1990, *Le Français parlé* : *études grammaticales*. Paris : éditions du CNRS.

BLASCO-DULBECCO, Mylène, 1999, *Les constructions disloquées en français contemporain*. Paris : Champion (Collection “les français parlés : textes et études”).

BYBEE, Joan and HOPPER, Paul, 2001, *Frequency and the Emergence of Linguistic Structure*. Amsterdam: Benjamins.

CAPPEAU, Paul, et SEIJIDO Magali, 2005, Les corpus oraux en français.

Inventaire. Document pour la Délégation Générale à la Langue Française et aux Langues de France.

CRESTI, Emanuela (a cura di), 2000, *Corpus di Italiano parlato,* volumes I e II. Firenze : Accademia della Crusca.

CRESTI, Emanuela e MONEGLIA, Massimo, (eds.), 2005, *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: Benjamins.

DEULOFEU, José, 1986, “Syntaxe de *QUE* en français parlé et le problème de la subordination”, *Recherches Sur le Français Parlé* n° 8, 79-104.

DEULOFEU, José, 1999, “Questions de méthode dans l'étude du morphème *que* en français contemporain”, *Recherches Sur le Français Parlé* n° 15 pp. 149-158.

DINCXU SHI, (2000) “Topic And Topic-Comment Constructions In Mandarin Chinese”, *Language*. Volume 76. Number 2.

DUVALLON, Outi,, 2003, *Les chaînes anaphoriques : étude d’un corpus de finnois parlé*. Thèse présentée à l’Ecole Pratique des Hautes Etudes, Paris.

FLAUX, Nelly et VAN DE VELDE, 2000, *Les noms en français: esquisse de classement.* Paris : Ophrys.

FOX, Barbara and JASPERSON, R, 1995, “A Syntactic exploration of repair in English conversation”, in DAVIES, Ph.D. (ed.), *Alternative Linguistics. Descriptive and Theoretical Modes*. Amsterdam: Benjamins.

FRANCARD, Michel, 1990, “Le français parlé des corpus oraux”, *Travaux de Linguistique* 21, 53-63.

FRANCARD, Michel et GERON, Geneviève, 2002, “La banque de données VALIBEL : des ressources textuelles orales pour l’étude du français en Wallonie et à Bruxelles”, in PUSCH, C.D. & RAIBLE, W. (eds.), *Romance Corpus Linguistics*. Tübingen : Gunter Narr, 71-80.

GREVISSE, Maurice 1987, *Le Bon Usage. Grammaire française.* Treizième édition, revue et refondue par André GOOSSE.. Louvain : Duculot.

HABERT, Benoît, NAZAREBCO, Adeline et SALEM, André, 1997, *Les linguistiques de corpus.* Paris : Colin.

HALLIDAY, M.A.K., 1985, *Spoken and Written Language.* Oxford: Oxford University Press.

KERBRAT-ORRECCHIONI, C., 1992, *Les Interactions verbales,* I, II. Paris: Colin.

LAMBRECHT, Knud, 1994, *Information Structure and sentence form: Topic, focus and the mental representation of discourse referents*. Cambridge: Cambridge University Press.

LAMBRECHT, Knud, 2001, “Framework for the analysis of cleft constructions”, *Linguistics* 393, 463-516.

LECLERE, 2005, “The lexicon-Grammar of French Verbs”, in Takagaki, Zaima, Tsuruga, Moreno-Fernandez an Kawaguchi (eds.), *Corpus-Based Approaches to Sentence Structures*. Amsterdam: Benjamins 29-45.

LE GOFFIC, P., 1994, *Grammaire de la phrase française*. Paris : Larousse.

LINDQVIST, Cristina, 2001, *Corpus transcrits de quelques journaux télévisé français*. Uppsala : Uppsala Universiteit.

MARTIN, Philippe, 1999, “L’intonation en parole spontanée”, *Revue Française de Linguistique Appliquée*, vol IV-2, *L’Oral spontané*, dirigé par M. Bilger, 57-76.

MERTENS, Piet, 1986, *Unités intonatives et structure de l’énoncé.* Leuven : KUL Preprint Nr 104..

MORENO, Antonio, G. de la MADRID, M. ALCANTA, A. GONZALEZ, J.M. GUIRAO & R. de la TORRE, 2005, “The Spanish Corpus”, in CRESTI, Emanuela e MONEGLIA, Massimo, (eds.), 2005, *C-ORAL-ROM, Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: Benjamins, 135-161.

MOREL, Mary-Annick, 2002, “Intonation et gestion du sens dans le dialogue oral en français”, in H.L. ANDERSEN et H. NOLKE, 119-139.

MOREL, Mary-Annick et L. DANON-BOILEAU, 1998, *Grammaire de l’intonation. L’exemple du français.* Paris : Ophrys.

ONO, T. & S. A. THOMPSON. 1995 “What can conversation tell us about syntax? ”, *Alternative linguistics*, ed. by Philip W. Davis. Amsterdam & Philadelphia: Benjamins.

POPLACK, Shana, 2001, 405-428, “Variability, Frequency and productivity in the irrealis domain of French”, in Joan BYBEE and Paul HOPPER, *Frequency and the emergence of Linguistic Structure.* Amsterdam: Benjamins.

SABIO, Frédéric, 1995, “Micro-syntaxe et macro-syntaxe : l’exemple des compléments antéposés”, *Recherches Sur le Français Parlé* n° 13, 111-156.

SABIO, Frédéric, 2002, “L’opposition de modalité en français parlé : étude macro-syntaxique”, *Recherches Sur le Français Parlé* n°17, 55-78.

SCHMALE-BUTON, Elisabeth, 1984, *Französisch 1. Conversations téléphoniques*. Bielefeld: Universität Bielefeld.

SINCLAIR, John, 1991, *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

TOGNINI-BONENELLI, Elena, 2004, *Corpus Linguistics at work*. Amsterdam: J. Benjamins (Studies in Corpus Linguistics).

TOMASELLO, Michael, 2003, “Some surprises for Psychologists” in TOMASELLO (ed.), *The new Psychology of Language*. New York: Elbaum, 1-13.

TSURUGA, Yoichiro, 2005, “A correspondence between N0-V-N1-de-N2 and N0-V-N2-Loc-N1 in French. The case of *Planter*”, in Takagaki, Zaima, Tsuruga, Moreno-Fernandez an Kawaguchi (eds.), *Corpus-Based Approaches to Sentence Structures*. Amsterdam: Benjamins, 213-232.

YATSIV-MALIBERT, I., 2004, *Etude syntaxique sur des enregistrements d’hébreu moderne parlé.* Thèse présentée à l’Ecole Pratique des Hautes Etudes : Paris.

WILLEMS, Dominique, 1991, “Les Francophones de Gand : éléments d’une analyse linguistique et sociolinguistique”, in *Variétés et variantes du français des villes : état de l’est de la France. Actes du Colloque Scientifique International du Centre de Recherches d’Etudes Rhénanes*, Paris/Genève : Champion/ Slatkine, 185-198.

WILLEMS, Dominique, 1998, “Données et théories en linguistique: réflexion sur une relation tumultueuse et changeante”, in BILGER, GADET et Van den EYNDE (éds.), *Analyse linguistique et approches de l’oral.* Louvain / Paris : Peeters, 79-87.

Challenges for English Corpus Linguistics in Second Language Acquisition Research[1]

Susan CONRAD

1. Corpus Linguistics, Second Language Teaching, and Language Acquisition Research

In the past decade, English as a Second Language (ESL) educators and researchers who work with corpus linguistics have begun to make substantial contributions to language teaching and learning. This relatively new area within applied linguistics uses computer-assisted techniques to analyze large, principled collections of written or transcribed spoken texts. The purpose is to describe how people use language in natural settings – for instance, describing the different distributions of grammatical structures and lexical items across registers (such as casual conversation, newspaper writing or academic prose), describing the associations between grammatical structures and particular lexical items, comparing language use by learners of different first language backgrounds, and a variety of other aspects of language use. With the help of computers, more language, more participants, and more interacting variables can be analyzed than is feasible when analyses are conducted by hand. (For more introduction to corpus linguistics, see, e.g., Biber, Conrad & Reppen 1998; Conrad 2005; Hunston 2002; Meyer 2002.)

Some of the greatest contributions of corpus linguistics so far have concerned ESL teaching and learning. Some work has focused on descriptions of native speakers' use of English, for example in the corpus-based *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad & Finegan 1999), which describes grammatical patterns in conversation, newspaper writing, fiction writing, and academic writing. Such descriptions are helpful to ESL/EFL teachers because they allow new factors to be considered in decisions about materials development and syllabus design: the frequency of use of features and the way that people actual use features (rather than relying on intuition or prescriptive rules).

[1] Parts of this paper were presented at the International Symposium on Corpus Linguistics – Perspectives for the Future, Internationales Wissenschaftsforum, Ruprecht Karls Universität, Heidelberg, Germany, in October 2004. I am grateful to that audience as well as the audience of the Second International Conference on Linguistic Informatics for their discussion and enthusiasm.

Another contribution is commercially available corpus-informed textbooks. For example, the *Touchstone* coursebook series (e.g. McCarthy, McCarten & Sandiford 2005) incorporates aspects of conversational language use that can be very important for face-to-face interactions but have been neglected in most traditional textbooks. The textbook *Focus on Vocabulary: Mastering the Academic Word List* (Schmitt & Schmitt 2005) focuses vocabulary practice on words found – through corpus study (Coxhead 2000) – to be frequent across a wide range of academic subjects. Teachers have also experienced success having students investigate corpora for themselves (see, e.g., Seidlhofer 2000; Yoon & Hirvela 2004; and the collections edited by Aston 2001 and Burnard & McEnery 2000).

Corpus linguistics is also beginning to have an impact on the study of language acquisition. Learner corpora – i.e., large, electronic collections of written and transcribed spoken texts produced by language learners – have been developed for a number of contexts in the past decade. The International Corpus of Learner English (ICLE) is perhaps the best-known learner corpus and is widely available for research (see Granger 1998a, 2003). It contains essays written by advanced-level EFL students of 14 different nationalities. Other large learner corpora have been compiled by faculty in certain institutions for their use; the Hong Kong University of Science and Technology Corpus of Learner English, for example, contains approximately 25 million words of language produced by students at that university (see studies in Flowerdew and Tong 1994). A number of smaller and more specialized corpora have also been compiled in recent years. Cheng and Warren (2000) discuss a corpus development project that includes spoken interactions of advanced non-native speakers and native speakers in Hong Kong.

These learner corpora are now being used for a variety of studies (see, e.g., the collections edited by Granger 1998b; and Granger, Hung & Petch-Tyson 2002; as well as Abe 2003; DeCock 2000; Hyland & Milton 1997; Nesselhauf 2004; Tono 2000; and a variety of other articles). Research covers a variety of topics, from morpheme acquisition orders to interlanguage contrastive analysis of vocabulary choices, and many others. Nevertheless, corpus-based studies are not well known in the field of second language acquisition (SLA), at least in the United States. When I looked over the last several issues of the best known SLA journals, *Studies in Second Language Acquisition* and *Language Learning*, I found that none of the last 41 research articles had used a corpus approach. When I checked the *International Journal of Corpus Linguistics*, only one of the last 53 articles were focused on second language acquisition. At least in the most visible circles, corpus linguistics is not having much of an impact on the field of SLA.

2. The Need for New Corpora for SLA Studies

There seems no question that well designed corpus-based studies could contribute valuable perspectives to SLA research. For one, as mentioned above, corpus studies can include data on more participants and the interaction of more variables than is usually feasible with other analysis techniques. The same participants can also be viewed from multiple perspectives – for instance, looking at the order of morpheme acquisition and studying sequences of negotiation with the same learners. Practically speaking, publicly available corpora also can save researchers a great deal of time; for example, it is far faster to use ICLE than to collect data from 14 different countries on your own.

The lack of impact of corpus work in SLA, then, does not stem from a lack of potential, but from several other reasons. One is that, compared with much work in SLA, corpus work is still very new, and techniques are being perfected. Early corpus studies sometimes did not consider all the variables that are important when examining learners' interlanguages. For example, some early studies using ICLE claimed differences in interlanguage based on first language background (e.g. Altenberg 1997). With further study, however, Ädel (2005) has found that the time variable (amount of time for writing the essay) has more effect than first language background. This is not a surprising finding given previous research, and the fact that some research did not take it into account can make some SLA researchers question corpus techniques generally. However, the corpus description *does* include information about the time variable so that research can include it. This experience is a good reminder that the organization of a corpus (in this case by first language group) should not blind researchers to other variables that have not been controlled in the corpus and need to be considered.

There are other issues within corpus design that are even more important if corpus linguistics is to have a large impact on SLA research, especially in the United States with its large ESL programs. Some of the areas that are most important for research and for policy decisions are not included in current corpora. Prime among these areas are the following:

1) Adult immigrant learners. Most corpora have been compiled with academically-oriented ESL programs, but the area of most concern for policy makers is adult immigrant students. Ninety percent of immigrants to the United States come from non-English speaking countries (Center for Applied Linguistics 2003), with the

result in local community colleges that ESL classes often have long waiting lists. For adults, attaining language proficiency to fulfill basic needs and obtain jobs are concerns of both the immigrants themselves and society at large, with considerable resources spent on the classes for these students.

2) Spoken language of lower-proficiency learners. Most learner corpora contain written tasks because these are far easier to collect. However, many new ESL immigrant students are not literate or are semi-literate, and would thus be excluded from written corpora. When spoken language has been collected (e.g. Cheng and Warren 2000), it is most often with academically-oriented intermediate or advanced-level students. Policy makers as well as teachers, however, are concerned with the spoken language skills that enable immigrant learners to interact in a larger society from the lowest proficiency levels.
3) Classroom language. Although naturalistic SLA has been the focus of many studies, instructed SLA is an important area of research and, thus far, corpora have not focused on classroom language. In fact, some have argued (S. Granger, personal communication) that it is impossible to compile a principled corpus of classroom language because there are so many variables in task type and the structure and vocabulary provided by the teacher.
4) Non-verbal cues for disambiguating meaning. Especially with lower-level learners, understanding interactions and meanings can be difficult in a corpus which is not connected to the physical gestures learners use. Existing corpora take learner language out of its visual context, and learners' meaning can be difficult to ascertain.

Many researchers would simply end the discussion here, arguing that compiling a corpus of spoken language by relatively low-level immigrant adults, including classroom tasks, and in particular having it all linked to the visuals of the situation, is simply not possible – at least with current technology. On the contrary, I believe we simply have not been thinking creatively enough about corpus compilation. In fact, we have the potential to make such a corpus – working off of a larger collection of videotaped recording from the National Adult ESOL Labsite, described in the next section.

3. The National Adult ESOL Labsite

The National Adult ESOL Labsite Project at Portland State University is part of the National Center for the Study of Adult Learning and Literacy (NCSALL). NCSALL, a partnership between Harvard University, Portland State University, Rutgers University, the University of Tennessee and World Education, is funded by the Educational Research and Centers Program, U.S. Department of Education (Stephen Reder, Principal Investigator; Award Number R309B60002; see further Reder, Harris & Setzler 2003; http://www.labschool.pdx.edu; and *Focus on Basics* volume 8 at http://www.ncsall.net/fileadmin/resources/fob/2005/fob_8a.pdf). The ESOL Labsite, or "Lab School," is designed to facilitate high-quality classroom-based research and professional development in adult ESL. The Lab School is run jointly with the local community college that offers ESL classes to adults in the area, Portland Community College. Four levels of community college ESL classes are being videotaped over a five-year period. Each term, two classes at each of two levels are designated as Lab School classes. Two community college teachers teach the four classes, which typically have 15 to 30 students each.

The four classes designated as Lab School classes are videotaped each day with multiple cameras and microphones so that there is a complete audio and visual record of the classes. Students take turns wearing lapel microphones so that pair and group work (which is undecipherable from the ceiling microphones) is captured. There are multiple recordings for each student during each term that they attend classes. As of December 2005, there were about 4,000 hours of video covering 750 students. The vast database which results is stored digitally and accessed through software designed specially for the project, named "ClassAction."

The students in the classes are typically adults who are immigrants to the United States. They come from 69 different countries and 39 different first languages, with Chinese and Spanish the most common. About one-third of the students in the lowest level are not literate or only semi-literate in their first language when they begin classes. The four levels of classes (designated A – D) cover roughly beginning and intermediate levels of proficiency. Figure 1 displays examples of the level of language in the second level of the program (Level B). Demographic information about students, including their first language, amount of previous schooling, age and other details, is stored in a database.

xxx = word that transcribers were unable to understand
+ = short pause
(2) = pause in seconds

1. Pairs answering the questions “What do you do every morning/evening/day/month?”

Student1: I smoking every (+) evening.
Student2: oh. (+) you smoke?
Student1: ((writes))
Student1: huh? (+) I no smoking just lying to you.
Student2: every evening. (4) every month.
Student2: I everyday drink coffee. drink coffee.
Student1: everyday. ((writes)) everyday you drink ((writes)) and maybe eat?
Student1: oh yeah. ((writes)) coffee. (2) xxx (4) every morning what you do.
Student2: I brush my teeth.

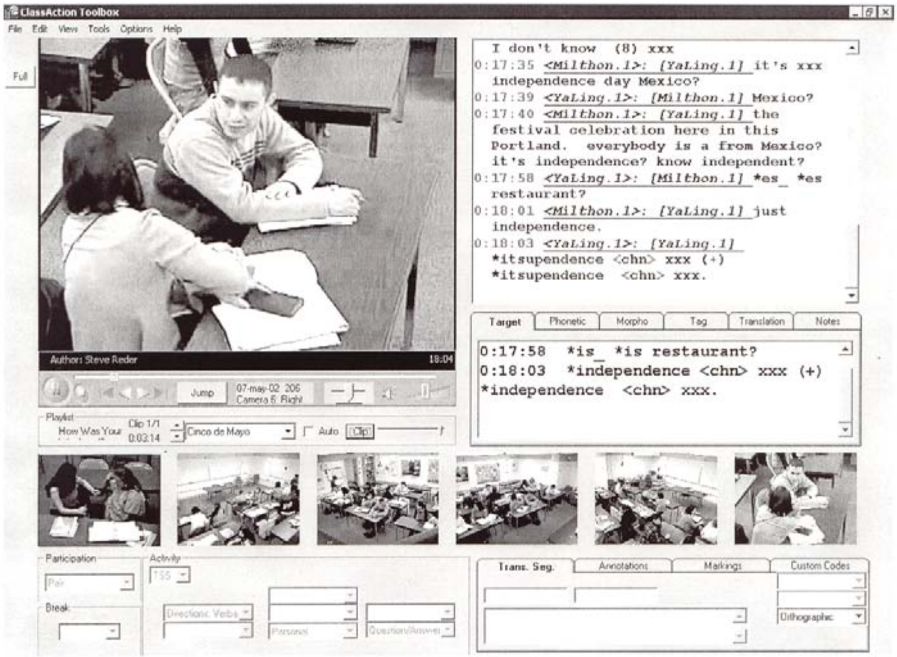
2. Partner has just asked “How was your weekend?”

very good. It was fun to em festival here in Portland and working xxx you know that festival xxx Mexican xxx five. five de mayo. may. you know may?

Figure 1. Examples of Lab School Learner Proficiency Level (Level B)

The ClassAction software developed for the Lab School includes the Toolbox program that allows viewers to watch the videotape along with any corresponding transcription (see Figure 2). All six camera shots are available at the bottom of the screen, and one shot is enlarged for main viewing. Only a small percentage of the thousands of hours of videotape have been transcribed, but for each class, about 20 minutes of pair work per class have been transcribed. (Pair work has been the focus of research in the lab school thus far, so pair work has been the focus of transcription.) When an interaction that has been transcribed is viewed, the transcription appears to the right of the video, as in Figure 2.[2]

[2] An additional part of the ClassAction software – the Query program – allows users to search for learners or class segments based on certain criteria. A user could specify, for instance, Level B pair work that took place between learners whose first language was Japanese. The Query program returns all the video clips that meet the criteria. The video clips can then be viewed with the Toolbox program.



```
  I don't know  (8) xxx
0:17:35 <Milthon.1>: [YaLing.1] it's xxx
  independence day Mexico?
0:17:39 <YaLing.1>: [Milthon.1] Mexico?
0:17:40 <Milthon.1>: [YaLing.1] the
  festival celebration here in this
  Portland.  everybody is a from Mexico?
  it's independence? know independent?
0:17:58 <YaLing.1>: [Milthon.1] *es_ *es
  restaurant?
0:18:01 <Milthon.1>: [YaLing.1] just
  independence.
0:18:03 <YaLing.1>: [YaLing.1]
  *itsupendence <chn> xxx (+)
  *itsupendence  <chn> xxx.
```

```
0:17:58  *is_ *is restaurant?
0:18:03  *independence <chn> xxx (+)
*independence  <chn> xxx.
```

Figure 2. ClassAction Software “Toolbox”

An additional aspect of the Lab School project is the associated “Labsite Student Study” in which cohort groups of students are individually interviewed in their homes each year, with new groups added each year for the first three years of the study. Currently, approximately 200 participants from seven different first language backgrounds are in the study. Participants continue in the Labsite Student Study even if they no longer are enrolled in classes, so longitudinal data are collected As part of the home interview, students are asked an open-ended question. For example, the first year students were asked, “Why did you leave your home country and come to the United States?”

As designed, the Lab School project facilitates several kinds of research in second language acquisition and teaching. Since the project’s inception four years ago, analyses have begun to contribute especially to our understanding of classroom interactions, addressing topics such as the dynamics of pair interactions in the classroom (e.g. Garland 2002; Hellermann in press-b), the negotiation of meaning that students engage in with different kinds of pair tasks (Harris in preparation), and students’ construction of classroom practices around literacy events (Hellermann in press-a). The data have also been used for teacher development resources

(Kurzet 2002, 2004) and reflections on the process of classroom research (Banke & Brillanceau in preparation; Reder and Brillanceau in preparation).

However, to date, studies with the lab school data have been intensive analyses of a relatively small number of learners or data. For example, Hellermann (in press-a & b) uses microethnographic methodology with a small number of participants (two and seven), Harris (in preparation) focuses on four pair interactions. Two morpheme acquisition studies (Disbrow-Chen 2004; Ouellette 2004) and a study of learner autonomy (Brillanceau 2005) have used a single case study approach. While these studies are extremely useful for addressing certain research issues, they do not address other important issues that require larger-scale analyses. In fact, the very strength of the Lab School database for an intensive microethnographic approach precludes studies that use a very large collection of data: the video. That is, the video is extremely valuable for detailed analyses of particular interactions but it is simply impossible to analyze a large amount of language without being able to look at transcriptions separately from the videotape. Furthermore, the large, complex nature of a multimedia database with specialized software presents substantial challenges to researchers in other places: learning to use the software and then troubleshooting problems without technical support staff at the same location, having effective internet connections for using the software and streaming video over the web, gaining access to databases in order to know what language samples are available in order to design a study (e.g. selecting participants for the best possible representation of demographic characteristics, such as first language, literacy level, and age). Furthermore, any research into language development requires analysis of transcriptions – to have a record of what students actually said – and there is currently no way for off-site researchers to obtain transcriptions without also watching the complete video, which is time-consuming.

The Lab School provides the perfect opportunity for creating a corpus that includes the characteristics absent from current corpora: relatively low-level speech of adult immigrant students, including classroom interactions. Since the video exists, corpus files could be indexed to the video, and when needed, researchers could view video to disambiguate meaning. Such a corpus could benefit SLA researchers and make the Lab School data more accessible for the research community. The corpus could be used for looking at traditional areas of SLA studies with more participants – e.g. analyzing the development of grammatical systems, analyzing error patterns, and making comparisons for first language groups or different levels of L1 literacy. Classroom-oriented studies would also be possible; for instance, to what extent do learners actually use the language frames and vocabulary provided by teacher and textbook prompts? What types of errors do learners make even when given clear language models? To what extent do students use the same language when they do the same classroom task? Learners' language use and language development could also be compared between the home interviews and the classroom setting.

However, such a corpus also requires facing some new challenges. The Lab School was not set up as a corpus linguistics project for SLA research; it was designed for research using videotape of the classroom with some basic indexing of certain classroom variables that are apparent from watching the video (e.g. pair work, teacher-fronted, etc.). Traditional corpus compilation projects have never dealt with this type of data.

4. Designing a Corpus from the National Adult ESOL Labsite

A corpus from the ESOL Labsite would combine familiar elements of current corpus design with new elements. The corpus would consist of transcribed learner language, compiled to facilitate cross-sectional research with numerous learners at each proficiency level (based on class level) and longitudinal research with individuals who have been recorded for numerous terms. Both a classroom component and a home interview component could be included. Like current corpora, this corpus could be downloaded from a server or distributed on a CD. An improvement over current corpora, however, would be that markers in the corpus would assure that the transcripts remain cross-referenced with the video. Thus, for disambiguation purposes or any other purpose, researchers who wanted to could gain access to the video via the Lab School internet site and see the visuals accompanying the language that they are studying (as well as hear intonation etc.).

Maintaining links to the video is not a difficult task. Codes can easily be put into the beginning of each corpus file so that the corresponding video can be found. However, there are two other aspects to the proposed corpus that are new challenges. The first concerns the importance of having samples that are comparable. That is, the speech that will be compared from different students or from the same student at different times must be produced under comparable circumstances. For the home interview component, this is not difficult because the participants all respond to the same question. However, given the variety of tasks used in classes and variation in the amount of language support provided by the teacher, this criterion is a challenge for the classroom component. A second concern is whether computer-assisted analysis techniques, as used in corpus linguistics, are truly of use for relatively low-level speech of learners: Can annotation systems that have been used with high level learner language or native speaker language be used with beginning level students?

The following sections address these concerns in turn, and then I consider the types of analyses that would be possible with the corpus.

4.1. Identifying Classroom Language Produced under Equivalent Conditions

Since pair work has been transcribed as part of the Lab School project, it makes sense (at least in initial stages) to take advantage of the transcription work that has been completed and make the classroom component of the corpus focus on pair work. For a principled corpus created from these classroom activities, it is crucial to know how much students are repeating language that has been supplied to them by the teacher (or textbook) versus creating language on their own

As part of the Lab School project, the classroom video is viewed and coded with a basic indexing system to reflect not only that an activity is done in pairs, but also the extent to which the activity is based on language provided by the teacher. As the coding system manual explains, "Language is coded with respect to the support or scaffolding that the teacher provides so that students can produce/comprehend the classroom language" (Bolstad, et al. 2002: 13-14).

The language codes result in six combinations that should describe the amount of support that students are given in an activity. These are:

1) TEACHER: ALL (and therefore Student: None)
2) STUDENT: ALL (and therefore Teacher: None)
3) TEACHER: LANGUAGE FRAME – STUDENT: TARGET ITEM
4) TEACHER: QUESTION – STUDENT: TARGET ITEM
5) TEACHER: LANGUAGE FRAME – STUDENT: QUESTION/ANSWER
6) TEACHER: QUESTION/ANSWER – STUDENT: QUESTION/ANSWER

The first two codes are relatively clear-cut because one participant provides virtually all of the language. In TEACHER: ALL the teachers provide language frames and specific vocabulary/expressions to use to complete the frames. For example, teachers told students to talk to their partners using the expression "How often do you…" with specific vocabulary such as *smoke, exercise, sing in the shower*, etc. They also provided specific words for students to use in answers: *always, usually, often,* and *never*. In STUDENT: ALL students usually answer open-ended questions. For example, students asked each other questions about their hometowns, or after reading silently for 20 minutes, were asked to tell each other about their book. The language of the responses is entirely student-generated.

The four other codes describe mixed levels of control. In all of these, teachers provide some language support – a frame to fill in or a question to answer – and the students respond with a target item or an answer from a limited set of choices. Typical activities include filling in a phone dialog, answering questions such as "Do you live in a house or an apartment?" or finding errors in a passage and correcting them.

Theoretically, it would be possible to design a corpus with the categories that are distinguished by the language codes. In a limited investigation to determine the feasibility of this approach in practice (Conrad 2004), however, I found – not surprisingly – that the last four codes were very problematic. Coding the amount of language support is clearly a very difficult task because the level of support varies along a continuum, making distinct categories hard to code consistently. Exactly how much language was provided by the teacher varied considerably even within a single code, so without further, time-intensive coding, it would be impossible to ensure that activities were produced under similar conditions.

The extremes of the continuum are clearest – language that is tightly teacher-controlled and language that is completely student-generated. The classroom corpus could therefore have two sections:

A) Student-controlled Language Tasks. This category would include tasks where students have not been provided with syntax or vocabulary to do a task or answer a question. Since explicit instructions or specific questions to answer are virtually always given in a classroom, the category would allow for one partner to ask a teacher-provided question to another student, as long as no language was provided for the answer. For example, if pairs are told to ask each other "What did you do this weekend?" the task would be included in student-controlled language. Of course, it would be impossible to know if particular structures or vocabulary that students use in their answers had been covered in the class sometime in the past, but there would be no instructions to students about the language to use. (A researcher could, however, look at the video from previous days to see if certain language the students used was practiced in class since this in itself is an interesting research question.)

B) Teacher-controlled Language Tasks. This category would include tasks where students are given language frames and items or expressions to use. They may be choosing between options for completing language frames, but they would not be creating their own language. Although some researchers may find this category irrelevant because it is not a fair representation of learners' own interlanguage, it provides useful data for anyone interested in language instruction. The extent to which learners actually use the language that they are instructed to use in an activity deserves more empirical investigation (and there is some evidence that learners speak differently when the teacher is nearby so teachers

> may not get a complete picture by listening in as students work, see Garland 2002). Being able to compare learners' language use across the very same teacher-controlled task would be useful for classroom-oriented research.

The Lab School corpus would indeed be a different type of learner corpus because it would not simply be trying to capture learners' interlanguage. Their interlanguage would be represented in home interviews, and also in the student-controlled language tasks. However, a far different set of research questions – related to teaching and to classroom second language acquisition – could also be addressed by such a corpus because of the classroom component.

4.2. Grammatically Tagging a Low-level Learner Corpus

A corpus developed from the Lab School multimedia data would be useful only if computer-assisted techniques used in corpus-based analyses can be used with the relatively low-level learner language, with all its ill-formedness and errors. In previous work with higher level second language learners and native speakers of English, corpus linguists generally have used a combination of automatic and interactive coding techniques (see e.g. Biber et al. 1999; Biber, Conrad & Reppen 1998; Granger 1998b; Granger, Hung, & Petch-Tyson 2002). Especially with spoken language and/or complex, potentially ambiguous language structures, interactive analysis is required because automatic coding is too inaccurate. In interactive analyses, human coders edit codes assigned by programs or assign codes themselves. Coding can include identifying and categorizing learner errors as well as grammatical structures or referents.

One of the most challenging sorts of annotation is "tagging" – that is, coding in information about the grammatical class of each word in the corpus. A tagged corpus is extremely helpful for research since it is then possible to search for structures (such as question words, pronouns or relative clauses) rather than search by word patterns. However, tagging by looking at each word individually is unacceptably time-consuming for a large corpus, so a relatively high level of accuracy for automatic tagging is needed. The level of language in the beginning level Lab School classes is below the level that is typically tagged.

In order to discover the feasibility of tagging the low-level language, I tagged ten sample transcripts from the second level, using a grammatical tagger developed by Biber (see description in Biber Conrad & Reppen 1998; and Biber 1988: Appendix II). In general, the tagger had only a little more difficulty with this learner language than with native speaker speech, and some of the problems could be fixed with adjustments to the tagging program to better fit the Lab School transcription conventions.

The majority of utterances by the learners were not necessarily grammatically well-formed in their clause structure, but they contained clear enough phrases – or strings of words – for the tagger to analyze. Example 1 in Figure 3 illustrates this type of situation. (The tagged texts are written out vertically, with the grammatical codes next to each word. For example, the code *^wrb+who+whq++* next to the first word "where" in Example 1 shows that it is a wh-word used in a wh-question.) In Example 1, the learner was confused about the question she is supposed to ask her partner – "Where did you last speak English?" She produces the following utterances:

> *Where is about eh eh speak English?*
>
> [Then appealing to the teacher for help:] *Teacher excuse me one moment oh the question is about uh speak English? or or the*

Both utterances are tagged accurately, despite ungrammatical structures, incomplete structures, and complete structures which do not have end punctuation (e.g. *excuse me one moment*).

Example 2 illustrates a more difficult sample, partly because there are several unclear words. A student is answering the question "How was your weekend?" His answer is:

> *very good. It was fun to em festival here in Portland and working xxx you know that festival xxx Mexican xxx five. five de mayo. may. you know may?*

Example 1

```
{speaker1
where ^wrb+who+whq++
is ^vbz+bez+aux++
about ^in++++
eh ^uh++++
eh ^uh++++
speak ^vb++++
English ^nn++++
? ^?+clp+++
{speaker2
uh ^uh+++??+
- ^-++++
huh ^uh++++
{speaker1
teacher ^nn++++
```

Example 2

```
{speaker1
very ^ql+amp+++
good ^jj+++??+
. ^.+clp+++
it ^pp3+it+++
was ^vbd+bedz+vrb++
fun ^nn++++
to ^to++++
em ^jj+atrb++??+
festival ^nn+nom+++
here ^rn+pl+++
in ^in++++
Portland ^np++++
{speaker2
um ^mm++++
```

excuse ^vb++++
me ^pp1o+pp1+++
one ^cd1++++
moment ^nn++++
oh ^uh++++
the ^ati++++
question ^nn+nom+++
is ^vbz+bez+vrb++
about ^in++++
uh ^uh+++??+
speak ^vb++++
English ^nn++++
? ^?+clp+++
or ^cc+cls+++
or ^cc++++
the ^ati++++

<filename= Speak2.asc>
<Date= 01-mar-02>

{speaker1
and ^cc+cls+++
(^(++++
+ ^&fo++++
) ^)+++??+
working ^vwbg+++xvbg+
xxx ^nn+++??+
you ^pp2+pp2+++
know ^vb+vprv+++
that ^tht+vcmp+++
festival ^nn+nom+++
xxx ^nn+++??+
Mexican ^jj+atrb+++
xxx ^nn+++??+
five ^cd++++
. ^.+clp+++
{speaker2
five ^cd++++
{speaker1
five ^cd++++
de ^at++++
mayo ^np++++
. ^.+clp+++
may ^np++++
. ^.+clp+++
you ^pp2+pp2+++
know ^vb+vprv+++
may ^np++++
? ^?+clp+++

<filename= Wkend2.asc>
<Date= 07-may-02>

Figure 3. Examples of Grammatically Tagged Texts (tagging errors in bold)

There are four mistakes in the tagging of this sample. Two of them can be eliminated with simple changes to the tagging program:

- *em* has been identified as an attributive adjective. Since it is never a word, it can automatically be identified with the same code as *um* and *mhm*.
- *xxx* (used for unclear words) has been tagged as a noun. Since it is impossible to know what grammatical category the word is, this tag can be changed to *xxx* to

signify "unclear transcription."

A third mistake concerns use of the student's first language: *de* is labeled as an article rather than a preposition in a foreign language. (Both *cinco* and *Mayo* are tagged correctly, as a number and proper noun, respectively.) Since *de* is not an article in English, this particular problem could be changed in the tagger, but the more important issue concerns learners' use of their first languages when the tagger is designed to analyze English. Fortunately, except in unusual cases where a phrase is considered to be in common use in English (such as *Cinco de Mayo*), transcribers simply designate use of the first language with the symbol <L1>, which can just be coded as L1 by the tagger also. Researchers interested in use of the first language could again use the video indexing to hear exactly what was said.

The final tagging mistake is a common mistake with native speaker data as well: *that* has been tagged as a verb complementizer when it is actually a demonstrative determiner. The tagging of *that* is virtually impossible to make highly accurate without an interactive checking program. Such programs are regularly used with native speaker corpora as well (see further Biber, Conrad & Reppen 1998: 257-262).

Overall, tagging a relatively low-level learner corpus appears feasible. Interactive checking of certain words would be necessary to ensure a high enough level of accuracy, but this is not an unusual requirement even for native speaker corpora. Analyses would be possible then with both untagged and tagged versions of the corpus.

4.3. Potential Analyses

One of the easiest types of corpus analyses is to investigate vocabulary, since that analysis relies on words, not grammatical categories. A variety of types of analysis appear feasible with a Lab School corpus. For instance, past studies have looked at lexical development and complexity in terms of type-token ratio (TTR) and associated measures which adjust for short text lengths (Daller, Van Hout & Treffers-Daller 2003; Yuan & Ellis 2003). With longitudinal studies it should be possible to see how students' vocabulary diversity increases over time. In cross-sectional studies differences in students' vocabulary use at different instructional levels could be investigated. With the home interview component, the vocabulary of learners who have continued in classes could be compared with those who stopped taking classes.

Another kind of lexical analysis that would be possible with the corpus concerns "lexical bundles" – fixed sequences of words that are used repeatedly across different texts within a register or genre. Research has shown that such bundles make up over a quarter of the words in conversation among native speakers of British and American English, and are also common in university class sessions and textbooks (Biber, Conrad & Cortes 2004; Biber et al. 1999: chapter 13). Looking at the repeated chunks that learners use and comparing them to native speakers' lexical bundles is a worthy area for investigation. A lexical bundle type of analysis is also relatively straightforward to conduct since it focuses on word forms. Fillers such as *um* and *mm* in a speaker's turn (common in the learner language) can easily be eliminated from the analysis.

Investigations of the development of learners' grammatical systems would also be possible. For example, the development of learners' verb tense/aspect system could be studied. Hypotheses such as the aspect hypothesis (e.g. Bardovi-Harlig 2000: chapter 4; Bardovi-Harlig & Reynolds 1995) – that learners' acquisition of tense and aspect morphology is influenced by the semantic properties of verbs – could be tested with a larger number of learners in the corpus, and a greater diversity of first languages and educational backgrounds.

This type of analysis would require intensive interactive coding of transcripts. In studying verb tense and aspect development, for example, an automatic program could identify verbs from their tags, and then researchers could code a variety of variables: the semantic category of the verb (e.g. stative vs punctual vs accomplishment); the obligatory (or acceptable) verb forms for the context; the verb form used and therefore the type of error if an error exists; whether the verb is regular or irregular in form; the specific verb; and aspects of the phonological environment, such as whether there is a final consonant cluster (cf. Adamson et al. 1996; Preston 1996). While more time consuming than a purely automatic analysis, this type of analysis allows a large number of variables to be considered at once, and can provide a more accurate picture of the complex conditions that shape learner language. Previously it was most often conducted with learner language using VARBRUL software (see Young & Bayley 1996) – a system of analysis developed for studying sociolinguistic variables – usually using fewer data than in corpus-based studies. With native speaker corpora, studies of multiple variables that affect language choices have been conducted for grammatical systems such as stance (Conrad & Biber 2000) and reference (Biber, Conrad & Reppen 1998: chapter 5) and for grammatical categories such as adverbials (Biber et al. 1999: chapter 10; Conrad 1999). Learner errors have also been interactively coded and then studied (e.g., Dagneaux, Denness & Granger 1998; Housen 2002; Milton and Chowdury 1994).

The basic procedure of interactively coding data seems feasible with a learner corpus and allows a variety of studies. Learner language always presents some additional challenges over native speaker data, such as in interpreting obligatory or appropriate choices for the context when learners' intended meaning is not clear. As noted above, however, being able to consult the video is likely to allow for more disambiguation than with other corpora.

5. Conclusion: Looking ahead at corpus development

Decades ago, Tarone (1979) characterized interlanguage as a chameleon, affected by the task, interlocutors, topics, and a variety of other factors that all need to be considered at once. The methods of corpus linguistics have proven themselves effective in considering multiple variables at once in native speaker speech. It is unfortunate that these techniques are so far having less impact on the field of SLA research.

The corpus that I have discussed here would counter a number of limitations SLA researchers may see in current corpus-based work, especially for its context in the United States. It provides a corpus for learning about the type of learners and level of language that is central in many policy decisions within the United States. Furthermore, it provides a corpus for investigating issues of concern in instructed language acquisition with a classroom component in the corpus. Furthermore, the corpus transcripts remain indexed to a videorecording, so that researchers can see the context and hear the participants if they wish.

A great deal of work needs to be done to design and compile this corpus. This work includes additional transcribing of the recordings of targeted students so that there are enough samples for both longitudinal and cross-sectional studies. Checking of the language codes and teacher prompts is also necessary, to ensure consistency within the student-controlled and teacher-controlled language tasks. The transcriptions must be made into free-standing files, separate from video but containing markers for the video. The corpus will then be more useful if it is grammatically tagged and the tags are checked to ensure accuracy.

The amount of work to be done can sound daunting, and the corpus is moving into uncharted waters by including different types of classroom tasks as well as home interviews, including quite low proficiency students, and having transcription linked to video. However, if we do not expand our thinking about corpus design and do not consider how to integrate corpus linguistics with other language acquisition projects, we limit the potential for corpus linguistics to contribute to SLA research.

References

Abe, M. 2003. "A Corpus-based Contrastive Analysis of Spoken and Written Learner Corpora: The case of Japanese-speaking learners of English". *Proceedings of the Corpus Linguistics 2003 conference (CL 2003).* [Technical Papers 16]. Archer, Rayson, Wilson & McEnery 2003. 1-9.

Adamson, H.D., B. Fonseca-Greber, K. Kataoka, V. Scardino & S. Takano. 1996. "Tense Marking in the English of Spanish-speaking Adolescents". *Second Language Acquisition and Linguistic Variation*, Bayley & Preston 1996. 121-134.

Ädel, A. 2005. "Involvement Features in Writing: Do time and interaction trump register awareness?" Paper presented at the combined conference ICAME 26/AAACL 6, Ann Arbor, Michigan, May 2005.

Altenberg, B. 1997. "Exploring the Swedish Component of the International Corpus of Learner English". *PALC 97: Practical applications in language corpora*, Lewandowska-Tomaszcyk & Nelia 1997. 119-132.

Archer, D., P. Rayson, A. Wilson & T. McEnery (eds.). 2003. *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003).* [Technical Papers 16]. Lancaster University: University Centre for Computer Corpus Research on Language.

Aston, G. 2001. *Learning with Corpora*. Houston, Texas: Athelstan.

Banke, S. & D. Brillanceau. In preparation. "Living an Educational Classroom Intervention: The Teachers' Voices."

Bardovi-Harlig, K. 2000. *Tense and Aspect in Second Language Acquisition: Form, meaning, and use.* Malden, MA: Blackwell.

Bardovi-Harlig, K., & D. Reynolds. 1995. "The Role of Lexical Aspect in the Acquisition of Tense and Aspect". *TESOL Quarterly* 29. 107-31.

Bayley, R. & D. Preston (eds.) 1996. *Second Language Acquisition and Linguistic Variation*. Amsterdam: John Benjamins.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., S. Conrad & V. Cortes. 2004. "*If You Look At…*: Lexical bundles in university teaching and textbooks". *Applied Linguistics* 25. 371-405.

Biber, D., S. Conrad & R. Reppen, 1998. *Corpus Linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *The Longman Grammar of Spoken and Written English*. Harlow, Essex: Longman.

Bolstad, B., L. Boyd, E. Davila, R. Disbrow-Chen, J. Garland, K. Harris, et al. 2002. *Portland State University Adult ESOL Labsite Classroom Coding Event Coding System*. Portland, Oregon: Portland State University.

Brillanceau, D. 2005. Spontaneous Conversations: A window into language learners' autonomy. *Focus on Basics* 8(A). 22-25. [Available via the Web at http://www.ncsall.net/fileadmin/resources/fob/2005/fob_8a.pdf]

Burnard, L. & T. McEnery (eds.). 2000. *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang.

Cheng, W. & M. Warren. 2000. "The Hong Kong Corpus of Spoken English: Language learning through language description." *Rethinking Language Pedagogy from a Corpus Perspective*, Burnard & McEnery 2000. 133-144.

Center for Applied Linguistics. 2003. *Immigrant Education*. Retrieved June 25, 2002 from the Worldwide Web: http://www.cal.org/topic/immigrant.html.

Conrad, S. 2005. "Corpus Linguistics and L2 Teaching." *Handbook of Research in Second Language Teaching and Learning*, Hinkel 2005. 393-409.

Conrad, S. 2004. *The Feasibility of Designing a Corpus for Second Language Acquistion Research from the ESOL Labsite Database*. Unpublished technical report. Portland, Oregon: Portland State University.

Conrad, S. 1999. "The Importance of Corpus-based Research for Language Teachers". *System* 27. 1-18.

Conrad, S. & D. Biber. 2000. "Adverbial marking of stance in speech and writing". *Evaluation in Text*, Hunston & Thompson 2000. 56-73.

Coxhead, A. 2000. "A New Academic Word List." *TESOL Quarterly* 34. 213-238.

Dagneaux, E., S. Denness & S. Granger. 1998. Computer-aided error analysis. *System* 26. 163-174.

Daller, H., R. Van Hout & J. Treffers-Daller. 2003. "Lexical Richness in Spontaneous Speech of Bilinguals". *Applied Linguistics* 24. 197-222.

DeCock, S. 2000. "Repetitive Phrasal Chunkiness and Advanced EFL Speech and Writing". *Corpus Linguistics and Linguistic Theory: Papers from the twentieth international conference on English language research on computerized corpora (ICAME 20), Freiburg im Breisgau 1999*, Mair and Hundt 2000. 51-68.

Disbrow-Chen, R. 2004. *Morpheme Acquisition in Relation to Task Variation: A case study of a beginning-level ESL learner.* Unpublished M.A. thesis. Portland, Oregon: Portland State University.

Flowerdew, L. & K. Tong, K. (eds.) 1994. *Entering Text*. Hong Kong: The Hong Kong University of Science and Technology.

Garland, J. 2002. *Co-Construction of Language and Activity in Low-level ESL Pair Interactions.* Unpublished M.A. Thesis. Portland, Oregon:

Portland State University.

Granger, S. 2003. “The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research”. *TESOL Quarterly* 37. 538-546.

Granger, S. 1998a. “The Computer Learner Corpus: A versatile new source of data for SLA research”. *Learner English on Computer,* Granger 1998b. 3-18.

Granger, S. (ed.) 1998b. *Learner English on Computer*. London: Longman.

Granger, S., J. Hung, & S. Petch-Tyson (eds.) 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Harris, K. In preparation. “Meaning Negotiation in Beginning Adult ESL Class Activities.”

Hellermann, J. In press(a). “Classroom Interactive Practices for Developing L2 Literacy: A microethnographic study of two beginning adult learners of English”. *Applied Linguistics* 27.

Hellermann, J. In press(b). “The Development of Practices for Action in Classroom Dyadic Interaction: Focus on task openings”. *Modern Language Journal* 91.

Hinkel, E. (ed.) 2005. *Handbook of Research in Second Language Teaching and Learning*. Mahwah, NJ: Lawrence Erlbaum.

Housen, A 2002. “A Corpus-based Study of the L2-acquisition of the English Verb System”. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching,* Granger, Hung & Petch-Tyson 2002. 77-116.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. & G. Thompson (eds.). 2000. *Evaluation in Text.* Oxford: Oxford University Press.

Hyland, K. & J. Milton. 1997. “Qualification and Certainty in L1 and L2 Students’ Writing”. *Journal of Second Language Writing* 6. 183-205.

Kurzet, R. 2004. “Professional Development in Action: Connecting adult ESOL teachers to research-in-progress”. Paper presented at the 2004 meeting of the American Association for Applied Linguistics, Portland, Oregon.

Kurzet, R. 2002. “Teachable Moments: Videos of adult ESOL classrooms”. *Focus on Basics* 5(D). 8-11.

Lewandowska-Tomaszcyk, B. & P. Melia (eds.). 1997. *PALC 97: Practical applications in language corpora*. Lódz, Poland: Lódz University Press.

Mair, C. and M. Hundt (eds.) 2000. *Corpus Linguistics and Linguistic Theory. Papers from the twentieth international conference on English* *language research on computerized corpora (ICAME 20), Freiburg im Breisgau 1999.* Amsterdam: Rodopi.

McCarthy, M., J. McCarten & H. Sandiford. 2005. *Touchstone 1*. Cambridge: Cambridge University Press.

Meyer, C. 2002. *English Corpus Linguistics*. Cambridge: Cambridge University Press.

Milton, J. & N. Chowdury. 1994. "Tagging the Interlanguage of Chinese Learners of English". *Entering Text*, Flowerdew & Tong 1994. 127-143.

Nesselhauf, N. 2004. "Learner Corpora and Their Potential in Language Teaching". *How to Use Corpora in Language Teaching,* Sinclair 2004. 125-152.

Ouellette, S. 2004. *Making the Effort: A study of one student's communication strategies in an ESL classroom*. Unpublished M.A. thesis. Portland, Oregon: Portland State University.

Preston, D. 1996. "Variationist Perspectives on Second Language Acquisition". *Second Language Acquisition and Linguistic Variation,* Bayley & Preston 1996. 1-45.

Reder, S. & D. Brillanceau. In preparation. "Teaching, Learning and Control in Educational Practice and Research: An experiment doing an experiment".

Reder, S., K. Harris & K. Setzler. 2003. "A Multimedia Adult Learner Corpus". *TESOL Quarterly* 37. 546-556.

Schmitt, D. & N. Schmitt. 2005. *Focus on Vocabulary: Mastering the academic word list*. White Plains, New York: Longman.

Seidlhofer, S. 2000. "Operationalizing Intertextuality: Using learner corpora for learning." *Rethinking Language Pedagogy from a Corpus Perspective,* Burnard & McEnery 2000. 207-223.

Sinclair, J. (ed.) 2004. *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins.

Tarone, E. 1979. "Interlanguage as Chameleon". *Language Learning* 29. 181-191.

Tono, Y. 2000. "A Computer Learner Corpus Based Analysis of the Acquisition Order of English Grammatical Morphemes". *Rethinking Language Pedagogy from a Corpus Perspective,* Burnard & McEnery 2000. 123-132.

Yuan, R. & R. Ellis, R. 2003. "The Effects of Pre-task Planning and On-line Planning on Fluency, Complexity and Accuracy in L2 Monological Oral Production". *Applied Linguistics* 24. 1-27.

Yoon, H. & A. Hirvela. 2004. "ESL Student Attitudes toward Corpus Use in L2 Writing". *Journal of Second Language Writing* 13. 257-283.

Young, R. & R. Bayley. 1996. "VARBRUL Analysis for Second Language

Acquisition Research”. *Second Language Acquisition and Linguistic Variation*, Bayley & Preston 1996. 253-306.

C-ORAL-ROM — Prosodic Boundaries for Spontaneous Speech Analysis —

Massimo MONEGLIA and Emanuela CRESTI

1. The C-ORAL-ROM Corpus

The C-ORAL-ROM multilingual resource provides a comparable set of corpora of spontaneous spoken language of the main romance languages, namely French, Italian, Portuguese and Spanish. The resource is the result of the C-ORAL-ROM project, which has been undertaken by an European consortium, co-ordinated by the University of Florence and funded within the Fifth EU framework program.

C-ORAL-ROM consists of 772 spoken texts and 123:27:35 hours of speech. Four comparable recording collections of Italian, French, Portuguese and Spanish spontaneous speech sessions (roughly 300,000 words for each Language) have been delivered respectively by the following providers:

University of Florence (LABLITA, Laboratorio linguistico del Dipartimento di italianistica);

Université de Provence (DELIC, Description Linguistique Informatisée sur Corpus);

Centro de Linguística da Universidade de Lisboa (CLUL);

Universidad Autónoma de Madrid (Departamento de linguistica, Laboratorio de Lingüística Informática).

C-ORAL-ROM is published in two multimedia forms:

1. In a form devoted to speech laboratories and to multiple users, through the ELDA catalogue, in 9 DVDs where files are non-compressed and non-encrypted.
2. Through Benjamins Publishing company, which presents the resource in compressed and encrypted format in only one DVD accompanied by an explicative book (Cresti & Moneglia 2005). This form, that does not allow copying, is designed for wide distribution in the linguistic community.

Each recorded session of the C-ORAL-ROM corpus is stored in acoustic files (wav Windows PCM, 22.050 Hz, 16 bit- ELDA edition; MP3 files in the Benjamins edition) and is delivered with the following main annotations:

a. Session metadata (in CHAT and IMDI format)
b. The orthographic transcription, (in CHAT format; Mac Whinney, 1994) enriched by the tagging of terminal and non terminal

prosodic breaks (in txt files)

c. The text-to-speech synchronization, based on the alignment to the acoustic source of each transcribed utterance, in .xml files.
d. Textual resource with Part of Speech (PoS) and lemma tagging of each form

C-ORAL-ROM is integrated by the Win Pitch Corpus speech software (© Pitch France) that allows the direct and simultaneous exploitation of the acoustic and textual information. In the Benjamins edition C-ORAL-ROM is also integrated with a concordance software (Contextes © Jean Veronis).

The main objective of C-ORAL-ROM is to allow linguistic studies and natural language technologies to face challenging language resources which testify spontaneous speech in real environment. To this end C-ORAL-ROM "aims to represent the variety of speech acts performed in everyday language and to enable the induction of prosodic and syntactic structures in the four romance languages, from a quantitative and qualitative point of view".

The collection of a spoken corpus, its transcription, annotation and text/sound alignment is a very heavy and expensive work; so its size is a necessarily limited if compared with that of written corpora (billions of words. See Leech et al. 2001). The corpus design must ensure general representativeness and internal balance (Sinclair 2005), which is not easy considering the limited dimension of spoken resources. Moreover in a multilingual collection the corpus design of each language resource must be comparable, if we want to use corpora for comparative studies. We believe that despite its small size C-ORAL-ROM has the general structure of a reference corpus and can therefore be held as a model for larger corpora. The corpus design has been presented in another paper of this book (Moneglia in this volume) and therefore the reader can refer directly to that chapter for an adequate report on the sampling strategy adopted.

The collection of files of the C-ORAL-ROM resource is also conceived to allow maximum and easy exploitation of the linguistic information recorded in the corpus. To this end, as it is widely recognized, spoken resources strictly require a link between a proper linguistic annotation and the access to the acoustic source (Oostdijk et al. 2004, Sinclair 1996). The two issues are strictly linked in our view.

The linguistic annotation of spoken language performances presents relevant problems. For this reason in C-ORAL-ROM special attention is devoted to the annotation of the 'reference units' of analysis for spontaneous speech. In this chapter we will concentrate on the annotation scheme used to identify the reference unit ranking above the word level in the speech continuum and we will focus on the relevance that prosodic annotation has with regard to this. The assumptions featured in C-ORAL-ROM will be supported by the generalizations induced by the annotation scheme at cross-linguistic level. We will show that the identification of the reference unit is the key information which allows a synchronization of the transcripts to the acoustic source that is suitable for linguistic analysis and easy to be obtained.

2. The C-ORAL-ROM framework for the annotation of reference units of spontaneous speech

The identification of the units of reference is the main added information for the linguistic analysis of a spoken corpus. This information is crucial for the understanding of peculiar properties of speech, but can hardly be identified through the same syntactic and semantic cues used for written resources (Blanche-Benveniste 1997, Biber et al. 1999, Cresti 2000, Miller & Weinert 1998, Izre'el 2005). The main problem is that syntax does not provide enough evidence for the identification of the linguistic unit ranking above word level. Almost 1/3 of speech events (according to C-ORAL-ROM and the Longman Grammar) do not have a verb and therefore do not show a clear syntactic structure; this means that configurations that are not clauses may be reference units in the speech flow. However in our view the reference unit of spoken language is not underdetermined if the pragmatic and prosodic features of speech are taken into account.

In the C-ORAL-ROM approach the reference unit for spontaneous speech is identified with the term 'utterance', that is defined following the pragmatic tradition (Austin 1962). The utterance is *the minimal linguistic entity such that can be pragmatically interpreted*; i.e. the linguistic entity that is 'concluded' and 'autonomous' from a pragmatic point of view.

Although this definition may sound familiar (Quirk et al. 1985, Cresti 2000) the annotation procedure that allows to parse the speech continuum into utterances in the C-ORAL-ROM approach is quite new. In C-ORAL-ROM the utterance is identified through an heuristic that allows its annotation as a function of prosodic properties, and more specifically it is based on the perception of *prosodic breaks* (Cresti & Firenzuoli 1999, Cresti 2000, Cresti & Firenzuoli 2002).

It is assumed that each utterance has a profile of *terminal intonation* (Karcevsky 1931, Crystal 1975) and therefore the presence of *terminal breaks* in a string is a cue for the detection of utterance boundaries. The speech flow and its transcription are so divided into reference units which rank above the word level taking into account this prosodic feature, that is considered the more easily detected formal property of the utterance. Each unit ending with a terminal break is considered an utterance.

Prosodic tagging is accomplished in C-ORAL-ROM on the basis of perceptual evidence in accordance with the following annotation scheme (Moneglia 2004):

- Prosodic tagging specifies each perceptively relevant prosodic break in the speech continuum
- All positions between two words are considered possible positions to be fitted with a prosodic tag. No within-word prosodic breaks are marked in C-ORAL-ROM
- Prosodic breaks are distinguished in accordance with two main qualities: *terminal* vs. *non-terminal*
- Each between-words position necessarily has one of the following values with respect to the prosodic tagging of the resource:
    - (O) no break
    - (T) terminal break , marked with a double slash "//" or "?"
    - (N) non-terminal break, marked with a single slash "/".

- Prosodic breaks are always tagged and reported according to perceptual judgments of the transcribers, within the process of corpus transcription and revision
- Prosodic tagging is part of the transcription and is reported within the text lines.
- The criterion for the segmentation of the speech flow into utterances is prosodic. Each prosodic break qualified as terminal defines the utterance limits in the speech flow

Concept	*Definition*
Prosodic break	Perceptively relevant prosodic variation in the speech continuum such that it causes the parsing of the continuum into discrete prosodic units.
Terminal prosodic breaks	Given a sequence of one or more prosodic units, a prosodic break is known as terminal if a competent speaker assigns the quality of *concluding such sequence* to it.
Non-terminal prosodic breaks	Given a sequence of one or more prosodic units, a prosodic break is known as non-terminal if a competent speaker assigns the quality of being *non conclusive* to it.
Prosodic pattern (Utterance)	Each sequence of prosodic units (≥ 1) ending with a terminal prosodic break

The annotation of prosodic break is an exercise that has been accomplished by expert mother tongue annotators (Phd. and Phd. students in linguistics), all acquainted with the concept of speech act and at least with a minimal understanding of prosodic feature. However the validation of the annotation of the four romance corpora has shown that the perceptual evidence indicated by the annotation scheme is clear also to non experts.

It was already well known that competent speakers have a strong perception of prosodic boundaries (Buhmann et al. 2002, Sorianello forthcoming). The C-ORAL-ROM annotation and its subsequent validation show that, at least for what regards romance languages, competent speakers can also easily discriminate prosodic boundaries that have a terminal value from those boundaries that indicate that the utterance goes on ('Non terminal breaks').

Data regarding the level of inter-annotator agreement recorded in C-ORAL-ROM (Danieli et al. 2004) are quite clear to this regard. Two mother tongue evaluators for each romance language, hired by an institution external to the C-ORAL-ROM consortium, have challenged the C-ORAL-ROM tagging of a statistically significant portion of the C-ORAL-ROM corpus. The following table summarizes the general results of the validation in terms of percentage of total agreement recorded for each position in the corpus, and the K-index recorded.

	French	Italian	Portuguese	Spanish
Total Agreement on T	95.05%	97.14%	98.12%	94.84%
Total Agreement on N	86.56%	93.15%	98.38%	94.62%
Total Agreement on O	97.54%	95.1%	99.22%	98.28%
K Index (General)	0.952	0.928	0.980	0.946
K Index (Realistic)	0.776	0.807	0.920	0.827

The level of agreement recorded clearly shows that in all romance languages the perception of prosodic boundaries, especially terminal ones, is strong and inter-subjective. The annotation scheme is therefore very reliable and seems based on a clear perceptual evidence.

The definition of the utterance boundaries through their prosodic properties allows a proper analysis of the four spoken romance corpora and to carry out their general comparison from a linguistic point of view. More specifically, it provides the term of reference for the statistic evaluation of the main properties of spoken language performance in the multilingual corpus. (Moneglia in this volume).

In the following paragraphs, we will challenge this type of annotation of the reference units of speech with other concurrent methods. On the basis of actual spontaneous speech data, we will lead the reader to the following main conclusions: while the other methods cause the emergence of a strong under-determinacy in the detection of the reference units in speech, the detection of terminal breaks is reliable and linguistically relevant for two main reasons: 1) the reference units so identified are in one-to-one correspondence with speech acts; 2) the information conveyed by terminal breaks is necessary to determine the syntactic relations between the elements of the speech performance. Therefore, in the analysis of speech performance, syntax depends on prosody and not vice versa.

Finally, we will stress that the annotation of terminal breaks provides an easy and linguistically meaningful criterion for the text-to-speech alignment of large corpora, that would otherwise be left undefined: each unit of speech matches a speech act and is therefore simultaneously aligned to its linguistically relevant counterpart.

3. The unit of reference for spontaneous speech analysis

The product of the speech performance widely recognized as the reference unit for spontaneous speech is the ‘utterance’ (Biber et al. 1999), but the definition of this entity is a complex matter. A practical equivalence has often been proposed between the utterance and various speech events that can be recognized in the speech flow. The main property of a ‘linguistic event’ is that it must represent a ‘continuous stretch of talk within the speech flow’. A first alternative used to identify such a continuum is to consider as such an event a speaker’s turn within a dialogue; another alternative has been identified in a sequence between two silences (see. TEI guidelines).

A third, more linguistic, alternative is to link the definition of the speech event to syntactic-semantic properties, thus enabling its identification through clauses, or propositional structures (a C-Unit in the Longman Grammar’s lexicon).

A fourth alternative is to try to identify the utterance by means of a semantic-pragmatic analysis. A stretch of talk is an utterance if it can be recognized to perform a speech act in accordance with an annotation scheme that lists the set of possible speech acts.

In the following sections, on the basis of evidence taken from real corpus analysis, we will show that: a) the *turn* is a too weak notion to identify the utterance, as many speech events can occur within a turn; b) the *timing* of an utterance is not linguistically significant, as it is, at the same time, too weak and too strong to determine the utterance’s boundaries in spontaneous speech corpora; c) the *syntactic structure* appears strongly underdetermined in spontaneous speech, while its definition is rather a function of prosodic cues.

The examples below present typical stretches of spontaneous face to face dialogues, as they are transcribed in the C-ORAL-ROM implementation of the CHAT format.[1] Each turn is introduced by the speaker's label in capitals, preceded by "*" and followed by ":". The sequence of speech turns is ordered vertically. The speech events which occur in the turn are reported horizontally. The double slashes "//" or the interrogative "?" indicate the terminal breaks, which identify the utterance limit, while the single slash "/" refers to the non-terminal prosodic breaks in the speech flow. Each utterance, that is also marked by a ranking number in brackets, is aligned to the acoustic counterpart.

IT1 (ifamdl15)
A*EST: [1] o vieni / dai //   [come on then]
B*CLA: [2] a patire //   [to suffer]
C*EST: [3] no // [4] ascolta / qui sopra ? [5] sì //
        [no // listen / (what about) up here ? yes //]
D*CLA: [6] qui ? [7] sì //   [here ? yes]
…..
X*EST: [8]... lei / prima veniva tutte le settimane //
        [she / used to come every week once //]

IT2 (ifamdl18)
A *ALE: [1] eh / io invece / mi sono preso una chitarra //
        [and instead / I got a guitar //]
B *IDA: [2] ah / la chitarra //   [ah/ a guitar]
C *ALE: [3] acustica // [4] bella //   [acoustic // beautifull //]
D *IDA: [5] e l' obiettivo ?   [and the lens ?]

FR1 (ffamdl 08)
B*JEA: [1] c' est la carotte quoi // [2]carotte devant le nez du lapin //
        [it is the carrot then // carrot in front of the nose of the rabbit //]
C*EMI: [2']elle le prend comme un gamin quoi // #
        [she consider him like a little boy then]
D*JEA: [3] ouais // [4]la carotte / ouais //   [yes // the carrot / yes //]

[1] Speech data can be downloaded by the reader from http://lablita.dit.unifi.it/Speech_Examples/. For citation purposes, each stretch is identified here by the language id. plus a number and the source file in C-ORAL-ROM in parenthesis; each dialogic turn within the dialogue is marked with a capital letter; each utterance within the dialogic turn is marked with a number in square brackets. The rank number refers to the set of utterances reproduced here and therefore differs from the rank in the C-ORAL-ROM corpus, where each utterance is identified by an ordered pair of characters, consisting of the a serial number of the utterance in the aligned text and a filename.

FR2 (fpubdl03)
A*EMA: [1] ça c' est clair // # [2] de plus en plus // #
[this is clear // more and more]
B*JUL: [3] quels sont vos rapports avec les clients en général ? #
[what are your relationships with customers in general ?]
C*EMA: [4] très bons // # [5] sur la base / très bons // #
[very good // basically / very good //]

PT1(pfamdl14)
A*JOS: [1]esteve na Baixa / recentemente ? [have you been in the Baixa / recently ?]
B*NOE: [2]não //$ [3] há muito tempo que não vou //[4]
acho que está muito gira // [5] não é //
[no // it’s been a long time since I last went // I know that it is very alive // isn’t it?]
C*JOS: [6]está a ficar muito bonita // [it’s getting very nice]
D*NOE: [7] estive no Chiado / há pouco tempo //
[I have been in the Chado / recently]
E*JOS: [8] hum <hum> // [uhm //]

SP1 (efammn06)
MAF: … ya saben / todas esas cosas // que generalmente les ofrecemos // …
[They already know / all these things // that we generally offer them]

Starting from the previous typical examples of spontaneous speech we will see that, as opposed to the alternative hypothesis considered, the identification of the utterance through the C-ORAL-ROM annotation scheme provides a much more effective performance.

3.1. Dialogic turns

The easier way to determine a reference unit of speech ranking above the word level in face to face dialogues is to observe speakers’ turns in a session.[2] The speech turn is a stretch of talk by the same speaker that constitutes an event in the speech session, and whether it is a fragment or real linguistic information, it is also an independent entity to which we can refer. The set of turns of a session is a linear order of continuous pieces of talk, each one produced by a different speaker.

[2] See for example the original CHAT annotation system where there is one to one correspondence between dialogic turns and utterances.

Even considering overlapping and intersection phenomena, that frequently occur in speech, the annotation of speaker turns is not problematic for humans. However, despite the fact that the dialogic turn is a linguistically relevant object for language analysis it should be obvious that a turn is not a real continuum and that, within one turn, more than one speech event may occur. In other words the speaker's turn is not the *minimal* linguistic unit ranking above the word level. More than one of these entities can be linearly ordered within the turn itself. We will give this for granted. Therefore the problem is how the units ranking above the word level are determined in the speech flow within the dialogic turn.

3.2. Pauses as utterances boundaries

The 'from silence to silence' criterion is used for the detection of utterance boundaries, probably because the automatic recognition of pauses in the speech flow is a quite easy task to be pursued given the current technologies. Such criterion may be preferable to the prosodic criterion adopted in C-ORAL-ROM, as it is more objective, whereas the prosodic criterion may be considered arbitrary as it is based on perception.

The figures below show the acoustic signal in the multimedia format of the C-ORAL-ROM corpus.[3] Indeed the utterance boundaries frequently occur in speech together with significant wave interruptions. For example, in the transition between [4] and [5] in IT1, after "no" there is an interruption of around 600 ms (in yellow in Figure 1) that accompanies the beginning of the second utterance.

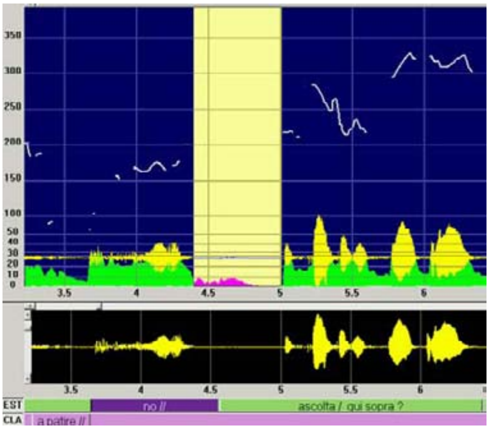


Figure 1.

[3] In the top window, the wave and the F_0 tracks are displayed. The acoustic signal is aligned to the transcription information, reported for each speaker in the bottom layers (Win Pitch Corpus speech software).

However C-ORAL-ROM data proof that this criterion is linguistically irrelevant, because it is both too weak and too strong in marking utterance boundaries of actual spontaneous speech data.

The criterion is too strong because an utterance may end, and another utterance can start with no need for pauses. The criterion is too weak because the occurrence of a pause is not a sufficient cue to infer the conclusion of an utterance For example the transition between IT1[6] and [7] (respectively marked in dark and bright in Figure 2) does not record any pause, but two distinct speech acts are perceived and a terminal break is identified.

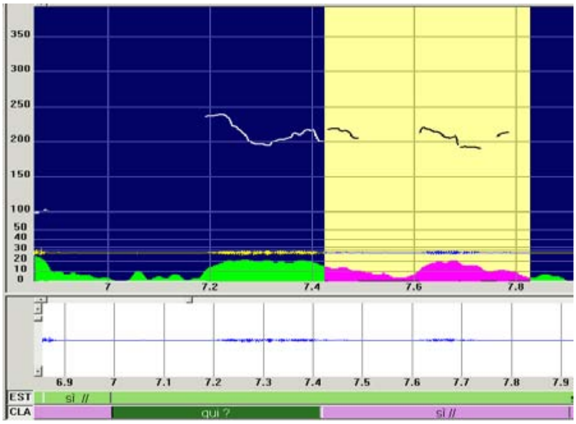


Figure 2.

We can notice here that although no pause split the two speech events the perception of prosody appears to be sensible to the 20hz discontinuity connected to the resetting necessary for the performance of the second utterance. Therefore the criterion is too weak to detect the utterance boundaries.

As for the ‘too strong’ side, we can see that a perceptively relevant prosodic break may be accompanied by a pause even if the break does not mark the end of the utterance. *Topic – Comment* structures are typical instance of the co-occurrence of a pause with a non terminal prosodic break as in IT1[8] and FR2[5]. In both utterances, regardless of the language, the first element has prefix intonation and is perceived as non concluded, while the second string is concluded. The bright part indicates quite a long pause of one second that does not determine the perception of the conclusion of the utterance.

*EST: ... [8] lei / prima veniva tutte le settimane // [she / used to come every week once //]

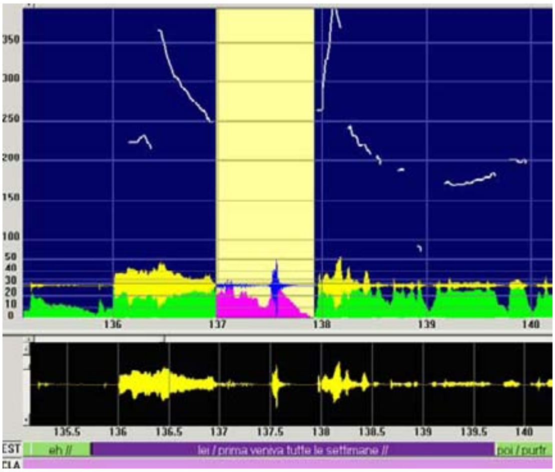


Figure 3.

C*EMA: … [5] sur la base / très bons // # [basically / very good //]

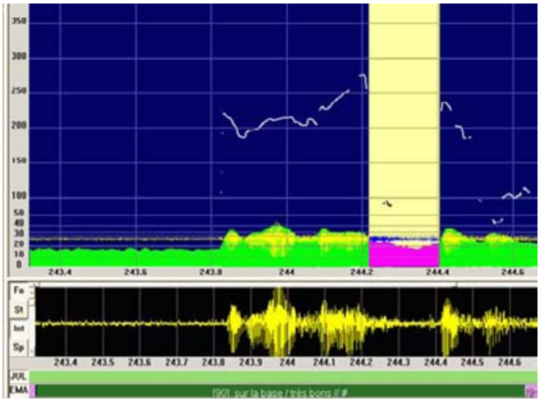


Figure 4.

Using the silence to silence criterion, the sequence will be wrongly considered a sequence of two distinct utterances. Therefore the concept of utterance as a sequence between two silences does not match the concept determined on a prosodic basis. It is at the same time too weak and too strong a notion.

Despite the theoretical demonstration of the inconsistency of this criterion, one could say that from a practical point of view it may be used for the annotation of spoken corpora, as the ‘silence to silence’ property may be considered ‘almost equivalent’ to utterance boundaries. This might be assumed considering that in many cases the end of an utterance is accompanied by pause, as it is in reading, and the occurrence pauses with 'non terminal pauses' is rare. But this is not the case and C-ORAL-ROM demonstrates this in a clear manner from a quantitative point of view.

The French section in C-ORAL-ROM has been tagged with both the temporal and the prosodic criteria. Pauses of more than 200 ms. have been detected automatically in the speech flow and annotated in the transcripts. At the same time the corpus has also been tagged with respect to all terminal and non terminal prosodic breaks, perceived by the expert operators who transcribed and tagged the corpus. On the basis of the results of this double tagging, we recorded that around 63% of sequences ending with a terminal break are accompanied by a pause, while 37% of sequences ending with a terminal break do not bear a pause. A strong under-extension.

On the other side it is also extremely relevant to note that around 42% of breaks that have been considered non terminal are also accompanied by a pause. A dramatic over-extension.

3.3. Syntax vs. speech acts and prosodic breaks

The hypothesis that the identification of the reference unit of spontaneous speech can be achieved observing the syntactic and semantic relations among categorized words that generate autonomous compositional elements is of course the basic assumption that has been held. From this point of view we can consider that the reference units may be 'sentences', or clauses, or whatever syntactic or semantic elements that can be judged autonomous from a semantic and syntactic point of view (Quirk et al. 1985).

However this assumption incurs an immediate problem. As we already pointed out in spontaneous speech, verb-less contexts, appear in around 30% of utterances (38% according to Longman Grammar). The measurements performed on the C-ORAL-ROM corpus show that verbless utterances are 38.1% in Italian, 24.1% in French; 37.23% in Spanish and 36.57% in Portuguese. In those cases the syntactic structure cannot be determined on the basis of the argument structure of the verb itself. Therefore it should not come as a surprise that the sequence of words occurring in a turn may be underdetermined by syntax and semantics. This occurs frequently in reality, in a variety of cases that are sketched in this paragraph.[4]

3.3.1. Nominal utterances

In spoken language nominal utterances frequently occur. This is the case for all C-ORAL-ROM languages and probably for spontaneous face to

[4] See Scarano to appear for more examples and discussion.

face dialogues in all languages. As for example IT2- [2] [3] [4] and [5] ; FR1-[4].

How and why one single noun or adjective can be a reference unit ranking above the word level? What are the criteria that allow to parse the speech continuum in reference units in this case? Many scholars, especially in the computational linguistics field, have tried the way of *ellipsis*; that is, information that is not present in speech, but that can be recovered from the linguistic or pragmatic context and that therefore provides evidence of a syntactic and semantic clause structure underlying each single element.

In our view it should be evident that, although in some cases a saturated predication structure can be in effect recovered from the context, as in IT2-[2] [3] "(I got a guitar that is)[5] *Acoustic // beautiful*", this step is totally arbitrary in most cases and generates more problems than solutions. Especially in the work of tagging corpora, the 'ellipsis hypothesis' does not generate a valid annotation scheme.

For example in IT2[5] the word '*objective'* may be also assigned to the same predicate occurring in the previous environment, but this is arbitrary. The intention of the speaker is vague and can be in principle also associated to an open set of different predicates (Did you get? Did you forget ? Where is…?). So the presence of a possible reference predicate in the context does not provide positive evidence for the selection of a predication structure.

IT2[3] and [4] present a third case. The contextual reference can be recovered —"*acoustic* (guitar)"; "*wonderful* (guitar)"— but the equivalence of the utterance with a predicative construction —"(the guitar is) *acoustic"*; "(the guitar) *is beautiful*"— cannot be established, because the predicative structure and the holophrastic utterance do not bear the same information. In other words holophrastic utterances cannot be substituted by copulative utterances with the same meaning. Indeed the above copulative utterances do not bear the modal value expressed by the corresponding holophrastic utterances, which are therefore not equivalent.

The presence of more than one term in a dialogic turn is the most interesting case. According to the gapping hypothesis, more than one possible structure may be in principle compatible with that syntactic information embodied in spoken texts. For example, on the basis of the possible contextual recovery IT2[4] and 5 could be assigned to two different structures —"(the guitar is) *acoustic*", "(the guitar is) *wonderful*"— or to just one structure —"(the guitar is) *acoustic* (and) *beautiful*". Does the word sequence in object have an ambiguous structure?

[5] The information that do not correspond to actual speech and it is recovered from context is reported among parenthesis in the examples.

In other words, once nominal utterances are foreseen as a possible language structure in a spoken text, the context provides little evidence regarding the constituent structure. This is evident in most cases when many words without a verb occur within a dialogic turn. For ex. In FR1[2] the verbless utterance *"carotte davant le nez du lapin //"* can be roughly considered the argument of a missing "(there is a)". However, as we said, this is arbitrary and inconsistent from a semantic point of view, we must underline that even this assumption does not provide real evidence for the annotation of the reference unit.

As a matter of fact there is no syntactic or semantic or pragmatic argument that could decide whether the sequence in object corresponds to one complex NP — "(there is a) *carotte davant le nez du lapin //*" — or to one NP followed by a PP — "(there is a) *carotte" "(It is) davant le nez du lapin*". Both possible inferences from the contextual information are allowed, with the same 'low' degree of validity. That is, the context does not determine the structure of the turn.

All hypotheses about possible gapping to be recovered through contextual information for the completion of a clause structure are highly speculative and are not supported by empirical evidence. Moreover, and this is more important for what concerns this paper, the 'gapping recovery' does not provide a valid annotation scheme for marking a linguistic information that can be otherwise easily recovered in the speech flow. In fact the ambiguity in the previous examples is a function of a wrong annotation scheme and does not emerge in practice.

No annotator would be uncertain in the previous cases. The reason is that ellipsis and gapping recovery do not play any role in this task, as the segmentation of the linguistic information in reference units is a function of other linguistic cues. In our view that is prosody and speech act performance.

What is essential is that examples like the previous ones, that Austin called "primitive speech acts", bear an illocutionary value (for example Expressive, in IT2 [2], [3] and [4], and Conclusion in FR1[2]) and for this reason they can be interpreted in a pragmatic context as a speech act and hence also considered autonomous events of the speech domain. Each piece of spoken text bearing illocutionary value can be pragmatically interpreted as a complete speech act regardless of its length and syntactic structure.

The pragmatic independence of the above examples can be recovered though the identification of terminal prosodic breaks, that always accompanies the performance of a speech act. That is all we need for parsing the speech flow into reference units. This information does not depend on syntactic considerations. Although many possible structures may be in theory assigned to a spoken turn or to a word sequence, there is no syntactic or semantic evidence that allows to decide what the actual one is. But on the contrary any syntactic analysis must be compatible with the information provided at the prosodic-pragmatic level.

This provide clear evidence to the supposed structural ambiguity that we just mentioned. The sequence in IT2[4] and [5] is not ambiguous at all, as it is perceptually evident that the two terms are two separate speech acts, each one autonomous, and separated from the other by a terminal break. All syntactic analyses must be consistent with this datum.

The opposite is true for FR1[2]. The instance that it can be parsed in two nominal utterances never arises, because the linguistic material is placed within only one prosodic envelope that is concluded by a terminal break, and only one speech act is performed.

All nominal utterances listed above are always accompanied by a terminal prosodic break that is necessary and sufficient to determine the segmentation of speech into autonomous reference units. Not considering this preliminary requirement of the syntactic analysis leads to the under-determinacy of the unit of reference in the speech flow, which on the contrary is completely determined.

3.3.2. Monorematic utterances and structural ambiguity

It is very important to underline that the problem with syntactic and semantic criteria for the segmentation of speech in units of reference goes beyond the problem of nominal utterances. Indeed, lexical entries, quite independently from their syntactic category, can be used as holophrastic utterances with no clause structure. These utterances can be nouns, or an adjectives as in the above examples, but also verbs IT1[2], pronouns IT1[6], adverbs FR2[2] —very frequently a sentential adverb (IT1[3] [7] [5]; FR1[3] PT1[2])— or interjections PT1[8].[6] This is not a proof of the defective nature of speech, but rather a universal feature of language, that is strongly based on the ontogenetic process of language acquisition (Moneglia and Cresti 2001).

The reader should not think that the probability of occurrence of such ambiguous contexts is rare, and that a syntactic conception of the reference units could be in general maintained in any case. The number of verbless utterances is high, and dialogic turns with more than one utterance are the majority. Therefore given the possibility of recording mono-rematic utterances in most language categories, if the prosodic structure is not taken into account, the syntactic under-determinacy of the unit of reference is a

[6] Although this is not the right place for a discussion on the matter, it should be clear that bound morphology and functional categories (conjunctions, prepositions, articles, auxiliaries) cannot occur as monorematic utterances.

highly frequent phenomenon in the analysis of spoken texts. This creates serious problems that can cause the low reliability of syntactic information and therefore a serious limit to the exploitation of spoken resources.[7]

This problem is mainly due to the freedom of most words for what regards their performance both as monorematic utterances or within a syntagmatic category. For example a sentential adverb like *si* [yes] it is not necessarily a monorematic speech act, as is frequently the case in spoken dialogues, but can be syntactically linked to a deictic pronoun like *qui* [here]. However a pronoun like *qui* [here] can be also an independent utterance. Therefore the turn IT1D could be ambiguous: a) two words of a single utterance; b) two independent monorematic utterances:

a) Qui si // [Here / ok //] adverbial predication

b) Qui // Si // [Here // Yes //] two utterances

But such ambiguity is again a consequence of the wrong theoretical model that defines the unit of reference as a function of syntactic considerations. The structure of the turn is not ambiguous. Only the first alternative is the possible structure, due to the fact that "qui" is a question and *si* is an assertion. Therefore they cannot be part of the same utterance. This is made clear by their general prosodic features and, for what regards this framework, from the specific prosodic property that allows to perceive a prosodic break with terminal value between the two elements.

The opposite occurs in FR14 that could be parsed again either in two independent utterances or as an utterance in which the 'carrot' and 'yes' are combined within the same linguistic unit.[8] Prosody shows that such an ambiguity is not a genuine linguistic datum given that the two elements together perform a single speech act.

[4] la carotte ouais // [the carrot yes //] Adverbial predication

[4] la carotte // ouais // [the carrot // Yes //] Two utterances

A similar ambiguity can arise with locative expressions, that can function both as a pronoun or as prepositions, as e.g. 'qui'[here] and 'sopra' [up]. On this basis, as in the example above we may assign to IT14 two possible structures according to the possibility that: a) "qui sopra" is considered as a

[7] The Spoken Resources delivered in the 90' suffer this kind of problem. See for examples the spoken part of the British National Corpus for English or the LIP (De Mauro et al. 1993) for Italian.

[8] The reader may notice that "the carrot yes" may hardly be considered a constituent. Although we cannot deal with this kind of compositional structure here, we can mention that it is a Comment — Appendix utterance according to the *Informational Patterning Theory* (Cresti 2000 and, for the main references Moneglia in this volume).

possible deictic multiword adverb (‘up here’); b) the turn is a sequence of monorematic deictic expressions (‘here’ and ‘up there’).

a)	Qui sopra //	[up here //]
b)	Qui // sopra //	[here // up here //]

Again the theoretical ambiguity does not exist, given that the first alternative is the reality and only one speech act is accomplished.

The parsing of the flow of speech may also present more complex cases. In some languages, such as Portuguese, a sentential adverb like the English ‘no’ have the same form of the corresponding negative verbal modification. As a consequence PT1 [2] and [3] may give rise to an alternative interpretation :

*NOE: não // há muito tempo que não vou // …
[no // it’s been a long time since I last went]
*NOE não há muito tempo que não vou // …
[* it’s not been a long time since I last went]

Again, the identification of the utterance boundaries through terminal prosodic breaks rules out the ambiguity. Only the first alternative is possible, given that [2] is concluded by a terminal break and performs an utterance. As far as one word carries out one act, then it cannot be a structural part of a larger speech act, and the possibility in b) does not arise.

3.3.3. Prosodic boundaries and syntactic relations

The possibility to find strict syntactic relations between words in the speech performance is strongly determined by prosodic boundaries which mark the utterance boundaries. Also for this reason the underdeterminacy of syntactic structures in spontaneous speech is not only linked to the absence of verbs. When a verbal utterance may be figured out from the speech data, this not always provides the actual structure. For example ‘adverbials’ can be interpreted in principle both as independent ‘added clauses’ or as ‘adverbial clause’, depending on a verbal construction. The turns FR2A and PT1A, reported below in a bare transcription without prosodic tagging, have the same superficial structure, that is a verb followed by an adverb.

A*EMA : ça c’ est clair de plus en plus	[this is clear more and more]
A*NOE : estive no Chiado há pouco tempo	[I have been in the Chado recently]

A verbal nucleus with an adverbial extension as in the first structural alternative (a) may indeed constitute a well formed utterance from both a syntactic and a semantic point of view. However also the second alternative (b), that is a sentence followed by a monorematic utterance, is a possible structure that is coherent with the semantic and pragmatic information in the given context:

a) EMA: [1] ça c' est clair de plus en plus // #
[this is more and more clear //]

b) EMA: [1] ça c' est clair // # [2] de plus en plus // #
[this is clear // more and more //]

Therefore, it is necessary to point out that, according to the speech performance, only the second is the real structure. Each utterance is complete and autonomous from a pragmatic point of view and is marked by a terminal prosodic break. All competent speakers will agree that the turn must be divided into two utterances, each one accomplishing one speech act. Therefore in no circumstances will the adverb modify the verb, as it gives rise to an independent reference unit.

Of course also the opposite can occur. On the basis of purely syntactic considerations, the Portuguese example PT1A shows the same alternative, also coherent on syntactic and semantic ground, making both the following structures in principle possible:

a) *NOE: estive no Chiado // há pouco tempo //
[I have been in the Chado // recently //]

b) *NOE: estive no Chiado / há pouco tempo //
[I have been in the Chado / recently //]

However, it is the access to the prosodic information that determines the structure. Only one utterance is performed and the adverbial does not give rise to an autonomous utterance. Provided the prosodic break between the two constituents is not terminal the input data are not ambiguous and the adverb modifies the verbal clause.

The knowledge of utterance boundaries is essential when the syntactic structure must be determined. Two words cannot be part of the same constituent structure when they are performed in two distinct utterances. For example a relative pronoun necessarily changes its syntactic value according to its position. Let's consider SP1 in its bare transcription:

MAF: … ya saben todas esas cosas que generalmente les ofrecemos …
[They already know all these things that we generally offer them]

The syntactic analysis of this piece of talk does not seem problematic. A restrictive relative seems to modify the head 'cosas'. But this analysis is misleading. Once the audio signal is provided the listener knows that this text is divided in two utterances and the restrictive interpretation is not allowed. In other terms the relative pronoun still refers, from a pragmatic point of view, to ‘cosas’, but it does not modify the syntax of the previous utterance and therefore does not establish a restrictive relation:

MAF: … [1] ya saben / todas esas cosas // [2] que generalmente les ofrecemos // …
[They already know / all these things // that we generally offer them]

Another case arises frequently in spontaneous speech. A verb can occur in spoken language both projecting its argument structure or not. The ambiguity can be complicated even more by those verbs that are also used as *discourse markers*. For example let’s consider IT1[4] again in its bare transcription:

Ascolta qui sopra [Listen up here.]

The verb *ascolta* [‘listen’ in its third singular imperative mode] could be considered: a) a plain verb with a locative reference argument (*qui sopra*) giving rise to a *directive* utterance; b) as a monorematic *conative* utterance, followed by a second (*directive*) nominal utterance; c) as a discourse marker with the value of *phathic*.

The first two alternatives are not consistent with the prosodic performance. Listening to the acoustic source it is easy to verify that the verb does not give rise to an autonomous utterance, as it might have been when followed by a terminal break (like in b).

a) Ascolta qui sopra // [Listen up here //]
b) Ascolta // qui sopra // [Listen // up here //]

This confirms our assumptions. If the presence of a terminal break determines the utterance boundaries, the absence of a terminal break prevents considering a stretch of speech as a speech act. However, quite surprisingly, the first alternative does not apply for the listener. In order to give rise to a verbal clause, the verb and its reference argument should have been performed within the same prosodic envelope as in a). On the contrary the verb is followed by a non terminal break, as in c), and it is really hard to split a verb from its direct argument:

c) Ascolta / qui sopra // [Liste / up here //]

Therefore only the third alternative is presented to the listener and the verb *ascolta* is interpreted as a discourse marker, that does not project an argument structure.

Non-terminal breaks can also be exploited in the linguistic analysis of speech. As many studies on the topic have shown (Cresti 2000, Frosali 2005, Panunzi et al. 2004), in order to be able to interpret a word as a discourse marker, this must be directly followed by a non-terminal break.

3.4. Speech acts and terminal breaks

In the above discussion we have identified speech acts in a continuum through their prosodic boundaries. We have seen that if a terminal break is perceived in the speech flow, then a speech event occurs which can be pragmatically interpreted as a speech act. We have also seen that the speech act boundaries so obtained specify the linguistic environment in which syntactic relations hold. This practice, however, goes in parallel with the identification of the corresponding speech acts, eventually in accordance with a *Speech act annotation scheme*. The annotation of speech acts should in principle be equivalent to the tagging of prosodic breaks as regards utterance boundaries.[9] For example IT1 could have been parsed into utterances, assigning an illocutionary label to each relevant string of text, as in the following:[10]

IT1
A*EST: [1] o vieni / dai // [come on then]
%ill: [1] invitation
B*CLA: [2] a patire // [to suffer]
%ill: [2] ironical assertion
C*EST: [3] no // [4] ascolta / qui sopra ? [5] sì //
[no // listen / (what about) up here ? yes //]
%ill: [3] reassurance; [4] question, introduced by a *phathic*
[5] self answer
D*CLA: [6] qui ? [7] sì // [here ? yes]
%ill: [6] question; [7] self answer

In other words, the annotation of speech acts does not strictly require a prosodic constraint. The annotators might have reached the same result as a function of their recognition of speech acts, with no reference to prosodic boundaries. Indeed, current annotation schemes do not require a preliminary prosodic analysis.[11]

Under these premises it is important to recognize that, although they should lead to equivalent results, the two tagging activities are not equivalent.

[9] The first scheme for dialogue acts was worked out by Sinclair and Coulthard 1975, for the analysis of classroom conversations, then developed in Stenström 1994. The main general schemes have been developed for Map Tasking (Carletta et al. 1997) and DAMSL (Discourse Representation Initiative). See also. Usami 2005 and Cresti in this volume for the LABLITA approach and discussion

[10] Speech act labels for each ranked utterance are reported in the dependent line “%ill:” in accordance with the LABLITA annotation scheme. See Cresti in this volume.

[11] Of course this is not our position. In our approach speech act recognition is significant only afterwards

While speech act recognition is a matter of ‘categorization’, the annotation of prosodic boundaries is a matter of ‘perception’.

This has some consequences on tagging. The various coding schemes for dialogue acts provide the annotator with a closed list of possibilities (ranging from 20 to 100 options), but it may still be difficult to categorize the actual act performed in a speech act. The replicability of the coding scheme is, as a matter of fact, one of the main problems for the annotation of dialogue acts, even in restricted domains such as *map tasking* (Isard & Carletta 1995). These difficulties are even more sensible if open-domain spontaneous speech performances are taken into account, given that the variety of speech acts increases the complexity of the scheme.

The degree of uncertainty regarding the categorization of speech acts is surprisingly high for humans. For example, [6] is easily categorized as a question, but the illocutionary value of [3] and [7], despite their absolutely normal form, is not easy to recognize and there is no obvious agreement among annotators on these values.

On the contrary, as the validation of the C-ORAL-ROM prosodic tagging shows, there is little uncertainty regarding the presence of terminal breaks, given that no choice of categorization is foreseen.

The C-ORAL-ROM method and the annotation of dialogue acts identify two independent levels of analysis of the linguistic performance: the identification of a stretch of speech as a speech act on one side, and the understanding of its specific illocutionary value on the other. This distinction strictly corresponds to the corpus annotation experience: while it is true that [3] and [7] are hardly categorized as specific type of speech act, there is no uncertainty in marking [3] and [7] as two distinct speech acts.

Therefore, in practice, we can verify that the recognition of utterance limits, which is strongly determined, is independently motivated, with respect to the categorization of illocutionary force, that has little determinacy. This confirms C-ORAL-ROM’s assumptions: the definition of utterance limits is a matter of direct perception of prosodic boundaries, while the assignment of a specific value to a dialogue act is a categorisation issue.

For this reason, although the recognition of utterance boundaries through prosodic cues and the recognition of illocutionary acts indeed go hand in hand, the two activities should not be equivalent from a corpus annotation point of view. Once the relation between prosodic cues and speech act performances is recognised, the parsing of the speech flow into discrete speech events does not rely on the recognition by the labeller of a specific performed action. It is rather the reverse: the recognition of the speech act type performed depends on the previous definition of the utterance boundaries. Future research will confirm whether the preliminary identification of utterance boundaries can improve the performance of coding schemes for speech act recognition.

Conclusions

Concluding, the prosodic structure highlighted by terminal and even non terminal prosodic breaks is the index which determines the choice of the possible structure, not the opposite. The structural ambiguity that emerges in the analysis of speech on the basis of the sole syntactic consideration is a not a genuine datum, but rather a consequence of a wrong choice regarding the nature of the reference unit of spontaneous speech.

Speech act-based reference units feature a clear and easy cue that may be used for their detection in the speech flow. The analysis of false structural ambiguities in speech gives an important theoretical confirmation to the hypothesis that terminal breaks are one of the most reliable indexes of speech act accomplishment: if there is no terminal break, then the interpretation of a stretch of speech as a speech act is not allowed. If there is a terminal break, a speech act interpretation is allowed and no syntactic relation can be held beyond the utterance boundaries.

Many scholars presently agree that the access to speech data is essential in order to eliminate trouble in the syntactic analysis of speech. All the problems seen above do not arise if the speech source is accessible. However people are also reluctant to mark the transcripts with prosodic breaks, as this can be considered uncertain due to its relying on perceptual judgements. But why should we avoid to mark an essential information?

The consequence of not marking terminal breaks is that the contribution of this information to the syntactic analysis cannot be recovered and that therefore its resulting identification by the corpus annotators is less reliable. On the contrary, the validation of the C-ORAL-ROM prosodic tagging shows that it is very reliable.

Finally, this information is also necessary for linking the acoustic information and the linguistic annotation in a way that is uniform and significant from a linguistic point of view. If the unit of reference is not defined, the consequent alignment is arbitrary, and the speech performance’s acoustic information cannot be linked to the relevant linguistic information. Therefore, not marking the utterance boundaries causes an inconsistency that is a very serious concern for the exploitation of the acoustic information. The utterance-based synchronization developed in C-ORAL-ROM solves this problem in a uniform and easy manner, testified by the text-to-speech alignment of 134,000 utterances in the multilingual corpus: *each transcribed utterance is aligned to its acoustic counterpart*.

References

Austin, L.J., 1962. *How to Do Things with Words*. Oxford: Oxford University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. 1999. *The Longman Grammar of Spoken and Written English*. London and New York: Longman.

Blanche-Benveniste, C. 1997. *Approches de la Langue Parlée en Français*. Paris: Ophrys.

Buhmann, J., Caspers, J., van Heuven, V., Hoekstra, H. Martens, J-P., Swerts, M., 2002. "Annotation of prominent words, prosodic boundaries and segmental lengthening by no-expert transcribers in the spoken Dutch corpus". In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2002)*, M. C. Rodriguez and C. Suarez Araujo (eds), 779-785. Paris: ELRA.

British National Corpus http://www.natcorp.ox.ac.uk/

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. and Anderson, A. 1997. "The reliability of a dialogue structure coding scheme". *Computational Linguistics* 23(1):13-31.

CHAT <http://childes.psy.cmu.edu/manuals/CHAT.pdf>

Contextes http://www.up.univ-mrs.fr/veronis/logiciels/Contextes/index-en.html

Cresti, E. 2000. *Corpus di Italiano Parlato*. Firenze: Accademia della Crusca.

Cresti, E. In this volume. Some comparisons between UBLI and C-ORAL-ROM.

Cresti, E. and Firenzuoli, V. 2001. *Illocution and intonational contours in Italian*, in htpp://lablita.dit.unifi.it./preprint/preprint-01coll04.pdf. Also published in *Revue Française de Linguistique Appliquée* IV(2): 77-98.

Cresti, E. and Firenzuoli, V. 2002. "L'articolazione informativa topic-comment e comment-appendice: Correlati intonativi". A. Regnicoli (ed.) *La Fonetica Acustica come Strumento di Analisi della Variazione Linguistica in Italia*. (Atti delle XII Giornate del Gruppo di Fonetica Sperimentale), Roma:Il Calamo 2002.153-160.

Cresti, E. & Moneglia, M. (eds.). 2005. *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Language*s. Amsterdam:John Benjamins.

Crystal, D. 1975. *The English Tone of Voice*. London: Edward Arnold.

Danieli, M, Garrido, J. M.; Moneglia, M.; Panizza, A, Quazza, S., Swerts, M. (2004) "Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech "C-ORAL-ROM" in M.T Lino, M.F. Xavier, F. Ferraira, R. Costa, R. Silva (eds) *Prococeedings of the 4^th^ LREC Conference*, ELRA, Paris, vol. 4 pp.

1513-1516.

De Mauro, T., Mancini, F., Vedovelli, M. and Voghera, M. 1993. *Lessico di Frequenza dell'Italiano Parlato.* Milano: ETAS.

Frosali, F. (2005), *Le unità di informazione di Ausilio dialogico: valori percentuali, caratteri intonativi, lessicali e morfosintattici in un corpus di italiano parlato* (C-ORAL-ROM), Master Thesis, Florence: Università degli studi di Firenze.

IMDI <http://www.mpi.nl/IMDI/>

Isard, A. and Carletta, J. 1995. "Replicability of transaction and action coding in the Map Task corpus". In *Empirical Methods in Discourse Interpretation and Generation: Working Notes of the AAAI Spring Symposium Series*, M. Walker and J. Moore (eds), 60-66. Stanford, Calif.: Stanford University.

Izre'el, S. 2005. "Intonation Units and the Structure of Spontaneous Spoken Language: A View from Hebrew". Cyril Auran, Roxanne Bertrand, Catherine Chanet, Annie Colas, Albert Di Cristo, Cristel Portes, Alain Reynier and Monique Vion (eds.), Proceedings of the IDP05 International Symposium on Discourse-Prosody Interfaces. <http://aune.lpl.univ-aix.fr/~prodige/idp05/actes/izreel.pdf>

Leech, G., Rayson, P. & Wilson, A, 2001. *Word Frequencies in Written and Spoken English*, London:Longman.

Karcevsky, S. 1931. "Sur la phonologie de la phrase". *Travaux du Cercle Linguistique de Prague* IV: 188-228.

MacWhinney, B. 1994. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Miller, J. and Weinert, R. 1998. *Spontaneous Spoken Language*. Oxford: Clarendon Press.

Moneglia, M. 2004. *Specifications of the C-ORAL-ROM corpus.* Paris:ELRA.

Moneglia, M. In this volume. Units of Analysis of Spontaneous Speech and Speech Variation in a Cross-linguistic Perspective.

Moneglia M. and Cresti, E. 2001. "The value of prosody in the transition to complex utterances. Data and theoretical implications from the acquisition of Italian". Almgrem, B. Barrena, M. J. Ezeizabarrena, I. Idiazabal, B. Mac Whinney (eds.) *Proceedings VIIIth International Congress IASCL, (12-16 luglio 1999, S. Sebastian),* Chicago: Cascadilla Press 2001. 851-873.

Oostdijk, N., Kristoffersen, G. Sampson, G. (Workshop organizers) 2004. "Compiling and Processing Spoken Language Corpora". M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva (eds), *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 563-566. Paris: ELRA.

Panunzi, A. Picchi, E., and Moneglia, M. 2004. “Using Pitagger for lemmatization and PoS tagging of a spontaneous speech corpus: C-Oral-Rom Italian”. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva (eds), 563-566. Paris: ELRA.

Quirk, R. S. Greenbaum, G. Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Sinclair, J.M. and Coulthard, R.M. 1975. *Towards of Analysis of Discourse: The English Used by Teachers and Pupils.* London: Oxford UP.

Scarano, A. To appear. “Prosodic annotation in speech resources. The C-ORAL-ROM corpus in linguistic research and the teaching of languages”. *Corpus Linguistics Studies*.

Sinclair, J. 1996. Preliminary Recommendations on Corpus Typology. EAGLES Document EAG-TCWG-CTYP/P, May 1996.

Sinclair, J. 2005 *The history of a corpus*. International Summer School of Corpus Linguistics, Tuscan Word Centre, Pontignano, Siena. Ms.

Sinclair, J.M. and Coulthard, R.M. 1975. *Towards of Analysis of Discourse: The English Used by Teachers and Pupils.* London: Oxford UP.

Sorianello, P. Forthcoming. “Per una definizione fonetica e fonologica dei confine prosodici”. *La Comunicazione Parlata* (International Congress. Naples 23-25 February 2006).

Stenström, A.-B. 1994. *An Introduction to Spoken Interaction*. London: Longman.

TEI <http://www.tei-c.org/Guidelines2/ >

Usami, M. 2005. “Why do we need to analyze natural conversation data in developing conversation teaching materials? Some implications for developing TUFS language modules”. Kawaguchi et al. *Usage-Based Linguistics Informatics*. Amsterdam Benjamins 2005. 279-294.

Win Pitch Corpus <http://www.winpitch.com/>

Conclusion

Susumu ZAIMA

Ladies and gentlemen!

First, I would like to express my sincere gratitude to all the participants of the Second International Conference on Linguistic Informatics. I would especially like to thank those participants who presented a wide range of valuable research reports.

As you are already aware, the goal of our university's COE Program "Usage-Based Linguistic Informatics" is to integrate the fields of linguistics, applied linguistics, and computer sciences and to create a new discipline called "linguistic informatics."

The goal of this international conference has been to clarify the theoretical foundation of this new discipline of linguistic informatics.

I think it is not only impossible but also premature to summarize the extent to which this conference has succeeded in clarifying the theoretical groundwork for linguistic informatics. Therefore, in closing this conference, I would instead like to say a few words on the importance that this COE Program holds for our university.

Research and education at our university covers over 50 languages, 26 of which are taught as major languages. From the nature of our university, what we seek from our linguistic research is not the pursuit of the universality of language but analyses of the actual "language use" of each language and the application of these results to language education.

Based on this kind of awareness, till date, we have been collecting data on language use by using cards and other such means.

This primitive state of affairs has been transformed by the recent progress of IT. The advancements in IT have enabled us not only to collect massive amounts of linguistic data and apply the results to language education but also to develop a new language education system, an e-learning system.

Our long-standing dream has been the integration of linguistics with language education, and amid the progressive trend of IT, this dream may become a reality through the fulfillment of this COE Program. The fulfillment of this project is undoubtedly a crucial issue related to the very existence of our university. Therefore, I would like to extend a sincere request to our guests from overseas to help us in this purpose in the future as well.

Please allow me to briefly express my thoughts on the use of language.

Over the past few decades, a wide range of linguistic theories have been put forward and a considerable amount of active research has been conducted. However, I would say that the characteristic trend at present is an emphasis on language use. Even in generative grammar, language use is now clearly positioned as a research subject. Furthermore, cognitive linguistics is a study that attempts to explain language use from the perspective of human cognition.

A language has existential value as a language only if it is used, and analysis of a language is possible only if there exists data on its use. Therefore, the importance of corpus analysis is only natural; however, this importance is not only of the spoken language but also of the written language.

Most of the presentations made during this conference on corpus analysis were by guests from overseas. I hope that our next conference will focus on presentations made by our own researchers and will be an opportunity for us to present to our guest from overseas the results of the research conducted at our university.

While believing that our dream will be realized, I would like to conclude my remarks by extending my sincere appreciation to the faculty (especially Mr. Kawaguchi and Mr. Takagaki) and students for their efforts in preparing and conducting this international conference, and expressing once more my sincere thanks to our guests from overseas.

Thank you.

2.
Workshop on Spoken Language Corpora
—C-ORAL-ROM and UBLI—

Introduction

Toshihiro TAKAGAKI *(Tokyo University of Foreign Studies)*

The present workshop was realized thanks to some fortunate coincidence and friendly cooperation. In 2004, I was on a sabbatical leave and had the opportunity to conduct research at the Universidad Autónoma de Madrid (UAM) in Spain. During the same year, the first Japanese department in Spain was launched at the University, which led to an interchange agreement between the University and Tokyo University of Foreign Studies. From the following year, a student exchange program was begun. TUFS also started sending young researchers to the Japanese department of the Universidad Autónoma de Madrid. Thus, close educational and scientific ties were established. During the same period, in the UBLI framework, two of my graduate students and I carried out field research to record more than thirty hours of spontaneous conversations in order to construct spoken Spanish corpus. Dr. Antonio Moreno-Sandoval of UAM was kind enough to give us full support for collecting spoken Spanish data. It was during a friendly conversation with him that I became familiar with the C-ORAL-ROM project and found a strong similarity of interest between UBLI and C-ORAL-ROM. I became aware of the importance of establishing a cooperative relationship between the two research projects. In this respect, I want to express my gratitude toward Dr. Moreno who acted as an invaluable mediator in helping us to get in contact with the other members of the C-ORAL-ROM. In this way, it was possible for us to plan and organize the collaborative workshop between UBLI and C-ORAL-ROM. This chapter contains nine contributions presented at the Workshop held at TUFS on December 10, 2005. The papers of the C-ORAL-ROM members are presented in the first part, and the papers of the UBLI members consist the second.

1. C-ORAL-ROM

Emanuela Cresti’s contribution, “Some Comparisons between UBLI and C-ORAL-ROM,” is dedicated to comparing the approach to spoken language adopted by the UBLI Center and C-ORAL-ROM-LABLITA, which is very similar with respect to the key issues. The paper focuses on the comparison of the reference units for the analysis of spoken language chosen by UBLI (/function/ and /discourse-sentence/) and those chosen by LABLITA and C-ORAL-ROM (/speech act/ and /utterance/) and highlights the differences that are largely due to the accent given to the prosodic features in the C-ORAL- ROM-LABLITA approach.

In “Units of Analysis of Spontaneous Speech and Speech Variation in a Cross-linguistic Perspective,” Massimo Moneglia presents various measurements of speech performance in the Romance languages, derived from the C-ORAL-ROM corpus and focuses on the strong variability of linguistic behavior. His paper identifies the cross-linguistic correlations between the general language properties of speech and the specific features of the context in which the performance takes place. This leads him to the conclusion that essential spoken language behaviors are determined by contextual factors.

José Deulofeu and Claire Blanche-Benveniste, in their “C-Oral-Rom —French Corpus—,” present some contributions for the French topics included in the C-ORAL-ROM. Setting up major distinctions of registers, they find interesting oppositions among the linking particles: *parce que* is closer to coordinators like *et* and *mais,* while *que* is radically distinct for its clear prosodic integration of the subordinate. The infrequency of the verbless utterances in spoken French is discussed in both quantitatively as well as qualitatively. Finally, based on the “macrosyntactic” approach, they select some frequent marco-syntactic patterns observable in nucleus and multinucleus utterances.

In “Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus —Methodology, Tools and Evaluation—”, Antonio Moreno-Sandoval, and José M.Guirao summarize the experience of the LLI-UAM group in tagging one of the largest spontaneous Spanish speech corpora now available (over 300,000 transcribed words). First, they describe some tagging problems particularly relevant in spoken corpora. The tagging procedure and the tool developed for helping human annotators in the process are then introduced. Finally, an evaluation of the precision rate provided by the tagger is calculated based on a “gold standard” corpus of 150,000 words.

In “The Role of Spoken Corpora in Teaching/Learning Portuguese as a Foreign Language—The Case of Adjectives Intensification—,” Maria Fernanda Bacelar do Nascimento and José Bettencourt Gonçalves insist on the importance of a corpus-based analysis in order to furnish the learners of Portuguese as a foreign language with authentic spontaneous speech data. Further, as an example of corpus analysis, they present the adverb ending with *–mente.* From a statistical viewpoint, they remark that on the one hand, its collocational characteristics, i.e., *devidamente* “duely,” are used exclusively with participles, and on the other hand, its strong association with specific thematic domain, i.e., *altamente confidenciais*, is used with administrative or political domain. In case the collocational relationship becomes maximally fixed, the pluriverbal unit becomes completely lexicalized and memorized as a single unit.

Maria Fernanda Bacelar do Nascimento and Amália Mendes and Sandra Antunes, in their "Typologies of MultiWord Expressions Revisited —A Corpus-driven Approach—,[1]" claims that multiword expressions (MWEs) have been and are still a challenge in linguistic analysis, lexicography, and natural language processing. In fact, several typologies of MWEs have been proposed taking into account several parameters, for example, their degree of cohesion, internal variation, and compositional nature. They add that the definition of an MWE is still controversial, and it appears that typologies based on discrete categorization fail to describe a phenomenon with such variation. In this paper, they plan to revise some typologies of MWEs by using a corpus-driven approach and analyze corpus findings and their relation to MWEs categorization.

2. UBLI

In "Usage-Based Approach to Linguistic Variation —Evidence from French and Turkish—," Yuji Kawaguchi states that the concept of norm at the early stage of linguistics can be considered as an issue entrusted to future development. Although a linguist like Hjelmslev underestimated norm, the importance of norm in the study of linguistic usage should not be dismissed. Using French and Turkish examples, he overviews the problems of linguistic variation at every level of language structure, i.e., phonetic, morphological, lexical, and syntactic. Usages are composed of the massive habits of language users. The possible changes in usage constitute a very interesting domain for the analysis of dynamic synchrony, which describes the ongoing variation in a given language community. Usages should be described quantitatively as well as qualitatively and he claims that the approaches for the analysis of linguistic variation should be corpus based.

In "Viewpoint and Postrheme in Spoken Turkish," using several modern Turkish corpora, Selim Yılmaz observes and analyzes the syntactic and semantic productions of the modo-enunciative proceedings involved by viewpoint markers and the postrheme, which shows the discursive strategy of the speaker in spoken dialogue. The analysis leads him to postulate subjectivity and assertion as modal values, whereas assuming what is said by the speaker, either in a consensus with the listener or in an egocentered position, will be considered as an enunciative, a value which defines the

[1] Since she was unable to attend the Workshop, Maria Fernanda Bacelar do Nascimento sent us her paper for the workshop.

position of the enunciator in the interaction. The enunciative position of the speaker is clearly defined by the type of use and the intonation of viewpoint markers and the postrhemes, according to the context. This confirms as well the explicit function of these modality marks in terms of enunciation.

According to what type of referent Malay demonstrative pronouns refer to, Isamu Shoho, in his "Nonreferential Use of Demonstrative Pronouns in Colloquial Malay," finds the following three uses for them: (1) referring to objects in the real world, (2) referring to what has been said before or will be said later, and (3) referring to mental images. Apart from these three uses, there can be another use. In this fourth use (he refers to this as the nonreferential use henceforth), what has been pointed to by the demonstrative pronouns is vague, or they do not refer to anything. The function fulfilled by demonstrative pronouns in nonreferential use is that of expressing feelings accompanying the sentence. He divides this nonreferential use into eight categories including the use for rhyming purposes.

2.1.
C-ORAL-ROM

Some Comparisons between UBLI and C-ORAL-ROM

Emanuela CRESTI

1. Foreword

The presentation of the Centre for Usage-Based Linguistic Informatics (UBLI) in the introductory article (Kawaguchi 2005), in the first of the two volumes dedicated to the Proceedings of the First International Conference on Linguistic Informatics, allowed us to become acquainted with the imposing work carried out by the Centre.[1] As a result, we came into even closer contact with this truly innovative and interesting foreign-language teaching system. We were pleased to discover that the experimental research direction such as that adopted in our laboratory LABLITA[2] which have been developed within the C-ORAL-ROM Project, was inspired by principles similar to those used by UBLI, and have led to results which are both similar and comparable.

In this contribution we will focus on the comparison between the reference units for the analysis of spoken language chosen by UBLI (*function* and *discourse-sentence*) and those chosen by LABLITA and C-ORAL-ROM (*speech act* and *utterance*). Beyond the similarity of their nature, some differences emerge mostly due to a different accent given to prosodic features, which are one of the most relevant parameter employed by LABLITA in the processing of spoken language.

1.1. A shared Approach

The article by M. Usami (2005)[3] points out the need for “natural

[1] The article “Centre of Usage-Based Linguistics Informatics (UBLI)” and other contributions from several collaborators at the Centre illustrate some foundational aspects of the UBLI system: these are our terms of reference and the object of our comparison.

[2] The Linguistic Laboratory of the Department of Italian (LABLITA) is devoted to the study of Italian spoken language on the basis of corpora of spontaneous speech. It develops studies on intonation of Italian, according to specific theoretical and experimental methods. It archives the most important spoken Italian resource; its database is made up of open corpora of spontaneous adult spoken language, radio and television collections and of large longitudinal corpora of Italian acquisition. See the web site http://lablita.dit.unifi.it

[3] The article “Why do we need to analyze natural conversation data in developing conversation teaching materials? some implications for developing TUFS language modules” will be one of our main references for comparison.

conversation data in teaching materials”. Actually, when the teaching and learning of a foreign language truly aims towards a real usage capability, it cannot do so unless it involves comparisons with spontaneous spoken interaction. However, perhaps, the observation we most agree with, is the emphasis placed by the author, and in general by UBLI, on the need for learners to become familiar with and to acquire the linguistic behaviour of a new language, rather than limit themselves to learning a series of “inert linguistic formulae”. Here we want to stress the theoretical importance of the conception of language as a process over its structural and configuration aspects, as it is typical of traditional studies based on competence.

Apart from theoretical – but also commonsense – considerations about the need to resort to spontaneous language in teaching, we can also mention two aspects which have been identified on the basis of teaching and research experiences:

a) the diversity of any real linguistic interaction whatsoever with respect to artificially reconstructed models;
b) the variation of exchange typology in different diaphasic contexts.

Particularly enlightening is Usami’s accurate comparison between the D-Module Skit – “Making a reservation”, and the recorded conversation of a comparable situation, a phone call for information about a reservation (Usami 2005:286-289). It points out several fundamental linguistic differences between the skit – which apparently seems to have been conceived of in an impeccable manner – and the real event. In particular, the following points have been observed in the recorded conversation: it is longer, has a high number of fillers and discourse markers, forms of self-repair, the same expression is used for multiple functions, unexpected archaic forms, and unexpected forms of linguistic interaction. These points make the recorded conversation substantially different from that of the module. Naturally, this does not at all mean that language teaching should abandon invented skits; rather, it highlights their complementary function when compared with natural conversation data.

With regard to the different performances of the same interaction depending on diaphasic contexts, as observed by Usami, linguistic behaviour has different realisations: typical forms and direct strategies rather than ellipses, or absence of typical forms, or indirect strategies, depending on the type of relation which exists among speakers (e.g., friends vs. strangers), that is to say in relation to primarily diaphasic traits.

Similar observations also emerged during the course of LABLITA research-studies, which are documented in C-ORAL-ROM.[4]

Summarizing, with regard to UBLI, it seems possible to conclude that linguistic behaviour is considered as a *relational usage of language*, which can be represented by a corpus if based on a diaphasic variation. The corpus design on which the four Romance languages have been sampled is sufficient to account for their systematic diaphasic variation, so a corpus like C-ORAL-ROM seems particularly suited to the objectives of UBLI, as it provides learners with a richness of resources which guarantee the representation of the social and relational usage of spoken language.[5] On the other hand, a series of research studies carried out by LABLITA, specifically on spoken Italian, has found aspects which, in our opinion, can also be of interest in language teaching, and we hope that they can complement the new L2 teaching strategies adopted by UBLI.

2. The Reference Units in 'Talk That Works' (TTW) and in LABLITA

Among the various contributions in the cited volume, we have considered those which, from different viewpoints, describe noteworthy aspects of UBLI, and have enabled us to form an idea of the general approach adopted in teaching. Apart from the overall agreement, as already stated, which is linked to a conception of language seen as linguistic behaviour, and, hence, the interest shown for language as a process, we were struck positively by the similarity of the reference units chosen for the analysis of discourse processing.

Usami (2005) when presenting the TTW training kit,[6] emphasises that, besides offering a fresh perspective on the nature of conversation, it also allows us to focus on some central aspects of spoken language, such as discourse processing, pragmatic features, politeness strategies and repair strategies. Indeed, we regard discourse processing and pragmatic features as general traits of spoken communication and it is with them that we want to begin a comparison with the research carried out by LABLITA.

The entities at the basis of discourse processing have a pragmatic nature, and can be identified in what TTW defines as *functions* as well as in the

[4] On the results on variation within the C-ORAL-ROM corpus design see Moneglia (in this volume).

[5] See the corpus design presented by Moneglia in Chapter I of Cresti & Moneglia (2005) and Moneglia (in this volume). Both contributions adopt a viewpoint on the problem of representativeness for spontaneous spoken language and should also be read together with the arguments raised towards C-ORAL-ROM by Moreno Fernandez in "Corpus of spoken Spanish language – The representativeness Issue"(2005).

[6] The TTW training kit has been developed at the University of Wellington and is used in teaching by UBLI.

corresponding linguistic units, namely the *discourse sentences*. For a description of them, we also refer to the article "An analysis of teaching materials" (Suzuki, Matsumoto and Usami 2005:300). The corresponding entities for LABLITA are *speech acts* and *utterances*, and they also have a clear pragmatic definition.

With regard to the *functions* in TTW and the *speech acts* in LABLITA, their common theoretical ground in an Anglo-Saxon tradition of studies is evident.[7] Precisely, the work of Holmes (2005),[8] which mentions – more than once – the concept and the term "speech act", confirms our impression.[9] Moreover, it must also be added that one of the definitions of speech act used by the author literally reports the title of Austin's most important work on this subject: "How to do things with words".[10] Perhaps, Holmes's explicit assertion on the synonymy of the terms "speech acts" and "functions" is even more important.[11]

For what concerns LABLITA, it is possible to verify that the theoretical reference to Austin constitutes the basis of our research activity.

In conclusion, we maintain the substantial equivalence of the *functions* of TTW and the *speech acts* of LABLITA. These are concepts with which we intend the same type of linguistic activity which is at the basis of discourse processing.

In the previously-mentioned article by Suzuki, Matsumoto and Usami (2005) it is stated that *discourse-sentences* are those expressions that perform a function and that fulfil a *substantive function in the conversation*.[12] As regards utterances, LABLITA traces its first definition back to Cresti (1987),[13] which has been enriched over the years to reach this version:

> "any expression which can be pragmatically interpreted is an utterance …the utterance is identified by the accomplishment of an illocutionary act and it is based on the hypothesis that an equivalence exists between units of the domain of human actions (acts) and linguistic units (utterances)" (Cresti & Firenzuoli 2001)

[7] Cfr. Austin (1962) and Searle (1969).

[8] See "Socio-pragmatics aspects of workplace talk" which refers to the project "The Wellington Language in the Workplace (LWP)" which that developed TTW.

[9] See for instance the various locutions used in the article: "affective speech acts"; "speech acts of refusal"; "this range of different speech acts";.

[10] See "the skills we develop in learning how to do all these things with words" (Holmes 2005:196)

[11] See "the ways in which different speech acts or functions of speech are appropriately expressed in different culture" (Holmes 2005:197).

[12] Suzuki, Matsumoto and Usami (2005:299) footnote 6.

[13] The concept was taken up in Italy by Fava (1995). An illocutive definition of the utterances can be found also in Biber et al. (1999). For a similar theoretical approach, see also Jacobs (1984).

Also in this case, the nature of the concept is the same: linguistic units which are defined on the basis of their semantic-pragmatic functionality.

We are mostly interested in trying to determine the extent of the similarities between the Centre of UBLI and LABLITA; nevertheless, a point of divergence seems to arise, precisely in relation to the identification of linguistic units, *functions* and *speech acts, discourse-sentence* and *utterance*.

2.1. Functions and speech acts compared

In this section, we deal with the comparison and a more detailed analysis of the issues of discussion, starting with functions. The list in Yuki, Abe and Lin (2005),[14] shows that 50 types of functions have been identified, extracted from the Japanese language materials, and 71 types have been extracted overall from the other foreign languages (17) studied and taught by the UBLI Centre.

A reduced table, derived from the combination and comparison among the previous ones, has been proposed and it corresponds to 40 types. The analysis carried out on the basis of our Italian corpora has led to the identification of about 70 *speech acts*. We here report the table for UBLI as published in Yuki, Abe , Lin (2005) and for LABLITA in Cresti e Firenzuoli (2001):

Table 1. 40 Functions (Yuki, Abe & Lin 2005)

Greetings	Thanking	Attracting attention	Introducing oneself	Apologizing
Giving	Saying goodbye	Asking Information (price)	Asking Information (experience)	Telling one's plan
Asking Information (degree)	Asking Information (time)	Asking Information (number)	Saying how and why	Asking skill and ability
Asking information (existence and place)	Asking Information (attribute)	Saying one's opinion	Saying one's taste (thing)	Saying one's taste (behaviour)
Stating procedure and order	Asking what one is	Saying how one acts under certain circumstance	Comparing (comparative and superlative degree)	Suggesting
Explaining why	Asking	Exemplifying	Compromising	Asking for permission
Confirming duty/negating	Prohibiting	Instructing	Asking for unacceptable thing	Confirming duty / affirming
Inviting	Advising	Demanding	Stating one's hope	Introducing someone

[14] See Table 8 (p. 349), in the article "Development and assessment of TUFS dialogue module" also reported below.

Table 2. Speech acts taxonomy (Cresti & Firenzuoli, 2001)

REFUSAL	verification	denominative	doubt	permit	condolences
ASSERTION	hypothesis / supposition	announce	expr. of intent	derision	compliments
weak assertion	narration	direct speech reference	admission/ attenuation	provocation	condemnation
answer	story telling	total question	giving up	reprimand	promise
commentary	description	partial question	exclamation	hint	bet
explanation	list	alternative question	expr. of surprise	encouragement	baptism
declaration	quotation	request of information	expr. of fear	assurance	decl. of legal value
definition	comparison	request of action	expr. of relief	warning	
inference	hypothetical period	request of confirm.	expr. of wish	pity	
identification	DIRECTION	prohibition	expr. of uncertainly	RITE	
confirmation	distal recall	instruction	claim	thanks	
conclusion	recall in proximity	EXPRESSION	regret	greetings	
objection	distal deixis	expr. of contrast	complaint	apologies	
approval/dis.	deixis in proximity	expr. of disbelief	imprecation	wishes	
agreement/dis.	presentation of event	irony	insinuation	congratulation	

The points of similarity between the two repertories are many. From a quantitative point of view, the number of linguistic activities is comparable: 40 for UBLI and 67 for LABLITA, but it must be noticed that 90% of the Italian occurrences is covered by only 30 most common speech act types.[15] Both *functions* and *speech acts* are conventionally recognised and repeatedly adopted by various speakers. Besides, in several cases, UBLI functions and the LABLITA *speech acts* appear to coincide (*asking, answering, saying, ordering, stating, comparing, confirming, instructing, suggesting*, etc.).

As a matter of fact, a more important aspect exists that is common to both approaches: *functions* or *speech acts* have not been identified based on theoretical principles, as usually happens in the literature on the subject. The most important example is, undoubtedly, from Searle's taxonomy which, on the basis of a "translatability" principle establishes a correspondence between each *speech act* and a predicate, the head of which is a verbal, performative lexeme, which expresses the action in question. Thus, the

[15] Anyway the Italian list is open, and we go on discovering new types, carrying out our research.

universe of Searle's *speech acts* is constructed in a logic-and-lexicon-based way. However, when carrying out an experimental analysis of Italian corpora, we had the opportunity, on the one hand, to verify the inadequacy of such a taxonomy: the richness of the actions really carried out is not contemplated, and, moreover, several – even very common – *speech acts* are not identified in the taxonomy since they have no equivalent verbal performative (*refusal, deixis, call, introduction*). On the other hand, importance is given to linguistic actions that never occur in our corpora or, in any case, seem to be quite rare.

Unlike in the logical-philosophical tradition, LABLITA has identified *speech acts* through the analysis of the available Italian corpora, and, thus, it seems to us that the same has been done by UBLI with Japanese and other language corpora. Only at this point, after having pointed out what the common general characters for UBLI and LABLITA are (conception, extension, experimental identification), we can move on to the detailed analysis of the differences. As already pointed out, they seem to be related to the criteria for the identification of *speech acts* and *functions*.

Firstly, for LABLITA, as can be deduced from the table, there are five classes within which speech acts can be grouped (*refusal, assertive, directive, expressive, ritual*).[16] The five typological classes of *speech acts* are based on the attitude-related characteristics of human actions, and, in the final analysis, on the relational dynamics. The basic assumption is that any activation of someone as a speaker, is necessarily based on the affection which emerges in relation to the interlocutor. This is an unconscious aspect, completely free and not regulated by conventions.

Secondly, given this affective basis, the specific act performed by the speaker shows pragmatic characters (social, professional, knowledge-related, semiological, etc.). These characters enable us to identify the act in a conventional manner.

On the contrary we do not know whether or not a hierarchy should also exists for UBLI *functions*. The 40 UBLI *functions* seem to derive from a combination and comparison, because they are, actually, more general and seem to summarise the previous repertories, but they do not appear as hierarchically structured.

But the most relevant difference regards the way to identify the *function* type, which seems to us content-dependant. The function *asking information,* for instance, corresponds to several functions distinguished by their content: *asking price, experience, degree, time, number, existence and place, etc.* Another example is the function *saying*, which is distinguished depending on

[16] Reported in capital in Table 2.

whether it is used to say: *good-bye, how and why, one's taste, one's opinion, etc.* A given type of activity is considered more than once, taking in account its semantic sub-distinctions.

The article by Suzuki, Matsumoto and Usami (2005), reports a detailed analysis on 7 of the most frequent *functions*, as extracted from a corpus of conversation in English.[17] Their definitions link together both pragmatic traits – which identify the actual performed action – and semantic traits – which typically specify the restriction of the action in issue –. For example, when defining the function <ASKING FOR INFORMATION (*about attributes*)>, the part concerning the action is explained as "the speaker asks", and only the part concerning the action restriction (*for information about attributes*) adds information, but it is defined in ontological terms such as *attributes of a person or an object*. In turn, the attributes are explained as *quality that can be found usually … and which does not include a temporary state*. One of the most frequent functions such as <STATING AN OPINION> is explained in relation to the action part as *the speaker makes an assertion*, and the specific semantic restriction *statements do not involve the speaker's judgment* etc.

According to the LABLITA classification, we first consider what the speaker's attitude towards his/her interlocutor is, and, consequently, in what class the specific act should be inserted. Then, we determine the action-related characteristics in order to assess its conventional value. For example, the assertive class contains a speech act identified as *confirmation*, and, we have seen that UBLI also identifies the function <CONFIRMING>. However, for us a *confirmation* exists as a type of action and it takes part in the assertive class, specifically to weaken an assertion, because of its attitudinal features. So it is defined on the basis of an affective attitude which is common to all assertive speech acts (assertion: *an attitude of self-confidence which, on the basis of own realisations of thoughts, enable one to express a judgment or a knowledge as a new object in the world*). It is a *confirmation* independently from its topic, and, therefore, we do not operate further distinctions within this act when, for example, the *confirmation* concerns the existence of something, or a past action, or a price.

A very frequent speech act such as the *answer* is again defined on the basis of its affective attitude as an assertive act, but it can be distinguished from other assertive acts, for instance *conclusion*, *narration*, etc .., because it has specific action-related traits. For example an *answer*: a) is a cognitive

[17] The corpus analysed is 11 minutes and 25 seconds long and, as illustrated, is made up of 291 discourse-sentences (Suzuki, Matsumoto & Usami 2005:300-301).

operation with verbal outcome (while, for example a *conclusion* is the result of a comparison or experimental operation); b) necessarily occurs in connection with environmental traits such as *opening channel activated, attentional horizon present* and *common nature of attention focus* (whereas in a *conclusion* all these traits are negative or missing);[18] c) its semantic content is not relevant; d) its textual length it is not relevant (whereas it is relevant in a *narration*, which also belongs to assertive class, but requires a certain extension to be considered as such).

Hence, the *answer* is a specific speech act, which is conventional due to its action-related characteristics, but which belongs to the assertive class due to its attitude-related characteristics.

In our opinion, the research in this field should be extended and investigated in depth; it does not yet appear sufficiently explained. It could be interesting to develop a join research program on how actions are defined within discourse processing, how the pragmatic and semantic traits of their definition are connected and more generally, what should be the theoretical framework in which to elaborate a shared taxonomy.

2.2. Discourse-Sentences and Utterances Compared

As regards discourse-sentences, what we could draw from the reading of the various UBLI contributions reveals both explicitly asserted aspects and some implications that, in our opinion, can be derived thereof. The positive assumptions regard, on the one hand, the request that discourse-sentences have a syntactic nature and, on the other hand, the previously mentioned observation that they fulfil a “substantive function in the conversation”.

Another point which seemed very interesting to us, is the empirically-based annotation reporting that, from a configuration viewpoint, a discourse-sentence may correspond to a single-word sentence, to an incomplete sentence, and/or to a structurally complete sentence.

If discourse-sentences are syntactic entities they ought to have a formal definition which is rules-based and which must explain their configuration. The observation that they can be fulfilled either by a complete sentence or an incomplete one, or by a one-word phrase, is not explicative. Indeed, many languages have sentences which are complete and are made up of a single word (for instance *atmospheric verbs*). The problem is to know whether or not a noun phrase (*a dog*) or adjectival one (*dear*) or adverbial one (*sincerely*) can be defined as a sentence (could they be elliptic sentences?), or

[18] The presence of such environmental conditions allows for the occurrence of an *answer* even in the absence of an explicit *question* on the part of the interlocutor.

not and, still remaining phrases, if they can carry out one of those substantive functions. In this case the definition of discourse-sentences cannot be considered syntactic.

Nonetheless, an important contradiction arises when it is maintained that even an expression which is syntactically a complete sentence, can also not be a discourse-sentence. For instance locutions that are complete, but do not carry out a substantive function (*let's see*), are not evaluated as discourse-sentences. This choice is understandable, but it leaves the syntactic definition inapplicable: any expression can either be or not be a discourse-sentence.[19]

Apart from the possibly different terminologies used (*sentence*, *discourse-sentence*, *clause*, *C-clause*, *utterance*) in our knowledge, this is the common result reached by all researchers who analyse a spoken language corpus intending to determine the syntactic features of the units which have communicational value.[20] However, the implication which derives from these assumptions leads one to say that any type of configuration either may or may not be a discourse-sentence. Such a definition, however, is not acceptable from a syntactical viewpoint. More specifically, the previous one cannot be a syntactic definition that offers formal identification criteria for discourse-sentences in a spoken corpus, as it both includes and excludes any configuration.

So a true syntactic definition seems to be missing and what enables the identification of a given expression as a discourse-sentence seems to be the mere fact that it plays a “substantive function in the conversation”. That is to say, its pragmatic character defines a discourse-sentence.

Any requirement regarding the syntactic nature of the reference units entails a formal requirement. According to the UBLI approach, criteria other than this, be they pragmatic or intonation-based, are vague and contradictory.[21] Nevertheless, the fact is that – unless we have misunderstood – the identification criterion for discourse-sentences clearly appears to be pragmatic, based on the recognition of discourse-processing functions.

3. Intonation as an identification criterion of a discourse-sentence

As can be understood, the identification criteria of the utterance or discourse-sentence are of extreme importance since, as already stated, the syntactic definition is not empirically consistent, but the pragmatic definition too runs the risk of resulting not yet mature and trustworthy. An independent

[19] See footnote 6 in Suzuki, Matsumoto and Usami (2005:299).

[20] See for instance Blanche-Benveniste (1997), Blanche-Benveniste et al. (1990), Biber et al. (1999), Miller and Weinert (1998), Cresti (2000).

[21] See Suzuki, Matsumoto and Usami (2005:299).

and "external" criterion is therefore required.

The impasse concerning the definition of discourse-sentences and their identification criteria push us to consider the speech continuum and its features. This involves the fact that it is not always easy and/or clear to identify where a word sequence performing a function or speech act begins and/or ends. Specifically, in informal dialogic spoken language, the textual structure of the language is simplified and fails to mark the habitual connections between words. This is a widespread characteristic of Italian, but also of the other Romance languages.

In several publications we have tackled this issue, which appears extremely relevant in the prototypical context of spoken language; that is, family-private dialogues.[22] Here, we report an example which corresponds to a dialogic-exchange by a speaker who performs the sequence quickly and without pauses:

*SUS: *lei gliene serve una anch'a lei una in più o no no lei ha questa*
[you need one more too or not no you have this one]

However, the exchange is not at all ambiguous for the listener, nor was it ambiguous for the speaker who uttered it. Indeed, when the grouping of words – which is operated by intonation – is considered, the corresponding completed linguistic actions achieve a unequivocal interpretation:

*SUS: *lei /gliene serve una anch'a lei ? una in più / o no ? no // lei ha questa //*
[you / (do) you need one also for you ? one more / or not ? no // you have this one //]
%ill: [1] question; [2] alternative question; [3] self-answer ; [4] conclusion

Thus, it is possible to identify 4 discourse-sentences, each characterised by the completion of a function or speech act. Nevertheless, cases like this apparently cannot be interpreted on the basis of the transcription, and it is not possible to decide which functions are enacted and to which groups of words they should be attributed. Indeed, when considering the mere linear word sequence, no configuration pattern can be detected. The problem is about defining criteria to decide which words or phrases exactly perform a given function.

If the acoustic aspect is also considered, and the grouping provided by intonation is marked, the identification of how many and which functions are performed by the words is "naturally" guaranteed. Faced with the difficulties of mapping between expressions, or groups of expressions, and function

[22] The example below is taken from Cresti (2000:45). The argument is also analysed in detail by Moneglia (2005:19-24), Scarano (to appear) and Moneglia and Cresti (in this volume).

completions – which is typical of spontaneous speech – our observation of corpora has led us to choose a criterion of acoustic perception, the one which any speaker-listener uses to disambiguate and decode speech.

The criterion adopted by LABLITA for the identification of the boundaries of utterances, that is of those expressions which perform “substantive functions in conversation”, is intonation-based. We would like to underline the fact that this criterion does not imply the evaluation of the different intonation features and their categorisation (evaluation of the movement types, tones, levels, focal points), which is a very complex and, generally, debatable point. Our criterion is only based on perception: detection of terminal and non-terminal prosodic breaks.[23]

We know that intonation cues as terminal prosodic breaks are so prominent that they require almost no training to be recognised.[24] To take into consideration the terminal prosodic breaks simply means not to eliminate a datum which is intrinsically linked to the production of speech. The acoustic parameters which enable this recognition are numerous, the most important of which is related to the F_0 pattern and the forms which it can assume to accomplish a conclusive pattern; then lengthening of stressed and unstressed syllables at the end of the pattern, and intensity values take part in this recognition process. The recognition is above all possible, in a relational context, between two consecutive utterances, because the performance of two consecutive utterances is accompanied by a general prosodic reset. A transitional point marks the end of an intonational pattern and the beginning of a new one allowing an instant perceptive recognition of the so-called terminal break.

Thus, an independent, immediate, and perceptively-based /validated criterion (prosodic breaks) exists, which permits the identification of the main word-groupings within the verbal production flow. In our experience, this criterion is not an evaluation nor a classification of speech act types, which refers to tag sets based on theoretical principles, but rather, it is a function of perceptual evidence. Prosodic breaks are marked in the transcription by simple diacritic means: slash and double-slash.[25]

[23] Prosodic breaks must not be mixed up with pauses when looking at utterance boundaries. In around 60% of cases, pauses act as a re-enforcement of terminal prosodic breaks; however also around 40% of non terminal breaks are accompanied by a pause. See Moneglia (2005:24). Moreover, many pauses have different functions or are accidental.

[24] See Moneglia (2005) and Moneglia et al. (2005) on the validation of the C-ORAL-ROM prosodic tagging.

[25] Currently much research to identify the prosodic and phonetic cues which may enable the automatic identification of terminal prosodic breaks is underway. See Sorianello (forthcoming).

3.1. A comparison with Japanese

A core question is whether or not, in a language such as Japanese – which is so radically different to Romance languages – intonation carries out comparable functions with those found in our corpora. Are there terminal prosodic breaks in Japanese?

We are aware of the existence and widespread use of *interactional particles,* as for instance *ne* and *sa,* more or less comparable to what the literature refers to as ‘discourse markers’, but also, of *functional particles* which signal concepts such as *subject* (*ga*) and *topic* (*wa*), and whose equivalent does not exist in Romance languages. So has Japanese, perhaps, developed morphological instruments, like said particles, which are in part absent or not of such relevance in Romance languages, to mark some functions of discourse processing? And are these more important than an organisation which in other languages, such as Romance languages, appears to be based on intonation?

It seems that in any case such interactional and functional particles are always accompanied by salient prosodic markers. Morita (2005) stresses the relevance of such interactional particles.[26] Nonetheless, from this work, it seems possible to understand that both interactional and functional particles always occur at the end of a tone unit and carry intonational traits which mark tone unit boundaries. Thus, on the basis of this work, it would appear that in Japanese, the presence of particles, is necessarily accompanied by F_0 movements which mark tone unit boundaries. Do mixed markers, suitably intoned, exist in Japanese?

The matter is all-important because it is connected to the problem of the identification of utterances or discourse-sentences within the speech flow in an independent and specific manner. The existence of intonation and/or morphological-pragmatic markers allows to solve the problem.

We verified, first for Italian and then for the other Romance Languages, that the portions of speech marked by terminal prosodic breaks systematically correspond to the completion of speech acts or of what can be called functions. If, also in the case of Japanese, the sequences identified through/by intoned interactional particles perform functions of discourse processing, the identification of the linguistic entity of reference (discourse-sentence) is ensured in an independent manner.

Hence, a point of comparison with UBLI seems to be related to intonation and the marking of discourse-sentences in the speech flow. In any case, it seems to be an aspect which needs to be taken into account when

[26] Based on a corpus of 11 different recording sessions, roughly 7h and 30’ of dialogues, face-to-face conversations and phone-calls of Japanese speakers.

dealing with usage-based teaching of Romance Languages.

4. Mapping between *normal forms* and functions

In the articles cited by Usami (2005) and in the one by Suzuki, Matsumoto and Usami (2005), we learn that the relationship between functions and discourse-sentences has been investigated in depth leading to research on the correspondence between discourse-sentence lexical and morpho-syntactic filling and to the carrying out of a specific function (*form-function mapping*)[27]. This is referred to as “corresponding linguistic form” or “explicit linguistic form corresponding to a function”, or “a linguistic form which is considered to represent the function from its literal meaning or conventional usage”.[28] A linguistic form can be represented by a word (an adverb or a verb or a locution), or by a morphological character (a verbal mood)[29], which the authors assume to lead to the carrying out of that function.

From this point onwards we will identify these linguistic, morphological and lexical forms, which represent a function as proposed by the UBLI authors, with the term *normal forms*.

On the basis of a greater or lesser degree of correspondence of a function with *normal forms*, Usami specifies that the seven functions which appear most frequently in the TTW data have been classified in three types:[30]

a) Type -1 which is accompanied by a *normal form*;
b) Type -2 which is not accompanied by a *normal form*;
c) Type -3 in which *any of the corresponding linguistic forms is used, but the function itself is not realized.*

We cannot give a detailed analysis of the different degrees of mapping of functions, however, it seems to us that a very general conclusion, which is in line with the research results obtained by LABLITA, can be drawn: no one-to-one correspondence exists between *normal forms* and functions. Indeed, even though, in Japanese, some functions are preferably performed through a *normal form*, nonetheless, a significant percentage, more than a quarter of discourse sentences, does not carry out a specific function by using *normal forms*.[31]

[27] See Suzuki, Matsumoto and Usami (2005:312).
[28] See Usami (2005:283).
[29] See Suzuki, Matsumoto and Usami (2005:304).
[30] See Usami (2005:283).
[31] Usami reports that in Japanese 73, 9% of functions are of Type 1 (Usami 2005:283). Anyway different percentile values for Japanese are declared in Suzuki, Matsumoto and Usami 2005:303-305. For what concerns spoken English, TTW data shows that 57, 1% of the functions are performed without typical forms (Usami 200:282).

In any case, considering that a function can have more than one corresponding *normal form*, and that a form can perform more than one function, it is easy to understand that it is not possible to assume a one to one correspondence between *normal forms* and functions.

With regard to this aspect, we can notice that in day-to-day spoken language there is a basic freedom in the production of talk, which is only generally conditioned by social norms that may require the use of *normal forms*. For instance an order is not always accomplished by means of a *normal form*, such as the imperative mode of the verb (*close the door!*), but quite often it is done with a simple deictical adverb (*there, come on, up*), or with a noun phrase (*the door*). Moreover, a certain semantic content can be used in different situations, with the accomplishment of different functions (*the door*, performing an answer*; the door!* performing an order*; the door?* performing a question). Thus, from a theoretical point of view, the mapping of semantic content and the enacting of a speech act appears to be almost completely free.

In the case of functions realised without a *normal form*, the solution proposed by our Japanese colleagues is that the context, taken as a whole, allows the disambiguation of discourse-sentences and the recognition of the functions which they render.[32] Moreover, in the various articles by UBLI, a generic reference is made to intonation, as one of the features which are part of the context or which aid the correct interpretation of the function carried out by a discourse-sentence, especially when the latter is not realised using a *normal form*.

With regard to this, we would like to introduce some considerations which, from a theoretical point of view, lead to the placement of intonation outside the notion of 'context', as a specific linguistic device independent from it.

4.1. The illocutionary values of intonation

As verified in many experiments and studies by LABLITA, Italian intonation, besides its primary function of utterance marking in the speech continuum, has an illocutionary function, carrying specific actional–communicational values.

Intonational features are present in many languages with the role of explicitly distinguishing, for example, assertive from interrogative and imperative utterances. Even languages such as English, which have a well-established morphological and syntactic system for distinguishing these 'modalities', in day-to-day spoken language, they often omit them relying

[32] See Usami 2005:282-283.

simply on intonation. In Italian, intonation is probably used in a more constructive manner than in other linguistic systems in order to operate the illocutionary distinction of utterances.[33]

Research on our Italian spontaneous spoken corpora highlighted some thirty intonational forms with specific illocutionary values:[34] thus, in theory, any linguistic expression, if semantically complete, can enact any illocutionary force, as long as it is intoned in an appropriate form.

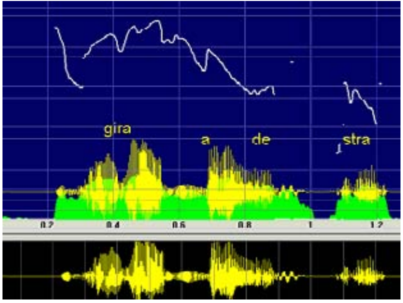
We would like to present below an example from the LABLITA experiments,[35] showing how the same word sequence (*gira a destra* [turn right]) can perform entirely different linguistic actions (the speech act reported on the right side of the page), if intoned with appropriate profiles.

1)	*gira a destra*	[(It) turns right]	Answer/Assertion
2)	*gira a destra*	[turn right!]	Order
3)	*gira a destra*	[the way of doing, it is to turn right]	Instruction
4)	*gira a destra…*	[it turns right…if we can really say this way?)]	Softening

Each figure below, shows the F_0 profile of utterances 1 to 4 performed within a verbal exchange that was placed in a context appropriate to each different speech act. The context has been filmed in a short movie (with sound); in this paper, this is roughly recalled by means of a dialogic context.

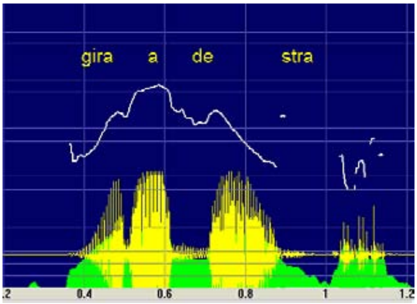
1') Answer / Assertion

- Does Viale Canova still continue after the Square ?
- It turns to the right *(Gira a destra)*



2') Order

– Turn to the right ! *(Gira a destra)*

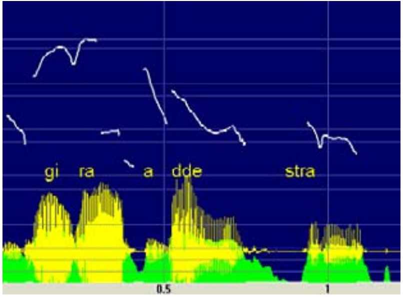


[33] See Cresti 2000; Cresti and Firenzuoli 2001; Firenzuoli 2003.

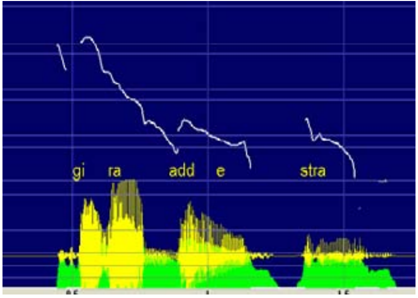
[34] See Firenzuoli 2003.

[35] These movies have been the object of a set of experiments driven in LABLITA. and can be downloaded by the reader from http://lablita.dit.unifi.it/Speech_Examples/

3’) Instruction
- where is the exam of Latin
- do you see the corridor ? Turn to the right *(Gira a destra)*



4’) Softening
- Also this news paper is “turning to the right” [i.e. getting more conservative]
- it turns to the right … *(Gira a destra)*



As its evident from figures the F_0 profiles are quite different. The table below reports the main prosodic characters (with regard to F_0 and Duration properties) which have been recorded in the study of the profiles having, respectively, an illocutionary value of ‘answer’, ‘order’, ‘instruction’ and ‘softening’.[36]

Table 3. F_0 Features

Illocution	**Nucleus**	**Structure**	**Range**	**Onset**	**Level**	**Alignment**
Answer	[1A] [D]	(Pre-nucleus) *Tail	Mid 100/200 Hz male 200/300 Hz female	Mid/ Low	Mid	Right side of the tone unit [1A] [D] on the focus
Order	[1A] sudden	(Pre-nucleus) (Tail [D])	Strong 150/250 Hz male 100/400 Hz female	Mid/ High	High	Left side of the tone unit [1A]
Instruction	[1A] [D] (Final Plateau)	(Pre-nucleus) *Tail	Strong 80/250 Hz male 150/300 Hz female	Mid	Mid/ High	Left side of the tone unit [1A] Right side [D] (Final Plateau)
Softening	[D] (Low)	*Pre-nucleus *Tail	Mid 100/200 Hz male 200/300 Hz female	Low	Mid/ Low	On the whole tone unit

[36] The movements of the nucleus of the tone unit are described according to the IPO terminology in ’t Hart, Collier, and Cohen A. 1990. The overall IPO system has been implemented in LABLITA’s research. See Firenzuoli (2003) for a more comprehensive framework.

Table 4. Duration Features

Illocution	**Length**	**Speed**	**Syllable L.**
Answer	+/- Short	Mid	Around 200ms. With lengthening of the tonic syllable
Order	Short	Mid / high	Around 200ms. Without lengthening of the tonic syllable
Instruction	+/- Long	Slow	Around 200ms. With lengthening of the tonic syllable in correspondence to the movement [D]
Softening	Not relevant	High	Between 100 and 200 ms.

However, it could be argued, as our Japanese colleagues propose, that it is the context and the situation that determine the correct illocutionary interpretation of utterances in our examples, even if they are in fact intoned with a specific and appropriate intonational profile. The sets of evaluation studies carried out in LABLITA, (based on audio and film material on which it is possible to operate replacements and changes) lead to a different conclusion.

The insertion of an utterance with an intonation which is not suitable in an ‘elicitation’ context is automatically judged by native speakers as not appropriate or not natural. For example, in the context which envisages that the speaker gives an *order* to the listener (as in cases where the latter does not have the necessary information or vision), and, instead, an *instruction*-intonated utterance is given, the control group evaluates the new version as strange or inappropriate. From a logical point of view, we would have expected the *instruction*-intonated utterance to have been considered equally fitting, as *order* and *instruction* both belong to the *directive* class, and, moreover, to the same subclass; but this is not the case, since the intonation competence is so strong.

In a different test, the control speakers, presented with the audio of an utterance with a certain type of intonation, quickly choose the correct (dumb) matching film of the ‘eliciting’ context, from among the different filmed contexts proposed to them. So it is on the base of intonation character that the right context is identified and not vice versa.

Our conclusion maintains that the context supports and elicits the interpretation of the intonation with an illocutionary value, but does not determine it. Every intonation profile has its own illocutionary value, not depending on the context. In Italian intonation must be considered a real linguistic parameter, with a conventional system coding and should not be confused among other environmental devices, such as the context. Moreover, it is the first and most important factor for clarifying the illocutionary value of the utterance.

The research carried out on the LABITA corpora with regard to the mapping of *normal forms* and intonation profiles with illocutionary value shows that in Italian different semantic contents can be intoned with the same intonation profile. This guarantees the correct pragmatic interpretation of the utterance. At the same time a given semantic content can be used in different situations, with the accomplishment of different speech acts, and, naturally, in Italian this is performed each time with the appropriate intonation. From this point of view, the mapping of semantic content and the realisation of a speech act appears to be almost completely free.

These observations lead us to a radical position for which whatever linguistic sequence, once it is intoned in an appropriate manner, can fulfil any illocutionary act.[37] Reciprocally, any intonation profile with illocutionary value can be applied to any semantically significant linguistic sequence. Hence, in Italian, while the mapping between intonation profiles and illocutions or functions is systematic, the mapping between illocutions or functions and *normal forms* seems quite limited.

We ask ourselves whether or not the described intonation characteristics, specific to spoken Italian, find any counterparts in Japanese. We must confess that due to our ignorance, we do not know what the importance of tone in contemporary spoken Japanese is, or what distinctive word value tones have. Could it be that in Japanese, systematic intonation profiles convey illocutionary values?[38]

We would like to point out that the evaluation of what act has been fulfilled with a specific utterance, identified by prosodic features, is a different type of operation compared to what we already illustrated above with regard to the perceptive recognition of terminal prosodic breaks. Indeed, the intonational evaluation of what profile has been used, except for some of the more evident cases, implies a real classification with attribution of an illocutionary value. Such an operation is not instant and involves a complete evaluation and classification of a datum.

Obviously, the research necessary to identify intonation profiles in a spoken corpus is a long and hard work, but we believe that this too could

[37] The only semantic condition is that the expression can be neither a free morpheme (determiner, preposition, auxiliary verb, conjunction, etc..), nor a bound morpheme (traits of gender, number, time, mood, etc..).

[38] The above cited book by Morita features the description of some contributions on the actional value of intonation in Japanese. For instance, she reports the work by Izuhara with observations on the intonation of the particle *ne* (1994), as well as work by Koyama (1997) and by Eda (2001) (Morita 2005:44-45). All these contributions seem to connect the actional value of intonation to the use of the particles already discussed in this paper too.

perhaps be a point for future development of the comparison between Japanese and Romance Languages.

4.2. Mapping of normal forms and information patterning functions

In Italian, functions or illocutions do not have a strict correspondence with *normal forms*; nonetheless, it is possible to find some relevant kinds of mapping between lexicon and functions at a different level of the utterance, i.e., that of its internal organisation.

In the research carried out by LABLITA, we discovered a level of organisation, within the utterance, which we named the level of *informational patterning*. Based on the wealth of the Literature throughout the 20th century, and on the analysis of our corpora, we observed that in the majority of cases utterances are not composed of a single word-grouping, signalled in the speech continuum by intonation, but correspond to a complex pattern composed of two or more word-groupings, again always marked by intonation.[39]

In the literature, one of the fundamental aspects of such a composition of the utterance has often been indicated by a simple dual functional opposition in the terms of *theme/propos* (Bally 1950), *theme/rheme* (Prague functionalism, Sornicola e Svoboda 1989), *topic/comment* (Hockett 1957), *topic/focus* (Chomsky 1971, Jackendoff 1972), *given/new* (Halliday 1976), *prefix/noyau* (Blanche-Benveniste 2003).

More recent studies have also focused on other components of the utterance: for example on *discourse-makers* and the many different functions they can carry out (Schiffrin 1987, Bazzanella 1994). But there is also research on the *parenthetical* and modal inserts of utterances (Tucci 2004), on the forms of *citation* and of *reporting* of the words and thoughts of others (Giani 2005).

However only few of these research have pointed out to the fact that all the functions are systematically signalled by intonation, which adopts specific strategies to highlight their informational value.[40]

We maintain that the previous information units give rise to *informational patterns*.

Every utterance corresponds to an *information pattern*, which is systematically signalled by an *intonation pattern* whose units are marked by

[39] In the Italian C-ORAL-ROM subcorpus, the percentage of utterances made up of groupings of more than one word is over 57%, but in the formal domain it is generally much higher. See Cresti & Moneglia (2005:220).

[40] As for Italian, see Cresti and Firenzuoli (2002). With regard to the problem of the value of tone units in Japanese, see Morita (2005). The third Chapter of her work is dedicated to this topic; however, in our opinion a specific solution has not been identified.

non-terminal prosodic breaks. The *information pattern* is interpreted in a tendentially isomorphic manner by a *intonation pattern.*[41]

The groups of words signalled by intonation, within the *information pattern* of an utterance, each have specific functional roles (*topic, comment, appendix, parenthesis, speaker introducer, opening, fathic, allocution, expressive, conative*). Thus, if the entire utterance (that is the entire *information pattern*) is characterised by the fulfilment of a speech act (a function of a superior order such as an *answer, narration, order, question, confirmation, introduction, suggestion, deixis,* etc..), the components of the utterance, or rather the information units that take part in it, are each characterised by the performance of lower level functions.

We can only outline these lower level functions which are all conceived with a pragmatic character. The *comment* unit is devoted to the accomplishment of the illocutionary act and is the only necessary and sufficient information unit type within the utterance. All the other functions are optional; for instance the information function of *topic* can be defined as the application field of the *comment*, that of *appendix* as the textual integration of *comment* and *topic,* that of *parenthesis* as a modal insertion and judgment over the hosting utterance, that of *speaker introducer* as the way to mark the reported speech and to introduce exemplification, listing, etc..

These first five functions (*comment*, *topic*, *appendix*, *parenthesis, speaker introducer*) can be considered as semantic components of the utterance; but the *informational patterning* of the utterance also comprises a second type of functions, i.e. dialogue-type functions. They are: *opening* dedicated to the general function of turn- taking, *fathic* for the regulation of the communication channel, *allocution,* for a direct calling of the interlocutor, *expressive* to emphasize the performance of the speech act, *conative* to push the interlocutor to do something or to avoid a certain behaviour.

In Cresti (2000), the main aspects of the *informational patterning* of the utterance are described: their functional aspect is defined, together with the identification of their basic intonational characters, and their distribution inside the utterance. Later corpus-based research investigations have been conducted by the LABLITA team and they enabled the discovery of levels of mapping between *normal forms* and specific information units; more specifically information functions of dialogue-type (*opening*, *fathic*, *allocution*, *expressive*, *conative*) . In these researches some morpho-syntactic

[41] The Informational Patterning has been introduced in Cresti (1987 and 1994) and developed in many publications by the LABLITA team. See also the debate on macrosyntax in Scarano (2003). For the ontogenesis of Informational patterning see Cresti & Moneglia (2001).

and lexical characters of each of the information units have emerged.

If all the semantic-types of function (*comment*, *topic*, *appendix*, *parenthesis, speaker introducer*) present a semantic fill-in freedom and are not linked to a mapping with *normal forms,*[42] the dialogue-types of function are in principle bound and restricted from a lexical and morpho-syntactic point of view. With regard to Italian, there are detailed qualitative and quantitative studies for all these functions, but we believe that many features must be the same as in other Romance Languages. Repertories of these expressions can be easily translated among Romance languages. The frequencies of dialogue-type functions are very high, more than 50% of utterances performed in spontaneous conversations present these kind of devices.[43]

For instances *allocutives*, lexically speaking, can be linked to proper nouns (*Mary, John*), or kinship/family names (*mom, daddy*) and social roles (*Sir, professor*), personal pronouns (*you*) and a few other expressions, mostly evaluative (*honey, stupid*).

Expressives are accomplished as limited stereotyped intercalations (*God, damned, boys, dash*).

O*penings* are very common in Italian, nearly 40% of utterances shares them, and are accomplished through typical expressions such as: *eh, no* (no) , *ma* (but), *perché* (because), *allora* (then), *sì* (yes), *vabbè* (well), *io* (I)

Fathics are represented again by *eh*, and *no*, or by *insomma* (in conclusion), but moreover we can find typical verbal forms *capito* (understood), *guarda* (look), *senti* (listen), *vedi* (you see), *non so* (I don’t know), *scusa* (excuse me), *diciamo, dico* (we say, I say), that are not used with different dialogic functions.

Conatives, which are less common and are employed only within peer to peer conversations, corresponds to forms like: *aspetta* (wait), *un momento* (a moment), *ascolta* (lissen), *dai* (give), *via* (out), *vai* (go) *avanti* (straight on), *su* (...)

We can add to the restricted lexical repertories which regard dialogue-type functions, some corpus based observations which enable to foresee at least some trends in the lexical and morphological fullfilment of the semantic-type functions. Thus, for example *speaker introducers* are 95% *verba dicendi* and of these, only two verbal entrances are used: *dire* (to say), *fare* (to do), of which one, in only few forms: *dice*((s)he/it says), *ho detto* (I said)[44].

[42] See Cresti (2000), Ferri (2003), Signorini (2005), Scarano (2004).

[43] See Frosali (2005).

[44] See Giani (2005).

Parentheses are made up of about 50% 'modalised' adverbs, and the other 50% by verbal clauses with preferred lexical and syntactic features (modal verbs, belief verbs, saying verbs, conditional moods, future tense).[45]

Topics do not have by lexical restrictions, however 85% of them are filled by noun-phrases, adverbial or prepositional phrases, and only 15% correspond to a clause.

Summarising, we believe that distinguishing two functional levels of the utterance, the primary illocutionary one, and an internal *information patterning* one, both signalled by intonation, gives the issue of mapping with *normal forms* the relevance it is due.

Only expressions that carry out lower dialogue-type functions, on an internal level within the utterance, correspond to *normal forms*, while this is not true in the case of information units that deliver the semantic information of the utterance.

We do not know if such a distinction could be useful in the analysis of Japanese. We only know that, for example, in cases such as those cited by Suzuki, Matsumoto and Usami (2005)[46], an expression which is syntactically a sentence, such as *let's see*, but which – quite correctly – is not considered a discourse-sentence by the authors, can be in a coherent manner analysed and the reasons why it is not a discourse-sentence can be explained.

In fact, an evaluation of the intonation, commonly found in Italian with similar meaning and function, demonstrates that they are realised more quickly, with less phonetic detail and with less intonational salience, when compared to the same expressions employed, on the contrary, as discourse-sentences. So we can argue that the syntactic sentence *let's see*, if performed with a dialogue-type function and consequently intoned, does not result in an illocution, or a high-level function, and cannot be considered as an utterance or a discourse-sentence, but merely a phatic information unit.

5. Conclusion

We are pleased to find a similarity in the approach to spoken language carried forward by the UBLI Centre and by LABITA, regardless of its purpose, be it teaching or research. This similarity forms the basis of a possible comparison.

First of all, we wish to reaffirm the validity of a corpus design such as that of C-ORAL-ROM, created above all on the choice of diaphasic traits, rather than on diatopic and diastratic criteria.

Secondly, let us point out the convergence towards spoken language

[45] See Tucci (2004).

[46] See. Footnote 6 in Suzuki, Matsumoto and Usami (2005:299).

considered as a process, linked to usage, rather than as a configurational system. This common point of view is at the origin of the choice of reference units for the analysis of speech.

In both UBLI and C-ORAL-ROM/LABLITA, regardless of the designation terms (*speech acts* or *functions*), the units of reference are ‘activity units’ whose definition refers to the Anglo-Saxon tradition. Moreover, the classification of such units is based on the analysis of vast corpora, rather than on lexical taxonomies.

The linguistic units corresponding to such activity units – whatever the designation terms (*utterance* or *discourse-sentence*) – have been compared with regard to their definition and their identification in the speech continuum.

Given the difficulty of a syntactic definition of utterances, it seems appropriate to draw attention on the specificity of the LABITA approach, based on the perceptive recognition of terminal prosodic breaks, and in general on the attention given to intonation, that enables not only the identification of the boundaries of an utterance in the speech continuum, but also the attribution of a specific illocutionary value to them. To this end, an important comparison between the two languages is necessary. In Japanese, perhaps, the task of signalling the boundaries of the utterance relies more on morphological instruments, such as interactional particles, even when appropriately intoned.

Lastly, with regard to the issue of mapping between *normal forms* and functions, which represents an important resource for teaching, the fact that a systematic mapping between them is impossible to predict, appears to us as an objective datum for Romance Languages. Nevertheless, the possibility of identifying an internal level of *information patterning* of the utterance allows the discovery in such a domain of a consistent mapping between *normal forms* and dialogue-type lower information functions.

Fundamental issues arise with regard to the structural diversity of our languages and in particular concerning the role of intonation and/or interactional particles. It seems to us that, in any case, such observations can lead the way for the study of aspects which are not always fully appreciated in both the comparing and the teaching of foreign languages.

References

Austin, L.J. 1962. *How to Do Things with Words*. Oxford: Oxford University Press.

Bally, C. 1950. *Linguistique Générale et Linguistique Française*. Berne: Francke Verlag.

Bazzanella, C. 1994. *Le Facce del Parlare*. Firenze: La Nuova Italia.

Berruto, G. 1987. *Sociolinguistica dell'Italiano Contemporaneo*. Roma: La Nuova Italia Scientifica.

Biber, D., Johansson, S., Leech, G., Conrad and S., Finegan, E. 1999. *The Longman Grammar of Spoken and Written English*. London and New York: Longman.

Blanche-Benveniste, C. 1997. *Approches de la Langue Parlée en Français.* Paris: Ophrys

Blanche-Benveniste, C., Bilger, M., Rouget, Ch., Van den Eynde, K., Mertens, P. 1990. *Le Français Parlé: Études Grammaticales.* Paris: Éditions du C.N.R.S.

Chomsky, N. 1971. "Deep Structure, Surface Structure and Semantic Interpretation". D. Steimberg & L. Jacobovits (eds.) *Semantics: an Interdisciplinary Reader.* Cambridge:Cambridge University Press 1971. 183-216.

Cresti, E. 1987. "L'articolazione dell'informazione nel parlato". In AA.VV. *Gli Italiani Parlati: Sondaggi sopra la Lingua d'oggi,* Firenze: Accademia della Crusca 1987. 27-90.

Cresti, E., 1994. "Information and intonational patterning in Italian". *Accent, Intonation, et Modéles Phonologiques,* B. Ferguson, H. Gezundhajt and P. Martin (eds.) Toronto: Editions Mélodie. 1994. 99-140.

Cresti, E. 2000. *Corpus di Italiano Parlato,* voll. I-II, CD-ROM. Firenze: Accademia della Crusca.

Cresti, E. and Firenzuoli, V. 2001. *Illocution and intonational contours in Italian,* in htpp://lablita.dit.unifi.it./preprint/preprint-01coll04.pdf. Also published in *Revue Française de Linguistique Appliquée* IV(2): 77-98.

Cresti, E. and Firenzuoli, V. 2002. "L'articolazione informativa topic-comment e comment-appendice: Correlati intonativi". A. Regnicoli (ed.) *La Fonetica Acustica come Strumento di Analisi della Variazione Linguistica in Italia.* (Atti delle XII Giornate del Gruppo di Fonetica Sperimentale), Roma:Il Calamo 2002.153-160.

Cresti E. and Moneglia, M. 2005. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages.* Amsterdam:Benjamins.

Eda, S. 2001. "A new approach to the analysis of the sentence-finalparticles ne and yo an interface between prosody and pragmatics". Nakayama M. and Quinn, C. (eds.), *Japanese/Korean Linguistics* 9. 167-180.

Fava, E. 1995. "Tipi di atti e tipi di frase". L. Renzi, G. Salvi and A. Cardinaletti (eds.) *Grande Grammatica Italiana di Consultazione*. Bologna:Il Mulino 1995. 19-48.

Ferri, C. 2003. *Caratteristiche sintattiche, intonative e frequenza dell'appendice di comment in un corpus di italiano parlato (LABLITA).* Tesi di laurea. Firenze:Università di Firenze.

Firenzuoli, V. 2003. *Le Forme Intonative di Valore Illocutivo dell'Italiano Parlato: Analisi Sperimentale di un Corpus di Parlato Spontaneo (LABLITA).* PhD thesis. Firenze:Università di Firenze.

Frosali, F. 2005. *Le unità d'informazione di ausilio dialogico: valori percentuali, caratteri intonativi, lessicali e morfosintattici in un corpus di parlato italiano (C-ORAL-ROM).* Tesi di Laurea, Firenze:Università di Firenze.

Giani, D. 2005. *Il discorso riportato nell'italiano parlato e letterario: confronto tra due corpora*. PhD Thesis. Firenze:Università di Firenze.

Grice, H. 1975. Logic and Conversation, in Cole,P., Morgan,G. *Speech Acts. Syntax and semantics*, vol 3., Academic Press, New-York, pp. 41-58.

Halliday, M.A.K. 1976. *System and Function in Language: Selected Papers*. London: Oxford University Press.

Halliday, M.A.K. 1989. *Spoken and Written Languages*. Oxford: Oxford University Press.

Hockett, C. F. 1958. *A Course in Modern Linguistics*. New York: The Macmillan Company.

Holmes, J. 2005, "Socio-pragmatics aspects of workplace talk" in Kawaguchi et al., *Usage-Based Linguistics Informatics*. Amsterdam: Benjamins 2005. 196-221.

Izre'el S., Hary, B. and Rahav, G. 2001. "Designing *CoSIH*: The corpus of spoken Israeli Hebrew". *International Journal of Corpus Linguistics* 6: 171-197.

Izuhara, E. 1994. "Intonation of interjectional particles, insertion particles, sentence final particles *ne* , *nee*". *Nihongokyooiku* 83. 96-107.

Jackendoff, R. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge Mass:MIT Press.

Jacobs, J. 1984. "Funktionale satzperspektive und Illokuktionsemantik). *Linguistiche Bericthe* 91. 25-58.

Kawaguchi,Y. 2005. "Center of Usage-Based Linguistic Informatics (UBLI)". Kawaguchi et al. (eds.) *Usage-Based Linguistics Informatics*. Amsterdam Benjamins 2005:3-8.

Kawaguchi,Y., Zaima,S., Takagaki,T., Shibano, K, Usami, M. 2005, *Usage-Based Linguistics Informatics,* vol I°-II°, Amsterdam:Benjamins.

Koyama, T. 1997. "Bunmatsushi to bunmatsu intoneeshon". Onseibunpookenkyuukai (ed.) *Bunpoo to onsei,* Tokio: Kuroshoio Publisher 1997. 97-119.

Miller, J. and Weinert, R. 1998. *Spontaneous Spoken Language*. Oxford: Clarendon Press.

Moneglia, M. 2005. "The C-ORAL-ROM resource". Cresti E. and Moneglia, M. *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance*

Languages. Amsterdam:Benjamins 2005. 1-70.

Moneglia, M. In this volume. “Units of Analysis of Spontaneous Speech and Speech Variation in a Cross-linguistic Perspective”.

Moneglia M. and Cresti, E. 2001. “The value of prosody in the transition to complex utterances. Data and theoretical implications from the acquisition of Italian”. Almgrem, B. Barrena, M. J. Ezeizabarrena, I. Idiazabal, B. Mac Whinney (eds.) *Proceedings VIIIth International Congress IASCL, (12-16 luglio 1999, S. Sebastian),* Chicago: Cascadilla Press 2001. 851-873.

Moneglia, M. and Cresti, E. In this volume. “C-ORAL-ROM”.

Moneglia M., Fabbri M., Quazza S., Panizza A., Danieli. M, Garrido J. M., Swerts M. 2005. Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM corpus”. E. Cresti, M. Moneglia (eds.). *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam:John Benjamins Company 2005. 257-276.

Moreno Fernandez, F. 2005. “Corpus of spoken Spanish language – The representativeness Issue”. Kawaguchi et al. (eds) *Usage-Based Linguistics Informatics*. Amsterdam:Benjamins 2005. 120-145.

Morita, E. 2005. *Negotiation of Contingent Talk. The Japanese Interactional Particles* ne *an*d sa. Amsterdam: Benjamins.

Scarano, A. (ed.). 2003. *Macro-syntaxe et Pragmatique. L’analyse Linguistique de l’ Oral.* Roma: Bulzoni.

Scarano, A. 2004. “Enunciati nominali in un corpus di italiano parlato: Appunti per una grammatica *corpus based”.* In *Atti del Convegno Nazionale “Il Parlato Italiano”,* CD-ROM, F. Albano Leoni, F. Cutugno, M. Pettorino, R. Savy (eds), 1-18. Napoli: M. D’Auria.

Scarano, A. To appear. “Prosodic annotation in speech resources. The C-ORAL-ROM corpus in linguistic research and the teaching of languages”. *Corpus Linguistics Studies*.

Schiffrin, D. 1987. *Discourse Markers.* Cambridge: Cambridge University Press.

Searle, J. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge:Cambridge University Press.

Signorini, S. 2005. *Topic e soggetto in corpora di italiano parlato spontaneo.* PhD Thesis. Florence: Università di Firenze.

Sorianello, P. Forthcoming. “Per una definizione fonetica e fonologica dei confine prosodici”. *La Comunicazione Parlata* (International Congress. Naples 23-25 February 2006).

Sornicola R., Svoboda A..1989. *Il campo di tensione*. Napoli:Liguori.

Suzuki, T., Matsumoto, K., Usami, M. 2005. An analysis of teaching materials based on New ZealandEnglish conversation in natural settings —implications for the development of conversation teaching materials— in Kawaguchi et alii *Usage-Based Linguistics Informatics*, Benjamins, Amsterdam, pp. 279-294.

’t Hart J., Collier R. and Cohen A. 1990. *A Perceptual Study on Intonation. An Experimental Approach to Speech Melody*. Cambridge: Cambridge University Press.

Tucci, I. 2004. *L’inciso: caratteristiche morfosintattiche e intonative in un corpus di riferimento.* In Atti del Convegno “Il parlato italiano”, Napoli, 13-15 febbraio 2003. D’Auria M., Napoli. pp. 1-14.

Usami, M. 2005. “Why do we need to analyze natural conversation data in developing conversation teaching materials? Some implications for developing TUFS language modules”. Kawaguchi et al. *Usage-Based Linguistics Informatics*. Amsterdam Benjamins 2005. 279-294.

Yuki, K., Abe, K., Lin, C. (2005) “Development and assesment of TUFS Dialogue Module-Multilingual and Functional Syllabus”, Kawaguchi et al. *Usage-Based Linguistics Informatics*. Amsterdam Benjamins 2005. 316-333.

Units of Analysis of Spontaneous Speech and Speech Variation in a Cross-linguistic Perspective

Massimo MONEGLIA

1. Introduction

C-ORAL-ROM (Cresti & Moneglia 2005) collects four comparable corpora of Italian, Spanish, French and Portuguese which record spoken romance languages in a huge variety of contexts. This paper presents some very general lexical and syntactic qualities of spoken performance, which are derived from the measurements of this multilingual corpus. Beside this task the paper mainly focuses on the correlations between the context in which the speech performance takes place and the variations in speech quality that are recorded at cross-linguistic level.

Despite the freedom that characterizes the linguistic performance (Chomsky 1958), three kinds of cross-linguistic measurements of speech show regular context-bound variations. Such measurements regard: a) the distribution of Part of Speech in the speech performance; b) the weight of the utterance, in terms of length and speed and their correlations with the weight of the dialogic turn; c) the main structural strategies used by speakers to build up the utterance in spoken language.

All the above language variations are independent from the speakers and are rather bound to the same types of contextual variations in all romance corpora. This leads to the conclusion that crucial spoken language behaviors are required by contextual features and that therefore the representation of spoken language activity calls for an adequate representation of contextual variation, as proposed by C-ORAL-ROM and other large spoken corpora initiatives in order to capture relevant qualities of language use.

In 2. the contextual parameters used to build up the four romance resources will be briefly presented, and compared to other current approaches. Then in 3. the main annotations of the C-ORAL-ROM corpora are sketched. These annotations, accomplished in parallel in the four corpora, mainly regard the Prosodic and Morpho-syntactic levels and are used for the detection of both lexical and structural properties of the utterance in speech. Finally in 4. the paper focuses on the analysis based on those annotations, and presents the main correlations found in the spoken romance languages between contextual parameters and both lexical and structural variations.

2. C-ORAL-ROM. Sampling criteria for the representation of spontaneous speech

C-ORAL-ROM consists of four comparable resources of Italian, French, Portuguese and Spanish spontaneous speech sessions (roughly 300,000 words for each language). In total, 772 spoken texts corresponding to 121:43:07 hours of speech from 1,427 different speakers. The resource aims to represent the variety of speech acts performed in everyday language and to enable the induction of prosodic and syntactic structures in the four Romance languages, from a quantitative and qualitative point of view. This poses a problem of representation, that is common in Corpus Linguistics, but particularly sensible in the spoken domain. Speech performance varies consistently. For instance, a story told to a child and a row between husband and wife vary in language register, dialogue structure, topic, illocutionary force. Comparing a set of business transactions with a lesson given by a professor to his class, we register variations with respect to dialogical character, programming, style, task, etc.

In the representation of the ‘Spontaneous speech universe’ we must make at least the hypothesis that the linguistic properties of the speech events may vary in connection with non-linguistic variations. This assumption is evident, for example, at the level of frequency lexicons. High-frequency lexicon may be under-represented in specific pragmatic domains where on the contrary low-frequency lexical items score the highest ranking. In this paper we will see that the connection between non linguistic variation and linguistic variation goes beyond the frequency of lemmas, but that also crucial structural linguistic qualities of spoken texts regularly vary in accordance with the needs established by the context.

The setting up of spontaneous speech databases is therefore a complex task. C-ORAL-ROM sampling is based on the definition of a set of variation parameters that have been considered significant in many socio-linguistic studies (Berruto 1987; Biber 1988; De Mauro et al. 1993; Gadet 1996). The following is the set of contextual parameters used for sampling the spontaneous speech universe:

a. *Register variation*: sessions characterized by a formal language register vs. sessions characterized by informal language uses;

b. *Channel variation*: face-to-face interactions vs media productions vs telephone recordings;

c. *Dialogical structure variations*: speech events having a dialogue or a multi-dialogue structure vs. monologues;

d. *Social context variations*: interactions belonging to family and private life vs. interactions taking place in public;

e. *Domain of use variation:* domains such as law, business, research, teaching, church, etc. are represented

Given the above variation parameters each romance corpus represents the universe in a comparable way as far as it complies with the following matrix, which specifies the proportion of words and samples in each field.

Table 1. C-ORAL-ROM Corpus design matrix

Language register	**Social context**	**Structure of the communication event**
Informal 150,000 words	Family/private 124,500 words	Dialogue and Multi-dialogue 102,000 words
	Public 25,500 words	Monologue 48,000 words
	Channel	**Typical domain of use**
Formal	Natural context 65,000 words	Political speech Political debate Preaching Teaching Professional explanation Conference Business Law
Formal	Media 60,000 words	Talk shows Scientific press Reportage Interviews Sport News Weather forecast
Informal	Telephone 25,000 words	Private conversations 15,000 words
		Human-machine interactions 10,000 words

Sample length is determined in an uniform manner in the four corpora. Informal samples record about 1500 words while formal samples record around 3000. Therefore, given the above general figures, corpora record also a comparable number of samples.

As the matrix shows, C-ORAL-ROM adopts two different sampling strategies for formal contexts and informal contexts. In formal contexts the genre and the domain of application of the sessions is strictly defined in a closed list of typical domain of use, while this is not the case for informal contexts, where the domain and the text genre is left random. In the informal the variation is defined only for the social and the structural characters of the speech event.

Considering current practices in designing spoken corpora this feature is far from obvious, and must be highlighted because of its theoretical relevance. For example, *The Spoken Dutch Corpus* uses similar parameters in designing a spoken reference corpus. As in C-ORAL-ROM, the samples record in uniform manner *Channel variation, Dialogical structure variations, Social context variations, Register variation*, however the design of the Spoken Dutch Corpus tries to define as much as possible a closed set of text genre, both in the formal and informal sections. The formal section (identified as 'more or less scripted') records *Interview* and *Discussion, Description of pictures, New*s, *Report, Current affair, Programs, Commentary, Speeches, Lectures, Read about texts*. The informal section ('Unscripted') also records a closed set of genres: *Face to face conversation, Interview, Business transaction, Discussion, Debate, Meetings, Lectures, Spontaneous commentary*.

At least in principle, this strategy may cause a strong limitation on the probability of occurrence of the more representative spontaneous speech contexts. Indeed, in a given social-historical context, formal speech can be identified by listing the typical contexts of use where the formal use of language is preferred, but the same does not hold for the universe of informal speech. To the end of characterizing the informal use of language no context is more typical than another and the set of situations where informal language is used must be left open if we do not want to have 0 probability of occurrence for most of the contextual variation in the informal domain.

Despite the huge language variation that is represented in C-ORAL-ROM the sampling strategy has been criticized, because it ignores in the design schema the main sociolinguistic factors that characterize speakers; i.e. age, education, geographical origin, role of the speaker in society (Moreno-Fernández 2005). Indeed C-ORAL-ROM registers all these characters in the metadata, but remains unbalanced with respect to the quality of speakers that participate in the corpus.

This is not necessarily the case in the design of a spoken resource. For example the spoken part of the British National Corpus (BNC) dedicates almost half of its size to recordings provided by a significant sampling of the British population. Subjects were asked to record their conversations during a certain period of time, so testifying the actual use of spoken language in accordance with the variation caused by speaker's parameters.

More recently CoSIH (Izre’el et al. 2001) designed even more coherently a frame which applies this strategy. CoSIH was designed to integrate, in a random sampling, demographic and contextual criteria. Day-long recordings of 950 informants statistically representing all social and ethnic groups of the Israeli population have been planned over a one-year period. In this hypothesis informants are captured in recordings while they go through all the contextual and interpersonal situations that occur in the day, so ensuring speech data that are balanced at the same time both at sociological and contextual variation level.

Although it should be considered the best strategy, the CoSIH approach is not easy to be pursued, and does not eliminate the need to define relevant contextual variations. From a practical point of view the recording of many contexts of use requires the setting up of a recording apparatus beforehand, and those situations remain excluded if not planned. The strategy is also difficult to be applied in the legal frame of the European Union where the consensus of each intervenient in a recording is required beforehand and the recording of many situations is constrained by severe rules that go beyond the consensus of the speakers.

However, beside the practical constraints that always arise in the setting up of a spoken corpus, we must stress that the identification of the population as the best source of data, does not relieve the investigators from sampling and classifying the contexts in which the recordings occur.

It should be clear that providing data through a statistically significant sampling of the population does not imply that all linguistic variations in the corpora are due to the socio-linguistic qualities of the speakers. In other words a sociological sampling of the population is valid as far it also captures relevant context variations. So, also in the CoSIH approach, when passing from raw material to corpus sampling, the metadata of each session record not only the speaker’s data but also the relevant contextual qualities. This work is the premise to observe the correlations between language variation and non linguistic features.

This paper will show that the contextual variation is highly predictive of specific language variation and that therefore the sampling strategy of C-ORAL-ROM captures relevant facts of spoken language use.

3. Utterance and prosodic breaks

3.1. The study of language properties in a corpus is a function of its linguistic annotation. Each recorded session in C-ORAL-ROM is annotated with:

a. Session metadata
b. Orthographic transcription

c. Text-to-speech synchronization of the utterances of each speaker

d. Part of Speech (PoS) tag of each transcribed form

In accordance with the CHAT de facto standard (MacWhinney 1994) the transcription contains the representation of the main dialogue properties in a computable form:

e. *speaker's turns;*

f. the main occurring *non-linguistic* and *paralinguistic events*.

g. *prosodic breaks* in the speech flow of the turn

h. the segmentation of the speech flow of the turn into discrete speech events

3.2. The PoS annotation of the four language resources has been accomplished through automatic tagging using various taggers and comparable tag sets. This work allows the comparison of the occurrence of the main part of speech in the corpora and the extraction of frequency lists of lemmas. We will take this annotation for granted.[1]

The identification of the units of reference for the study of spontaneous speech is the main added information of C-ORAL-ROM. This information is crucial for the understanding of peculiar properties of speech, but cannot be identified through the same syntactic and semantic cues used for written resource (Izre'el 2005). The reader must consider that a great many speech events (almost 1/3 of those events in C-ORAL-ROM, as we will see below) do not have a verb and therefore do not show a clear syntactic structure. For example the transcription of the following C-ORAL-ROM examples report two dialogic turns of the French speaker JEA, while he was chatting with a friend about his child. The first turn is not a possible sentence, and, as far as the scope of the verb is not determined, its syntactic structure is underdetermined. The second does not bear any verb, ad its structure is also mysterious:

*JEA: c' est la carotte quoi carotte devant le nez du lapin
[it is the carrot then carrot in front of the nose of the rabbit]
*JEA: ouais la carotte ouais
[yes the carrot yes]

The problem is that syntax does not provide enough evidence for the identification of the linguistic unit ranking over the word level. In the C-ORAL-ROM approach the reference unit for spontaneous speech is identified with the term 'utterance', that is defined following the pragmatic

[1] See in Cresti & Moneglia 2005 the description of the taggers and tagsets used for PoS annotation in the four romance corpora and the evaluation of their results.

tradition (Austin 1962). The utterance is *the minimal linguistic entity such that can be pragmatically interpreted* and/or the linguistic entity that is ‘concluded’ and ‘autonomous’ from a pragmatic point of view.

Although this definition may sound familiar (Quirk et al. 1985, Cresti 2000) the annotation procedure that allows to parse the speech continuum into utterances is quite new. In C-ORAL-ROM the utterance is identified through an heuristic that allows its annotation as a function of prosodic properties, and more specifically on the perception of *prosodic breaks* (Cresti & Firenzuoli 1999; Cresti 2000; Cresti & Firenzuoli 2002). It is assumed that each utterance has a profile of *terminal intonation* (Karcevsky 1931; Crystal 1975) and therefore the presence of *terminal breaks* in a string is a cue for the detection of the utterance boundaries.

The speech flow and its transcription are so divided into reference units which rank above the word level looking to this prosodic feature, that is considered the property of the utterance that is more easy to be detected. Therefore each prosodic unit ending with a terminal break is considered an utterance.

The perceptual prominence of terminal breaks is strong in all romance languages. It was already well known that competent speakers have a strong perception of prosodic boundaries (Buhmann et al. 2002). The C-ORAL-ROM annotation and its validation shows that competent speakers can also easily discriminatẹ prosodic boundaries that have a terminal value (‘Terminal breaks’, marked with a double slash) from those boundaries that indicate that the utterance goes on (‘Non terminal breaks’, marked with a single slash).[2]

For example in the C-ORAL-ROM annotation both the above turns have been parsed in two utterances, identified through the terminal prosodic boundary, each one corresponding to a separate speech act:

111 *JEA: c’ est la carotte quoi // carotte devant le nez du lapin // (*ffamol08*)
[it is the carrot then // carrot in front of the nose of the rabbit //]
114 *JEA: ouais // la carotte / ouais // (*ffamol08*)
[yes // the carrot / yes //]

The structure of the two turns is not underdetermined. Competent speakers do not feel that there is any ambiguity. Their structure is specified by the speech act boundaries that are clear to perception as far as they go hand in hand with prosodic boundaries. Each speech act is pragmatically autonomous, complete and separate from the other. No subpart of each speech act shares those properties. We will see that the annotation of terminal

[2] The level of inter-annotator agreement has been evaluated by an external user (Danieli et al. 2004).

breaks provides the term of reference for the statistic evaluation of the main properties of spoken language performance in the multilingual corpus.

3.3. The non terminal breaks that occur within an utterance can also be exploited to highlight other crucial properties of spoken language. The hypothesis underlying the C-ORAL-ROM annotation concerning non terminal breaks points to the fact that prosodic breaks parse the utterance into prosodic units (also called ‘prosodic envelope’ or ‘tonal unit’, according with various terminologies) that are the main index of its ‘informational structure’.

The idea of a strict correspondence between ‘information units’ and ‘intonation units’ can be derived from Halliday (1976) and has been used within different frameworks (Benveniste et al. 1990, Lambrecht 1994, Brazil 1995; Cresti 2000, Scarano 2003, Simon 2004, Izre’el 2005). Therefore the prosodic annotation gives an important contribution to the study of this language structure, highlighting the chunk of texts that form informational units within the utterance.

According with the specific background assumptions of the C-ORAL-ROM prosodic tagging (Cresti 2000) the utterance may be structured through a verb, that defines its syntactic form, but also through *an informational pattern which is isomorphic with an intonation pattern* (‘t Hart et al. 1990). The informational pattern is made:

a) by a necessary and sufficient information unit: *comment*, which is devoted to the accomplishment of the illocutionary force; (Austin 1962).

b) by other optional information units, in one to one correspondence with prosodic units, which establish linguistic relations with the comment unit (*Topic* unit, *Appendix* unit, *Parenthetical* units, *Dialogic* units)

From this point of view an utterance can be ‘simple’ or ‘compound’. Simple utterances consist of a single prosodic envelope ending with a terminal prosodic break and necessarily feature a single informational unit of the ‘comment’ type (Cresti 2000; Hockett 1958).

Simple utterances generally correspond to a brief and syntactically simple linguistic sequence that may or not contain a verb:

*CIC: mille a voi // (*ifamcv14*)
[(there’s) your thousand (lire)]
*SAM: non ho capito la domanda // (*inatla03*)
[I haven’t understood the question]

Compound utterances consist of a number of linguistic chunks separated by at least one non-terminal prosodic break. They bear one comment unit, plus at least one supplementary informational unit. Their structure involve an informational relation between functionally distinct chunks of information, thus creating an utterance which may be fairly long and syntactically complex. A complex utterance may contain a verb, as the following one:

*DON:se tu non hai i soldi / rimani malato e muori // (*imedrp03*)
[if you haven’t got the money / you stay ill and die]

or may be verbless:

*LUC: sabato mattina / all’ undici / eccotelo // (*ifamcv22*)
[saturday morning / at eleven / there he comes]

The previous utterances, despite the fact that they contain or not a verb have both an informational Topic-Comment structure. The following verbal and verbless examples have a Comment-Appendix informational structure:

*ELA: poi non lo mangia / i’ biscotto // (*ifamdl02*)
[then he doesn’t eat it / that biscuit //]
*SND: belli / i jeans // (*ifamcv21*)
[nice / those jeans //]

This paper is not the right place to present the various types of informational relations, that are not marked in C-ORAL-ROM.[3] However the added value conveyed by non terminal breaks for linguistic investigation must be highlighted. The value of prosodic breaks for the annotation of the linguistic properties in spoken corpora goes beyond the marking of utterance limits. The presence of more than one prosodic unit within an utterance is an index that the utterance has an informational structure. In other words, the number of tone units that parse an utterance is a formal index of structural complexity, that goes beyond verbal predication.

In conclusion, on the basis of the annotation of word forms, lemmas, PoS, utterance limit, and information units within the utterance it becomes possible to investigate in the four romance languages very basic language properties of spoken language, and to highlight its variations faced with the contextual variations recorded in the corpus design.

[3] Both the illocutionary type of each utterance and the specific informational relations that occur in each utterance remain unspecified in the annotation scheme, as those properties are not object of simple detection based on direct perception, but rather their tagging relies on complex categorization schema. See Cresti 2000 for a discussion based on Italian.

4.1. Lexical distribution in the C-ORAL-ROM corpora

The presence of a large lexicon in a reference corpus is the first condition to ensure that the universe is sufficiently represented. The contextual variation of the C-ORAL-ROM corpora, by testifying many semantic domains, gives rise to a quite large vocabulary, if compared to the small dimension of each language corpus (only 300.000 graphic words). The table below records the number of lemmas for each language resource and the proportion of tokens and types according to the general distinction Closed vs. Open class lemmas:

Table 2. Token / type distribution in the C-ORAL-ROM corpus

		Open Class		Closed Class	
	Lemmas	tokens	types	tokens	types
ITALIAN	15,286	172000	13804	122505	365
FRENCH	11,801	130835	11124	120966	525
SPANISH	11,743	146350	11163	136649	328
PORTUGUESE	11,453	157172	9795	142713	751

The dimension of the C-ORAL-ROM corpora is absolutely not consistent with the needs of corpus linguistics for what concerns collocation and colligation (Tognini-Bonelli 2001). However the structure of each language corpus is minimally consistent with the properties that are required of a reference corpus for what concerns the fundamental lexicon of each language; that is the lexicon that more frequently gives rise to language structures.

The fundamental lexicon is the set of higher ranked lemmas that cover roughly 85% of tokens in a general corpus. For example in a huge reference corpus like the BNC, that records around 120,000 lemmas and 100,000,000 occurrences, the 6500 lemmas with higher rank total an occurrence of 85%.[4] C-ORAL-ROM offers enough information for the study of the core part of the fundamental lexicon of each language. The following curves represent the incidence of lemmas (ordered according to their rank) over the total tokens in each corpus. The curves shows that 90% of occurrence in each corpus is covered by a comparable set of high frequency entries (around 2000 in Italian and around 1500 in Spanish, Portuguese and French.

[4] We will use the BNC as reference corpus for what regards the properties of written language. Figures reported here comes from BNC (http://www.comp.lancs.ac.uk/ucrel/ bncfreq/flists.html), Kilgarriff web site (http://www.kilgarriff.co.uk/bnc-readme.html), Leech et al. 2001 and from various measurements accomplished in the LABLITA lab. The recorded lemmas covers 98% of the corpus and 45.000 items are Proper Names.

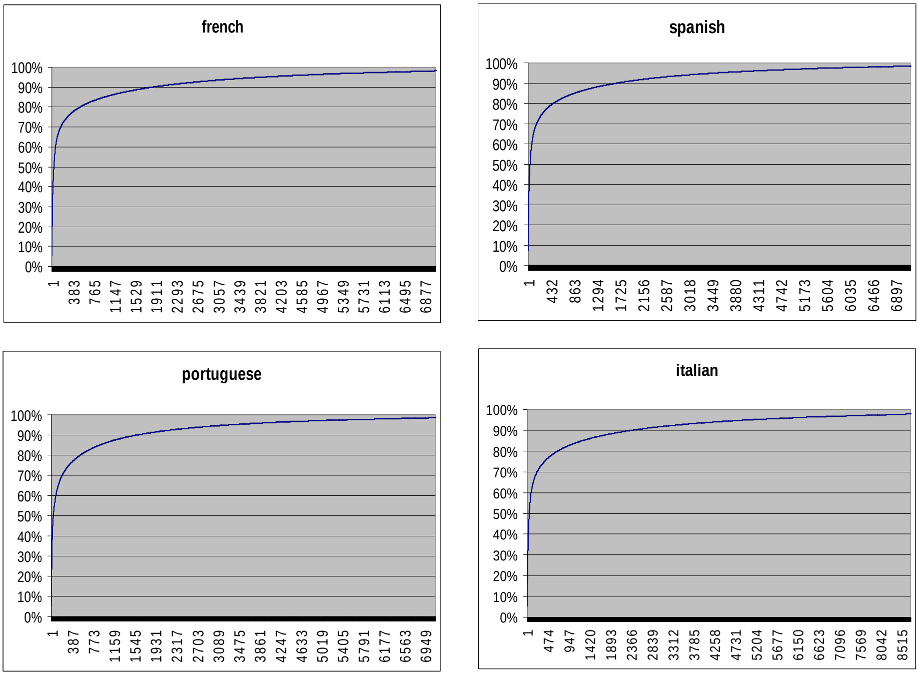


Figure 1. The fundamental lexicon in the C-ORAL-ROM corpora

The comparability and the significance in the structure of the lexicon in the four corpora can be better evaluated focusing on the frequency lexicon of each resource. Despite the rough level of precision of PoS assignment, we can note that the proportion between *open* and *closed class* forms is extremely constant in the four corpora.[5]

The strong cross linguistic consistency of the above proportion testifies on one side a feature of spoken language, that registers a lower percentage of open class lexical forms with respect to the written domain (60% of open class expressions in the BNC, according to our measurements), and on the other also a high level of comparability in the data provided by the four corpora.

[5] This evaluation relies on the performance of the automatic tagger of each corpus. The main discrepancy is found in the Italian sub-corpus and it reflects a strong overestimation of Nouns by the Italian tagger.

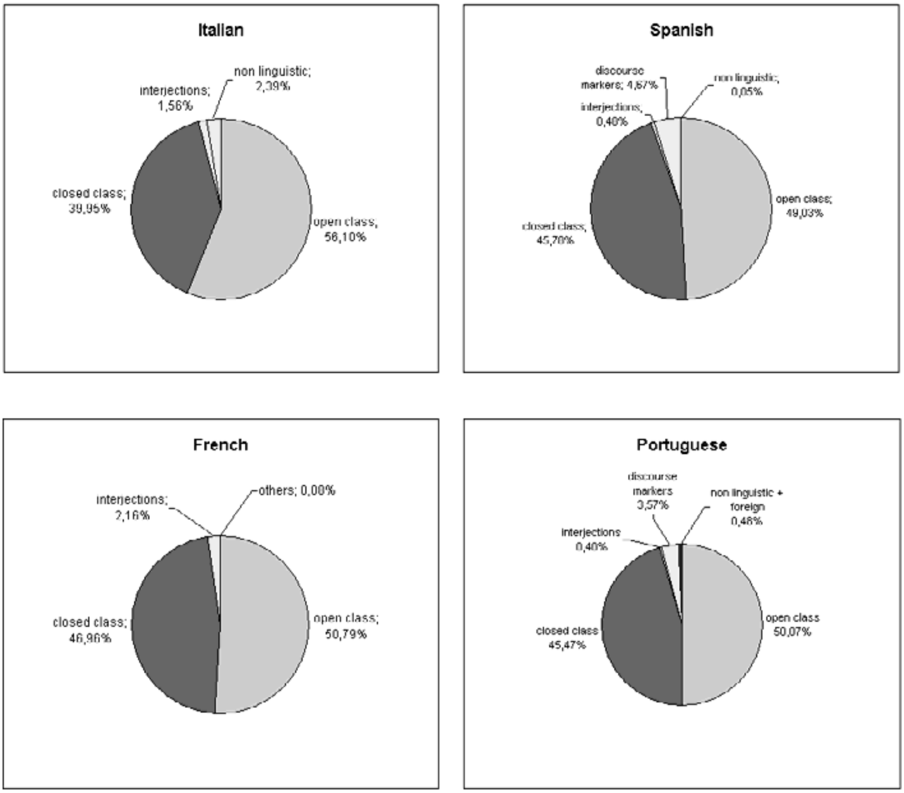


Figure 2. Percentage of Open class and Closed class forms in the C-ORAL-ROM corpora

The meaning of such a huge amount of closed class forms in spoken corpora must be considered carefully. This lexicon, that has mainly a grammatical value, occurs with a limited number of lemmas, each of these recording a large number of tokens. This is confirmed by the figures of the fundamental lexicon. The table below shows that in each corpus the closed class part of the fundamental lexicon covers around 50% of the total lemmas of this class.

Table 3. Open and Closed class lemmas in the fundamental lexicon

	TOTAL LEMMAS	Fundamental	Open Class	Closed Class
ITALIAN	15286	2390	2118	187
FRENCH	11801	1981	1778	178
SPANISH	11743	1749	1489	165
PORTUGUESE	11453	1684	1381	224

This property is important from the point of view of the exploitation of corpora for linguistic studies. Corpora where the closed class lexicon occurs with a sufficient number of tokens may in principle testify the main syntactic constructions that each lexical entry of this type conveys in each language.

4.2. The "lexical strategy" in speech

Measurements of the C-ORAL-ROM lexicon highlight relevant differences between spoken and written language for what regards the distribution of the main open class lexical categories. According to our measurements Nouns and Verbs record approximately 43% of tokens in the BNC, which is a higher percentage with respect to spoken language if we consider the measurements in Figure 3. However Nouns are the prevalent part of speech in written language (Biber 1988; Biber et al. 1999; Giordano & Voghera 2002; Halliday 1989). This is confirmed also by our measurements of tokens in the BNC (around 26% of Nouns and 17% of verbs), while in spoken language the use of verbs is on the contrary more frequent. As Figure 3 shows the proportions of verbs and nouns is cross linguistically consistent in C-ORAL-ROM, therefore the 'verbal' lexical strategy of speech is confirmed at cross-linguistic level.

The distribution of part of speech shows cross-linguistic consistency not only for nouns and verbs, but also for what concerns the main word classes. As Figure 4 shows, the low percentage of Nouns in spoken language is mirrored in the consistent number of pronouns (almost one to one in all corpora). This feature testifies the deictic character of reference in spoken performance.

The number of Adjectives is also constant at cross linguistic level (around 4%), and is marked by a very low percentage with respect to written language (7.5% in the BNC), while the category of Adverbs (comprising subordinative, sentential and non subordinative adverbs) has a double frequency.[6]

Moreover the main free morphology types (i.e. Preposition, Conjunctions, Articles) are distributed almost in equal proportions in the four languages, with a tendency to record a higher percentage in Prepositions.

Finally, although the tag-set adopted for Italian and French does not allow uniform measurements regarding the so-called Discourse Markers, the overall distribution shows that the specific spoken language lexicon (Discourse Markers, Interjections, Non Linguistic Forms) characterizes speech by a lesser percentage (only from 2 to 4%).

[6] The discrepancy of the Spanish corpus is caused by the tagset choice

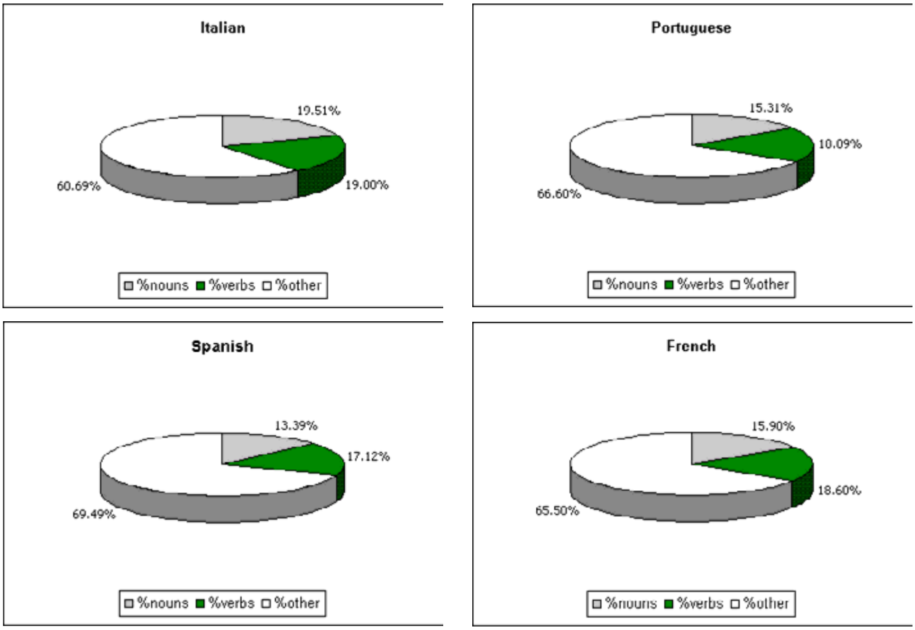


Figure 3. Percentage of Nouns and Verbs in the C-ORAL-ROM corpora

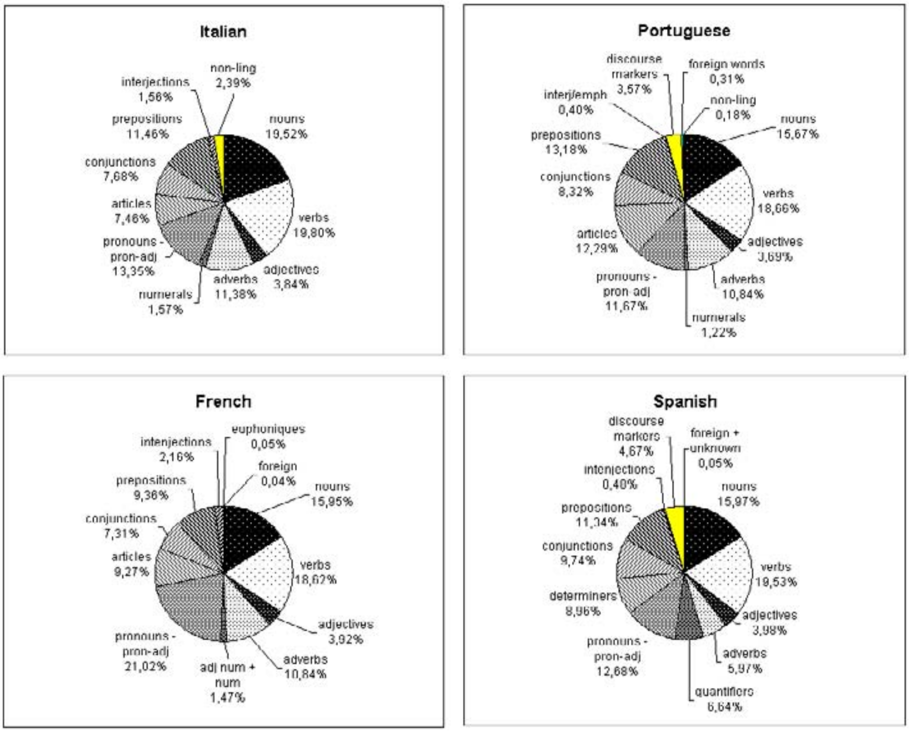


Figure 4. Percentage of PoS in the C-ORAL-ROM corpora

However the lexical distribution in spontaneous speech is even more significant if the values recorded are considered along with the contextual variation parameters of the corpus design.[7] The variation of the ratio between nouns and verbs in the romance corpus is regular and strictly corresponds to contextual variation. The line diagrams in Figure 5 show a regular increase in nouns from Informal Dialogues to Media and Formal Monologues, with a very marked drop in the Telephone node, and, in a complementary way, a decrease in verbs from Informal Dialogues to Formal Monologues.

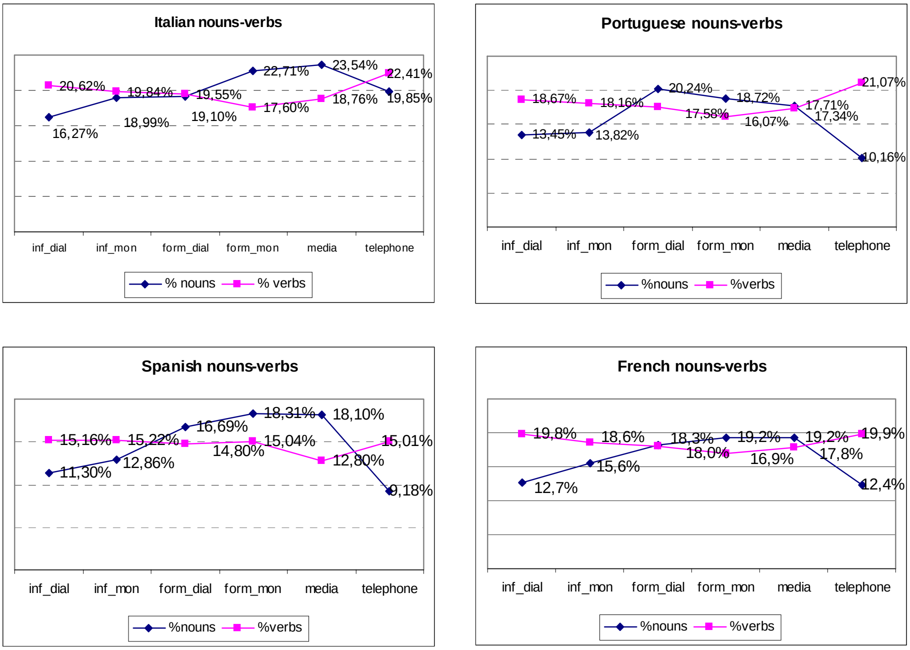


Figure 5. Variation of the percentage of Nouns and Verb in the main contexts of the design

Summarizing, a decrease in verbs from Informal to Formal nodes and an increase in nouns from Informal to Formal is recorded at cross-linguistic level. The feature of Formality required by the context seems to be the main one responsible for the progressive increase of nouns through the corpus structure and the proportional decrees of verbs. Therefore in accordance with the data provided the ratio of nouns and verbs, that is the main feature of the

[7] This measurements regard only Verb an Nouns, that have a sufficient number of occurrence for an evaluation in the C-ORAL-ROM sub-corpora. The line of variation of the other PoS requires bigger corpora.

lexical strategy in speech, changes as a function of context variation.

4.3. Cross-linguistic variations of the utterance

The following two sets of comparative values highlight the constraints that operate on the speech performance at the constructive level of the utterance. The first set regards the very general properties of speech performance in the four romance, that will be considered in terms of length and speed. In the second series we will focus on some more qualitative measurements related to the construction strategies of the utterances in spontaneous speech, to their complexity, and variation lines.

4.3.1. General measures of the spoken performance and context variation

In spoken language the utterance limits specify the domain of the main linguistic relations. For instance, argument structure, constituency, head dependency, and chunking relations, hold among elements of a same utterance. The mid-length of the utterance is therefore a quantitative value that reflects the complexity of speech performance.

The spoken language domain shows a strong variability with respect to this parameter, but certain variation tendencies can be clearly foreseen. The data show that in all Romance Languages the Mid-length of the utterance varies according to the structure of the communicative event, the sociological domain of use, and the channel; this means that the length of the linguistic object which is the outcome of the speech act is bound to non linguistic factors. It is:

- much higher in formal language
- significantly higher in monologues
- variable in accordance with the channel (lower in media, with respect to formal in natural context, and in telephone, with respect to informal dialogues)

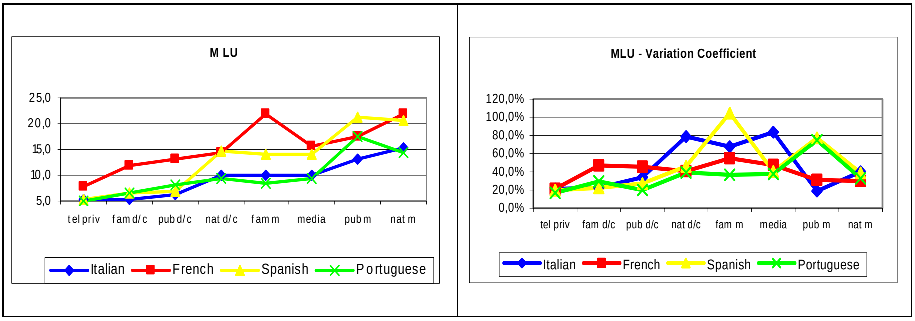


Figure 6. Mid-Length of utterance and its variation coefficient in the main contexts of the design

The range of variations appears quite predictable in *informal dialogues*, that is the prototypical domain of application of spontaneous speech. In informal dialogues the values are cross-linguistically recorded within two ranges (5-7 words per utterance in Italian, Spanish and Portuguese, around 10 in French). So the restriction on the number of words that belong to the same utterance in informal dialogic contexts is severe.

For all languages the tendency to have much longer utterances is cross-linguistically verified in connection with two features: a) monologic structure of the linguistic event; b) context requiring a formal use of language.

The Variation Coefficient with respect to MLU in the informal part is much lower and testifies the significance of the correlation in the prototypical context of spontaneous speech. The variation is especially high in media speech.

The MLU has a positive correlation with the length of the dialogic turn (MLTw). As the previous line diagram has shown, Monologues, which have only one long turn, always have a higher MLU value. Moreover the line diagram in Figure 7 also shows that in dialogue structures (telephone, family private, public informal and formal) MLU and MLTw co-vary in the four collections. In spontaneous speech the longer the turns of a text are the longer each utterance is.

This is a natural rather logical constraint on spoken performance. In principle the opposite could also be reasonable; that is, the more the linguistic performance is complex, the more each piece of information may be simpler. Apparently this theoretical alternative does not apply to natural languages and the opposite tendency is verified cross linguistically and independently from the speaker.

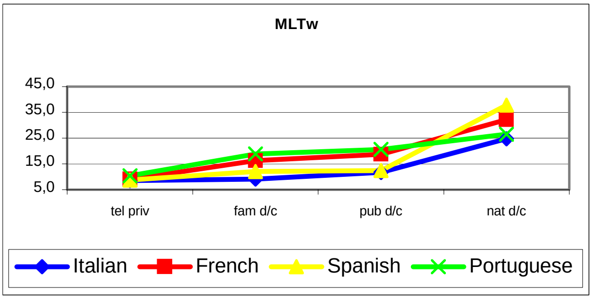


Figure 7. Mid-length of the turn in the main contexts of the design

The Speed registered in the speech performance also varies in accordance with both context and language-specific factors. Within the Romance family Speed turns out to be a character proper of each language, but the difference in Speed is mainly recorded in the informal dialogic part, where French and Spanish have clearly higher values. In the informal dialogic sub-corpus (Telephone - Family – Public) each language records a different speed: lower in Italian (2.7 w/s <), constantly over 3.5 w/s in French, between 3.5 and 3 in Portuguese and Spanish. This is confirmed as a genuine language specific datum by the low Variation coefficient.

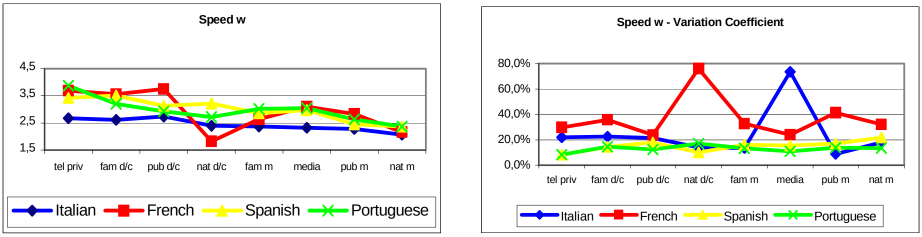


Figure 8. Speed (words/second) and its variation coefficient in the main contexts of the design

On the contrary cross-linguistic differences are lower in the formal part, where the speed decreases in all languages for reasons presumably linked to the task of the speech performance.

As a consequence of the previous correlations between context variation and speed variation the results regarding the Speed of speech performance are quite interesting when compared with MLU and MLTw. S*peed* turns out to have an inverse correlation with MLTw and with MLU. The longer the turn the slower the flow of speech. The longer the utterance the slower the flow of speech. Given this general tendency the variation in speed between the informal part and the formal part is more sensible for those languages with a higher speed.

In all the languages in object the length of the Utterance and the length of the turn co-vary in accordance with contextual parameters. The more the context is formal the more the linguistic task requires a long turn and the more each resulting utterance is long and structured. The opposite is true for speed. It is therefore interesting to see that the Length of the tone unit is a measure that on the contrary does not vary in accordance with contextual features.

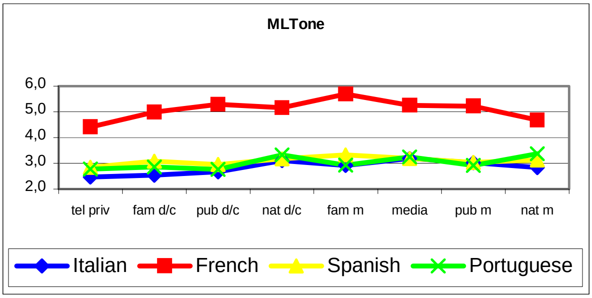


Figure 9. Mid-length of the tone unit and its variation coefficient in the main contexts of the design

This is true at cross-linguistic level. In all the languages under consideration the length of the tone unit appears independent from the contexts of use, and is fairly constant in all languages. MLTone has strong upper limits (due to breath constraints that force a low, predictable, average value.) More specifically, the similarity between Italian, Spanish and Portuguese turns out evident, while French strongly diverges, probably due to a different syllabic weight of words.[8]

4.3.2. The variability of the utterance structure in spontaneous speech

The following notes on the four spoken Romance languages are devoted to capturing the main constructive strategies used in spoken language for structuring the utterance and its main variation lines faced with the contextual variation. We will see that specific contextual variations strictly correlate with systematic variations in the constructive strategies of the utterance found in all language corpora.

The primary index of the utterance structure is of course the presence of a finite verbal form. When a verb appears in an utterance this can be roughly considered also the main structuring element, but, although verbs are proportionally more frequent than in writing, it is frequently the case that a verb is absent in spoken utterances. C-ORAL-ROM strongly confirms the consistency of verbless utterances in speech, that has been already claimed for English in the Longman Grammar (38%). As the following figure shows the percentage of verbless utterances is high in all romance corpora, with a remarkable similarity in values among the romance languages, and a significant variation in French.

[8] The word weight in terms of number of syllables in French is probably at the origin of this variation. The syllabic reduction of words in speech, with respect to their graphic counterpart, is systematic in French. Therefore speakers can in principle produce more words within the breath unit.

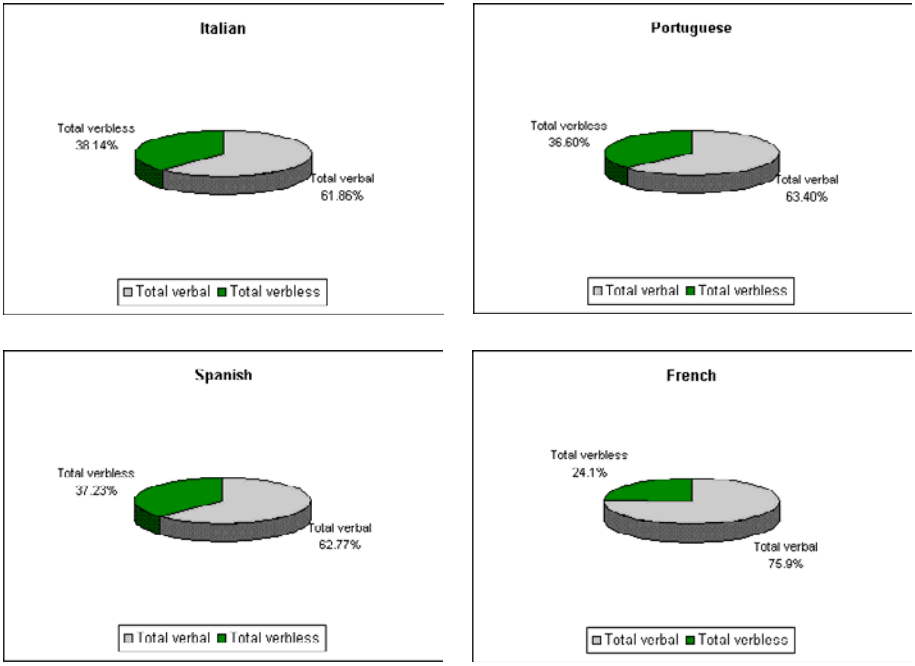
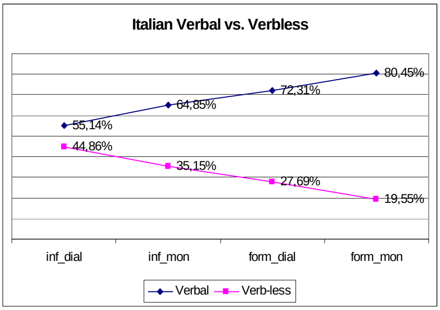
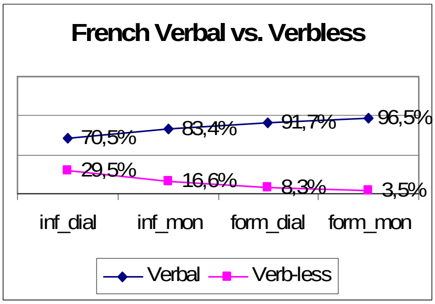


Figure 10. Distribution of verbal and verbless utterances in the C-ORAL-ROM corpora

However it must be also clarified that verbless utterances strongly vary in incidence in the various contextual situations recorded in the corpus. Moreover the line of variation that influences the amount of verbless utterances in spoken language turns out parallel in the four romance languages and is bound to specific contextual features.





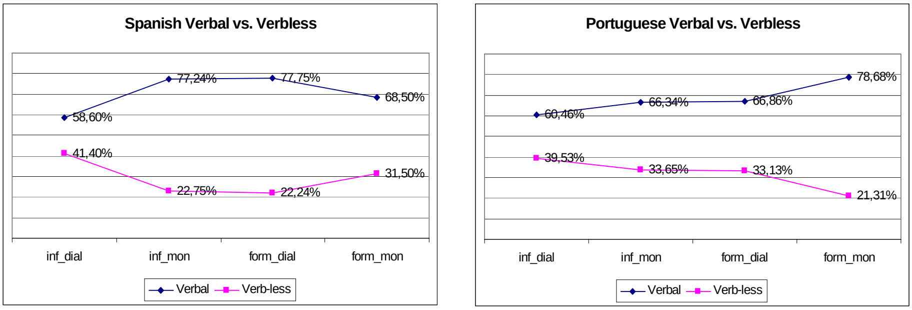


Figure 11. Variation of the distribution of verbal and verbless utterances through main contexts of the design

As the line diagrams in Figure 11 show the ‘verbal structuring strategy’ is positively influenced in the four romance languages by the monologic structure of the communicative event, according to the relation: monologic > more verbal. While the distribution of verbless utterances is complementary. Therefore it is safe to say that the main linguistic character of spoken language performance is bound to context requirements, that strongly influence the probability of occurrence of verbless utterances.

The variation in percentage of verbless utterances in connection with the dialogical structure of the linguistic event occurs at cross-linguistic level and looks like a genuine variation parameter of spoken language. The prosodic annotation reported in the C-ORAL-ROM corpora, however, allows to enrich the dichotomy between verbal (structured) utterances and verbless (unstructured) utterances with a further basic dichotomy of speech: utterances bearing ‘informational relations’ vs. utterances that are ‘simple’ from an informational point of view.

The measurements on utterances made possible by C-ORAL-ROM come from the intersection of the two criteria: ‘Verbal syntactic strategy’ and ‘Prosodic informational strategy’. The intersection of the two criteria foresees four main possible cases that are relevant to identify the utterance’s level of complexity:

(1) Simple verbless; (2) Simple verbal; (3) Compound verbless; (4) Compound verbal.

Figure 12 shows the percentage of each type in the C-ORAL-ROM corpora. The distribution of the utterance types is surprisingly constant in the four romance corpora. The measure of the four structural types reveals that compound verbal utterances are the most common structural type. However it is also important to highlight that more than half of the spontaneous speech utterances do not show an informational structure.

The ‘compoundness’ of an utterance is in most cases associated with the presence of a verb, while on the contrary the ‘simple strategy’ does not show a preference for a verbal filling rather than a verbless one.

We already noticed the consistency of verbless utterances in spoken language, now we can see that the proportion between those verbless that correspond to simple speech acts and those having an informational structure varies from 1/3 to 1/4.

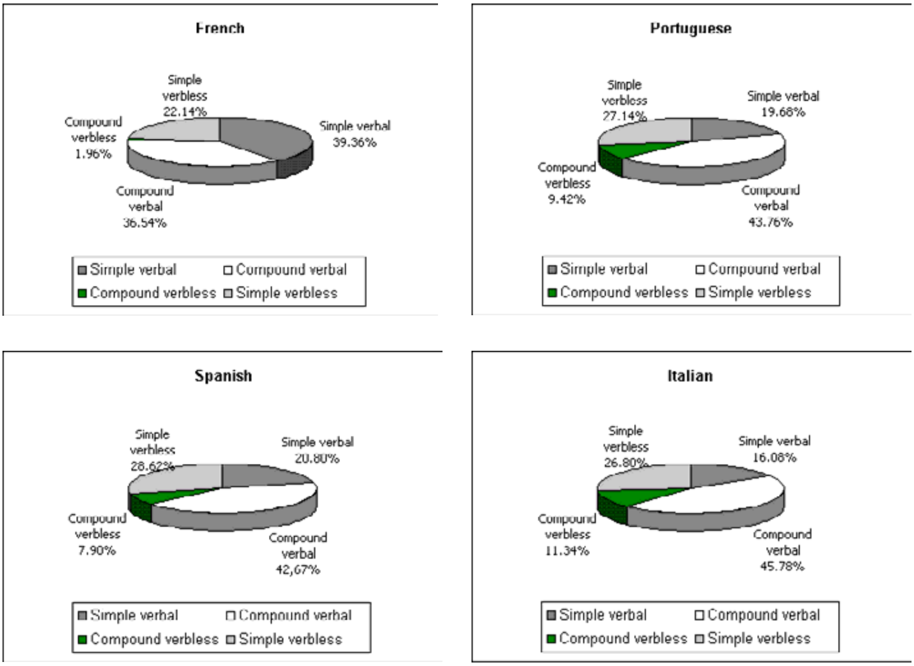


Figure 12. Distribution of the four types of utterances in the C-ORAL-ROM corpora

Following these points, primary structural strategies of speech can be established. At cross-linguistic level we can identify that the percentage of the four types is extremely consistent for what concerns Italian, Spanish and Portuguese, while is quite divergent for French, which records comparable values only in the incidence of the more frequent types (simple verbless and compound verbal).

As a whole, the peculiarity of structural strategies of speech records not only the incidence of verbless utterances, but also, on one side a balance between simple and compound utterances, and on the other the relevance of the simple utterance strategy, that is also extremely significant when compared with the written domain.

3.3.2. Variation of the Utterance types in the four languages

Examining the four structural types in the corpus's contextual variation, however, reveals data that are even more significant. More specifically the C-ORAL-ROM multilingual corpus show the principles of construction that may be considered a constant feature of spoken language and on the contrary other construction strategies change significantly their weight according to the contextual variation.

According to the area diagrams in Figure 13, that specify the proportion of the different types of utterances in each corpus node, we can argue that a spoken text is always characterized by the complementary balance between compound verbal utterances and simple verbless ones. The other two remaining types, simple verbal and compound verbless, represent a constant 'belt' (nearly 27%).

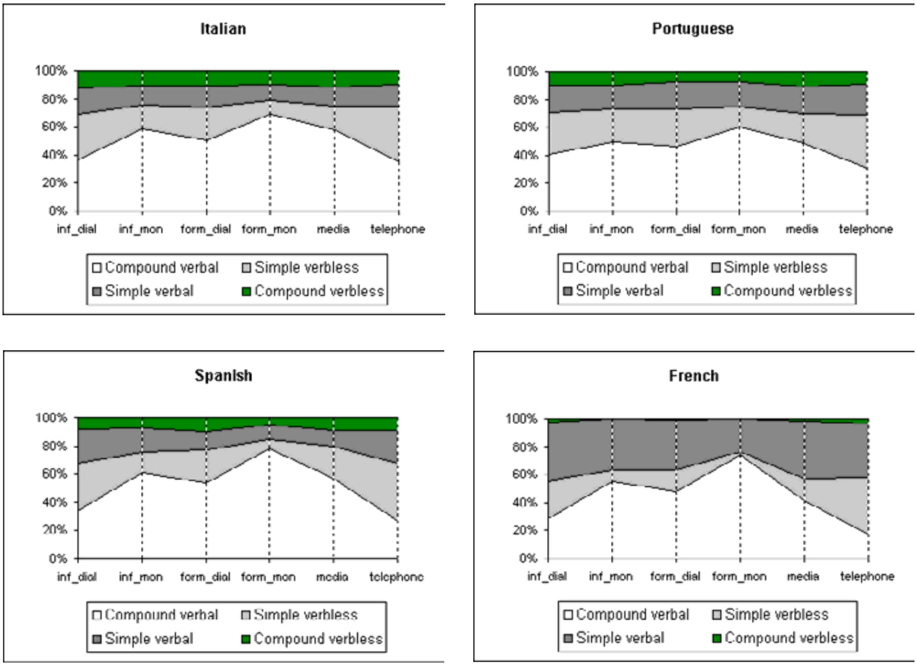


Figure 13. Distribution of four types of utterance in the main contexts of the design

In other words, although spoken language is characterized by the existence and a given amount of various peculiar types of utterances (verbless utterances, simple utterances, compound verbless utterances) spoken language varies only for what regards the ratio between two types: that is the amount of simple verbless with respect to the amount of compound verbal utterances.

The two types are complementary in speech. More specifically the variation of the two types is strong and can be foreseen on the basis of two features of contextual variation, respectively belonging to the structure of the linguistic event and to the language register: the more the spoken language is dialogical the more it records simple verbless utterances and in parallel a lower percentage of compound verbal utterances (this variation is more sensible if the register is formal). On the contrary the more the structure of the communicative event is monological, the more it records a large amount of compound verbal utterances and only a few verbless utterances (also this variation is more sensible if the register is formal).

5. Conclusions

In conclusion spoken romance languages vary with respect to general linguistic properties in connection with contextual variations. This happens at different levels of the language performance: at the lexical level, at the level the utterance’s weight, and, at the more qualitative level of utterance construction. The low range of variation among figures at cross linguistic level allows to argue that the correlation between linguistic and non linguistic factors concerning the context in which the speech events take place is strong and predictive and is therefore independent from the individual difference in speech performance that may occur among speakers.

For what concerns the lexicon the main strategy that characterizes the four romance languages in the distribution of PoS in speech is the well known dominance of Verbs in the Verbs/Nouns ratio, that sets it apart from written texts. The variation line of this ratio shows, however, that this strategy is not a constant feature of spoken language. The number of Nouns varies positively in correlation with the level of formality required by the context. The change of register is the main contextual parameter to which lexical properties are sensible.

The properties of the utterance have been investigated in two sets of cross-linguistic measurements. The variation in length and speed of the utterance is inversely correlated with all parameters of the corpus design (structure of the event, channel and formality). The utterance is longer in monologues especially in formal contexts. It is short and predictable in informal dialogues. Speed is slower and similar in all languages, when the speech performance is more complex, while the utterance is longer and also less predictable in those contexts. Speed varies consistently from language to language mainly in the informal dialogic context, where the length of the utterance is constantly very low. The length of the tone unit is on the contrary a constant of spoken language performance; i.e. it is a property of speech performance that is not sensible to context variation.

The cross-linguistic correlation between language variation and context variation occurs also at the more qualitative level of the utterance structure, where the predicative way of construction, based on the verb, is linked to the informational relations between information units, conveyed by prosody. In this respect a strong correspondence between contextual parameters and the utterance complexity shows up in uniform manner in the four romance corpora. The very high percentage of *verb-less utterances* is one of the main cues of the spoken language variety at cross-linguistic level, but this percentage rises dramatically in informal dialogic contexts. The structure of the linguistic event is the contextual parameter that has more impact on the structural variation of the utterance.

The C-ORAL-ROM tagging strategy also shows that spoken language is characterized by a constant percentage of *simple verbal* and *complex verbless* utterances, that occur with constant figures, not depending on the context. The type of strategy used for structuring the utterance goes hand in hand with contextual features for what regards the percentage of *simple verbless* and *compound verbal* utterances. These two strategies appear complementary in all corpora with respect to the *dialogue* vs. *monologue* and *formal* vs. *informal* contextual parameters and their dominance can be foreseen in the spoken performance as a function of context variation.

References

Austin, L.J., 1962. *How to Do Things with Words*. Oxford: Oxford University Press.

Berruto, G. 1987. *Sociolinguistica dell'Italiano Contemporaneo*. Roma: La Nuova Italia Scientifica.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Blanche-Benveniste, C. et alii (1990), *Le français parlé; ètudes grammaticales*, Editions du CNRS, Paris.

Brazil, D. 1995. *A grammar of Speech*, Oxford:Oxford University Press.

British National Corpus http://www.natcorp.ox.ac.uk/

Buhmann, J., Caspers, J., van Heuven, V., Hoekstra, H. Martens, J-P., Swerts, M., 2002. "Annotation of prominent words, prosodic boundaries and segmental lengthening by no-expert transcribers in the spoken Dutch corpus". In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2002)*, M. C. Rodriguez and C. Suarez Araujo (eds), 779-785. Paris: ELRA.

Chomsky, N. 1959. "B.F. Skinner. *Verbal Behaver*", *Language*, 35. 26-58.

CoSIH CORPUS http://www.tau.ac.il/humanities/semitic/cosih.html

Cresti, E. 2000. *Corpus di Italiano Parlato,* voll. I-II, CD-ROM. Firenze: Accademia della Crusca.

Cresti, E. & Firenzuoli, V. 1999. "Illocution et profils intonatifs de l'italien". *Revue Française de Linguistique Appliquée* IV(2). 77-98.

Cresti, E. & Firenzuoli, V. 2002. "L'articolazione informativa topic-comment e comment-appendice: Correlati intonativi". In *La Fonetica Acustica come Strumento di Analisi della Variazione Linguistica in Italia. Atti delle XII Giornate del Gruppo di Fonetica Sperimentale (XII GFS),* A. Regnicoli (ed.), Roma: Il Calamo. 153-160.

Cresti, E. & Moneglia, M. (eds.). 2005. *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Language*s. Amsterdam:John Benjamins.

Crystal, D. 1975. *The English Tone of Voice*. London: Edward Arnold.

Danieli, M., Garrido, J. M., Moneglia, M., Panizza, A., Guazza, S., Swerts, M. 2004 "Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech C-ORAL-ROM" in M.T Lino, M.F. Xavier, F. Ferraira, R. Costa, R. Silva (eds.) *Prococeedings of the 4th LREC Conference* , Paris: ELRA, vol. 4. 1513-1516.

De Mauro, T., Mancini, F., Vedovelli, M. and Voghera, M. 1993. *Lessico di Frequenza dell'Italiano Parlato.* Milano: ETAS.

Dutch Corpus http://lands.let.kun.nl/cgn/doc_English/topics/project/pro_info.htm

Gadet, F. 1996. "Variabilité, variation, variété". *Journal of French Language Studies* 1: 75-98.

Giordano, R. and Voghera, M. 2002. "Verb system and verb usage in spoken and written Italian". In *Proceedings of 6th International Conference on the Statistical Analysis of Textual Data (JADT 2002),* A. Morin and P. Sébillot (eds), 289- 299. IRISA/INRIA: Université de Rennes.

Halliday, M.A.K. 1989. *Spoken and Written Languages*. Oxford: Oxford University Press.

Halliday, M.A.K. 1976. *System and Function in Language: Selected Papers*. London: Oxford University Press.

't Hart J., Collier R. and Cohen A. 1990. *A Perceptual Study on Intonation. An Experimental Approach to Speech Melody*. Cambridge: Cambridge University Press.

Hockett, C. F. 1958. *A Course in Modern Linguistics*. New York: The Macmillan Company.

Izre'el, S. 2005. *"Intonation Units and the Structure of Spontaneous Spoken Language: A view from Hebrew"*. Cyril Auran, Roxanne Bertrand, Catherine Chanet, Annie Colas, Albert Di Cristo, Cristel Portes, Alain Reynier and Monique Vion (eds.), Proceedings of the IDP05 International Symposium on Discourse-Prosody Interfaces. <http://aune.lpl.univ-aix.fr/~prodige/idp05/actes/izreel.pdf>

Izre'el, S., Hary, B. and Rahav, G. 2001. "Designing *CoSIH*: The corpus of spoken Israeli Hebrew". *International Journal of Corpus Linguistics* 6: 171-197.

Lambrecht, K. 1994. *Information Structure and Sentence Form*. Cambridge: Cambridge University Press.

Leech, G., Rayson, P. & Wilson, A, 2001. *Word Frequencies in Written and Spoken English,* London:Longman.

Karcevsky, S. 1931. "Sur la phonologie de la phrase". *Travaux du Cercle Linguistique de Prague* IV: 188-228.

MacWhinney, B. 1994. *The CHILDES Project: Tools for Analyzing Talk*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Moneglia, M. 2004. "Measurements of Spoken Language Variability in a Multilingual Corpus.Predictable Aspects" in M.T Lino, M.F. Xavier, F. Ferraira, R. Costa, R. Silva (eds) *Prococeeding of the 4th LREC Conference,* ELRA, Paris, vol 4 pp. 1419-1422.

Moreno-Fernández, F. 2005. "Corpora of Spoken Spanish Language – The rapresentatveness Issue" in Kawaguchi Y., Zaima S., Takagaki T. Shibano K. Usami, M.(eds.) *Linguistic Iformatics. State of the Art and the Future,* Amsterdam:John Benjamins.120-144.

Quirk, R. S. Greenbaum, G. Leech and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Scarano, A. (ed.). 2003. *Macro-syntaxe et Pragmatique. L'analyse Linguistique de l' Oral.* Roma: Bulzoni.

Simon, A. C. 2004. *La Structuration Prosodique du Discours en Français. Une Approche Multidimensionnelle et Expérientielle.* Berne: Peter Lang.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at work,* Amsterdam:John Benjamins.

C-Oral-Rom — French Corpus —

José DEULOFEU and Claire BLANCHE-BENVENISTE

1. Text Units and Segmentation with the "Illocutionary" Criterion

C-Oral-Rom contains a Spoken French Corpus oriented in a Romance contrastive perspective (Cresti 2005 p. 209). It can be very useful for second language teaching and for passive inter-comprehension. Here are some contributions for main French topics contained in *C-Oral-Rom* and some more developments that could be added.

The *C-OraL-Rom* corpus provides a segmentation of the text in text units based on prosodic and pragmatic cues. We can consider as texts units the segments included between two major pragmatic-prosodic boundaries : // , for assertions, //? for questions. Those boundaries encompass units characterized by prosodic autonomy and independent illocutionary force. The units separated by a minor prosodic boundary, having no terminal contours can be considered as sub-text units. Among the subunits we can distinguish the sub-unit which bears the illocutionary force of the whole text unit. It will be called the "nucleus". The nucleus can stand by itself as an autonomous utterance, which is not the case of the sub-units without illocutionary force. Utterances can be further classified according to the grammatical composition of their nucleus. The main distinction is between utterances with a tensed clause as nucleus ([+verbal]), and other types of nucleus ([- verbal]).

Oral performances can be characterized as a whole by the high rate of non-verbal utterances. And the various sub register of oral corpus can be further grouped as follows : there is a basic opposition between telephone register and formal register (nat), the other registers ranging gradually in between. The presence of a verb and the complexity of the utterance behave in the same direction, as it appears from the two tables below. The situation of media register is specific and should be further clarified.

Therefore *C-Oral-Rom* allows us to study how this parsing matches with grammatical units and compare the results with those evidenced in Croft (2002) for English and many other languages, This can help us to deal with some important linguistics issues regarding the structure of spoken language utterances and specially the grammatical composition of text units**.**

Some main points may be developed, to enlarge *C-Oral-Rom* descriptions:

— fragmented utterances and linking particles
— The ratio of verbal and non-verbal utterances
— Frequent macro-syntactic patterns built around a verbal nucleus and a final illocutionary frontier

2. Fragmented Utterances and Linking Particles

A corpus based account of several French conjunctions makes the description much easier. First, it is not true that the whole array of what is usually called “conjunctions” do act as subordinators. Second, in spoken usages, some conjunctions are pretty more frequent than other ones. Instead of teaching the whole list of usual conjunctions, the corpus makes it possible to sort out the most frequent ones, in such and such positions, and to show some of the most frequent types of embeddings. It enables learners to use syntactic complexity in a very early stage.

Some statistical results:

Table 1. simple and complex units
+simple (that is not including the subunit mark /)

cat	simple %	simple-verbless %	simple-verbal
fam	64.6	36.9	55.0
pub	64.2	33.5	56.1
tel	79.3	51.3	69.0
nat	36.9	23.1	31.3
med	56.9	27.0	49.9
TOTAL	61.5	36.0	51.9

By combining the two utterances features, we can characterize quite relevantly the linguistics registers of the corpus. The ratio verbless / verbal observed in the French corpus is inferior to those observed in other corpuses and is specially relevant to study the influence of the register. In some registers, telephone and natural context, the proportion is very high This leads us to develop in point 3 a qualitative study of verb-less utterances.

The ratio of simple / complex utterances can be interpreted as an index of fragmentation of the information inside the utterances. From this point of view, we may emphasise that telephone reveals as the most fragmented register.

We should be very cautious in commenting the data summarized in the tables of this section. The main reason is that they rely on the tagging, which is far from perfect, and furthermore they should be in many cases supplemented by a fine grained syntactic analysis in order to be really relevant. Nevertheless, some interesting comments to be further checked can be made.

The statistics above show that there are intricate interrelations between syntactic, morphologic, prosodic and pragmatic "integration" of clauses in one text unit. Based on observations on *que, parce que, pour que / pour inf*, it can be hypothesized that there is a tendency towards a parallel integration of subordinate into main clause within a text unit (the more syntactically integrated, the more prosodically and pragmatically integrated) But the tendency can be reversed due to information "packaging" factors ("light" subordinates can be topicalized or focalized).

An other way of putting the things is to notice the analytical way of packaging information: in informal styles, one descriptive semantic unit per text unit (utterance). In formal styles, more than one (verbs plus descriptive nouns).

Beyond the raw features found in the tables below, it is possible to draw some qualitative conclusions. For instance we can compare tables for coordination morphemes, *et*, *mais* (table 2) to those for subordination, *que* (table 3) and *parce que* (table 4). The tables provide the ratio of occurrences of these morphemes, as first tone unit (after / tag), as first utterance unit afer // or as first unit of a turn.

Table 2. Coordinators (units beginning by *et* or *mais)*

Cat	1st.tone Unit %	1st.utter %	1st.turn %
fam	33.0	16.2	8.3
pub	34.1	12.4	8.0
tel	56.2	25.0	17.1
nat	20.7	3.3	1.0
med	35.8	14.3	6.2
TOTAL	32.4	13.5	6.9

Table 3. subordinators : units beginning by *que*

cat	C/P	*que Conjunction*			*que Pronoun*		
		1st.tone U %	1st.utter %	1st.turn %	1st.tone U	1st.utter	1st.turn
fam	2.37	7.6	1.1	0.9	5.9	1.4	0.7
pub	3.08	9.5	0.5	0.3	11.8	5.9	5.9
tel	3.13	8.8	0.0	0.0	12.8	4.3	2.1
nat	2.74	12.6	0.9	0.1	16.0	0.4	0.0
med	2.94	8.1	0.9	0.9	12.6	5.0	2.7
TOTAL	2.66	9.1	0.9	0.6	10.3	2.4	1.5

Table 4. parce que

cat	1st.tone U	1st.utter	1st.turn
fam	42.8	11.9	5.7
pub	38.1	7.1	5.3
tel	62.5	18.1	13.9
nat	43.3	9.4	1.6
med	52.8	9.7	6.9
TOTAL	45.0	11.3	5.9

The figures are directly relevant to the question of “integration” of clauses into utterances. Basically, constructions can be linked either by discourse principles based on mere concatenation (juxtaposition) or by grammatical relations (subordination or coordination). A first correlation can be hypothesized between syntactic and morphologic levels. Special classes of morphemes are supposed to mark different types of links : zero or discourse particles for juxtaposition, conjunctions for subordination and coordination. This correlation can be phrased, as far as clauses are concerned: grammatical relations between clauses are marked by grammatical means and discourse relations by discourse particles. If we restrict to clauses, a further correlation between prosodic and syntactic integration has often been suggested : the more syntactically integrated, the more prosodically integrated. The tables above bring interesting empirical data in this regard. The figures on table 3 does not totally support the first correlation. Indeed, in more than 20 % of the cases coordinatives *et* and *mais* function between discourse connected units (first turn and first utterance unit), and this proportion is certainly higher in the case of clauses because many “coordinations” internal to one tone unit may concern non clausal coordination. So the figures confirm that coordinative conjunctions can play the part of discourse or interaction connectives. The second correlation is also disconfirmed by the figures: more than 50% of coordinations are not integrated in the same tone unit. The situation is radically distinct for subordination illustrated by *que*. The default case is clearly prosodic integration of the subordinate (about 90% of occurrences), with very few cases strictly interpretable in terms of discourse connective : less than 2%. The main function of utterances initial *que* conjunction is to introduce dislocated arguments, which excludes analysing it as a discourse connective. But the strong link that *que* bears with arguments of verbs makes it special among subordinators. So we decided to compare *que* case to the figures for an adjunct oriented subordinator, namely the most frequent of them, *parce que*. The choice was also motivated by the fact that previous studies pointed out that *parce que* clauses in spoken styles almost never occurred in preverbal position, which means that an utterance initial *parce que* is necessarily a discourse marker. Table 4 shows that the behaviour of *parce que* is closer to coordinator than to *que* subordinator. *Parce que* is still less prosodically integrated than *et* or *mais*. So our corpus confirms that some subordinators can function in marking either grammatical or discourse links.

3. Verbal vs Verbless Units

We have characterized utterances as text units by the feature [+/-verbal], that is utterances in which the sub-unit bearing the illocutionary force (nucleus) is either a tensed "main" verb [+verbal] or not [–verbal].

C-Oral-rom gives us quantitative data as shown in table 5 about the proportion of verb-less utterances [-verbal]. The results are specified by linguistic gender :

Table 5. Verbless versus verbal utterances in the French corpus

cat	verbless %	verbal
fam	25.9	74.1
pub	23.8	76.2
tel	44.1	55.9
nat	9.2	90.8
med	16.7	83.3
TOTAL	24.1	75.9

Italian, Portuguese and Spanish have nearly 37% of verb-less utterances, whereas French has a peculiar difference, with only 24%. We give below a framework for a typology of these structures.

Verb-less utterances, functioning as autonomous tone units encompass different phenomena that we have to distinguish at the structural level: utterances with one component or two components, with or without an explicit dependence to any previous verbal utterance.

3.1. One component verb-less utterances

Such utterances can be interpreted by themselves, without any necessary lexical link with anterior segments. They have several autonomous prosodic patterns.

(a) One specific type acts as a presentation sketch, sometimes as a title, frequently made of a nominal phrase:

(1) La Patagonie
'Patagonia'

(b) Nominal phrases realized with a suspended intonation, have exclamatory effects :

(2) Toutes les têtes des muguets étaient toutes penchées ! *Oh ! la crise de nerfs ! (Choix*83)
'Lily of the valley with all the heads bent down. Oh the fit of hysterics !

(c) Polite phrases and requests, marked with performative force, have specific prosodic patterns:

(3) *Merci à vous*
'thank to you'

(4) *Bonjour.* Alors je vais vous poser quelques questions (*Cierger* 1,1)
'Good morning Then I am going to ask you some questions'

(5) *S'il vous plaît* [1]! Je voudrais bien une, vous me faites une belle poupée (*Contes* 1,8)
'please ! I would like a…you give me a nice doll'

(d) In another type, the one-component utterance is prosodically autonomous but, having a negative or affirmative modality, *oui, non, pas du tout, bien sûr* (yes, no, not at all, of course) it refers to a previous discourse:

(6) L1.- C'est pas péjoratif ? – L2. – *Non. Pas du tout* ! (*Apo* 58,1)
'isn' it derogatory ? - No. Not at all'

(e) An important type, made of adjective, participial and adverbial phrases, acts as a comment on the ongoing speaking interaction, like *bon* (well):

(7) voilà, *bon*, à nous deux
'here we are, well, it is up to both of us'

Some, such as *juré, promis. bien fait* (sworn, promised, well done), act as strong performative elements, applying to a previous utterance. A very frequent one, *vrai* (true), enhances an earlier assertion

(8) C'est sur sa tête que ça tombe ! *Vrai* !
'it is on his head that it falls down. True'

(9) Il ne m'a pas écouté. *Bien fait pour lui !*
'he did not listen to me. Well done for him'

(10) C'était au sujet de l'alcool. *Bien entendu.*
'it was about alcohol. Of course'

They often behave as modalities do and they occur in the same positions, as can be seen for *oui* and *promis* (yes, promised) acting as responses in the same way:

[1] Dans "s'il vous plaît", il n'y a plus de verbe "plaire", comme on le voit dans la prononciation rapide, "siouplaît".

(11) L1 - Tu en prends bien soin ! - L2. - Euh, oui, oui ! *Promis ! promis* ! (Lesfil 1,9)

‘ you take care – Yes yes. Promised, promised

This type of participles and adjectives is strongly restricted by semantic and syntactic limitations. Participles cannot appear in this construction if they cannot be interpreted as stative. For instance *écouté, regardé* (listened to, looked at) are excluded, while *entendu, vu* (heard, seen) are frequently used. Verbs of feeling, belief, thought, act, do not meet the requirements and their past participles never occur as autonomous verb-less utterances:

(12) aimé, cru, prononcé, ….

‘loved, believed, pronounced’

Adjectives and participles compatible with this construction share another specific feature: apart from this verb-less status, their most frequent occurrence is with verbal locution *c’est* (it is), which indicates a kind of neutral meaning

(13) c’est vrai, c’est possible, c’est grave, c’est dommage

‘ it is true, it is possible, it is serious, it is a pity’

More than 50% of their occurrences are with [c’est -] and they very seldom occur in epithetic or predicative positions. Such are, for instance, *dommage, entendu, bien fait, fini, juré, compris, pas grave, pas possible, terminé, bien vu, vrai, merveilleux, superbe, tant mieux, tant pis* (pity, heard, well done, finished, sworn, understood, not serious, not possible, over, seen, true, marvellous, superb, all the best, too bad).

These restrictions clearly show that the use as verb-less utterances cannot be explained only by the virtue of elision.

(f) The verb-less utterance construction is a very ordinary pattern for answering questions in informal conversations. For instance, when Speaker 1 asks *à qui est-ce qu’elle cause ?* (who does she speak to ?), Speaker 2 answers with an utterance corresponding to the questioning part, *à qui* (whom…to) and he does not repeat the whole clause:

(14) L1. - A qui est-ce qu’elle cause, elle ? – L.2. *A son chien* (P95Div 13,11)

‘ Who does she speak to, she ? – To her dog’

The responding utterance usually contains the exact morpho-syntactic marks which were used inside the questioning clause. For instance, when the question was *d’où est-ce que vous êtes ?* (where are you from ?), with preposition *de* (from), the answer is introduced with the same preposition *de: de Poitiers* (from Poitiers):

(15) L1. – Vous, hein, d’où est-ce que vous êtes ? - L.2. Euh, *de Poitiers* (FRPRI 1295)

‘you, hey, where are you from ? – eh from Poitiers’

The same happens with negative answers (here preposition *en*) :

(16) L.1. –Vous y allez *en* quelle saison ? – L2 - seulement *en* été (Cl76,7)
‘ you go there in which season ? –only in summer’

The same happens when a speaker brings in a contradiction to a previous clause given by his interlocutor. In (17) Speaker 1 produced the clause *d’aller soigner les soldats* (to go and to tend the soldiers) and Speaker 2, a child, expresses and refutes an opposite interpretation, *pas de les tuer* (not to kill them):

(17) L 1. - Ça me plairait beaucoup par exemple d’aller soigner les soldats – L2. Ah ! *Mais pas de les tuer !* (Puget VI, 27)
‘That would really please me for instance to go and to tend the soldiers – Ah, but not to kill them

The verb-less utterance is then an exact copy of what was said before.

(g) Another well-known type of autonomous verb-less utterance is the one functioning as epexegis (cf. 4.1.2.) Although it has an independent prosodic pattern, the utterance keeps a grammatical link with a previous verb. It is often the case in conversation, when speech-turns break a verbal construction into various autonomous pieces. For instance, in (18), Speaker 1 uses the verb *vous avez commencé* (did you begin ?) in his first turn-taking; he then lets Speaker 2 answer the question. Then he comes to another turn-taking in which he brings in a complement to his precedent verb,, *avec la dame que je vous ai* dit (with the lady I told you):

(18) L.1. Vous avez commencé, alors ? – L.2. Ouais ! - L.1. Ah ben, c’est bien ! *Avec la dame que je vous ai dit* ? (*Choix* 120)
‘Did you begin, then ? – Yes I did – Ah, then it is good ! With the lady I told you ?’

Speakers very often use this possibility in order to bring in a contradictory element, as it happens in example (19), in which Speaker 3 denies a possible locative complement, *pas à Marseille* (not in Marseille) relative to an event mentioned earlier by Speaker 1, *il y avait eu un tremblement de terre* (there was an earthquake)

(19) L1. - Une fois je les ai entendus dire vaguement qu’il y avait eu un tremblement de terre, mais je crois que j’étais même pas née, voyez – L.2. Oui – L.3. *Mais pas à Mars-, à Marseille* (FP 9,13)
‘1: Once I heard them say vaguely that there had been an earthquake, but I think I was not even born, you see – 2: Yes –3: But not in Mars-, not in Marseille’

Speakers also use this type of opposite statement when they address their own discourse. For instance they first state an event, for instance, in (20), *ça m’arrivait* (it happened to me), and then, in another prosodic unit, ,they give a complementary adjunct, *mais pas toutes les semaines* (but not every week):

(20) Ça m'arrivait. *Mais pas tou- toutes les semaines* (*Femmes* 135,1)
'it happened to me – But not ev-, not every week'

(21) Elle était belle – *Mais pas comme j'aurais euh voulu* (*Femmes* 76,4)
'she was beautiful – But not the way I wanted'

These verb-less utterances have strong grammatical links within the preceding context and cannot be taken into account in the same way as utterances which have none.

3.2. Two-component verb-less utterances

In these utterances, the first component is usually the pre-nucleus, marked with a non-conclusive intonation, and the second component, bearing the conclusive intonation, is the nucleus (also said "comment", "predicate").

One type can be considered as completely independent, having no grammatical or lexical link with a previous clause. The nucleus part expresses an existential statement, whereas the pre-nucleus part contains precisions regarding time location (*le lendemain*: next day), or space location (*au centre du pays*: in the centre of the country):

(22) Le lendemain, bombe
'next day, bomb'

(23) Voilà, Juin 1940 : l'exode (Choix 145)
'here we are, June 1940 : exodus'

(24) Le matin, plus de valise
'next morning : no more suit-case

(25) Au centre du pays, une véritable guerre
'in the centre of the country, a real war'

In another type, the verb-less utterance is semantically and pragmatically linked to the previous context and it is interpreted as being part of previous meaning. For instance, in (26), Speaker 2 says *//pour le personnel / bien sûr//* (for the staff, of course). The first component is the theme, already mentioned by Speaker 1 in the previous context, with the same wording; the nucleus, *bien sûr,* is a positive modality applied to of the preceding clause, *c'est dur* (it is hard)

(26) L1- et c'est dur donc aussi moralement alors pour les pour le personnel – L2 – pour le personnel bien sûr (Choix 100, 77)
'and it is hard, so, also, morally, then for the for the staff ? – //for the staff / of course//

The opposite ordering, with the comment coming first in the nucleus and then theme coming second in the post-nucleus, is less frequent in ordinary conversations, although it is often quoted as a normal possibility in

French:

(27) Comment était l'émission ? Super, l'émission (Garcin 51,7)
'how was the broadcast ? - Superb, the broadcast'

Lexical, semantic and syntactic parallelisms often emphasize the links with the context. For instance in (28), there is a clear parallelism between two temporal indications, one in the previous verbal utterance, *il y a quinze ans* (fifteen years ago) and one in the verb-less construction which follows, *maintenant* (now)

(28) Il y a quinze vingt ans, je partais en vélo ; maintenant, terminé
' fifteen years ago, I used to go on bike ; now, finished'

Several topics can thus be given successively, as in (29) : the question produced by Speaker 1 in a verbal form, bears on the time when christening was given to a newborn child; the answer produced by Speaker 2 in two successive verb-less utterances, considers two topics, *moi* (me) and *il y a des gens* (some people):

(29) L1- et alors le baptême, il se faisait dans les combien de temps en général ?
L2 – oh moi, va, trois jours après.(…) il y a des gens, huit jours (*Sca* 81, 4)
'and then the christening, it happened how long after, generally ? - oh me, three days after (….) some people eight days'

However, apart from the situation of questions and immediate answers, the gapping of a precedent verb is rather rare in usual conversations. It only occurs in formal situations, for instance, as in (30), produced in a radio-talk :

(30) Le recueil est divisé en livres et chaque livre en poèmes (Choix 112)
'the anthology is divided into books and each book into poems'

Such gappings of the verb have been considered are highly grammaticalized expressions, because they can be embedded into a verbal construction,

(31) Il disait que le recueil est divisé en livres et chaque livre en poèmes
'he said the anthology is divided into books and each book into poems'

or they can be placed under the dependence of a conjunction:

(32) Il le fait pour que le recueil soit divisé en livres et *chaque livre en poèmes*
'he did it in order that the anthology be divided into books and each book into poems'

But no clear example is attested in all our data. Such heavy restrictions on the possibilities of realizing verb-gappings are to be taken into account. The only "natural" gapping of the verb happens when the nucleus part is the expression of a modality. For instance, in (33), Speaker 2 answers a question with a verb-less utterance in which the pre-nucleus part is a topic already mentioned earlier, *le patois d'Aussois / ceux d'Aussois* (the dialect from A. / those from A.) and the nucleus part is an affirmative modality, *oui* (yes)

(33) L.1. – Est-ce que tu comprendrais le patois euh d'Aussois, de Lanslebourg, des

villages, euh aux alentours ? - L.2. *Ceux d'Aussois, oui* (*Savoie* 2,16)
'Would you understand le dialect eh from A., L., villages all around ? – Those from Aussois, yes'

Affirmative modality can be combined with an additive effect, by using *aussi* (too, also)

(34) Peut-être que nous nous allons souffrir. *Nos enfants aussi* (*Chraibi* 64,3)
'perhaps we, we shall suffer. Our children too'

(35) Sept huit francs, dix francs, on avait un superbe cyclamen. *Un azalée aussi*, hein (*Choix* 94)
'for seven or eight francs, you could by a nice cyclamen. An azalea, too'

Negative modalities are expressed by *non, jamais, peut-être pas* (no, never, maybe not). For instance in (37), a topic *les autres*, contrasting with a precedent one *certains* (some), is given in the pre-nucleus part, and a negative modality *non* (no), linked to the earlier verb *surmonter* (get over), figures in the nucleus component:

(36) Certains les ont bien surmontés ; *les autres, non* (*Bonnet* 11,9)
'some could get over easily ; others, no'

Oppositions of the same type occur in next examples:

(37) Tu vois cette mer elle se lave deux fois par jour ; *la Méditerranée, jamais* (CF 4,8)
'you see this sea, it washes twice a day ; Mediterranean sea, never'

(38) Le client maintenant est exigeant. Il veut tout de suite, maintenant. *Avant, non* (*Choix* 62)
'the customer now is demanding. He wants it at once, now. Beforehand, no'

(39) Elle nous attire. *Vous, peut-être pas*. Mais moi je suis attirée (*Barian* 12,16)
'It attracts us. You, maybe not. But I am attracted'

Negative modalities are often given with several lexical variations, enhancing particular interpretative versions of the negation, for instance not a penny, not a word:

(40) J'ai un peu appris à parler l'allemand ; *le polonais, pas un mot* (Conseil 7,4)
I learned a little how to speak German; Polish, not a word'

It would be very interesting to describe the specific features which could explain why French has less verb-less utterances than the other three Romance languages quoted by *C-Oral-Rom.* In all four languages, verb-less utterances tend to appear in informal conversations more than in public speech, and more in dialogs than in monologs. But French grammar seems to include specific restrictions, which apply in particular to two-components utterances. A contrasting study, extending to other languages such as English and German, would certainly prove very useful.

4. A Typological Sketch of Fragmented Verbal Utterances

On quantitative grounds, fragmented verbal utterances represent two thirds of the total verbal utterances. On qualitative grounds, they show a wide range of patterns, far beyond the basic sentence structures given in reference grammars of French. Their common feature is that they are composed of one basic verbal text unit, which can function as an autonomous utterance. We called it the "nucleus". Around the nucleus are grouped various non autonomous text units, according to some determinate patterns. We will use a typology of minimal text units based on criteria adapted from the "macrosyntactic" approach (Blanche-Benveniste [1999, 2002]).

4.1. One nucleus utterances

We will first address the cases of subunits included within // boundaries. Two types of subunits can be distinguished, according to their position before or after the nucleus.

4.1.1. Subunits preceding the nucleus (prenucleus)

They are characterized by a raising non terminal contour. An important feature is that these segments are devoid of illocutionary force. We find patterns with numerous concatenated prenuclear units as in the following example (nuclear subunit is in bold):

(41) // mais bon honnêtement / ça me enfin je sais pas / moi / bord de mer comme ça / Cannes / tout ça / **c' est pas c' est une ville de vieux quoi** // (ffam11)
'but well honestly/ it me in fact I don't know/me/ seaside like this / Cannes / all this /it is not it is a town of old people, isn'it'

The nucleus can stand as an autonomous utterance:

c' est une ville de vieux quoi

On the contrary, the prenuclear part will not be accepted as an autonomous utterance :

? ? mais bon honnêtement / ça me enfin je sais pas / moi / bord de mer comme ça / Cannes / tout ça
/'but well honestly/ it me in fact I don't know/me/ seaside like this / Cannes / all

The prenuclear units fulfill various pragmatic functions : link with the preceding context, evaluation, stance, modality, spatio-temporal frame, topic setting. Utterances with numerous concatenated prenuclear units appear as a peculiarity of spontaneous spoken French. Some cases of prenucleus-nucleus patterns are interesting, for instance :

— "hanging topic", as in the following example:

(42) il y a plein de trucs tu **les** vois après en fait / **les défauts** (ffamcv 11)

‘ there are many things, you see them later on / the defects’

il y a introduces a pre-nuclear unit, with no co-referential link within the nucleus, while the pronoun *les* is co-referential with *les défauts*.

— wide scope prenuclear units. In some cases, the pre-nuclear unit extends its scope above several tensed clauses:

(43) il y a des personnes vous allez les voir elles pleurent pendant deux ou trois heures elles vous pleurent dans les bras et vous repartez elles pleurent encore
‘there are people you go and see them they cry for two or three hours they cry in your arms and you leave them they are still crying’

— sub-groupings of prenuclear units. A very frequent grouping in prenuclear position consists of first person pronoun *moi* and a temporal or locative adjunct, both forming a global unit, which can be paraphrased as : ‘since I am in Master:’

(44) moi en maîtrise / je peux faire une demande (ffamcv01)
‘me in Master/ I may make an application’

4.1.2. Subunits following the nucleus

There are two types of subunits following the nucleus, appenix and suffix.

4.1.2.1. Appendix

The first type is exemplified in (42). The subunit *les défauts* is characterized by a specific “flat” contour determined by the contour of the nucleus. As for functional aspects those units are interpreted as afterthoughts. They do not constitute an illocutionnary act per se.

4.1.2.2. Suffix

We have further identified larger text units which extend beyond a // boundary. They are composed of an obligatory nucleus and of one or more units following the nucleus. The units following the nucleus share with the nucleus the property of bearing an illocutionary force, but they cannot constitute an autonomous utterance. This is the reason why we call these units “suffixes”, as they appear as syntactic units obligatorily linked to a preceding nucleus (as it is the case for morphological suffixes always dependent on a lexical “root”). They are introduced by various connectives (mainly coordinative and subordinative conjunctions). This framework allows us to give theoretical status to the special behaviour of the clauses identified in section 1, which are introduced by a subordinative conjunction but follow a main prosodic boundary. Two subcases must be distinguished on descriptive grounds.

— Suffix without grammatical link. In the first subcase the “suffix”is not linked to the nucleus by any dependency relation:

(45) Moi je peux vous dire que c'est dans les mines qu' on a une plus grande camaraderie // **car** vraiment euh sur le plan s [/] solidaire on est tous euh [/] tous ensemble quoi //

‘me I can tell you that it is in the mines that you have the greater friendship//for really euh on the level of solidarity we are all together’

The clause introduced by *car,* a coordinative conjunction, cannot function as a nucleus, as shows the inacceptability of (46)

(46) ?? Moi **car** vraiment euh sur le plan &s [/] solidaire on est tous &euh [/] tous ensemble quoi//

Nevertheless the presence of adverb *vraiment* and discourse particle *quoi* shows that the clause has the illocutionary force of an assertion.

The same pattern can be observed with subordinative conjunction *quoique* (although). In (47) it follows a non verbal nucleus *(ouais)* The independent assertive illocutionary force of the *quoique*-clause is underlined by the discourse feedback particle *hein*. Furthermore, there is a morphological mark of this “macrosyntactic” use of *quoique*.: indicative mood (*c’est*) instead of the subjunctive mood, which appears in subordinative devices:

(47) L1 - et surtout que si tu es classé en deuxième ou troisième position tu as encore une chance tu vois / c’ est pas terminé tu vois / surtout deuxième //

L2 - ouais // **quoique** / bon tu sais si c’ est le local qui se bat / euh **c’ est** rare qu’ il s’ en aille hein //

‘ and above all if you are in second or third position you still have a chance you see/ it is not finished you see/ specially second

Yeah// although / well you know if it is the local candidate who competes, he never resigns

In the following example, *parce que* introduces a suffix with interrogative force:

(48) c’est quoi que tu aimais **parce que** tu le voyais comment ?

‘what was it that you liked because you figure him how ‘

If *parce que* was a subordinative conjunction, it would be impossible to have it in an embedded wh question clause.

The forms of suffixes are far less restricted than those of nuclear subordinate clauses, as we can see in example (49) in which *puisque* introduces a long concatenation of sentences :

(49) c’est à partir de ce moment là que la carrière d’un arbitre se développe considérablement **puisque** s’il a des aptitudes il commence à arbitrer en 3eme division de rugby s’il a des aptitudes il monte en seconde et s’il a des aptitudes il

finit en première (Orlandi, 12,9)
'it is from this time on that the progression of a referee develops considerably because if he has got talents, he begins in third division in rugby, if he has got talents he goes up to second division and if he has got talents he ends with first division'

— Suffix with grammatical link (épexégèse). In this case, a unit which is a syntactic dependent of a noun or of a verb, functions as an autonomous text unit, not grouped with the nucleus of the governing category. In example (50) an adjective modifier is separated by a main prosodic boundary from a head noun :

(50) ce qui me préoccupe c'est l'investissement au niveau du travail // **mais physique** (ffamcv06)
'what bothers me is the investment at the level of work // but physical investment'

In example (51) the same situation happens with a verbal modifier:

(51) tu retournes dans ta famille // à Toulon non ? (ffamcv11)
'do you go back to your family // in Toulon, isn't'

All these fragmented utterances seem to support the "Preferred-Clause Construction" hypothesis (Lambrecht 1986), according which the canonical form of utterances in spontaneous speech consists of only one tensed clause with preferably one lexical phrase. But, interestingly enough, other structures in our typology evidence an opposite tendency in that they present a grouping two tensed clauses in a single nucleus.

4.1.3. Two autonomous tensed clauses combined in a unique nucleus

It is a case of concatenation of two tensed clauses without any marker: two clauses are subsumed under a unique illocutionary force and form a semantically unique event (not a complex of events). There is no prosodic boundary between the two clauses, which form a unique nucleus (between brackets). The second verbal construction, *I was a teenager*, is treated as an adjunct for the first one, although there is no subordination marker. It behaves semantically exactly as a semantic modifier of the verb.

(52) Moi [j'y suis allée j'étais adolescente] /
'me I went there I was a teenager'

There are cases in which a prosodic boundary matches exactly with an interpretation of the second clause as modifier

(53) / bon elle fait un bruit sur l' autoroute / on dirait un un avion (ffamcv11)
'well it makes a noise on the highway you would say it is an an airplane'

Such examples show that further investigation is needed to clarify the prosodic pragmatic interface.

4.2. Multinucleus utterances

Nuclei may combine into one single utterance, according to various patterns. Here are some examples.

4.2.1. Parallel groupings of nuclei

A typical parallel grouping relies on the use of correlative morphemes, such as *the more… the more*:

(54) **autant** il y a beaucoup de gens euh au Népal qui vont faire du trekking et des choses comme ça **autant** en Inde très peu non (voy, 43)
‘as much as there are many people euh in Nepal going trekking and things like that, as much there are very few in India, no’

In such cases, both components are mutually dependent and no one can form a possible utterance.

4.2.2. Parenthesis

A nucleus can be quite freely inserted into another one, with a back-grounding function :

(55) J’aimais mieux l’autre route parce que avant (maintenant ça a changé) c’était une petite route de montagne
‘ I preferred the other road because before (now it has changed) it was a little mountain road’

4.2.3. Nucleus grafted in a construction

(56) c’est tellement peu important / **que oh allez / à la poubelle / avec le reste** (fnat pd OI)
‘it is so unimportant that oh go, to the garbage with the rest’

Instead of the expected tensed verbal clause, the complementizer *que* is followed by a non-verbal nucleus. There are numerous occurrences of grammatical constructions completed by syntactically unrestricted forms which can function elsewhere as independent nuclei.

Conclusion

A major challenge for a corpus based linguistic is to complete the typology of utterances in spontaneous speech. In order to meet it, we need to increase the number of available corpora. The *C-Oral-Rom* model, is a positive starting point, in as much as it allows to establish morpho-syntactic and prosodic descriptive generalisations.

References

ANDERSEN, Hanne-Leth et NØLKE, Henning, (éds.), *Macro-syntaxe et macro-sémantique, Actes du colloque international d’Aarhus, 17-19* *mai 2001*. Berne : Peter Lang (Sciences pour la communication).

BEGUELIN, Marie-José, 2002, “Routines macro-syntaxiques et grammaticalisation : l’évolution des clauses en *n’importe*”, in H.L.ANDERSEN et H. NØLKE, 43-70.

BERRENDONNER, Alain, 1990, “Pour une macro-syntaxe”, *Travaux de Linguistique* n° 112, 31-49.

BERRENDONNER, Alain, 2002, “Morpho-syntaxe, pragma-syntaxe, et ambivalences sémantiques”, in H-L ANDERSEN et H.NØLKE, pp.23-42

BILGER, Mireille, 1998, “Le statut micro et macrosyntaxique de ET”, in M. BILGER, F.GADET, et K. van den EYNDE (éds.), *Analyse linguistique et approches de l’oral. Recueil d’étude offerts en hommage à Claire Blanche-Benveniste*. Louvain/Paris : Peeters, 91-102.

BLANCHE-BENVENISTE, Claire, 1999, *Approches de la langue parlée en français.* Paris : Ophrys.

BLANCHE-BENVENISTE, Claire, 2002 “Macro-syntaxe et micro-syntaxe : les dispositifs de la rection verbale”, in A.L. ANDERSEN et H. NØLKE, 95-118.

BLANCHE-BENVENISTE, Claire, à paraître, “Phrase et construction verbale”, in M. Charolles, Le Goffic et M.A. Morel (éds.), *Y a-t-il une syntaxe au-delà de la phrase ?* Colloque de Paris-3.

BLANCHE-BENVENISTE, Claire, BILGER, Mireille, ROUGET, Christine et Van den EYNDE, Karel, 1990, *Le Français parlé : études grammaticales*. Paris : éditions du CNRS.

BLASCO-DULBECCO, Mylène, 1999, *Les constructions disloquées en français contemporain*. Paris : Champion (Collection “les français parlés : textes et études”).

CRESTI, E & MONEGLIA, M. (eds) 2005, *C-ORAL-ROM Integrated Reference Corpora for spoken Romance Languages*, Amsterdam, John Benjamins (Studies in Corpus Linguistics 15).

CROFT, William, “Intonation units and grammatical structure in Wardaman and English”

DEBAISIEUX, Jeanne-Marie, 2002, “Le fonctionnement de parce que en français parlé : étude quantitative sur corpus”, in Claus D. Pusch, Wolfgang Raible (eds.), Romanistische Korpuslinguistik - Korpora und gesprochene Sprache, Romance Corpus Linguistics, Corpora and Spoken Language, Gunter Narr Verlag Tübingen

DEULOFEU, José, 1986, “Syntaxe de *QUE* en français parlé et le problème de la subordination”, *Recherches Sur le Français Parlé* n° 8, 79-104.

DEULOFEU, José, 1988, “Les couplages de constructions verbales en français parlé”, *Recherches Sur le Français Parlé* n° 8, 79-104.

DEULOFEU, José, 1999, “Questions de méthode dans l’étude du morphème *que* en français contemporain”, *Recherches Sur le Français Parlé* n° 15

DEULOFEU, José, 2001, “La notion de construction corrélative en français : typologie et limites”, *Recherches sur le Français Parlé* n° 16, 103-124.

DEULOFEU, José, 2002, “Cadres pour une typologie syntaxique des prédications en français”, in S. LEROY et A. NOWAKOWSKA (éds.), *Aspects de la prédication*. Université P. Valery : *Praxiling.*, 199-220. pp. 149-158.

HOPPER, Paul. 1988. “Emergent grammar and the A Priori Grammar Postulate”. *Linguistics in context: Connecting observation and understanding,* ed. by Deborah Tannen, 117-134.Norwood, NJ, Ablex.

IWASAKI, Sh., ONO Ts. 2001, ‘Sentence in spontaneous spoken Japanese language”, in Bybee and Noonan (ed) *Complex sentences in Grammar and Discourse*, 176-202Amsterdam, John Benjamins

LAMBRECHT, K., 1986, *Topic, focus, and the grammar of spoken French*. Unpublished PhD dissertation, University of Calfifornia, Berkeley.

MARTIN, Philippe, 1999, “L’intonation en parole spontanée”, *Revue Française de Linguistique Appliquée*, vol IV-2, *L’Oral spontané,* dirigé par M. Bilger, 57-76.

MITHUN, M. (2005) “On the assumption of the sentence as the basic unit of syntactic structure” in *Linguistic diversity and language theories*, Zygmund Frayzingier (ed) Studies in language Companion series, Amsterdam, John Benjamins

ONO, T. & S. A. THOMSPON. 1995 “What can conversation tell us about syntax? ”, *Alternative linguistics,* ed. by Philip W. Davis. Amsterdam & Philadelphia: Benjamins

MOREL, Mary-Annick, 2002, “Intonation et gestion du sens dans le dialogue oral en français”, in H.L. ANDERSEN et H. NOLKE, 119-139.

MOREL, Mary-Annick et L. DANON-BOILEAU, 1998, *Grammaire de l’intonation. L’exemple du français.* Paris : Ophrys.

MULLER, Ckaude, 2002, “Schèmes syntaxiques dans les énoncés longs : où commence la macro-syntaxe ? ”, in H.L. ANDERSEN et H. NOLKE, 71-94.

SABIO, Frédéric, 1995, “Micro-syntaxe et macro-syntaxe : l’exemple des compléments antéposés”, *Recherches Sur le Français Parlé* n° 13, 111-156.

SABIO, Frédéric, 2002, “L’opposition de modalité en français parlé : étude macro-syntaxique”, *Recherches Sur le Français Parlé* n°17, 55-78.

Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus — Methodology, Tools and Evaluation —

Antonio MORENO-SANDOVAL and José M. GUIRAO

1. Problems in spontaneous speech morpho-syntactic tagging

This paper summarises the experience of the LLI-UAM group in tagging one of the largest spontaneous speech corpora now available (over 300.000 transcribed words). To begin with, we will describe some tagging problems especially relevant in spoken corpora and we will introduce the tagging procedure and the tool developed for helping human annotators in the process. An evaluation of the precision rate provided by the tagger, as well as the general reliability of the expert human annotators are calculated based on a 'gold standard' corpus of 150.000 words.

The goal of any morpho-syntactic annotated corpus is to provide the most appropriate tag with morphological, part-of-speech and lexical information for every token in the text. An interesting question that will be addressed here is whether or not there is any difference in tagging written and spoken texts. The evaluation and comparison of different taggers will be also discussed.

1.1. Basic concepts

Morpho-syntactic tagging is the assignation of the most appropriate tag (that is, a descriptive symbol) with grammatical and lexical information for every token in the text. A tagset is the definition of all possible tags and the criteria for assignation to wordforms. Different tagsets vary regarding the information they provide. The basic, obligatory annotation includes:

- Part-of-Speech
- Lemma
- Grammatical features

The design of a tagset for a given language critically depends on 1) the morphological characteristics of the language; 2) the automatic tagger; and 3) the kind of texts to be annotated.

With respect to language, the morphological complexity of an inflecting or fusional language like Italian, Spanish or Russian implies many more possible tags than the equivalent set for an isolating or analytic language such as English or Chinese. For example, the EAGLES Guidelines (EAGLES 1996) provide two tagsets with equivalent information for English and Italian. Just for verbs, we found 21 tags for English and 84 tags for Italian. Morphology is the linguistic level that most shows surface variation across languages. In accordance with this typological fact, it is difficult to define a universal tagset size. If we wanted to use a similar number of tags for English and Italian, for example, it is clear that it would be overspecification (in the case of English) or underspecification (for Italian). That is, additional information or lack of information in the morpho-syntactic annotation of a corpus.

The tagset size affects directly to the automatic tagger performance. A large tagset (bigger than 100 tags) may potentially be most discriminative than a small tagset (around 30 tags) but more tags also lead to sparse data problems for statistical taggers. To ensure good performance, those taggers must be trained on a sufficient number of examples. With small tagsets, the training data will contain a large number of each word sequence, but on the other hand the discriminative ability will be worse. Discrimination is achieved by allowing a large tagset that can model individual cases. There is an inevitable trade-off between robustness (a small tagset) and fine-grained discrimination (a large tagset). Finding out the best tagset is a hard problem and most researchers approach it empirically: setting a series of experiments using various tagsets and choosing the one with better results for a given tagger and training set.

The last factor is the type of texts. "Corpora drawn for different communicative activities differ greatly one from another in respect of word frequencies, construction frequencies, and so forth" (Gazdar 1996:19). This is especially true when tagging written and spoken texts. A contribution of this article is to quantify the increased ambiguity in morpho-syntactic tagging of spontaneous speech compared with the same task with written corpora of Spanish. These facts (the ambiguity in spoken vs. written texts, and in different languages) are relevant when considering a comparison between taggers since, as is currently accepted in the field, the problem of POS tagging is to resolve the ambiguities, choosing the proper tag for the context (Jurafsky & Martin 2000). In the evaluation section, we will discuss the problem. Next, we will present other relevant differences in tagging spoken with respect to written texts.

1.2. Multi-word and amalgam recognition

In order to assign properly the lemma, the meaning must be taken into account regardless the number of orthographical words that the lemma consists of. The language evolution or the writing conventions make the lexical unit not always corresponding with the graphical word, understood as a string between blank spaces. As in other linguistic levels, we face the significant-signified asymmetry (many to one or one to many):

- Multi-words, such as 'Buenos días' (*Good morning*) or 'es decir' (*that is*).
- Amalgams, such as 'del' (a preposition fused with an article, 'de' and 'el') or 'dámelo' (a combination of an imperative wordform plus two clitics, 'da' 'me' 'lo')

In the first case, several significants represent a single signified. In the latter, a graphical word combines several independent meanings. In Spanish, as in most languages, multi-words occur more frequently than amalgams.

Multi-word and amalgam recognition is a tokenization problem both for written and spoken taggers of Spanish, but they are especially frequent in speech, since there is an extensive use of imperatives with clitics and multi-word discourse markers. For instance, four multi-words rank among the top 100 frequent lemmas in the Spanish C-ORAL-ROM corpus. But more important, multi-word detection is necessary for getting a correct annotation of many word sequences, as in Table 1.

Table 1. Multi-word annotation

Correct tagging	Incorrect tagging
o_sea/DM *that is*	o/C + sea/V
en_lugar_de/PREP *instead of*	en/PREP + lugar/N + de/PREP
por_ejemplo/DM *for example*	por/PREP + ejemplo/N

As a consequence, a tagger that cannot analyse multi-words properly will produce poor results for a spoken corpus. This fact is relevant for comparing taggers.

1.3. A tag for Discourse Markers

The main annotation guidelines (EAGLES, XCES) do not recommend a tag for discourse markers, although they are very often described in the linguistic literature on spoken language. A discourse marker is a linguistic unit that guides, according to their morpho-syntactic, semantic and pragmatic features, the inferences which take place in communication (Portolés & Martín Zorraquino 1999).

Discourse markers are much more frequent in spontaneous speech than in written texts: 12 DMs are in the top-100 Spanish C-ORAL-ROM list. As in the case of the multi-words, DMs tagging is a must for the correct annotation of many words, as shown in Table 2, where the same token can be given at least two different tags:

Table 2. Examples of ambiguous Discourse Markers

bueno/ADJ	bueno/DM
Juan es bueno Juan is *good*	bueno / espero que te guste *well* / I hope you like it
hombre/N	hombre/DM
Juan es un hombre bueno Juan is a good *man*	hombre / no te enfades Don’t be mad / *man*.

Most tagsets and taggers do not include a specific tag for DMs, they are treated as ADVs, ADJs, INTs or Ns instead. In the last section, an empirical estimation of the percentage of DMs occurring in spontaneous speech will be given.

Some partners in the C-ORAL-ROM project decided not to tag them as DMs, on the basis that “there is no widely accepted analysis concerning these words” and “they also are very often homographs of adjectives, pronouns or adverbs. This renders automatic tagging very difficult.” (Campione, Véronis & Deulofeu 2005:118).

As for the difficulty of deciding whether the proper tag is DM or another PoS, from our experience we found that intonation and the pragmatic context help the trained annotator in most cases.

The second argument by Campione, Véronis and Deulofeu is correct: DM ambiguity is responsible for a residual uncertainty, which is almost impossible to handle by the current taggers. We prefer, however, to reduce our precision rate but to improve the quality of the annotation, adding a finer analysis of these particles, which are essential for guiding any spontaneous dialogue or conversation.

In the Table 3, a comparison of equivalent units in French, Italian, Spanish and Portuguese for the English words ‘for example’ and ‘then’/’well’ are given with the total number of occurrences in each corpus. French and Italian tagsets do not have a tag for DM, using ADV instead. Spanish and Portuguese distinguish between ADV and DM, and the Portuguese has the additional distinction between Discourse Marker (DM) and Discourse Locution (LD)[1].

Table 3. Comparison of discursive particles in the four corpora

	French	Italian	Spanish	Portuguese
For example	par exemple ADV 156	per esemplio B 126	por ejemplo MD 256	por exemplo LD 139
Then / well	alors ADV 958	allora CC 581	entonces P 759	então ADV 365
		allora B 430	entonces MD 236	então MD 316

As the data reflect, there are different interpretations for the appropriate POS tag assignation. In the case of ‘for example’, the French and Italian teams consider it as an ADV while the Spanish and Portuguese annotate it as a discursive particle. With respect to ‘then/well’, there is no agreement between groups, but at least three teams distinguish among two different tags.

1.4. Tokenization

Normally, the presence of whitespace surrounding a single character or a group of characters defines an explicit token. In NLP, tokenization is extended to all the previous text processing that takes place before the lexical look-up and the morpho-syntactic tagging. It is the segmentation of the text in lexical units (words, both single and multi-words), punctuation units, and textual units (sentences, paragraphs, etc.).

Tokenizing lexical units includes recognition of multi-words and amalgams (already mentioned in 1.3.) as well as recognition of Named Entities (NE), term introduced by R. Grishman in the MUC-6 (the Sixth Message Understanding Conference). Originally, the NE recognition consisted “of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are “unique identifiers” of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages).” (Grishman 1996). Although NE can appear in spoken language, they are typical of the written texts, as used in Information Extraction and Retrieval. Proper names are quite frequent in spoken texts, but they are not a problem for the tagger, since the transcription

[1] CC stands for “Coordinative conjunction” and B for “Adverb” in the Italian tagset. ‘Entonces’ (*then*) is tagged as Pronoun in the Spanish corpus because of its deictic nature. See more details in section 2.1.

convention permits only proper names to be written with uppercase. Therefore, every token starting with an uppercase character is annotated as NPR.

The tokenizer must be adapted to the conventions of the punctuation and textual units in spoken corpora. Table 4 shows the differences between both types of texts.

Table 4. Differences in tokenization

Spoken texts	Written texts
Turns, utterances	Paragraph, sentences
Prosodic marks (tone units, retracting, overlapping, disfluencies marks)	Punctuation marks

1.5. Disfluencies and prosodic breaks

Speech disfluencies are elements of speech which are not generally recognized as containing formal meaning, usually expressed as filled pauses such as *uh* or *er*, but also extending to repairs or retractring (also called, false starts), and paralinguistic elements such as laugh or cry. All these phenomena are quite frequent in spoken texts, and they are a potential source of problem for the taggers. The way this problem is handled in C-ORAL-ROM is marking disfluencies with special symbols:

- & speech fragments, example, *&eh, &ah, &bue* (fragment for *Buenos días*)
- hhh paralinguistic elements, such laugh.

A real problem for the taggers are the prosodic breaks, which are considered in C-ORAL-ROM the most relevant cue for determining utterance boundaries (Moneglia 2005). Utterances for the spoken language are somehow the equivalent to sentences in written texts. Using prosodic boundaries as tone unit markers allows the division of the speech into information units. There are two kinds of prosodic breaks: terminal and non-terminal. The first type has the quality of concluding a speech sequence (equivalent to the accomplishment of an illocutionary or locutionary act). The latter does not have the conclusive quality. In C-ORAL-ROM each sequence ending with a terminal break is considered an utterance. Table 5 shows the prosodic break types annotated in C-ORAL-ROM:

Table 5. Prosodic break types

Terminal breaks		Non-terminal breaks	
Symbol	Description	Symbol	Description
//	Conclusive prosodic break	/	Non conclusive prosodic break
?	Conclusive with interrogative value	[/]	Non conclusive, a false start with repetition
…	Conclusive, intentionally suspended by the speaker	[///]	Non conclusive, retracting with no repetition
+	Conclusive, interrupted		

For the automatic taggers, the problematic prosodic breaks are those when the utterance is not 'properly' concluded: suspension (…), interruption (+), and the repairs, with and without repetition of linguistic material. In general, automatic taggers are trained on written corpora, where sentences are finished and unnecessary repetition ('la la la muchacha era alta,' *the the the girl was tall*) is avoided. 'Agrammatical' sequences are common in spoken language and precision provided by written-trained taggers is reduced. In C-ORAL-ROM, the French tagger was capable of detecting repetitions, and according to the French team (Campione et al. 2005: 120) that fact explains to a large extent the very high results obtained (98.75 % precision). In the Spanish corpus, we have not yet applied a similar strategy and our estimation is that at least 1% of the total precision depends on the proper treatment of these phenomena.

1.6. Unknown words

"Vocabulary increases with corpus size. No matter how big one's lexicon, previously unknown words will always be encountered" (Gale & Church 1990, cited from Gazdar 1996). The famous problem of the sparse data in corpus, firstly noted by Chomsky, has its counterpoint in the lexicon: the best lexical competence (human or machine) will always lack of complete knowledge of every word in a given language.

Moreno & Guirao (2003) conducted an experiment before starting the annotation process of the corpus in order to estimate the percentage of words unknown for the tagger. The result was 8% of the test corpus (22747 tokens), divided in four categories:

- Foreign words.
- Words typically or exclusively of the spoken register.
- Errors in transcription.
- Neologisms, mostly derivatives.

Handling the first two types is just a matter of adding those words to the lexicon. Errors in the source transcription texts were corrected, and then tagged by the program. The last type was more effort-consuming: a new set of rules for handling derivative morphology was added to the analyser. At the final evaluation, only 117 tokens (out of 313.504) were not given a tag by the program. Most of them were disfluencies (fragments), strings of alphanumeric codes, acronyms and residual tokens.

2. Methodology and tools for annotating the Spanish C-ORAL-ROM corpus

2.1. The tagset and some specific characteristics

In Spanish linguistics, as in any other language grammar studies, the POS problem is still far from being solved. We wanted to define a neutral and general tagset, but also one that reflected the peculiar features of the spoken language. Both the functionalist and the formal approach are typically syntactic-oriented in the definition of the POS tags. To our knowledge, this is so because of the clear orientation to study written samples. In the written register, the sentence is well-defined, and it constitutes the core linguistic unit. Segmenting the sentence in smaller units on the basis of structural arguments is a natural consequence. On the contrary, in the spoken register the core unit is not yet clearly defined. In C-ORAL-ROM, we support the *utterance*, on the ground of prosody and speech acts, as the spoken unit. In any case, the dialogic nature of the spoken language needs a slightly different approach that reflects its distinctive features. In particular, each spoken sample is a communicative act with the basic purpose of transmitting information. According to this, the meaning and the semantic criterion has been favoured with respect to the syntactic and morphological points of view, when assigning a tag.

The full tagset consists of 129 tags[2], including one for each Spanish verb wordform (46) and each auxiliary verb wordform (46). This distinction between V and AUX is one of the main sources of ambiguity for the automatic tagger. The tagset organised by categories is given in the Table 6.

[2] The complete tagset can be consulted in the DVD accompanying the book (Cresti & Moneglia 2005) and also in the home page of the Madrid C-ORAL-ROM (http://www.lllf.uam.es/coralrom/).

Table 6. Spanish C-ORAL-ROM tagset

Category	Main Tag	Number of different subtags per category
Noun	N	5
Proper Noun	NPR	1
Adjective	ADJ	9
Article	ART	4
Possesive	POSS	1
Demonstrative	DEM	1
Quantifier	Q	1
Pronoun	P	9
Relative Pron.	REL	1
Verb	V	46
Auxiliary verb	AUX	46
Preposition	PREP	1
Adverb	ADV	1
Conjunction	C	1
Discourse Marker	MD	1
Interjection	INTJ	1

There is no tag for Punctuation, necessary in the written corpora tagsets, since the equivalent spoken prosodic marks are not considered part of the morpho-syntactic annotation.

A specific tag for discourse markers (MD) is one of the distinctive features of this tagset. Other tagging decisions are:

- A string of words is considered a multi-word when: 1) there is absence of compositional meaning; and 2) no insertion of words is possible.
- All the traditionally classified as pronominal adverbs (Kovacci 1999) are annotated here as pronouns, on the basis of their deictic behaviour. We adopt the semantic criterion in the definition of those POS such as ‘ahora’ (*now*), ‘ayer’ (*yesterday*), ‘entonces’ (*then*), etc. These words behave semantically as pronouns because they are open entities whose referent is not fixed nor kept constant, but it changes with speaker, listener or space and time coordinates. This referential value is typical of pronouns.

2.2. The tagger: GRAMPAL

GRAMPAL was originally developed as a morphological processor of Spanish (Moreno 1991, Moreno & Goñi 1995) for written texts. In order to annotate C-ORAL-ROM, new modules were specifically developed for spoken Spanish: a tokenizer, disambiguation modules and an unknown words recogniser. (Moreno & Guirao 2003, Moreno et al. 2005:143-146).

The GRAMPAL lexicons currently consist of approximately 50.000 entries of stems, endings and multi-words (Table 7). New entries can be easily added or current entries can be modified or removed, thanks to a tool developed to help in the annotation process (see next section).

Table 7. Entries in the GRAMPAL lexicons

Endings lexicon		Stems lexicon		Multi-words lexicon	
		N	25,426	ADV	507
Noun morphs	4	ADJ	11,290	PREP	349
Verb morphs	179	V	10,568	C	91
		ADV	189	INTJ	70
		P	109	FOREING-W	30
		MD	74	Q	24
		PREP	40	ADJ	15
		C	26		
		POSS	26		
		REL	16		
		ART	5		
Total	183		47,769		1086

The tagging procedure involves several phases:

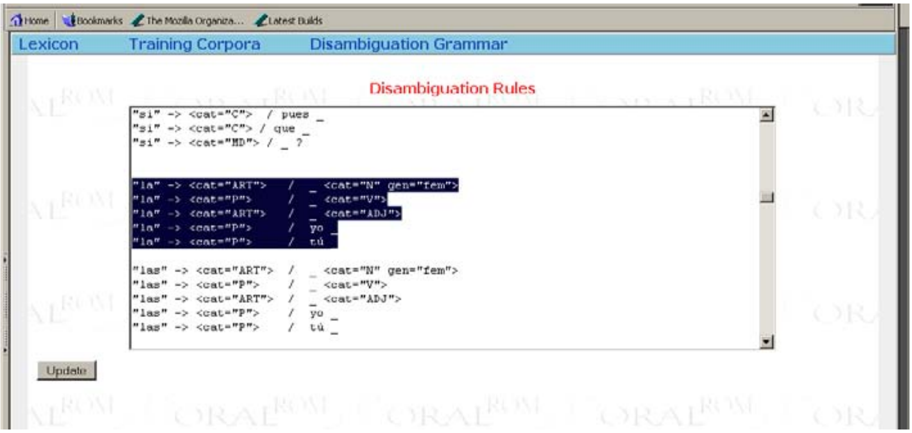
1. Lexical pre-processing: once the tokenizer has segmented the transcription in tokens, the programme splits the fused words (amalgams and verbs with clitics).
2. Multi-word recognition: the text is scanned for candidates for multi-words, looking up a lexicon compiled from printed dictionaries and corpora. After the process, any non-ambiguous multi-word is annotated.
3. Single word recognition: every single word is given all the possible tags, according to the morphological rules and the general lexicon. Approximately 52% of the single words have more than one possible analysis.
4. Unknown words recognition: the remaining tokens pass first through the derivative morphology rules. If any token remains still un-tagged, they are held until the last phase (6), when the most likely tag is given.
5. Disambiguation phase 1: a feature-based Constraint Grammar resolves some of the ambiguities.
6. Disambiguation phase 2: A statistical tagger (the TnT tagger, Brants 2000) resolves the remaining ambiguous analyses, and the unknown words.

2.3. Tools

Moreno and Guirao have developed some tools for assisting human annotators in two different tasks: lexicon and disambiguation grammar management (the *developing tool*) and annotation revision (the *revision tool*).

During the annotation process, which took more than six months, new lexicon entries were added and the disambiguation grammar was written. Several runs and tests were conducted to improve the tagger for spoken language. In order to have better control of the different tasks, a web-based tool was created (Guirao & Moreno 2004), consisting of three editors:

- *Training corpus editor*: the most basic tool is an editor-concordancer that allows searching for problems and wrong analyses. This concordancer is used to edit by hand incorrect tags, which were not properly assigned by the tagger. This way a first training set of 50000 tokens was finished in order to train the statistical module.
- *Lexicon editor*: this option edits the GRAMPAL lexicon, allows introducing, modifying and deleting entries, and saves the enriched lexicon.
- *Disambiguation grammar editor*: in the interactive process of revising the annotated texts, the grammar writer wants to add new rules. This editor allows the human expert to edit the grammar file, compile it and try an utterance. This option is useful for checking new rules without running the whole tagging process on a text (see Figure 1).



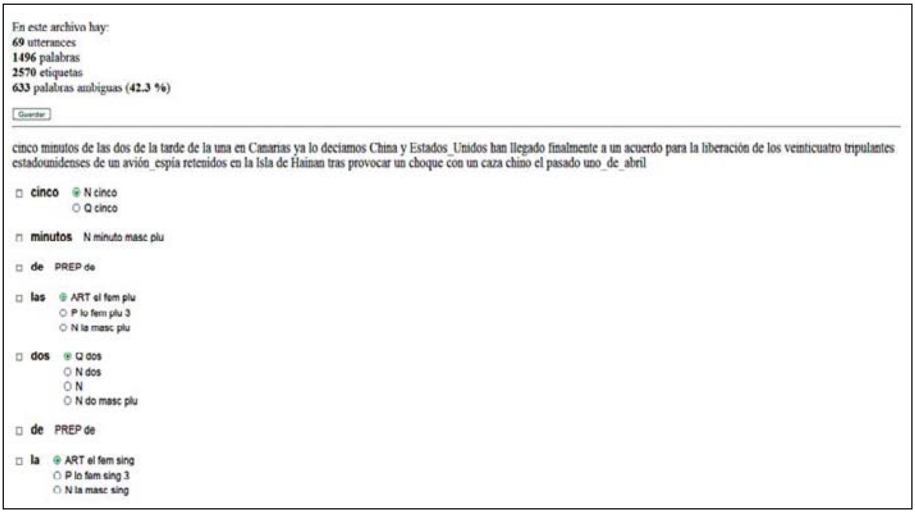
```
"si" -> <cat="C">  / pues _
"si" -> <cat="C"> / que _
"si" -> <cat="MD"> / _ ?


"la" -> <cat="ART">    /  _ <cat="N" gen="fem">
"la" -> <cat="P">      /  _ <cat="V">
"la" -> <cat="ART">    /  _ <cat="ADJ">
"la" -> <cat="P">      /  yo _
"la" -> <cat="P">      /  tú

"las" -> <cat="ART">  /  _ <cat="N" gen="fem">
"las" -> <cat="P">    /  _ <cat="V">
"las" -> <cat="ART">  /  _ <cat="ADJ">
"las" -> <cat="P">    /  yo _
"las" -> <cat="P">    /  tú _
```

Figure 1. Disambiguation grammar editor

After the automatic tagging, expert annotators revise and correct the tagged corpus. Recently, the originally *Tagged Corpus Editor* has been improved and modified to allow several annotators checking the tags, and to conduct automatically an inter-annotator comparison. In the previous version, each token was followed by its tag in a horizontal fashion. In the new version, the annotated text is segmented by utterances, and every lexical unit is presented in a column layout with all the possible tags for each unit. Every POS tag has a toggle button. The chosen tag by the tagger is marked (Figure 2).

The annotator’s task is to verify one by one the different options and, in case of finding an annotation error, to mark the proper tag. After saving the change, the editor will modify the output. The file can be revised and modified as many times as desired. When the annotator has introduced all the changes, the revised annotation is saved for comparison. Typically two annotators correct each text, and their revised versions are compared automatically producing a version with differences (Figure 3). This time, a third annotator checks the ‘diff’ version and selects one of the alternatives (“dos” as a Q, or “dos” as a N). The output is the final tagged version of the text. In the run, the automatic tagger and three trained annotators have participated.

The new editor has proven very useful. It is a time-saving user-friendly tool, assures coherence in the output, and produces a compared version. In the short term, a counter of inter-annotators agreement will be added to the tool.



En este archivo hay:
69 utterances
1496 palabras
2570 etiquetas
633 palabras ambiguas (**42.3 %**)

cinco minutos de las dos de la tarde de la una en Canarias ya lo deciamos China y Estados_Unidos han llegado finalmente a un acuerdo para la liberación de los veinticuatro tripulantes estadounidenses de un avión_espía retenidos en la Isla de Hainan tras provocar un choque con un caza chino el pasado uno_de_abril

□ **cinco** ● N cinco
○ Q cinco

□ **minutos** N minuto masc plu

□ **de** PREP de

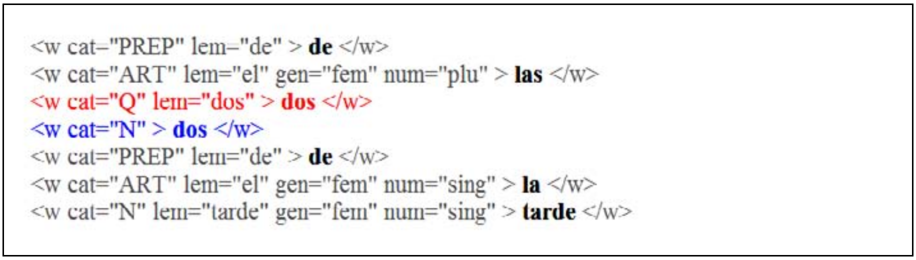
□ **las** ● ART el fem plu
○ P lo fem plu 3
○ N la masc plu

□ **dos** ● Q dos
○ N dos
○ N
○ N do masc plu

□ **de** PREP de

□ **la** ● ART el fem sing
○ P lo fem sing 3
○ N la masc sing

Figure 2. The new Tagged Corpus Editor



```
<w cat="PREP" lem="de" > de </w>
<w cat="ART" lem="el" gen="fem" num="plu" > las </w>
<w cat="Q" lem="dos" > dos </w>
<w cat="N" > dos </w>
<w cat="PREP" lem="de" > de </w>
<w cat="ART" lem="el" gen="fem" num="sing" > la </w>
<w cat="N" lem="tarde" gen="fem" num="sing" > tarde </w>
```

Figure 3. The ‘version with differences’ screen

3. Evaluation

The final goal of our annotation practice is to get the full corpus revised by linguists. For the official release of the C-ORAL-ROM corpora, a fragment of 50000 tokens and 44144 lexical units were verified in March 2004. By the end of 2005, the revised section is almost half the corpus: 145.466 tokens and 138.702 lexical units. With the experience and the new annotation tool, we expect to complete the revision in 2006.

The arguments for having a complete revised annotation are:

1. To produce an accurate list of lemmas and forms.
2. To maximize the quality of search in the corpus, with finer analyses.
3. To deliver a larger ‘gold standard’ for training and evaluation of Spanish taggers.

The last application will be explored in this section. In current resources evaluation praxis, a ‘gold standard’ is the expert-produced version used for comparison against the machine version. Assuming that the gold standard represents, as accurately as practicable, the best results in the morpho-syntactic annotation of a given corpus, the number of matches with the machine version will be the precision rate of the tagger.

According to this methodology, in Table 8 we summarize the precision rates and corpus size for the three evaluations of GRAMPAL conducted until now.

Table 8. Main features of the three evaluations

Evaluators	Date	Precision	Test corpus size in tokens	Evaluation features
Moreno & Guirao	Sept. 2003	98.3%	22747	No multi-word units, no Discourse Markers
C-Oral-Rom	Mar 2004	95.6%	± 50000	Training and Test corpus are the same.
Moreno & Guirao	Dic 2005	95.3%	14321	Separation of the training corpus from the test corpus. Test corpus selected randomly

The precision rate has decreased in every new evaluation, despite the continuous improvement of the lexicons and the disambiguation modules. The reason for this paradoxical fact is that precision decreases when the analysis (and the evaluation itself) gets finer.

In the first evaluation, when we run directly our written-trained tagger, an outstanding precision was obtained (98.3%, in the border line of the best written taggers). But neither multi-word units nor Discourse Markers were considered. As we insist in previous sections, both units are clearly necessary in spoken annotation. The non-inclusion of those units in the annotation produces poorer descriptive quality.

In the second evaluation, both multi-words and DMs were annotated and the precision lowered almost 3 points, and it could be worse if many changes and improvements in the lexicon and grammar have not been added. A similar decrease of precision was reported by the Portuguese team from 91.5 % (when the DMs were not tagged as such) to 88 %:

> Taking into account the errors regarding PoS tag, as already expected (since the tagger did not include this tag) a high percentage of the errors occurred in the annotation of discourse markers (18% of the total errors [including both lemma and tag errors]; 25 % of tag errors; 2% of the corpus). The annotation of locutions[3] also increased the error rate, since they represent 29% of the total errors (39% of tag errors; 4% of the corpus). It is important to underline once again the fact that most of the locutions consist of discursive locutions, which are extremely frequent in oral discourse and particularly difficult to predict and, therefore, to automatically tag. […] if we excluded these errors from the error rate we would have a success rate for the lemmatiser tool of 96.8 % and a success rate for the tagger of 96.7 %" (Bacelar et al. 2005: 186-187).

Between the second (March 2004) and the third evaluations (December 2005), the gold standard augmented its size from 50000 to almost 150000 tokens. During the process, improvements in the lexicon[4] and in the disambiguation rules were incorporated. The precision, however, decreased again, this time only 0.3 points. The reason was due to significant changes in the evaluation procedure:

- The gold standard was splited into a training set (approximately 90%) and a test set (the remaining 10%). In previous evaluation, the whole corpus was used for training and evaluation. Since part of the disambiguation is based on a statistical tagger, a clear bias was introduced in the evaluation.
- The test corpus was selected randomly from the gold standard, choosing a whole utterance instead of isolated tokens. This sampling is roughly 5% of the whole corpus (14321 tokens).

The current approach is more accurate to evaluate the tagger performance, while the approach used by the French and Italian teams is focused on the

[3] 'Locutions' is the term that the Portuguese team uses for multi-words.

[4] Many tagging errors were due to the lack of the proper category tag. For instance, 'rector' as a Noun (*President of a university*) was not in the lexicon, although 'rector' as an ADJ (*governing*) existed. Therefore, any instance of rector was tagged as an ADJ. The addition of new entries has caused more ambiguity.

annotation itself [5]. When the whole corpus will be manually revised (that is, being a gold standard), the precision of the annotation will be virtually 100% and the tagger precision will be tested again. At the present, the rate seems to be stable around 95%.

In the rest of the section, we will show the details of the training and test corpora, and the main ambiguity factors will be analysed. Table 9 shows the figures for the gold standard.

Table 9. Figures of the Gold Standard corpus

	Gold Standard (total)	Training corpus	Test corpus
Tokens	145466	131135	14321
Lexical units	138702	125012	13690

In order to evaluate the difficulty of the annotation task is necessary to estimate the ambiguity rate because of its impact in the precision. The procedure for calculating the error rate (applied on the Test corpus) is:

1. Extract all the lexical units, both single words and multi-words (13690).
2. Look them up in the lexicon and assign each one all the possible tags.
3. Count the number of lexical units with only one tag (i.e. the non-ambiguous) and the number of lexical units with more than one tag (the ambiguous).
4. Calculate the percentage of ambiguous / non-ambiguous with respect to the total.

The results are shown in Table 10. In addition, the ambiguity rate for PoS was calculated. The most (potentially) ambiguous categories in spoken Spanish are verbs, nouns and DMs, in absolute figures. In relative figures, around two thirds of Vs and AUXs are ambiguous, while only a third of Ns are. ADJs and Ps are even less ambiguous (a fifth and a sixth, respectively). 3 out 7 discourse markers are ambiguous and none proper noun (2.6 % of total) has more than one analysis in C-ORAL-ROM[6]. Those statistics will be compared against the error figures obtained from the tagger in order to relate potential ambiguity with actual errors.

[5] Compare the two different strategies in the C-ORAL-ROM evaluation. The French and Italian teams used a random sampling of 1/100 tokens picked out of the whole corpus. Those isolated tokens were manually revised and errors were classified in different types (mainly in tag and in lemma). The Spanish team decided to evaluate the tagger on the basis of the gold standard. To annotate properly, the tagger is to be fed with a whole utterance, in order to have the necessary contextual information for disambiguating.

[6] Because only PRN are transcribed with first letter in uppercase.

Table 10. Ambiguity rate in the test corpus and in the UAM Spanish Treebank

	C-ORAL-ROM	UAM Treebank
Total number of lexical units	13690	22015
Multi-words	561 (4.10 %)	866 (3.93 %)
Non-ambiguous lexical units	7175 (**52.41** %)	14336 (**65.12** %)
Ambiguous lexical units	6515 (**47.59** %)	7679 (**34.88** %)
Lexical units with 2 tags	4886 (35.69 %)	5501 (24.99 %)
Lexical units with 3 tags	1346 (9.83 %)	2123 (9.64 %)
Lexical units with 4 tags	283 (2.07 %)	55 (0.25 %)
Ambiguity per POS		
NPRs	294 (2.15 %)	1708 (7.76 %)
MDs	741 (5.41 %)	0 (0.00 %)
Ambiguous MDs	469 (3.43 %)	0 (0.00 %)
Ns	1772 (12.94 %)	5035 (22.87 %)
Ambiguous Ns	658 (4.81 %)	1519 (6.90 %)
ADJs	468 (3.42 %)	1485 (6.75 %)
Ambiguous ADJs	168 (1.23 %)	496 (2.25 %)
Vs	2212 (16.16 %)	2696 (12.25 %)
Ambiguous Vs	1393 (10.18 %)	853 (3.87 %)
AUXs	495 (3.62 %)	342 (1.55 %)
Ambiguous AUXs	331 (2.42 %)	189 (0.86 %)
Ps	1662 (12.14 %)	818 (3.72 %)
Ambiguous Ps	409 (2.99 %)	293 (1.33 %)

Another interesting comparison is the proportion of ambiguous / non-ambiguous lexical units in spoken and written texts. Using the same procedure, the potential ambiguity rate of the UAM Spanish Treebank (Moreno et al. 2003) was calculated. This corpus can be considered as a gold standard of written texts. It is similar to the spoken test corpus in size (20000 lexical units) and, more important, in tagset: the only difference is the non-inclusion of Discourse Mark as a tag in the treebank.

The distribution of POS in written (journalistic) texts is remarkably different. Nouns and adjectives appear much more frequently than in spoken texts. With respect to the ambiguity rate, the low ambiguity in verbs (one out four) is especially significant.

After setting that, at least for Spanish, spoken texts are more ambiguous than written corpora, four experiments were run in order to evaluate the difficulty of reaching precision rates similar to those reported for written corpora. The experiment design was divided in two parts: firstly, we wanted to know the baseline (i.e. the precision rate obtained randomly[7]) and the success rate when the most frequent tag is chosen, regardless the context. Secondly, we wanted to test whether a combined employ of disambiguation rules and statistics performs better than only statistical disambiguation.

The results of the four experiments are given in the following tables and in Figure 4:

1. Random disambiguation: 74.4%
2. Disambiguation with the most frequent tag: 86.2 %
3. Disambiguation with the statistical model (TnT): 94.9%
4. Disambiguation with Rules + TnT: 95.3%

The performance is similar to the other taggers: the best precision is obtained with a combination of statistics and detailed linguist rules, as the French tagger (Cordial, 98.75%). Statistical taggers alone as the PiTagger, the Italian tagger, get worse results (90.36%). Rule-based alone taggers, as the Brill tagger (the Portuguese) also get around 91%.

Table 11. Four tagging strategies

	Random	Most frequent tag	TnT	Rules + TnT
Successful tagging	10686	12378	13631	13675
Fail in one-token words	3628	1274	409	365
Fail in multiwords	42	141	316	316
Unknown tokens	0	563	0	0

In the case of the Spanish tagger, the difference between the TnT and the combined rules and TnT is only 0.4%. The explanation could be in the training set (131135 words). Statistical taggers need as many data as possible, but also the maximum accuracy[8] and adaptation to the register[9]. In future experiments, with more training data, we will check whether the difference is stable or it disappears.

[7] We must bear in mind that, from the ambiguity estimation, to get a precision of 75% is direct: 52.4 % of the test corpus only has one tag (no error possible) and almost half of the remaining (48%) only has two tags. The probability of choosing a correct tag is virtually 50%. Therefore, 52.4 + 23 = 75.4%

[8] Many training sets are not carefully hand-annotated or revised. Our annotation procedure warranties at least three different expert annotators.

[9] Most taggers are trained only with written texts, since there are not any tagged spoken corpora available for training.

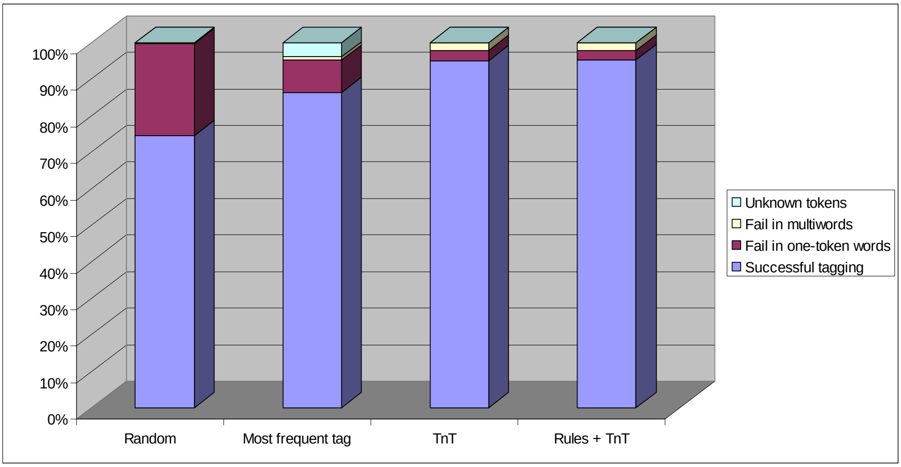


Figure 4. Graphical representation of the four experiments

Using the annotation produced by the last experiment, we focused in the main errors (Table 12). The percentage is calculated with respect to the total number of lexical units in the test corpus. Here again we can observe the influence of the multi-words and discourse particles. If we remove the DMs from the tagset, the precision would raise at least 2%, with a rate comparable to the French tagger. As we seen in the ambiguity rates (Table 10), multi-words represent the 4.10% of the corpus. Discourse Markers represent 5.41% of the total, and 3.63% of the ambiguity. This is the most problematic remaining error: the tagger resolves only 1.6%, less than the half of cases.

Table 12. Main tagging errors

Total errors	681	4.7%
Multi-words errors	316	2.2%
Discourse Marker errors	286	2.0%
Confusion ‘que’ P / C	78	0.5%
Confusion V / AUX[10]	76	0.5%

[10] Analyzing the errors in the third and the fourth experiments, the only difference is in the confusion V / AUX. With the TnT tagger alone, 117 errors were produced, while with the combination tagger, only 76 errors were detected. The specific contextual rules for disambiguating V / AUX are responsible.

4. Conclusions

In this article, we have addressed two related topics:

1. Morpho-syntactic tagging of spontaneous speech is different from the same task in written corpora:
   - Different tokenization.
   - New and different lexical units.
   - A tag for Discourse Markers is needed.
   - Over all, much more ambiguity: the ratio is 52/48 (in spoken) against 65/35 (in written texts).

2. Morpho-syntactic tagging evaluation and comparison is highly dependant of:
   - The number of tags in the tagset.
   - The annotation of multiwords as lexical units.
   - The inclusion of a tag for Discourse Markers.
   - Separating the training set from the test set.

As a general remark, precision scores can vary between 8-10% if some of these factors are missing. Due to a higher ambiguity rate, taggers perform worse than in written corpora. The best scores are obtained by taggers with a combination of statistics and rules (this result is similar to the written texts).

We want to stress that this is not a final evaluation report on the tagging and the tagger: when the whole corpus will be hand-revised, a new evaluation will be conducted, as well as an inter-annotators agreement one.

References

Bacelar, M.F., Bettencourt, J., Veloso, R., Antunes, S., Barreto, F. & Amaro, R. 2005. The Portuguese corpus. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. E. Cresti & M. Moneglia (eds.). 163-208.

Brants, T. 2000. "TNT – a statistical part-of-speech tagger." *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP 2000)*, Seattle, USA. 224-231.

Campione, E., Véronis, J., & Deulofeu, J. 2005. The French corpus. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. E. Cresti & M. Moneglia (eds.). 111-133.

Cresti, E. & Moneglia, M. 2005. C-ORAL-ROM: *Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam: John Benjamins.

EAGLES, 1996. *Recommendations for the Morphosyntactic Annotation of Corpora*. http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html.

Gazdar, G. 1996. Paradigm merger in natural language processing. In Robin Milner & Ian Wand, eds., *Computing Tomorrow: Future Research*

Directions in Computer Science, Cambridge: Cambridge University Press, 88-109.

Grishman, R. 1996. MUC-6. [http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html].

Guirao, J.M., & Moreno, A. 2004. “A “toolbox” for tagging the Spanish C-ORAL-ROM corpus”. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004) – Workshop “Compiling and Processing Spoken Language Corpora”. Paris: ELRA. 28-32.

Jurafsky, D., and Martin, J. 2000. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ, Prentice Hall.

Kovacci, O. 1999. “Adverbios”. *Gramática descriptiva de la lengua española*. I. Bosque & V. Demonte (eds.). Madrid: Espasa.

Leech, G., & Smith, N. 2000. *Manual to accompany the British National Corpus (Version 2) with Improved Word-class Tagging*. [available at http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2postag_manual.htm]

Moneglia, M. 2005. The C-ORAL-ROM resource. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. E. Cresti & M. Moneglia (eds.) 1-70.

Moreno, A. 1991. *Un modelo computacional basado en la unificación para análisis y generación de la morfología del español.* Ph.D. dissertation. Univ. Autónoma de Madrid.

Moreno, A., & Goñi, J.M. 1995. “GRAMPAL: a morphological model and processor for Spanish implemented in Prolog .” *Proceedings of Joint Conference on Declarative Programming (GULP-PRODE 95*). Salerno: Palladio. 321-331.

Moreno, A., & Guirao, J.M. 2003. “Tagging a spontaneous speech corpus of Spanish”. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2003).* Borovets, Bulgaria. 292-296.

Moreno, A., de la Madrid, G., Alcántara, M., González, A., Guirao, J.M. & De la Torre, R. 2005. The Spanish corpus. *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. E. Cresti & M. Moneglia (eds.) 135-162.

Portolés, J. & Martín Zorraquino, M.A. 1999. “Los marcadores del discurso”. *Gramática descriptiva de la lengua española*. I. Bosque & V. Demonte (eds.). Madrid: Espasa.

XCES: Corpus Encoding Standard for XML. Vassar College. [http://www.cs.vassar.edu/XCES/]

The Role of Spoken Corpora in Teaching/Learning Portuguese as a Foreign Language — The Case of Adjectives Intensification —

Maria Fernanda Bacelar do NASCIMENTO and José Bettencourt GONÇALVES

1. Introduction

The choice of this subject for this paper is inspired by the low degree of exploitation of Portuguese corpora, both in the preparation of didactic materials and directly in the classroom, for teaching and learning Portuguese as a foreign language.

In the last twenty years, the development of linguistic corpora, of computer science and of tools for Linguistic Resources exploitation have increased hugely and now-a-days, any language teacher or student can access large amounts of data available on-line, as well as tools for its exploitation. The access to spoken corpora, is nevertheless more difficult but crucial for learning and teaching a foreign language since it provides data on spontaneous speech language, produced by native speakers, talking face-to-face, in a particular context, in formal or everyday language. Those materials reflect authentic usage and they can dispel myths and distortions perpetuated in grammars, dictionaries and course books (Cf. Johns, T. 1991:30). The direct access to authentic data actually offers a variety of stimulating inductive strategies to teach and learn a foreign language.

Spoken corpora can be used for teaching purposes in many different ways: we will speak first of the importance of listening to authentic spoken discourse and then of the extraction of information from corpora: concordances, frequencies and statistical data

The oral data we will use in this presentation is extracted from the spoken subcorpus of CRPC, and its extension is of about 1.300.000 tokens, raised from the 70s until now, comprising the C-ORAL-ROM Portuguese corpus. The Center of Linguistic of Lisbon University (www.clul.ul.pt) has available a corpus, Corpus de Referência do Português Contemporâneo (CRPC), of 334.711.788 words, including geographical varieties of Portuguese: Portugal, Brazil, Angola, Cape Verde, Mozambique, Guinea-Bissau, Sao Tome and Principe, Macao and Goa (represented in different domensions). The dimension of the written subcorpus is about 332 million words from Books, Newspapers, Magazines, Parliament Sessions, Supreme Court Verdicts, Pamphlets, Correspondence and Miscellaneous including fiction, techno-scientific, didact and general discourse. The dimension of the spoken subcorpus is about 2.5 million words, and includes informal and formal discourse.

2. Spoken corpora

The spoken corpus – particularly collections like the C-ORAL-ROM corpus, which provides the acoustic data and the orthographic transcription with text to speech syncronisation based on the alignment of each transcribed utterance are a valuable resource for the indispensable exposure of learners to spontaneous dialogues and conversation in a wide range of real life situations. In fact, in what concerns the Portuguese didactic materials and teachers' behaviour they use audio materials fabricated artificially for the classroom, sometimes even read by actors or professional speakers. Thus, the learners are not exposed to spontaneous speech which have typical phonetics, morpho-phonology and prosodic properties, and also typical syntactic and discourse organization.

The published DVDs *C-ORAL-ROM*, (Cresti E. & Moneglia, M. (eds) 2005) show the evidence and the excellence of this kind of materials.

3. Information extraction: the association tendency (the case of intensifier adverbs: concordances and statistical data

In order to display the importance of information extraction (frequency and statistic data and concordances) from corpora, for Portuguese language teaching, we will refer to the case of some adverbs ending in *–mente*, that function as adjectives intensifiers.

The study follows an empirical and descriptive approach and all examples are taken from European Portuguese corpora.

We intend to show that intensification is mainly employed to achieve expressivity and that it is very closely linked to semantic change, once the intensification adverbs are progressively delexicalised, in some cases to the extent of complete grammaticalization. In this order, we will follow the proposals of Lorenz, G. (2002).

3.1. Concordances

All data are taken from the Reference Corpus of Contemporary Portuguese and the Portuguese C-ORAL-ROM corpus. We mainly used the spoken subcorpus, but we used a larger subcorpus of written language of about 50 million words for confirmation of statistic and quantitative data.

We will refer, in first place, the research based on concordances of sequences of adverb with adjectives that can be seen as a data-driven learning material

The observation and analysis of concordances of a given adverb, i.e. the context in which the adverb co-occurs with an adjective, allows both students and teachers to cooperate as participants in an inductive research in which the linguistic phenomena are easily observable. The concordances and the quantitative data are the ideal context where one can find very relevant contributions for attaining a good domain of a language, understood as the use of a great deal of "prefabricated" language in appropriated situation (Cf. Nattinger, J. R. & DeCarrico, J. S. (1992:XV).

We can observe that intensifiers, taken as "a lexical category, or a member of this category, whose members typically function as modifiers of an adjective or adverb and express the degree to which the quality expressed by that item is present" (Trask, 1993:75), are an heterogeneous set, comprising a closed-class such as *muito* 'very', *bastante* 'quite', *bem* 'well' and an open class formed by adverbs ending in *–mente*, such as *absolutamente* 'absolutely', *completamente* 'completely', *realmente* 'really' etc.

The closed-class of intensifier adverbs is traditionally described by grammars and dictionaries, contrary to what happens with the open-class of adverbs in *-mente*. Thus, we have fixed our attention in this open-class of adverbs in *–mente*, particularly those which, due to their high frequency of use progressively loose their denotation meaning, developing grammatical functions. At the same time it is possible to observe a tendency for the occurrence of strong associations formed by some of these adverbs with certain adjectives, in different thematic domains:

The association tendency of the adverbs in *–mente* that we have observed are of different nature:

3.1.1. Morphosyntactic combinations

Some adverbs intensifiers, for instance, show a preference for co-occuring, mainly in passive constructions, with adjectives of verbal basis (participles). Such is the case of DEVIDAMENTE 'duely', which occurred exclusively with participles.

Examples

632-07-R04-003- 3 E... NÃO... E OS CLUBES NÃO SEREM **devidamente apoiados** VISTO QUE

767-01-N11-002- IROS NA SECÇÃO ONDE ESTOU; DEPOIS DE **devidamente apreciados** PELOS NOSSOS TÉCNICOS E PELOS SER

767-01-N11-002- LICENÇA POIS, DEPOIS DOS PROJECTOS **DEVIDAMENTE APRECIADOS**, NÃO É,

OP27P4C0091XPP2	fazer, ha mas que em portugal não é	**devidamente**	**compensado** e recon
O-0037-T-E-P-Li	a altamente modernizada e que... e	**devidamente**	**equipada**, e portan
O-0031-T-A-P-Li	baixar os braços até isto estar...	**devidamente**	**esclarecido**. L25: a
OPXXX0008L	simplesmente não estão ainda a ser	**devidamente**	**explorados** por ins
O-0028-R-O-P-Li	eito por escrito. o doente deve ser	**devidamente**	**informado** para pod
284-07-A00-015-	E ALGUNS PROFESSORES QUE NÃO ESTÃO	**DEVIDAMENTE**	**INFORMADOS** SOBRE ES
1046-18-A00-103	DADE, DE, DAS PESSOAS QUE NÃO ESTÃO	**DEVIDAMENTE**	**INSTRUÍDAS** SOBRE, A
798-03-B00-003-	DEVIA DE HAVER REALMENTE INDIVÍDUOS	**DEVIDAMENTE**	**PREPARADOS**. AQUI IS
108P098	X: PORQUE, EM REGRA, AS PARTES ESTÃO	**DEVIDAMENTE**	**PROPORCIONADAS**, E Q
OP13T3S0041XPP2	lutas dos professores e que não são	**devidamente**	**remunerados** x: sim,
128-04-J00-001	NTRO DUMA CAIXA, A CAIXA É	**DEVIDAMENTE**	**SELADA**, ABREM UM B

3.1.2. Semantic combinations with intensifying effect

In order to describe the intensifier effect created by the combination of adverbs in *–mente* plus adjectives, we followed the proposal of classificassion of G. Lorenz (2002:147fol).

Semantic roles of adverbs in *–mente*.

- Scalar : Adverbs that express the notion of degree, scaling an adjectival quality

Examples

1094-04-TA0-005	PORQUÊ... ISTO É	**ABSOLUTAMENTE SEGURO**, QUER 'absolutely safe'
O-0037-T-E-P-Li	genetista, porque seria	**altamente custoso**. então c 'highly expensive'
-T-Ci-P-L	dapretas, eh, e... tem que ser	**perfeitamente autónomo**. 'perfectly autonomous'
943-08-B00-008-	TO. E ENTÃO SENTIU-SE LÁ	**PLENAMENTE FELIZ**. VEJA LÁ 'completely happy'
O511	oriente antigo fosse uma via / outra	**extremissimamente sugestiva** 'very extremely suggestive'

- Semantic feature copying : Adverbs that achieve their intensifying effect by copying a substantial part of the adjectives denotation, that is to say, in the adjective exists already, in part, the notion that the adverb intensifies

Examples

0008-P	contece aqui que estou muito	**intimamente ligado** 'innerly attached'
460-14-A00-011-	TRABALHADORES, E A LISTA ESTÁ	**AMPLAMENTE ALARGADA** P 'widely wided'= 'widely enlarged'
O-0029-R-O-P-Li	a antram, que, eh, está	**directamente relacionada** c 'directly related'

• Evaluative : Adverbs that besides scaling their focus, express a judgemental notion on the part of the speaker

Examples

O-0034-T-Cu-P-Lda biografia do pintor **precocemente desaparecido**.
‘precociously disappeared’
= too early dead’

O-0034-T-Cu-P-L e desnecessárias, **francamente desnecessárias**. L
‘honestly unnecessary’

726-25-P00-00 SO UM BOCADO, SOU **EXAGERADAMENTE SENSÍVEL**
‘too sensitive’

O-0035-T-A-P-L ora isto é que eu acho **excessivamente grave**
‘too serious’

• Comparative : Adverbs that achieve intensification by comparing the referent with its rivals or equals

Examples

OP23M6G0048XPP1 poderíamos dizer num outro **igualmente bonito** li
‘equally beautiful’

O-0040-R-E-P-Li bilizador. mas são também i**gualmente necessárias**
‘equally necessary’

129P10493MONDEGO QUE TRAZIAM AQUI HORTALIÇA **RELATIVAMENTE BARATA** E
‘relatively unexpensive’

1236-24-A00-006 POSTA LÁ NO FUNDO FOR **SUFICIENTEMENTE RESISTENTE** À INFILTRAÇÃO, OS BURACOS NÃO
‘suficiently resistent’

• Modals : Adverbs that express the extent to which a speaker is willing to attest the truth of a proposition

Examples

O-0029-R-O-P ancia zero fico **obviamente muito feliz** porque
‘obvioulу very happy’

O-0032-T-Ci-P d uma actividade **aparentemente solitária**, está
‘apparently’ solitaire’

O-0007-P-LIS uma inadaptação, **possivelmente momentânea**, procurámos
‘possibly momentaneous’

O-0038-T lmento, eh,muito **provavelmente acusado** de homicídio
‘probably charged of homicide’

3.1.3. Specific domain combinations

It can be observed that strong associations in a thematic domain or in a specific situation can be week associations in other domains or situations. In order to exemplify these cases, we will present next some types of situations in which the combinations occoured .

Combinations observed in administrative, political or economic domains:

Examples

`O-0011-P-LIS natureza, porque são coisas` **`altamente confidenciais`** `para as empresas LM - pois.`

`OP44B9B0005XPP2 uma zona onde, a pecuária é` **`altamente lucrativa`** `- dissemos que quanto à estrutura fu`

Combinations observed in the clothing domain:

Examples

`O-0004-R-A-P-Li alguns desses homens também` **`impecavelmente vestidos`** `e de smoking, trazem por cima um`

`O-0004-R-A-P-Li homens estão de facto` **`impecavelmente vestidos`**`, eh, calças pretas, camisa branca`

3.2. Frequency and statistical data

“The more grammaticalised an intensifier, the more it will loose its lexical restrictionss and increase in frequency. At the same time, its collocates and contexts of occurrence will change in relation to its own semantic change” (Lorenz, G. 2002:144)

Frequency and statistical data become significant only in large corpora. Thus, in order to confirm “the associative strengh of some adverbs in *–mente* with adjectives, frequent in the spoken corpus, by means of statistical information, we decided to run CLUL’s program CONCOR-CB for multiword units extraction on a written corpus of 50 million words. The fact that this written corpus is very diversified (literary, journalistic, scientific, technical texts, etc) allows to confirm that one is not dealing with occasional co-occurences, but with strong associations in language, which form ‘prefabricated’ groups developed in the real use of the language. They are at the heart of language acquisition.

In order to confirm this associative tendency observed in the spoken corpus, we chose the groups displayed in the following table:

Oral *corpora*		
GROUP	CRPC *subcorpus* 1.300.000 words	C-ORAL-ROM 300.000 words
completamente diferente	69	17
totalmente diferente	31	7
extremamente importante	37	8
propriamente dito	24	6

We will show now the results of the same groups in COMBINA-PT corpus (50.000.000 words) and their Combinatory Index (CI), i. e. the CI between the word X and the co-occurent word Y is the calculated index between the co-occurrence probabilities composed by the pairs of words X and Y and their independent occurrence (Cf. Pereira, L. A. S. (1994:149)).

As studies carried out for Portuguese language have shown, when the IC is equal or superior to 7 it becomes relevant. (Cf. Bacelar do Nascimento (2002:38-51) and Pereira, L.A. S. & Mendes, A. (2002).

The data in the table below are selected from the data authomatically extracted from the COMBINA-PT *corpus*:

Group	eg	og	Ic	fg	fe	N
completamente diferente	(2)	(227)	(9,200001)	(227)	(3.521;5.515)	(50.310.890)
completamente diferentes	(2)	(89)	(7,121774)	(89)	(3.521;9.131)	(50.310.890)
extremamente importante	(2)	(106)	(7,928022)	(106)	(1.890;11.586)	(50.310.890)
propriamente dito	(2)	(51)	(9,222288)	(51)	(1.205;3.565)	(50.310.890)
totalmente diferente	(2)	(68)	(7,815896)	(68)	(2.753;5.515)	(50.310.890)

Legend

eg = number of the combinatory elements

og = number of the combinatory occurrences

ic = combinatory index

fg = group frequency

fe = frequency of each combinatory element in the *corpus*

N = number of words of the *corpus*

4. Conclusions

In the data extracted, many interesting sequences of adverb with adjective were found; as it is the case of “view point adverbs”, “manner adverbs”, “time adverbs” and so on, which will be object of posterior studies. The objective of these studies is to analyse different types of word sequences, from the point of view of the degree of collocational relationships that are established by use, leading to the formal fixedness of the sequence, together with a semantic fixedeness; when this process achieves maximum fixedness, the result is a pluriverbal unit, totally lexicalized, i. e., with strong morphosyntactic and syntactic fixedness (sometimes also phonological) of its elements and with a unitary meaning, memorized as an individual unit. We intend to provide resources of fundamental importance for the development of description and teaching materials for researchers, teachers and students.

References

BACELAR do NASCIMENTO, M. F. & MOTA, M. A. (2002). “Associations lexicales: du corpus aux dictionnaires” In MELKA, F. and AUGUSTO, M. C. (eds.) *De la Lexicologie à la Lexicographie /From Lexicology to Lexicography,* Utrecht, Utrecht Institut of Linguistics (OTS), pp.38-51.

CRESTI, E. & MONEGLIA, M. (eds.) (1985), *C-ORAL-ROM, Integrated Reference Corpora for Spoken Languages*; Amsterdam / Philadelphia, John Benjamins Publishing Company.

JOHNS, Tim (1991) “From printout to handout: grammar and vocabulary teaching in the context of data-driven learning” in JOHNS, Tim & Philip King, Classroom Concordancing, ELR Journal, vol.4. Birmingham: Centre for English Language Studies – University of Birmingham.

LORENZ, Gunter (2002) “Really worthwhile or not really significant? – A corpus-based approach to the delexicalization and grammaticalization of intensifiers in Modern English” in WISCHER, Ilse and Gabrielle Diewald (eds), Amsterdam-Philadelphia: John Benjamins Publishing Company.

NATTINGER, James R. & Jeanette S. DeCarrico (1992) Lexical Phrases and Language Teaching, Oxford: Oxford University Press.

PEREIRA, L. A. S. (1994) *Como se combinam as Palavras? Contributo para um Dicionário de Combinatórias do Português*, MA. Dissertation. Lisboa: Faculdade de Letras de Lisboa.

PEREIRA, L. A. S. & A. MENDES (2002), “An Electronic Dictionary of Collocations for European Portuguese: Methodology, results and applications” *in proceedings of the 10th EURALEX International Congress*, Copenhagen, Denmark.

TRASK, R. L. (1996) *A Dictionary of Gramamatical Terms in Linguistics*; London, Routledge.

Typologies of MultiWord Expressions Revisited — A Corpus-driven Approach[1] —

Maria Fernanda Bacelar do NASCIMENTO, Amália MENDES and Sandra ANTUNES

1. Introduction

In the 50's, Firth (1955) firstly introduced the concept of collocation, defining it as the characterization of a word according to the words that typically co-occur with it. The increasing interest in the study of the lexicon (particularly the description and classification of lexical categories according to their different and possible meanings) allowed the development of several studies that showed that the lexicon does not consist mainly of simple lexical items but appears to be populated with numerous chunks, more or less predictable, though not fixed.

> "On the one hand, *bank* co-occurs with words and expressions such as *money*, *notes*, *loan*, *account*, *investment*, *clerk*, *official*, *manager*, *robbery*, *vaults*, (...). On the other hand, we find *bank* co-occurring with *river*, *swim*, *boat*, *east* (...)" (Hanks, 1987: 127, *apud* Church & Hanks, 1989: 76).

It became notorious that natural languages follow complex regular associative patterns and that the identification of such patterns would give important information on the meanings of the word and its actual uses (Sinclair, 1991). Once they start to be frequently repeated, these word associations tend to correspond to a conventional way of saying things, turning out to be an important aspect in the lexical structure of the language.

> "Several nouns are frequently qualified by the adjective *hard*. We talk of *hard luck*, *hard facts* and *hard evidence*. We can also talk about *strong evidence* but are unlikely to use *strong facts* or *strong luck*; *tough luck* but not *tough facts* or *tough evidence*; *sad facts* but not *sad luck* or *sad evidence*. Of course, it is always possible to depart from the normal patterns of English, so it is not claimed that *sad evidence* can not occur – just that it's not worth following as a pattern.
>
> Note that in the above examples of *hard*, there are two rather different meanings. In *hard luck*, *hard* means *unfortunate*, but in *hard facts* and *hard evidence* it means

[1] This study was undertaken under the Project Word Combinations in Portuguese Language (COMBINA-PT), developed at the Centre of Linguistics of the University of Lisbon under a research grant of the Portuguese Ministry of Science (POCTI/LIN/48465/2002).

> *unlikely to be proved wrong*. Despite this, the patterns of collocation show that the near-synonym *strong* goes only with *evidence*. So, the patterns of collocation are not governed by meaning." (Sinclair (1987), Introduction to the Cobuild Dictionary, apud Krishnamurthy, 1997: 44-45)

These lexical associations may present different degrees of cohesion, ranging from totally frozen groups, semi-frozen groups or just sets of favoured co-occurring forms. We will use the term Multiword Expressions (MWEs) to refer to this range of different word associations. A number of typologies of these MWEs have been proposed taking into account several parameters, like, for example, their degree of cohesion, internal variation or compositional meaning. However, the exact definition of a collocation or of a MWE is still controversial. While some authors clearly distinguish the phenomenon of collocations from other types of word associations and syntagmatic relations (Hausmann (1979) and Mel'cuk (1984)), others have a broader perspective (Sinclair, 1991). In section 2 we will present these different definitions of MWEs; in section 3 some typologies based on discrete categorization will be reviewed; section 4 will address the corpus-driven methodology; section 5 will discuss the corpus data and how it follows or challenges MWEs typologies.

2. Reviewing some definitions of MWEs

One of the criteria used by some authors to define a MWE relies on its meaning. In this way Hausmann (1979) and Mel'cuk (1984) define collocations as a conventional combination of words, whose meaning can not be predicted by the meaning of the words that compose it. In fact, for Hausmann (1979), a collocation is constituted by a base (*Basis*), that is semantically autonomous, and by a collocator (*Kollocator*) that needs the base in order to get its full meaning. For the author, collocations consist of affine combinations of striking habitualness and have limited combinatorial capacity. The author distinguishes 8 types of collocations according to the word class of its elements: (1) N + Adj (*célibataire endurci* 'confirmed bachelor'); (2) $N_{(subject)}$ + V (*la colère s'apaise* 'the anger wears off'); (3) V + $N_{(object)}$ (*tenir un journal* 'to keep a diary'); (4) V + Adv (*exiger énergiquement* 'to insist firmly'); (5) Adv + Adj (*gravement malade* 'critically ill'); (6) N + (prep) + N (*marché du travail* 'labour market'); (7) V + prep + N (*rougir de honte* 'to blush'); (8) Adj + N (*(dans un) proche avenir* 'in the near future').

Mel'cuk (1984) introduces the *Lexical Functions* (LFs) that describe the combinatory properties of lexical units (LUs) in a systematic way. In the process of text production, the speaker has to select lexical units to build his sentences. In this perspective, two types of LUs have to be distinguished: (i) LUs that are selected according to their meaning (*semantically-driven lexical choices*); (ii) LUs that are selected contingent on other LUs (*lexically-driven lexical choices*). This second type of choice is carried out along with two major linguistic relations: a *paradigmatic relation*, that subsume all substitution relations that may hold between lexical units in specific contexts (like the lexemes *young* and *tall*, that are paradigmatically related in the pairs of phrases *young student* and *tall student*), and a *syntagmatic relation*, that holds between lexical units that can co-occur in the same phrase or clause (like *boy* and *ran*, that are syntagmatically related in the phrase *the boy ran*).

"Lexical Functions (LFs) are a set of formal tools designed to describe, in a fully systematic and compact way, all types of genuine lexical relations that obtain between LUs of any language" (Mel'cuk, 1996: 38). Formally, LFs correspond to mathematical functions: f(x) = y (where x is the argument/keyword; y is the value).

Examples of LFs:

1. Adjectival LFs: **f** is intense/very; intensification → **Magn**
   a. **Magn**(*malade* 'ill') = *très* 'very', *gravement* 'critically'
   b. **Magn**(*dormer* 'to sleep') = *profondément* 'deeply', *comme une souche* 'like a log'
2. Verbal LFs
   a. **Oper$_1$**(*remarque* 'remark') = *faire* 'to make' [ART–]
   The keyword of **Oper$_1$** is its direct object (*faire un remarque* 'to make a remark')
   b. **Func$_1$**(*aider* 'help') = *vient* 'comes' [*de* 'from' N]
   The keyword of **Func$_1$** is its grammatical subject (*l'aide vient de qn* 'aide comes from someone')
   c. **Labor$_{12}$**(*note* 'note') = *prendre* 'to take' [N *en* 'in' –]
   The keyword of **Labor$_{12}$** is its indirect object (*prendre qc en note* 'to take note of something')

Also based in the meaning criterion, Cruse (1986) defines a collocation from a different point of view. For the author, "the term collocation will be used to refer to sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent" (Cruse, 1986: 37). The author exemplifies collocations with expressions such as *fine weather*, *torrential rain*, *light drizzle* and *high winds*.

However, the author also points out that collocations also have a semantic cohesion that "is the more marked if the meaning carried by one (or more) of its constituent elements is highly restricted contextually, and different from its meaning in more neutral contexts" (op. cit.: 37). That is the case of *heavy* in expressions like *heavy drinker/smoker/drug-user*. This sense of *heavy* requires narrowly defined contextual conditions and for this sense to be selected, the notion of 'consumption' seems to be a prerequisite. The author claims that we are still in the realms of transparent sequences, because each constituent produces a recurrent semantic contrast:

1. heavy/light (He's a ___ smoker) = heavy/light (They were ___ drinkers)
2. drinker/smoker (He's a heavy ___) = drinker/smoker (They were light ___s)

Another criterion used to define a collocation relies on its fixedness. That is the criterion used by Benson et alii (1986) that claim that there are many fixed, identifiable, non-idiomatic phrases and constructions that may be called recurrent combinations, fixed combinations or collocations. For the authors, collocations fall into two major groups: *grammatical collocations* and *lexical collocations*. A grammatical collocation is a phrase consisting of a lexical word (noun, adjective or verb) and a grammatical word (preposition, article or conjunction), like the expressions *account for*, *adapt to*, *agonize over*, *aim at*. The authors distinguish this type of collocations from what they call *free combinations*, that "consist of elements that are joined in accordance with the general rules of English syntax and freely allow substitution" (Benson et alii, 1986: ix), such as *after lunch*, *at three o'clock*, *in the library*, *on the boat*, that may have a limitless number of possible combinations. Lexical collocations, in contrast to grammatical collocations, are exclusively composed by lexical words, such as *warmest regards* (Adj + N) or *commit murder* (V + N). These are expressions with a high degree of cohesion, since, in the first case, we can not have **hot regards* or **hearty regards*, and, in the second case, the verb *commit* is limited in use to a small number of nouns meaning 'crime' or 'wrongdoing'. The authors also distinguish this type of collocations from *free lexical combinations*, in which the elements are not bound specifically to each other and may occur with other lexical items freely (the expression *condemn murder* is considered a free combination since the verb *condemn* may occur with an unlimited number of nouns, such as *abortion*, *abduction*, *abuse of power*, etc.).

Finally, Sinclair (1991) considers that in order to explain the way in which meaning arises from language text we have two principles of interpretation: *the open-choice principle* (where the speaker has a very large number of complex choices and the only restraint is grammaticalness) and *the idiom principle* (where the speaker has available a large number of semi-preconstructed phrases that constitute single choices, reflecting a natural tendency to economy of effort). In fact, it has been observed that the speaker actually uses his memory and routine, and that his discourse corresponds to single choices presented in the idiomatic principle. For the author, collocations illustrate the idiom principle. Words appear to be chosen in pairs or groups and these may not be necessarily adjacent. According to the author, a collocation is "the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening. Collocations can be dramatic and interesting because unexpected, or they can be important in the lexical structure because of being frequently repeated" (Sinclair, 1991: 170).

3. Reviewing some typologies of MWEs

As can be seen above, the different definitions of a collocation presented by different authors show that this is not a consensual topic and this controversy is also reflected in the different proposals of typologies. Hausmann (1989) proposes the typology presented in figure 1.

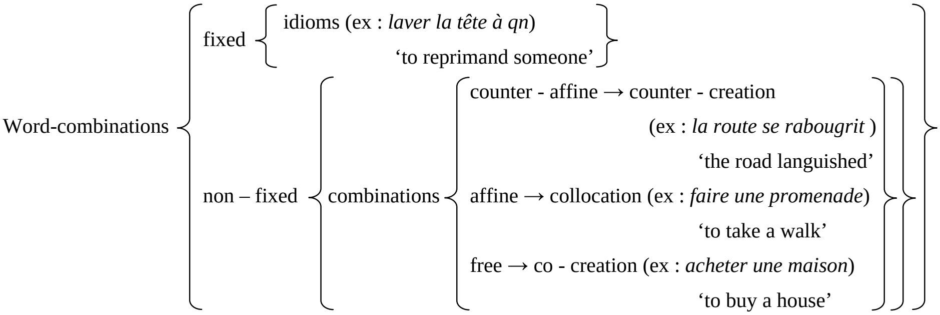

Word-combinations
fixed
idioms (ex : *laver la tête à qn*)
'to reprimand someone'
non – fixed
combinations
counter - affine → counter - creation
(ex : *la route se rabougrit*)
'the road languished'
affine → collocation (ex : *faire une promenade*)
'to take a walk'
free → co - creation (ex : *acheter une maison*)
'to buy a house'

Figure 1. Hausmann's classification of word-combinations

This typology relies essentially in the distinction between fixed and non-fixed expressions. Whether the first one only comprises the idioms, the second registers three types of non-fixed expressions ranging from counter-creations (or "poetic metaphors" (Lakoff, 1993)), collocations (cf. Hausmann's definition in section 2.) and co-creations (semantically motivated combinations).

A different approach is introduced by Mel'cuk (1996) who distinguishes between free combinations (relations that hold between lexemes in a phrase with a purely compositional semantics) and non-free combinations (relations that hold between lexemes in a phrase whose semantics has to be partially or entirely derived from the phrase as a whole). In what non-free combinations are concerned, the author also distinguishes those which definitely do not have a compositional meaning from what he calls 'pragmatemes', i.e., pragmatically constrained combinations where the phrases in question are semantically freely composable but unexchangeable in specific contexts by any other synonymous expression (ex: *best before*).

Returning now to non-free combinations with non-compositional meaning, these are called ‘semantic phrasemes’ and are subclassified by the author into ‘full phrasemes’, or idioms (whose semantics is completely opaque and its meaning can not be obtained from the meaning of the constituent lexemes (ex: [*to*] *cool one's head*; [*to*] *speal the beans*)), ‘quasi-phrasemes’ (whose semantics is partially obtainable from the meanings of its constituent lexemes, but contains, however, an additional meaning that can not be derived from those meanings (ex: *start a family*)) and ‘semi-phrasemes’, or collocations (cf. Mel’cuk’s definition in section 2.).

A general overview of Mel’cuk’s typology is presented in figure 2.

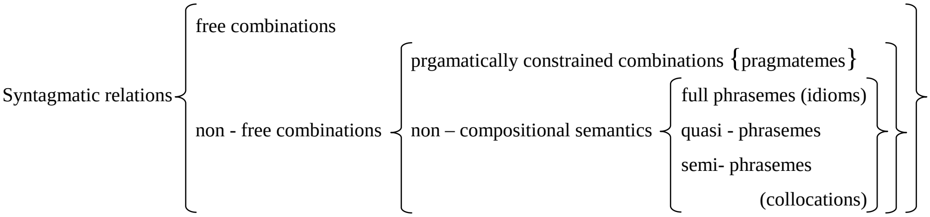


Figure 2. Mel’cuk’s typology of syntagmatic relations

Viegas et alii (1998) argue for a continuum perspective, ranging from free-combining words (totally compositional meaning) to semantic collocations, idiosyncrasies and idioms (non-compositional meaning):

— free-combining words (*a wonderful man*);
— semantic collocations (*a fast car*; *a long book* (cf. Pustejovsky (1995) account of such expressions by the use of a coercion operator));
— idiosyncrasies
  - restricted semantic co-occurrence (the meaning of the collocation is semi-compositional. “There is an entry in the lexicon for the base (...), whereas we cannot directly refer to the sense of the semantic collocate in the lexicon, as it is not part of its senses. We assign the co-occurrence a new semi-compositional sense, where the sense of the base is composed with a new sense for the collocate. (...) For instance (...), a *heavy smoker* is someone who smokes a lot, and not a ‘fat’ person. (...) we do not have in our lexicon for heavy a sense for ‘a lot’ (...)” (Viegas et alii, 1998:1329-1330);
  - restricted lexical co-occurrence (the meaning of the collocate is compositional but has a lexical idiosyncrasy behavior. “(...) there are entries in the lexicon for the base and the collocate, with the same senses as in the co-occurrence. (...) What we are capturing here is a lexical idiosyncrasy or in other words, we specify that we should prefer this particular combination of words” (op. cit.: 1330). It is the case of

expressions such as *rancid butter* or *sour milk*);

— idioms (*to kick the bucke*t).

Finally, a more complex typology, created from a natural language processing point of view in order to avoid overgeneration, idiomaticity and parsing problems, is presented by Sag et alii (2002), and it covers the following types of expressions:

1. Lexicalized Phrases

   Word combinations that present at least partially idiosyncratic syntax or semantic or contain words which do not occur in isolation. They can be subclassified into:

   a) Fixed Expressions – Immutable expressions that are fully semantically and syntactically lexicalized, like *in short*, and that do not undergo neither morphosyntactic variation (**in shorter*) nor internal modification (**in very short*).

   b) Semi-fixed Expressions – Expressions that present constraints on word order and composition, but undergo some degree of lexical variation. These expressions can be subclassified into:

      (i) Non-decomposable Idioms – Expressions that have a non-compositional meaning (*kick the bucket*) and that are not subject to syntactic variability (**kick the great bucket*). The only type of variation observable is inflection (*kicked the bucket*).

      (ii) Compound Nominals – Syntactically unalterable units that can inflect for number, like *car park* or *part of speech*.

      (iii) Proper Names – Expressions syntactically idiosyncratic where one of the elements may be optionally ellidable (*the Oakland Raiders* → *the Raiders*).

   c) Syntactically-flexible Expressions – Expressions that exhibit a much wider range of syntactic variability than fixed or semi-fixed expressions. These expressions can be subclassified into:

      (i) Verb-particle Constructions – Constructions that consist of a verb and one or more particles, such as *look up* or *fall of*. In some cases these verb-particle constructions may take a NP argument between or following the verb and particle(s) (*call Kim up*; *call up Kim*). However, other cases are compatible with only one realizations (*fall of a truck*; **fall a truck of*). Adverbs can often be inserted between the verb and the particle (*fight bravely on*).

      (ii) Decomposable idioms – Expressions that do not have a compositional meaning but tend to be syntactically flexible to some degree (*sweep under the rug*).

      (iii) Light Verbs – Constructions highly idiosyncratic, like *make a mistake* or *give a demo*, where is difficult to predict which light verb combines with a given noun (**do a mistake*; **give a demo*). These constructions are subject to full syntactical variability, like passivization (*a demo was given*), extraction (*how many demos did Kim give*) and internal modification (*give a revealing demo*).

2. Institutionalized Phrases

   Expressions that are syntactically and semantically compositional but statistically idiosyncratic, like *traffic light*, *fresh air* or *kindle excitement*. Given the strict compositionality, it would be expected the same concept to be expressible in other ways (like *traffic director* or *intersection regulator*). The idiosyncrasy of these expressions are statistical rather than linguistic in that they are observed with much more higher frequency than any other lexicalization of the same concept. As institutionalized phrases are fully compositional, they undergo full syntactic variability.

In order to provide a contribution for the study and classification of MWEs in Portuguese language, the project Word Combinations in Portuguese Language (COMBINA-PT), developed at the Centre of Linguistics of the University of Lisbon (CLUL), aims at the creation of a large lexical database of European Portuguese MWEs automatically extracted through the analysis of a large corpus of naturally occurring data, statistical interpreted with lexical associations measures and validated by hand. The availability of large amounts of textual data and corpus-driven analysis enables adequate descriptions of the concrete use of language, which would remain impossible if researchers only rely on introspection and native speaker intuition.

4. Corpus-driven methodology

For MWEs extraction, a corpus of 50M tokens was compiled, using a 330M tokens monitor corpus of Portuguese language developed at CLUL, the *Reference Corpus of Contemporary Portuguese* (CRPC)[2]. The COMBINA-PT corpus of 50M tokens is a balanced written corpus covering newspapers, books, magazines and journals and other documents (see Table 1 below).

[2] CRPC is a written and spoken monitor corpus compiled at CLUL since 1998 and comprises all the national and regional varieties of Portuguese (http://www.clul.ul.pt/english/sectores/projecto_crpc.html).

Table 1. Constitution of the corpus

CORPUS CONSTITUTION			
NEWSPAPERS			30.000.000
BOOKS	Fiction	6.237.551	
	Technical	3.827.551	
	Didactic	852.787	10.818.719
MAGAZINES AND JOURNALS	Informative	5.709.061	
	Technical	1.790.939	7.500.000
MISCELLANEOUS			1.851.828
LEAFLETS			104.889
SUPREME COURT VERDICTS			313.962
PARLIAMENT SESSIONS[3]			277.586
TOTAL			50.866.984

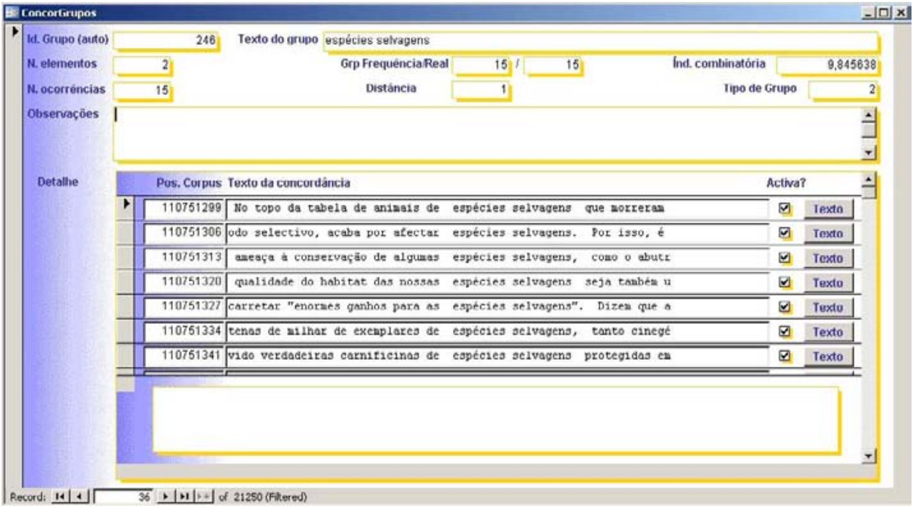
A program specifically developed to extract MWEs (CONCOR.CB) was then applied on the corpus in order to automatically extract all groups of 2, 3, 4 or 5 tokens. The following information is provided for each group:

- Number of elements of the group;
- Distance between the group elements: groups of 2 tokens can be contiguous or be separated by a maximum of 3 tokens, while groups of more than 2 tokens are contiguous;
- Frequency of the group at a specific distance;
- Total frequency of the group in all occurring distances;
- Frequency of each element of the group;
- Total number of tokens in the corpus;
- Concordances lines (KWIC format) of the MWE in the corpus, together with an index code pointing to its exact occurring position in the corpus.
- Lexical association measure: groups automatically extracted are statistically analysed using a selected association measure and are afterwards sorted. The tool allows the user to select which measure to apply, and was first run with Mutual Information (MI). MI calculates the frequency of each group in the corpus and crosses this frequency with the isolated frequency of each word of the group, also in the corpus (Church & Hanks 1989).

The large candidate list extracted from the corpus and the need of effective ways to reduce noise made it necessary to implement several cut-off options. With the first option we eliminated groups with internal

[3] Parliament sessions are considered written data since the spoken sessions undergo extensive revision when transcribed.

punctuation, while with the second we eliminated word pairs with first or final grammatical word using a stop-list (to rule out non-lexical associations). The third option eliminated groups under a selected total minimum frequency: 4 for groups of 3 to 5 tokens, and 10 for 2-token groups. The final candidate list obtained still comprises the considerable number of 1.751.377 MW units. A lexical database was designed in MySQL format so as to enable the representation of MW units and to offer a platform for user-friendly manual validation. An example of a record represented in the database is presented in figure 3. For more information on the extraction and validation process, see Mendes et alii (2006) and Antunes et alii (2006).



Id. Grupo (auto) 246
Texto do grupo espécies selvagens
N. elementos 2
Grp Frequência/Real 15 / 15
Índ. combinatória 9,845638
N. ocorrências 15
Distância 1
Tipo de Grupo 2
Observações
Detalhe

Pos. Corpus	Texto da concordância	Activa?
110751299	No topo da tabela de animais de espécies selvagens que morreram	☑
110751306	odo selectivo, acaba por afectar espécies selvagens. Por isso, é	☑
110751313	ameaça à conservação de algumas espécies selvagens, como o abutr	☑
110751320	qualidade do habitat das nossas espécies selvagens seja também u	☑
110751327	carretar "enormes ganhos para as espécies selvagens". Dizem que a	☑
110751334	tenas de milhar de exemplares de espécies selvagens, tanto cinegé	☑
110751341	vido verdadeiras carnificinas de espécies selvagens protegidas em	☑

Record: 36 of 21250 (Filtered)

Figure 3. Record for the collocation *espécies selvagens* ‘wild species’ in the database

5. Typologies meet corpus data: definitions of MWEs challenged

As we saw in Section 1, definitions of MWE are essentially based on two fundamental criteria: syntactic fixedness and semantic non compositionality, although typologies of MWEs present different views regarding the second criterion since compositional groups are also viewed in some literature as being a type of MWEs. A third criterion, also much discussed, is frequency of occurrence and statistical information.

5.1. Lexical and Syntactic fixedness

The more restrictive definitions consider that a MWE must present a certain degree of syntactic fixedness, but the exact degree of variation that a MWE can undergo without ceasing to be one has not been established and, when working with corpus data, we find high levels of lexical and syntactic variation.

The first and obvious variation in a highly inflected language like Portuguese, as well as in other romance languages, is inflected variation, together with contractions of prepositions and articles/pronouns. For example, in the group *estar atento a* ‘to be attentive to’, the verb can vary in person, number and time, the adjective varies in gender and number and the prepositional element can be contracted with articles and pronouns, giving a large set of possibilities (e.g., *estou atento à* ‘I’m attentive to_the[fem, sg]’, *estamos atentos ao* ‘we are attentive to_the[masc, sg]’, *estivemos atentos àquela* ‘we were attentive to_that_one[fem]’ - contracted elements are connected in our English translation). To cover all possible realizations of the MW expression lemma *estar atento a* implies recovering and organizing all different word forms that the group comprises.

MWE fixedness is usually related to contiguous realization of the group elements, but when we observe corpus data, it becomes obvious that, especially in the case of MWEs including a verb form, non contiguity is extremely frequent. In most cases, an adverbial element can be inserted, like the group *respire fundo* ‘breathe deeply’, that also occurs as: *respire bem fundo* ‘breathe very deeply’. In these cases, should we consider the existence of one MWE *respire fundo*, with possible variations, or of two independent MWEs? The question becomes even more difficult to answer when facing another group occurring in the corpus and clearly related: *respire profundamente* ‘breathe profoundly’.

With verbal expressions, contiguity is also challenged when verb complements occur inside the MWE: the MWE *pôr em causa* (‘to question’, literally: ‘to put in cause’) will require a direct object that will mostly occur in post-MWE position *pôr em causa* [*algo*] ‘to question [something]’, although it can be lexicalized inside the MWE (*pôr* [algo] *em* causa ‘to [something] question’) and pronominalized as well (as in the corpus occurrence *pô-lo em causa* ‘to question it’, literally: ‘to put it in cause’).

A similar process occurs in the case of the MWEs comprising possessive constructions, where the prepositional phrase expressing possession can be lexicalized as a possessive pronoun inside the MWE. For example, the following occurrences: *está nas mãos do governo* ‘(it) is in the hands of the government’, *está nas mãos da Assembleia* ‘(it) is in the hands of the Assembly’, *está nas nossas mãos* ‘(it) is in our hands’, *está nas vossas mãos* ‘(it) is in your hands’ are in fact all realizations of two abstract structures: *estar nas mãos de* [*X*] ‘to be in the hands of [X]’, *estar nas* [*POS*] *mãos* ‘to be in [POS] hands’, where the varying elements, the nominal phrase and the possessive pronoun, are expressed with placeholders. The two structures are obviously related and might be seen as corresponding to a single MWE with syntactic alternation, but it is also true that there is not always correspondence. For example, if the possessive element expresses the first or second person: *está nas minhas / tuas mãos* '(it) is in my / your hands', the structure with prepositional phrase is not available: **está nas mãos de mim / ti* '(it) is in the hands of me / you'. As these examples show, hands-on work with a high number of MWEs candidate list raises the difficult question of determining, when faced with high lexical and syntactic variation, which MWEs are in fact realized in the corpus.

5.2. Syntactic Alternations

Just like in the precedent example of possessive constructions, fixedness is also challenged by the syntactic variation of most MWEs comprising a verb, since most admit syntactic alternations like passive or relative constructions. For example, the same expression *pôr em causa* [algo] can undergo passivization of the direct object: *ser posto em causa* 'to be questioned'. The verbal expression *correr riscos* 'to take chances' can also undergo passivization, with elevation of *riscos* to subject position: *foram corridos riscos desnecessariamente* 'chances were taken unnecessarily' (in this example, subject will preferably occur in post-verbal position). Relativization of the word *riscos* also occurs in the corpus: *os riscos que correm* 'the chances that they take', and takes morpho-syntactic variation even further since the MWE *correr riscos*, with no article, is then obligatorily realized with a definite article *os riscos* 'the chances'. The MWE *correr riscos* also occurs with *riscos* in singular and preceded by definite article *correr o risco de* 'to take the chance of', so that three different lexical realizations are presented in the corpus: *correm riscos*, *os riscos que correm*, *correm o risco de*. While the second one is more directly related to the first, via relativization, and would be considered a variant of the MWE *correr riscos* (despite the insertion of a plural definite article), the third one will be considered a separate MWE *correr o risco de* [X], since *risco* occurs in singular form and is usually followed by a complement (prepositional phrase). While some variation corresponds to syntactic alternations of a MWE, other will point to the existence of another MWE, although clearly related to the first one.

5.3. Semantic patterns

Corpus data also shows cases where lexical variation in one specific position points to a specific semantic pattern that can be lexicalized as very different elements. For example, the verbal expression *revelar pormenores* 'to reveal details' always occurs in the corpus preceded by elements expressing a negative value, that can be a single adverb *não* 'no' or *sem* ‘without’ or complex sequences like *ainda é cedo para* ‘(it) is still early to’, variants of the structure [*NEG*] *revelar pormenores* ‘[NEG] reveal details’.

não	revelar pormenores	‘**not** to reveal details’
sem	revelar pormenores	‘**without** revealing details’
escusando-se a	revelar pormenores	‘**avoiding to** reveal details’
Ainda é cedo para	revelar pormenores	‘**(it) is still early to** reveal details’

Figure 4. Concordances of the expression [*NEG*] *revelar pormenores* ‘[NEG] reveal details’

These interesting patterns of semantic and syntactic co-occurrence go beyond lexical variation among the same morpho-syntactic category and point to the existence of MWEs that are a complex combination of fixed elements and a semantically constrained structural position.

5.4. Lexical variation

Definitions and typologies of MWEs usually associate fixedness and non compositionality as criteria for identifying MWEs. However, our corpus data show that MWEs considered as frozen, like idioms, can show a surprising level of lexical (and sometimes syntactic) variation. The following idiomatic expression *No poupar é que está o ganho* ‘In the saving is the profit/Profit is in saving’ forms a sentence that occurs 3 times in the corpus, while several other corpus occurrences show that one position inside this MWE allows large lexical variation:

No	**poupar**	é que está o ganho.	‘Profit is in **saving**.’
No	**anunciar**	é que está o ganho.	‘Profit is in **announcing**.’
No	**atacar**	é que está o ganho.	‘Profit is in **attacking**.’
No	**descontar**	é que está o ganho.	‘Profit is in **discounting**.’
No	**prejuízo**	é que está o ganho.	‘Profit is in **losing**.’
No	**esperar**	é que está o ganho.	‘Profit is in **wainting**.’
No	**provar**	é que está o ganho.	‘Profit is in **tasting**.’
No	**cooperar**	é que está o ganho.	‘Profit is in **cooperating**.’
No	**comparar**	é que está o ganho.	‘Profit is in **comparing**.’
No	**economizar**	é que está o ganho.	‘Profit is in **economizing**.’

Figure 5. Lexical variation of the expression *no poupar é que está o ganho* ‘profit is in the saving’

Although expressions like *No poupar é que está o ganho* are clearly frozen in our mental lexicon, corpus shows that speakers do substitute some parts of the expression when using it. This does not undermine the idiomatic nature of the expression since, when confronted to the non canonical versions, Portuguese speakers immediately acknowledge that it is a version of a frozen expression. However, it does challenge our conception of idioms as the MWEs showing the highest degree of fixedness, leaving the question of whether there exist a totally frozen type of MWEs, that typologies consider to be at one end of the continuum of fixedness. This lexical variation of even the most idiomatic expressions raises questions regarding automatic identification of MWEs in the corpus: as mentioned in Section 4, a threshold was established as a cut-off measure, eliminating groups under the minimum frequency of four, and thus eliminating the expression *No poupar é que está o ganho*. (This expression can be recovered by the smaller group *é que está o ganho*, occurring twelve times.)

This internal lexical variation shows clearly that this MWE, although perceived as a single unit, do have internal structure and is analysed as such by the speakers.

5.5. Non-compositionality

The task of determining whether a MWE has compositional or non compositional meaning is also not straightforward in many cases. Non compositional meaning would imply that the meaning of the expression is not equivalent to the sum of the words individual meanings. However, in cases like *preencher um vazio* ‘to fill emptiness [in a psychological sense]’, the MWE can be considered compositional if the meaning of *preencher* ‘to fill’ and *vazio* ‘emptiness’ are not assumed as being only physical, which they are not. Establishing the compositional nature of a MWE is thus a task that presumes that one knows what is the meaning or the meanings of each element of the group, not a smaller task.

Some expressions are still compositional but also gain a pragmatic value, like the case of *podes crer* ‘you bet’ (literally: (you) can believe) that really expresses that someone can believe what was previously expressed by another speaker, but that also expresses a subjective attitude from the speaker, an attitude of strong assertion in informal contexts of dialogue or conversations.

Since lexicalization is the result of a gradual process, a specific word sequence can present different degrees of cohesion, synchronically observable. For example, a sequence like *fazer a cama* can be: a free combination with compositional meaning (to built a bed); a fixed combination but still compositional since the meaning of the expression is deduced from the meaning of its elements (to make/arrange the bed); and a strongly lexicalized expression, with non-compositional meaning (to frame someone).

5.6. Frequency and statistical data

We mentioned above that frequency is a much discussed criteria for MWEs identification. When applied to MWEs like *no poupar é que está o ganho*, that we expected to be totally frozen but was not, the impact of frequency for the identification of this particular type of MWE is clearly negative, since low frequency of the group in its original form makes it non recognizable via frequency. However, in the case of MWEs that show a lower degree of lexical and syntactic fixedness as well as a compositional meaning, like the case of preferred co-occurring forms that correspond to a usual way of saying something, then frequency and statistical information is an important criteria to identify those lexical associations and is part of the definition of those units. Those MWEs tend to express semantic relationships: semantic domain sharing, like *insultos e ameaças* ‘insults and threats’, *críticas e acusações* ‘criticisms and accusations’, *competências e atribuições* ‘competences and atributions’; antonymy, like *ganhos e perdas* ‘profits and losses’, *fixos e móveis* ‘fixed and mobile’, *públicas e privadas* ‘public and private’; complementarity, like *trabalhadores e empregadores* ‘workers and employers’; or adverbial intensification with a specific adverb *absolutamente indispensável* ‘absolutely indispensable’.

Looking at MWEs occurring in the corpus also gives us important information on the most frequent types of MWEs. For example, in what concerns verbal expressions, two different kinds are extremely frequent: those involving a verb with its internal complement, like the MWE *correr riscos* ‘to take chances’, and those involving what is usually called a light verb, like *pôr em causa* ‘to question’, with the light verb *pôr* ‘put’. However, a very infrequent type of verbal MWE is the one involving a verb and its subject, like the examples *correm rumores* ‘rumours are flying around’ and *os exemplos abundam* ‘examples abound’.

6. Conclusion

Large corpus data gives us important information on MWEs since it makes visible lexical and syntactic variation that speakers are not always conscious of and challenge our intuitive native speakers’ beliefs on the total fixedness of at least certain types of MWEs. This corpus-driven and usage-based information has two important consequences to the study of MWEs: a revision of the fundamental criteria that define what constitutes a MWE: fixedness, non-compositionality and frequency; the study of their applicability to different subtypes of MWEs.

Besides the issues on fixedness degree and compositional meaning, the study of these MW expressions allows to identify associative patterns that characterizes a word according to: (i) co-occurrence patterns (systematic co-occurrence with particular lexical items in a contiguous or non-contiguous form); (ii) grammatical patterns (systematic co-occurrence with a certain verb class, with specific temporal verb forms or with certain syntactic constructions); (iii) paradigmatic patterns (hyperonymy, homonymy, synonymy or antonymy phenomena); (iv) discursive patterns (strong associations in one language register can be a weak association in another register).

The ultimate goal is to establish a proposal of a corpus-driven typology of MWEs for Portuguese language taking into account the three main criteria discussed above, as well as morphosyntactic properties of the expressions.

References

Antunes, S., M. F. Bacelar do Nascimento, J. M. Casteleiro, A. Mendes, L. Pereira, T. Sá (2006) "A Lexical Database of Portuguese Multiword Expressions" in VIEIRA, R. et alii (2006) *PROPOR 2006*, LNAI 3960, Berlin, Springer-Verlag, pp. 238-243.

Bahns, J. (1993) "Lexical collocations: a contrastive view", *ELT Journal,* 47:1, pp. 56-63.

Benson, M., E. Benson & R. Ilson (1986) *The BBI Combinatory Dictionary of English: a guide to word combination*, Amsterdam/Philadelphia, John Benjamins Publishing Company.

Braasch, A. & S. Olsen (2000) "Toward a Strategy for a Representation of Collocations – Extending the Danish PAROLE-lexicon", *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May – 2 June 2000, vol. II, pp. 1009-1016.

Butler, C. S. (1998) "Collocational Frameworks in Spanish", *International Journal of Corpus Linguistics*, vol. 3(1), pp. 1-32.

Calzolari, N. et alii (2002) "Towards Best Practice for Multiword Expressions in Computational Lexicons", *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands; Spain, 29 May – 31 May 2002, pp. 1934-1940.

Church, K. W. & P. Hanks (1989) "Word association norms, mutual information, and lexicography", *Computational Linguistics,* 16 (1), pp. 22-29.

Clear, J. (1993) "From Firth principles: Computational tools for the study of collocation", in Baker, M., G. Francis and E. Tognini-Bonelli (eds.) *Text and technology: In honour of John Sinclair,* Amsterdam, John Benjamins.

Cruse, A. (1986) *Lexical Semantics*, Cambridge, Cambridge University Press.

Evert, S. & B. Krenn (2001) “Methods for the Qualitative Evaluation of Lexical Association Measures”, *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, pp. 188-195.

Firth, J. (1955) “Modes of meaning”, *Papers in Linguistics 1934-1951*, London, Oxford University Press, pp. 190-215.

Firth, J. (1957) “A Synopsis of Linguistics Theory, 1930-1955”, *Studies in Linguistic Analysis*. Oxford Philological Society; reprinted in Palmer, F. (ed.) (1988) *Selected Papers of J. R. Firth*, Harlow, Longman.

Hausmann, F. J. (1979) “Un dictionnaire des collocations est-il possible?”, in *Travaux de Linguistique et de Littérature XVII*, 1.

Hausmann, F. J. (1989) “Le dictionnaire des collocations”, in Hausmann, F. J. et alii (eds.) *Wörterbücher: ein internationales Hanbuch zur Lexicographie. Dictionaires. Dictionaires*. Berlin/New-York, De Gruyter, pp. 1010-1019.

Heid, U. (1998) “Towards a corpus-based dictionary of German noun-verb collocations”, *Euralex 98 Proceedings*, Université de Liège, Belgique.

Kjellmer, G. A. (1994) *Dictionary of English Collocations*, Oxford, Oxford University Press.

Krenn, B. (2000a) “CDB - A Database of Lexical Collocations”, *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May – 2 June 2000, vol. II, pp. 1003-1008.

Krenn, B. (2000b) “Collocation Mining: Exploiting Corpora for Collocation Identification and Representation”, *Proceedings of KONVENS 2000*, Ilmenau, Deutschland.

Krishnamurthy, R. (1997) “Keeping good company: Collocation, Corpus and Dictionaries”, in *Cicle de Conferències 95-96*, Institut Universitari de Lingüistica Aplicada, Universitat Pompeu Fabra, Barcelona, pp. 31-56.

Lakoff, G. (1983) “The Contemporary Theory of Metaphor”, in Ortony, A. (ed.) *Metaphor and Thought*, Cambridge, Cambridge University Press, pp. 202-251.

Mackin, R. (1978) “On collocations: Words shall be known by the company they keep”, in *Honour of A. S. Hornby*, Oxford, Oxford University Press, pp. 149-165.

Mel’cuk, I. (1984) *Dictionnaire explicatif et combinatoire du français contemporain*, Les Presses de L’Université de Montréal, Montréal, Canada.

Mel’cuk, I. (1996) “Lexical Functions : A Tool for the Description of Lexical Relations in a Lexicon”, in Wanner L. (ed.), *Lexical Functions in Lexicography and Natural Language Processing*. Studies in Language

Companion Series (SLCS), Amsterdam/Philadelphia, John Benjamins Publishing Company, pp. 37-102.

Mendes, A., S. Antunes, M. F. Bacelar do Nascimento, J. M. Casteleiro, L. Pereira, T. Sá (2006) "COMBINA-PT: a Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions", *Proceedings of the V International Conference on Language Resources and Evaluation - LREC2006*, Genoa, May 22-28 2006.

Pearce, D. (2002) "A Comparative Evaluation of Collocation Extraction Techniques", *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Spain, pp. 13-18.

Pereira, L. A. S. & A. Mendes (2002) "An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications", in Braasch, A. & C. Povlsen (eds.), *Preceedings of the 10th EURALEX International Congress*, Copenhagen, Denmark, vol. II, pp. 841-849.

Pereira, L. A. Santos (1994) *Como se combinam as palavras? Contributo para um Dicionário de Combinatórias do Português*, M.A. Thesis, Faculty of Letters, University of Lisbon, ms.

Pustejovsky, J. (1995) *The Generative Lexicon*, Cambridge/Massachussets, The MIT Press, Massachussets Institute of Technology.

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger (2002) "Multiword Expressions: A Pain in the Neck for NLP", in Gelbukh, A. (ed.) *Proceedings of CICLing-2002*, Mexico City, Mexico.

Sinclair, J. & A. Renouf (1991) "Collocational Frameworks In English", in Aijmer, K. and B. Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Longman, Harlow, pp. 128-143.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Viegas, E., S. Beale & S. Nirenburg (1998) "The Computational Lexical Semantics of Syntagmatic Relations" in *Proceedings of the 17th International Conference on Computational Linguistics*, Montreal, Quebec; Canada, Volume II, pp. 1328-1332.

2.2.
UBLI

Usage-Based Approach to Linguistic Variation — Evidence from French and Turkish —

Yuji KAWAGUCHI

1. Introduction

There exists an idiomatic Japanese expression of “a forgotten umbrella.” The expression is derived from the legend of a famous architect who intentionally left his umbrella inside a construction that he had built in order to indicate the imperfection of his masterpiece. The question of linguistic *usage* versus *norm* constitutes one of the forgotten umbrellas in the history of modern linguistics. In the introduction of *Language*, Leonard Bloomfield professes the following:

> “(...) why, for example, many people say that *ain’t* is “bad” and *am not* is “good.” This is only one of the problems of linguistics, and since it is not a fundamental one, it can be attacked only after many other things are known.” (Bloomfield 1935: 22)[1]

Further, in the last chapter entitled “Chapter 28 Applications and outlook,” he again refers to the issue of *norm* with regard to the sociolinguistic diversity between *it’s I* and *it’s me* (Bloomfield 1935: 496ff). For Bloomfield, the question of linguistic usage versus norm was not an urgent and fundamental problem of linguistics; however, it remains as a question entrusted to the future development of this empirical science. The structuralists-functionalists’ framework is in an identical situation, and their researches have never been particularly concerned with the notions of usage and norm, as is clear from the following statement by André Martinet:

> “(...) But we deliberately disregarded these differences so as to not to complicate our exposition: the analysis of a supposedly uniform language is such a delicate task that one needs to simplify the data as much as possible. However, now that this analysis is accomplished, we must necessarily introduce into our examination all the facts which we provisionally set aside.” (Martinet 1964: 136)

For Martinet, after the completion of his structural sketch of the French

[1] See also Blanche-Benveniste (1997) pp.35–45 for discussions on grammatical errors and Gadet (2003) pp.17–23 for discussions on norm.

phonological system, he was thus not late in taking back his forgotten umbrella, i.e., the "diversity of data deliberately disregarded in his analysis," in order to make his phonology more elaborate. It was with this conviction that he, together with his collaborator Henriette Walter, published *Dictionnaire de la prononciation française dans son usage réel* in 1973; the main concern of the dictionary was to capture the diversity in phonological variation observable in a small language community of seventeen Parisians and to depict its situation as faithfully as possible. Based essentially on the same data, Walter also published *La dynamique des phonèmes dans le lexique français contemporain* in 1976. With regard to the French language, it would be safe to state that the first systematic analyses on linguistic usages began from the phonological level in the midseventies.

2. Norm versus usage

It is well known that general linguistics is deeply indebted to two linguists, Louis Hjelmslev and Eugenio Coseriu, for the notions of *usage* and *norm* [2]. Not content with the famous saussurian dichotomy of *langue* and *parole*, Hjelmslev was of the opinion that even if social reality is essential for *langue,* we could consider *langue* as having three distinct forms: (a) *schema* or the pure form, defined independently from its social realization and material manifestation; (b) *norm* or the material form, defined by social realization but independently from the detail in its manifestation; and (c) *usage* or the simple sum of the habits adopted in a given society and defined by observed manifestations.

For instance, the phoneme /r/ represents *schema* as an item of opposition in the French phonological system. For Hjelmslev, the phoneme /r/ represents a schema that is unrelated with its phonetic realization and is a kind of phonological prototype. Its sole raison d'être is to be distinguished from the other sounds except /r/.

However, once this phoneme is related to phonetic reality in French, the fricative property of /r/ is the *norm* in contemporary French because /r/ is realized in standard French as a fricative sound.

Finally, the phoneme /r/ can demonstrate the different phonetic habits of the French-speaking community, such as an alveolar trill and voiced or voiceless uvular fricatives, which are all *usages* in French.

Interestingly, immediately after Hjelmslev posited *norm* between *schema* and *usage*, he declared that the notion of *norm* is no more than a

[2] It is also indebted to Klaus Heger from the germanistic viewpoint. The scientific notion of *usage* can probably date back to a sixteenth-century French grammarian, Louis Meigret; see Glatigny (1982).

fiction or an artifact to depict *usages* and concluded that it would provide nothing but unnecessary complications for linguistic theory[3]. Consequently, in the above example of /r/, only two notions are left, i.e., *schema* as the phoneme /r/ and *usages* as different phonetic habits.

Along the same lines as Hjelmslev, Coseriu—who was not entirely convinced of the saussurian *langue* as a pure relational system—conceived the notion of *norma* as a social institution of *langue*. For instance, according to Coseriu's explanation, Spanish does not have the opposition between the closed /e/ and the open /ɛ/. At the level of *langue,* we thus have one phoneme /e/. However, this statement is justified only at the level of system, not of *norma,* because the vowel is closed in *queso* and *cabeza* but open in *papel* and *afecto* [4]. The realization of whether /e/ is closed or open is determined through the social institution of Spanish, and this is the reason why Coseriu emphasized the importance of *norma* in linguistic analysis. Although the phoneme is identified in the phonological system in question, the social aspect of *norma* plays an important role in determining its phonetic realization.

Both Hjelmslev's *usage* and Coseriu's *norma* may represent the latitude in phonetic realizations that a linguistic unit, here phoneme, can occur in any given language community.

Although Hjelmslev underestimated the notion of *norm* in his linguistic description, the importance of *norm* in the study of linguistic *usage* should not be dismissed. Examples of syntactic variation will help to clearly distinguish *usage* from *norm* and demonstrate the raison d'être of *norm*. Shana Poplack observed spoken Canadian French and gave some examples, on which she commented as follows:

(1) Ce soir, on *va* te *ramener* (PF) puis tu y *alleras* (IF) à soir à cinq heures. (071/584)
"Tonight, we'*re going to bring* you *back* and you'*ll go* there at five in the evening."

(2) Si mon petit *allait* (IMP) à l'école là, s'il *serait* (COND) à l'école puis qu'il *reviendrait* (COND) puis qu'il *dirait* (COND), "Un professeur m'a tapé dans la face là," il *aurait* affaire à moi. (037/437)
"If my kind *went* to school, if he *would be* at school, and he *would come* back,

[3] "La *norme,* d'autre part, est une fiction, — la seule fiction qu'on rencontre parmi les notions qui nous intéressent. L'usage, comprenant l'acte, ne l'est pas. Le schéma non plus. Ces notions représentent des réalités. La norme, par contre, n'est qu'une abstraction tirée de l'usage par un artifice de méthode. Tout au plus elle constitue un corollaire convenable pour pouvoir poser les cadres à la description de l'usage. À strictement parler, elle est superflue; elle constitue quelque chose de surajouté et une complication inutile." Hjelmslev (1971) p.88

[4] Coseriu (1981) p.57.

> and he *would say,* ‘a teacher slapped me across the face,’ he’*d have* to deal with me.”

> (...) The replacement of both the subjunctive by the indicative and the imperfect (IMP) by the conditional (COND) are thoroughly non-standard, while the incursion of the periphrastic variant (PF) into the domain of the inflected future (IF) is generally considered colloquial. (e.g., Poplack 2001: 407, underlined and annotated by Kawaguchi)

In sentence (2), the imperfect is replaced three times by the conditional—*serait, reviendrait,* and *dirait*—and in sentence (1), the simple future is replaced once by the periphrastic future *va ramener*. It is evident that both these syntactic variants belong to the usages of spoken Canadian French. A highly significant difference between these two usages lies in the fact that only the former is judged as nonstandard, i.e., it is not socially accepted from the prescriptive viewpoint. Further, only the notion of *norm* can explain the difference in the natures of these two usages, (1) and (2). As previously defined by Hjelmslev, normative judgment is independent of its material manifestation. Thus, the form *alleras* in (1) can be the norm at least in the spoken Canadian French quoted here, but never in that of French in France.

Such a divergence of norm between Canadian French and French in France is also attested in the usage of the subjunctive. Blanche-Benveniste clearly remarks on a significant difference in the usage of the subjunctive in spoken French between Canada and France.

> But, as far as we can rely on our spoken French corpus, the subjunctive is widely used in the present tense, with a rather large lexical dispersion, although there is an important lexical fixation on one verb governing the subjunctive, the verb *falloir* (must).[5]

It is true that some lexical effect seems to cause a high frequency of occurrences such as *il faut que (je/il) (fasse/aille/dise)* in spoken French in France. Nevertheless, the number of verbal and prepositional units governing the subjunctive is greater than fifty. Their lexical variety is sufficiently large. On the contrary, as Shana Poplack could show, in the large corpus of Canadian spoken French[6], two-thirds of the 2694 subjunctive forms are

[5] See the paper by Blanche-Benveniste, “Linguistic Analysis of Spoken Language—The Case of French Language,” in this volume, pp.35-66.

[6] “The corpus contained the speeches of 120 adult native speakers of French residing in Ottawa-Hull. The speakers were selected and interviewed according to standard sociolinguistic procedures, resulting in about 240 hours (3.5 million words) of naturally occuring speech”; see Poplack (1992) p.243.

subordinated to the main verb *falloir*. These two studies can eloquently attest to the existence of two different norms in the spoken French on both sides of the Atlantic Ocean.

As we have already seen, Martinet and Walter's dictionary is entitled *French Pronunciation Dictionary in Its <u>Real Usage</u>* (*Dictionnaire de la prononciation française dans son <u>usage réel</u>*). Their research on real usage in the pronunciation of French later developed into the study of dynamic synchrony. Dynamic synchrony is particularly important for structuralists because the phonological system is presented as a more or less static and fixed schema, i.e., a pure relational linguistic unit that is apparently lacking impetus for further evolution. As predicted by Martinet, dynamic synchrony was a future task after the structural description of the French phonological system, in which the diversity in phonetic realizations was deliberately disregarded. Dynamic synchrony was one of the perspectives for structuralists to elicit idiosyncratic or sociolinguistic tendencies or variations that can be observed in any given language community.

Based on the dictionary of Martinet and Walter, we have summarized in Table 1 the phonetic latitudes that seventeen Parisians were using around 1970 as a palatal nasal phoneme (Brandão de Carvalho et Kawaguchi 2002: 12).

Table 1.

		[ɲ]	[nj]	[ɲj]	[n]	[gn]
Predominant palatal nasal [ɲ]	*éloignement*	15	2			
	gagne-petit	14	2		1	
	hors-ligne	13	2			
	trépignement	12	5		1	
	châtaigne	10	7			
	charogne	10	7			
	peignure	10	4	3		
Fluctuation	*agneau*	8	9			
	beignet	8	9			
	daigner	8	9			
	saigner	8	9			
	peignier	5	8	4		
	récognitif	3	4		1	9
Predominant [nj]	*saignée*	7	10			
	gainier	6	11			
	panier	3	14			
	lainier	1	16			

The phonetic variation for French palatal nasal phonemes seems to be qualitatively divided into three usages:

1) the palatal nasal [ɲ], which is generally realized before a consonant (*éloignement, gagne-petit,* etc.) and in word-final position (*châtaigne, charogne*)
2) fluctuation, which occurs in intervocalic position (*agneau, beignet,* etc.)
3) [nj] is predominant in the ending *–nier* [nje]

These different usages are apparently conditioned by their phonetic environments. However, it is important to note that they are obviously distinct from the combinatory variants of a given phoneme because two variants—palatal nasal and [nj]—coexist in all the items of Table 1 in various proportions of occurrences. At the same time, it must be understood that these usages do not represent free variants because they cannot be exchanged in any situation and will not be used completely in accordance with the speaker's caprice.

As appropriately remarked by Hjelmslev, we are fully convinced that these usages are ontologically interpreted as the social phonetic habits of the French-speaking community. However, what do social habits imply in this French palatal nasal, given that no fewer than three different usages exist among only seventeen Parisians? In any case, it seems very difficult to uncover the real scale of these usages among the entire French-speaking community. Further, how will these so-called social habits be operative in linguistic analysis? We are rather interested in depicting these linguistic usages in order to better comprehend the dimension of linguistic variation in general. In other words, we want to examine here the kind of methodology by which such usages will be identified, the manner in which the usages can be integrated in the liguistic description, and the reason why the usage-based viewpoint is important for the analysis of linguistic variation.

3. Quantitative aspect of usage

It is not surprising that the existence of different linguistic usages is demonstrated in domains other than phonology. In fact, we can find a similar phenomenon in the morphology of contemporary Turkish. Some Turkish adjectives can derive corresponding emphatic forms through a special prefixation. Usage is fixed and stable for words that begin with a vowel. For example, the emphatic forms of *engin* "vast" and *olgun* "mature" are *ep-engin* and *op-olgun* respectively at any time and for any Turkish native. On the contrary, for other adjectives, usages seem idiosyncratic and variable. According to my questionnaire that was administered to sixteen Turkish university students in Istanbul, several usages are in current use.[7]

“very calm”		“very simple”	
sapsakin	100%	*bambasit*	63%
		basbasit	38%
“very clever”		“very big”	
zepzeki	80%	*yüpyüce*	63%
zemzeki	17%	*yüsyüce*	31%
		yümyüce	6%
“very bad”		“very wild”	
fepfena	81%	*vapvahşi*	50%
fesfena	13%	*vamvahşi*	31%
femfena	13%	*vasvahşi*	19%
“very clean”		“very natural”	
pampak	75%	*dopdoğal*	50%
paspak	25%	*dosdoğal*	43%
“very happy”		“very light”	
hophoş	69%	*haphahif*	44%
homhoş	19%	*hashafif*	19%
hoshoş	13%	*hamhafif*	10%

The above examples prove that Turkish intensive adjectives can be formed through the reduplication of the first consonant and vowel with a syllable-final inserted consonant /p/, /s/, or /m/: *sakin → sap-sakin, doğal → dos-doğal,* and *basit → bam-basit*. For *zeki, fena, hoş,* and *hafif,* the rules of insertion of the syllable-final /p/, /s/, or /m/ are not common for all Turkish natives. Further, for some adjectives, we can observe a lexical effect or pressure in determining the inserted consonant: *zeki* and *fena* for /p/ and *pak* and *basit* for /m/. It is interesting to note that the insertion of /p/ may furnish a schema in the morphology of intensive adjectives (Kawaguchi 1992: 321–322). The insertion of /p/ would be regarded as a kind of “zero degree of prefixation” and it would probably be safe to state that this schema with the inserted /p/ is omnipresent in the morphological variation of Turkish emphatic adjectives.

As suggested in the model of usage-based grammar posited in the framework of cognitive linguistics, we define type frequency as the number of different lexical items to which a particular morphological or syntactic pattern is applicable. Type frequency may become an important factor in

[7] Kawaguchi (1992). All my informants were born in Istanbul and their parents are also from the same city. Since the questionnaire contained multiple-choice questions, some of the scores are over 100%. The scores under 100% imply the presence of other minor variants.

determining the productivity of a given usage (Bybee and Thompson 2000, Poplack 2001). In Turkish emphatic adjectives, it is assumed that type frequency would be assigned to the schema with the inserted /p/.

Ronald Langacker offers a convincing argument regarding the following orientation. “We must recognize that language is a mixture of regularity and irregularity and deal with this fact in a natural, appropriate way. (...) Linguists have occasionally invoked suspicious devices to make things appear more regular than they really are” (Langacker 1987: 45). In fact, some linguistic schools overestimate the conception of language as a system of general rules and tend to ignore the irregular and idiosyncratic aspects of language. On the other hand, usage-based linguistic analysis attempts to comprehend regularity and irregularity in a more natural way and to understand social and idiosyncratic phenomena simultaneously.

As already demonstrated in Poplack’s investigation of subjunctive forms in spoken Canadian French, it is often the case that the functional load of a given usage is weighed by its token frequency. Her research reveals that nearly two-thirds of the total 2694 occurrences of the subjunctive include a single verb *falloir* “must.” She comments that “this imbalance is compounded by two other verbs, *vouloir* “want” and *aimer* “like” (Poplack 2001: 411–412). The frequency effect and the lexical strength of *falloir, vouloir*, and *aimer* therefore account for nearly three-fourths of all the subjunctive forms in her corpus. The result is far from the general sketch of the subjunctive in traditional grammar and structural description. Occurrences of the subjunctive are closely related to the “kind” of the verb.

The following is a final example to convince our readers regarding the discrepancy between a structural sketch and real usage. In French, there are two synonymous verbs of cognition, *savoir* and *connaître* “to know.” Structural analysis reveals two syntactic contraints for *connaître*: the verb cannot be followed by either an infinitive or a subordinate clause. On the other hand, semantic analysis focuses on the nature of the direct object that both verbs can take. In examples (3) and (4), their synonymy is no more than superficial. In fact, the semantic trait of *savoir* lies in the fact that the speaker directly holds the discourse process, which enables him to speak about Pierre’s profession—for instance, “*Il est professeur.* (He is a teacher.)” However, this is not the case with *connaître*. The verb *savoir* in (5) may represent deeper knowledge. The sentence “*Il sait la forêt.*” seems to be paraphrased into “*Il connaît l’âme de la forêt.*” Finally, semantic analysis reports that (6) and (7) are real synonyms.

(3) Je sais la profession de Pierre. “I know the profession of Pierre.”

(4) Je connais la profession de Pierre.

(5) Il sait la forêt. = Il connaît l’âme de la forêt.

"He knows the essence of the forest."

(6) Je sais l'anglais. "I know English."

(7) Je connais l'anglais.

These syntactico-semantic explanations seem irrelevant to the real usage of these two verbs. From my search of three spoken French corpora comprising over one million words, I found 160 tokens of *je sais* and 166 tokens of *je connais* [8]. The percentages of the possible constructions for both verbs are striking; see Figure 1. There is not a single context in which *je sais* and *je connais* followed by a direct object noun might be concurrent. The only verb that can be followed by a direct object noun is *connaître*. This implies that the above explanations regarding (3)–(7) are theoretically justified at the level of schema as items of opposition, but they are difficult to prove in real usage. The use of *savoir* is particulary exploited with a direct object infinitive or clause. Among others, it is important to note that indirect questions with *si* and subordinate clauses with *que* account for 35% of the total occurrences of *je sais*. The most significant result of this modest search, however, is the high frequency of *je sais* and *je connais* in their absolute constructions without any overt direct object noun. Of course, their high frequency has been predicted when exclusively selecting first-person singular forms. Nevertheless, the discrepancy between a structural sketch and real usage is evident to any reader of the following lines.

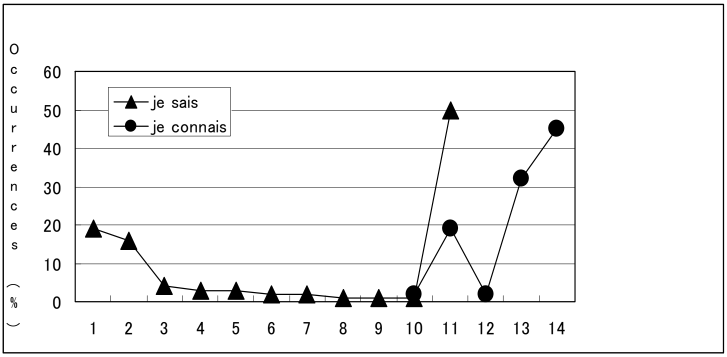


Figure 1. Direct object of *je sais* and *je connais*
Subordinate Clause or Inf.:
(1) je sais si..., (2) je sais que..., (3) je sais ce que..., (4) je sais quel..., (5) je

[8] The corpus was published at the site of the linguistics department of the Catholic University of Louvain: the spoken French corpora of Orléans, Tours, and Auvergne. However, in order that the occurrences may be equal for both verbs, I used the examples found only in the Tours and Auvergne corpora for *je sais*. cf. http://bach.arts.kuleuven.ac.be/lancom/

sais + infinitif, (6) je sais pourquoi..., (7) je sais où..., (8) je sais combien..., (9) je sais comment...

Preposed Object:

(10) N je sais/connais

Without direct object:

(11) je sais/connais Ø

Nominal Object:

(12) je sais/connais ça, (13) je connais inanimate N, (14) je connais animate N

The percentages of the occurrences of this absolute construction without any overt direct object are 50% for *savoir* and 19% for *connaître*. Such a high frequency seems surprising in these corpora representing a "relatively formal" register of an interview in French[9]. It must be an interesting fact that the frequent use of *je sais* causes not only the deletion of the overt direct object but also the emancipation from the original cognitive meaning of *je sais,* which serves almost as an interjection (Bybee 2001: 9).

4. Computer-assisted usage analysis

As shown in the previous section, the study of usage reveals the dynamic aspect of language, i.e., *dynamic synchrony* in structuralists' terminology and *emergent structure* in the current trends of cognitive linguistics. It is assumed that different usages represent an ongoing process in the resystematization of language. In most cases, the frequency effect is an important impetus in both innovative and conservative terms in the ongoing resystematization, and the process of constant restructuring of usages can be described quantitatively on the basis of relatively large linguistic corpora. Thus, linguistic researches in the analysis of real usage are inevitably assisted by informatics.

The standardization of regional dialects is a typical case of the resystematization of different usages toward a single standard. As confirmed by Yarimizu et al. (2005), simple statistics may clarify different degrees of standardization as tendencies in the different dialects of the Linguistic Atlas of Île-de-France and Orléanais (ALIFO), not far from Paris. However, if we

[9] The Orléans corpus is based on the field recording carried out from 1968–71. The total recording time is around 315 hours, of which approximately 80 hours are transcribed as the text, which contains about 900,000 words. The field methods used were of several types: very formal interviews (47 hours), secret interviews (27 hours), telephone interview (2 hours), discussions (32 hours), etc. The Tours corpus, which is a small corpus with 36,000 words, has been compiled from relatively formal face-to-face interviews that were conducted in 1974. It involved 193 informants from various professions and generations. Finally, in 1976, 17 hours of relatively formal round-table conversations were recorded for the Auvergne corpus. The corpus contains about 176,000 words.

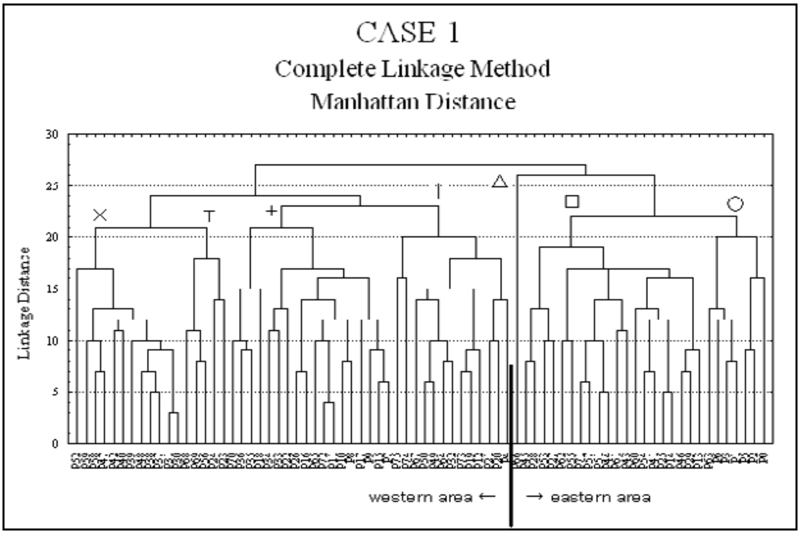
want to estimate the lexical effect of the standardization, e.g., the fact that some words are more susceptible to standardization than others, we should presuppose that even in the areas where standardization is relatively advanced, it is not always the case that the same words have been standardized at all the given points. Cluster analysis is one of the popular methods for grouping apparently heterogeneous data by measuring the distances between the elements.

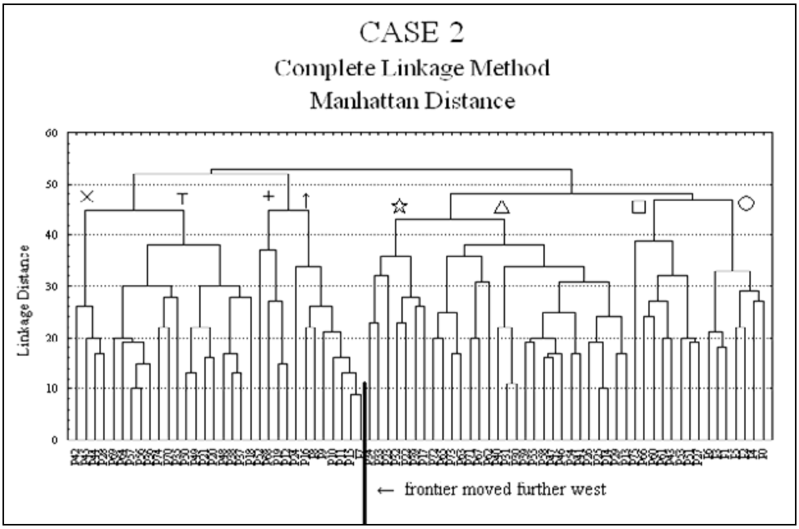
Kawaguchi (forthcoming) postulates three different cases to measure the distance between the geographical variants of ALIFO. In Case 1, the value "one" is assigned to the standard form and "two" is assigned to all other forms. The standard form is based on the description contained in Martinet and Walter's *Dictionnaire de la prononciation française dans son usage réel*. Case 1 strongly emphasizes the distinction between the standard and nonstandard forms. In other words, Case 1 neglects any further possibility of differentiation among the nonstandard forms. In Case 2, the ordinal scale from "one" to "four" is adopted; however, in some cases, we are obliged to assign the value "five"' so that several variants can be appropriately ordered. In Case 3, the same ordinal principle is adopted; however, an arbitrary and discrete value "ten" is assigned exclusively to the lexical variants. This special weighting is an attempt to mathematically separate the lexical variants from the other variants. The example of Map 62 *charrue* "plough" illustrates the values assigned to each variant.

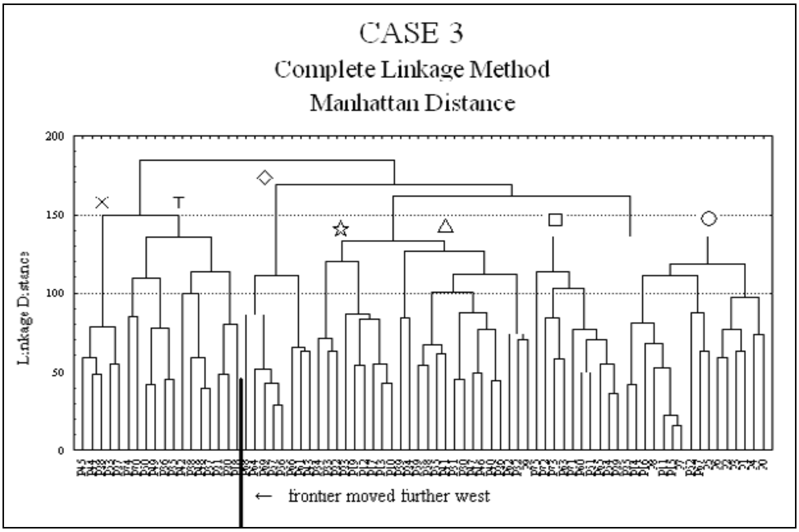
No.62 charrue		Case 1	Case 2	Case 3	
variant	category	value	value	value	occurrences
¢àrü	standard form	1	1	1	6
¢árü	phonological variants	2	2	2	59
¢á^{e}rü		2	2	3	1
¢árü-vèrswèr		2	3	4	2
¢árü$_{y}$		2	3	4	1
¢èrü$_{y}$		2	3	3	1
¢a^{o}rü		2	2	3	6
¢èrü		2	3	3	2
bràkér	lexical variants	2	4	10	1
kutèt		2	4	10	1
vèrswè		2	4	10	1
vèrswa^{è}r		2	4	10	1

Geographical variants are not numerical and the decision regarding the formal similarity or difference is sometimes equivocal and arbitrary. Thus, by combining the Manhattan distance and the complete linkage method, we attempted to separate as far as possible two apparently near variants and to calculate the farthest distance between two variants. We analyzed fifty-one different ALIFO maps by means of cluster analysis.

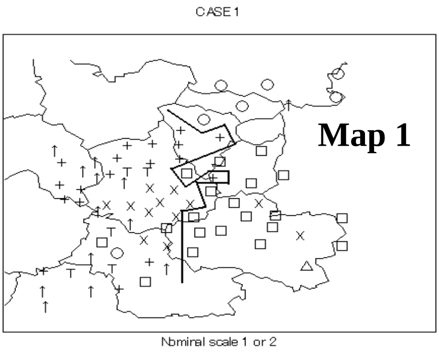
The following three dendrograms show two large dialect groups. See the bold perpendicular by which the points of ALIFO are divided into two main dialect groups, i.e., the group of ○, □, △, ✳, ◇ and that of X, T, +, ↾. Proceeding from CASE 1 to CASE 3, we see the borderlines moving from the right to the left.

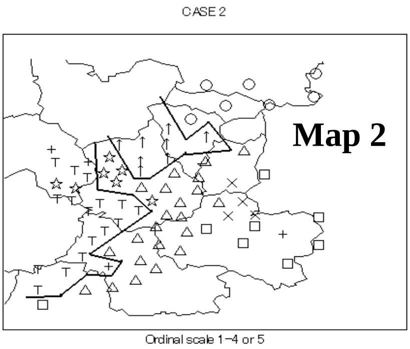


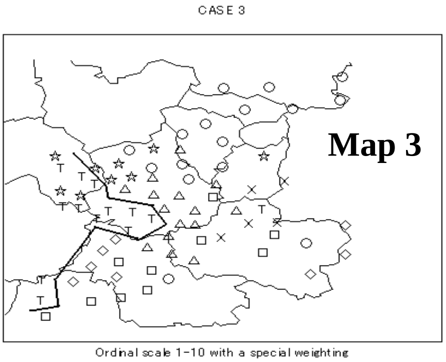




In order to interpret the transition of the borderlines and the two main groups in the dendrograms, we plotted these boundaries and the seventy-six research points on the linguistic atlas of ALIFO; see Maps 1 to 3. The boundary drawn on Map 1 (CASE 1) separates the eastern standardized area from the less strongly standardized western area. In Map 2 (CASE 2), the standardized area is extended toward the west. It is important to note that the area of the triangles and stars, which are distributed particularly in Eure-et-Loir, Loiret, and Loir-et-Cher prefectures, represents the area where we can observe morphophonological variants of the standard French. In Map 3 (CASE 3), the area of lexical variation is isolated particularly in Orne, Sarthe, and southwest Eure-et-Loir; see the area of T.







Eure
Paris
Orne
Eure-et-Loir
Essonne
Sarthe
Loiret
Loir-et-Cher
ALIFO

The three different maps clearly show three different dialect situations of ALIFO. The first standardized area is located around Paris; see the area of circles and rectangles in Map 1. The second area of morphophonological variants is marked by triangles and stars in Map 2. The third area of lexical variation is the area of T's and X's in Map 3. These three areas appear to reproduce the distances between the geographical variants. Maps 2 and 3 clarify the difference between morphophonological variants and lexical variants. The area delimited by the borderline in Map 2 is considerably more extended than the area demarcated by the borderline in Map 3. This implies that the morphophonological variation is considered as closer to the standard. It seems that multivariate analysis contributes not only to the definition of the linguistic distance but also to clarify the distinctions among different regional usages of the language.

It is surprising that we are today at the disposal of a tagged and parsed corpus of post-1990 written Turkish texts; this corpus has been developed by the Middle East Technical University (METU). From this METU Turkish corpus, whose total number of words was estimated at about 862,700 at the time of my analysis, I selected all the examples of two clause linkage suffixes, -DIK- and -mE-. Some determining factors for the choice of these two clause linkages are composed of the differences in the semantics and cognitive features of the main verb. The tripartition of verbs, i.e., (1) manipulation verb, (2) modality verb, and (3) perception-utterance-cognition verb, explains to some extent a general tendency or usage in the choice of the two clause linkages.

main verb	manipulation verb	modality verb	perception-utterance-cognition verbs
clause linkage	generally -mE- with some fluctuations	mostly -mE-	generally -DIK- with many twofold cases

It is well known that *söylemek* "to say," when it controls the -mE-complement, has a deontic meaning, which it does not have with the -DIK-complement. Here, the distinction among the clause linkages is accompanied by the semantic difference of the main verb.

(8) Sana *haber etmemi söylediler*: (00082133)
to-you news to-do(mE) they-said
"They told me to inform you (= that I should inform you)":

(9) Benimle de bir röportaj yapmak *istediğini söyledi.* (00065211)
with-me too a reportage to-make to-want(DIK) he-said
"He said that he wanted to make a reportage with me too."

In example (8), the verb *söylemek* will be interpreted as a manipulative verb, and this is the reason why the verb takes the suffix -mE- as a clause linkage. On the contrary, in (9), where the same verb is used as an utterance verb, it is accompanied by -DIK-. In perception, utterance, or cognition verbs, we can find many twofold cases, i.e., examples of both -mE- and -DIK- attested for a single verb. In twofold cases, the choice of -mE- or -DIK- often corresponds to the aspectual distinction: -DIK- expresses an action that has been accomplished before the moment of utterance, while -mE- represents a simultaneous or posterior action with respect to the moment of utterance. This analysis demonstrates the existence of linguistic usage for the choice between -mE- and -DIK-. On the one hand, the choice of -mE- or -DIK- is closely related to the cognitive category of the main verb: manipulation and modality verbs are generally linked with -mE-, while perception-utterance-cognition verbs are generally linked with -DIK-. On the other hand, the choice of -mE- or -DIK- also depends on the aspectual difference of the embedded verb. In brief, the Turkish clause linkages -mE- and -DIK- are balanced between two different pressures from both main and embedded verbs, and this dynamics will allow perception-utterance-cognition verbs to have twofold cases.

The METU Turkish corpus contains post-1990 written Turkish texts. We wonder if we can expect the same results from a spoken Turkish corpus. The answer will be negative because parataxis, and not hypotaxis, is a frequent syntactic device in spontaneous spoken conversation. However, this kind of generalization could always be dangerous. In fact, as appropriately remarked by Blanche-Benveniste, the opinion that written language is based on hypotaxis and spoken language on parataxis turns out to be extremely simplistic[10]. In any case, no study has demonstrated such register variation in

[10] Blanche-Benveniste (1997) p.59.

the syntactic usages of contemporary Turkish. Further, it is more important to recognize that even for the most globally studied language—English—only around ten years have passed since computer-assisted corpus studies of its various registers were carried out.

Susan Conrad and Douglas Biber (2001) proposed a challenging methodology to examine the register variation in English. Using factor analysis of the frequency counts, they analyzed the co-occurrence patterns among linguistic features and interpreted the factors functionally as underlying dimensions of variation. Finally, computing the dimension scores for each text with respect to each dimension, they compared the mean dimension scores for each register in order to recognize the salient linguistic similarities and differences among the registers analyzed. As a result of this so-called multidimensional methodology, they arrived at a significant confirmation that the relations among registers can be plotted as several characteristic texts ordered along a continuum of dimensions of variation[11]. For instance, Dimension 1 is defined as the primary purpose of the writer or speaker and the production circumstances. Registers along Dimension 1 with high positive scores represent face-to-face and telephonic conversations, while registers with negative scores are those of academic prose, press reports, and official documents. This multidimensional study for examining register variation will be considered as a typical computer-assisted usage analysis.

The significance of computer-assisted usage analysis is not limited to current languages; on the contrary, it is being generalized in the philological analysis of old written languages. As far as Old French is concerned, an international workshop was held in 2006[12]. Further, with regard to the written texts of a later period, an interesting contribution to the philological study of French pronunciation usages was published by Yves-Charles Morin, based on several pronunciation dictionaries from the eighteenth century. In contemporary French, in word-final position, only the closed vowel [o] occurs, e.g., *sot* and *saut* [so]. The neutralization of two back mid vowels had been recorded since the beginning of the nineteenth century, but became the norm in the twentieth century. The only invariable native word—*trop*—had an open vowel at the beginning of the nineteenth century. The evidence reveals that we can discern a morphophonological effect at the early stages

[11] Conrad and Biber (2001) p.14, 26–28.

[12] *The New Amsterdam Corpus*, a workshop organized by Pierre Kunstmann (LFA, Ottawa) and Achim Stein (ILR, Stuttgart) and financed by the Program Trans Coop of the Alexander von Humboldt-Stiftung, Lauterbad (Schwarzwald), Germany, February 23rd–26th, 2006. The main objective of the workshop was destinated to the construction of the tagged and parsed corpora of several Old French literatures and charters.

of this sound change; it also reveals that nouns and adjectives were modified first and invariable words were modified only later. As appropriately remarked by Morin, it would be very difficult to comprehend the process of this change without large computer corpora (Morin 1989: 194).

5. Conclusion

It will not be an exageration to state that one of the most popular current trends of linguistics is the construction of spoken language corpora and their analysis. The C-ORAL-ROM project presented us with an opportunity to hold a collaborative workshop[13]. The problem of linguistic usage is clearly included in the main objective of C-ORAL-ROM. Massimo Moneglia explains it as follows:

> The C-ORAL-ROM corpora have been collected in Continental Portugal, Central Castilian Spain, Southern France, and Western Tuscany (...), and are intended to represent *a possible standard usage* that occurs in these areas. (...) C-ORAL-ROM thus represents the language actually spoken in relevant national centers (namely, Madrid, Lisbon, Aix-Marseille, and Florence) and in their neighbouring areas (Cresti et al. 2002). (...) (italicized by Kawaguchi)
>
> The goal of the collection is therefore limited to ensuring the representation of the sole diaphasic and diastratic varieties of the language that is spoken in the region in question (...).

The bulk of the different linguistic usages attested in every single and small language community may always be a torment for linguists when constructing a linguistic corpus; this is because it is extremely difficult to know if the corpus can manifest a certain degree of the "representativeness" of the language community in question. However, after the predominance of sociolinguistic interests in linguistics, the hypothesis of the "homogeneity" of a language community—which has been "bullhorned" in the structuralists' framework—began to be reproached. Nevertheless, it cannot be stated that some reliable descriptive methods for sociolinguistic variations have been well established in current linguistics. It is only recently that linguists have begun to make much account of actual usages and linguistic interactions. The main interest of our *Center of Usage-Based Linguistic Informatics (UBLI)*[14] is to elucidate the realities of linguistic usages. Taking an example

[13] The collaborative workshop of C-ORAL-ROM and UBLI, *Spoken Language Corpora — its significance and application,* was held on December 9th, 2005, at the Tokyo University of Foreign Studies, Fuchu City, Tokyo, Japan.

[14] The 21st Century COE Program at the Graduate School of Area and Culture Studies, Tokyo University of Foreign Studies (TUFS), cf. http://www.coelang.tufs.ac.jp/english/index.html

from our spoken Turkish corpus, let us confirm some problematic aspects of linguistic usages in the spontaneous interaction of two Turkish natives.

A1 – evet. # mer(h)aba nasılsınız duygu hanım ↑ [Gülme][15]
"*Yes. Hello. How are you, Ms Duygu? [Laugh]*"

D1 – iyiyim sağolun siz nasılsınız <ahmet hamdi bey ↑ >
"*Fine thanks. And you, Mr Ahmet Hamdi?*"

A2 - <teşekkür ederiz> gerçi az evvel söylediğimiz gibi sizli konuşmuyca(ğı)z ama neyse öyle şimdilik başlangıç olsun # evet # ee <ben>
"*Thanks. It's true, as I said before, we will not talk using "siz," but anyway, like this at the beginning. Yes. Ee. I*"

D2 - <tatile> çıkmayı düşünüyo(r) musun ↑
"*Are you going to be away on vacation?*"

A3 - ben tatile çıkmayı tabi düşünüyorum {herkez gibi} ama, # bu sene imkanlarımız olacak mı bilemiyorum e tatil için # gitmemiz gitmeyi düşündüğümüz yerler arasında güney sahilleri var gerçi ama...
"*I will take a vacation of course like everyone, but this year I don't know if there will be possibilities, e, for a vacation. The southern beach is among the places where we want to go, it's true, but...*"

D3 - nereye gitmeyi düşünüyo(r)sun ↑
"*Where do you want to go?*"

This kind of interaction can often be observed at the beginning of face-to-face conversations between relatively intimate Turks with similar social backgrounds. The above example is of a conversation between two postgraduate students of the Marmara University at Istanbul. There is no doubt that this interaction represents a common form of everyday conversation. Overlapping occurs between their discourses; see the parentheses < >. The omission of some consonants or syllables is typical of spoken usage: *mer(h)aba konuşmuyca(ğı)z* and *düşünüyo(r)*. Dislocation is also frequent in spoken Turkish; see the braces { }. All these linguistic traits enable us to claim that the present interaction will be characterized as a register of face-to-face conversation. However, I suppose that this fact will require additional consideration.

First, the interaction commenced with a relatively high politeness strategy: *nasılsınız duygu hanım, siz nasılsınız,* and *teşekkür ederiz*. Then, the locutor A changed his strategy into a more familiar one, declaring the following to his interlocutor D: *sizli konuşmuyca(ğı)z* "we will not talk using *siz* (= polite form)." It is not necessary for us to assume that a single

[15] Conventional transcriptions: # = pause; < > = overlapping; ( ) = form reconstructed by transcriber; [ ] = nonverbal action; { } = dislocated elements; ↑ interrogation

interaction or written text is always composed of one fixed register. A register may be susceptible to a change in discourse strategy by both the locutor and the interlocutor.

Second, a discourse marker *gerçi* “it’s true” is attested twice only in the discourses of A. In order to show this in finer detail, in our spoken Turkish corpus[16], we have only two of five informants who used this discourse marker; see the table below.

Informant	A	B	C	D	E
Occurrences of *gerçi*	11	14	3	0	0

It is evident that the use of *gerçi* is more dependent on our informants than on the topic of interaction[17]. Linguistic usage may be related with the idiosyncratic linguistic habits of informants.

Hopefully, our readers will understand the significance of usage-based approaches for the analysis of linguistic variation. The possible changes in register or usage constitute an extremely interesting domain of the analysis of dynamic synchrony, which describes the ongoing variation in a given language community. Usages are composed of the massive habits of language users[18]. Some usages endowed with social prestige or prescriptive value may become linguistic norms[19]. Usages should be described quantitatively as well as qualitatively, and in the former viewpoint, they would be affected not only by the frequency effect but also by the lexical effect. The main part of the analysis of linguistic variation should be corpus based. In dynamic synchrony, as different usages conflict with each other, the frequency effect may provide some cues to explain the inner dynamism of variation and to predict the emergence of some predominant types that will later be recognized as standard usages.

[16] The duration of the total recording was 288 minutes at the beginning of 2006. The corpus, which has around 42,900 words, is composed of ten different interactions between two Turkish postgraduate students.

[17] Topics of interaction (Occcurrences of *gerçi*): tourist resort (7), homeland (6), holiday (5), cigarette (3), television (3), EU and Turkey (1), explanation on Turkey (1), free time (1), foreign language education, and Turkish (1)

[18] Cf. “la norme objective”; see Helgorsky (1982) pp.1–5, Gadet (2003) p.19.

[19] Cf. “la norme subjective ou fictive” ; see Gadet (2003) p.19. However, it seems very difficult for linguistic analysis to make a clear distinction between objective and subjective norms.

References

Blanche-Benveniste, Claire. 1997. *Approches de la langue parlée en français,* Gap: Ophrys.

Bloomfield, Leonard. 1935. *Language*, London: George Allen & Unwin.

Brandão de Carvalho, Joaquim et Yuji Kawaguchi. 2002. "Linéarité et variation : le cas du "*n* mouillé en français." *Flambeau* 28 : 1-20.

Bybee, Joan and Sandra Thompson. 2000. "Three frequency effects in syntax." *Berkeley Linguistics Society* 23: 378-388.

Bybee, Joan. and Paul Hopper. 2001. "Introduction to frequency and the emergence of linguistic structure." In: Joan Bybee and Paul Hopper (eds.) *Frequency and the Emergence of Linguistic Structure.* Amsterdam: John Benjamins. 1-24.

Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

Conrad, Susan and Douglas Biber. 2001. "Multi-dimensional methodology and the dimensions of register variation in English." In: *Variation in English : Multi-Dimensional Studies.* Essex: Pearson Education Limited, 13-42.

Coseriu, Eugenio. 1981. "Sistema, norma y habla." In: *Teoría del lenguaje y lingüística general.* 1973. Madrid: Gredos. Japanese translation. translated by M. Hara and H. Ueda. *"Gengo Taikei, Gengo KaNyo, Gen." Coseriu Gengogaku Senshu 2. Gengo Taikei.* Tokyo: Sanshusha. 3-95.

Gadet, Françoise. 2003. *La variation sociale en français.* Gap: Ophrys.

Glatigny, Michel. 1982. "La notion de règle dans la «Grammaire» de Meigret", *Histoire, Epistémologie, Langage* 4:2. 93-106.

Helgorsky, Françoise. 1982. "La notion de norme en linguistique." *Le Français Moderne* 50: 1-13.

Hjelmslev, Louis. 1971. "Langue et parole", first published in *Cahiers de F. de Saussure* 2: 1943. 29-44. reproduced In: *Essais Linguistiques.* Paris: Minuit. 77-89.

Kawaguchi, Yuji. 1992. "Sur les adjectifs intensifs en turc moderne." *Turcica* 24: 317-330.

Kawaguchi, Yuji. 2002. "Gengo-nitotte kihan-toha nani-ka [The notion of *norm* for language]." *Gogaku Kenkyusho Ronshu* [Journal of the Institute of Language Research] 7: 49-73.

Kawaguchi, Yuji. 2005. "Two Turkish Clause Linkages: -DIK- and -mE- —A pilot analysis based on the METU Turkish Corpus—." In: Toshihiro Takagaki. Susumu Zaima. Yoichiro Tsuruga. Francisco Moreno-Fernández and Yuji Kawaguchi (eds.) *Corpus-Based Approaches to Sentence Structures,* Amsterdam: John Benjamins. 151-177.

Kawaguchi, Yuji. forthcoming. “Is it possible to measure the distance between near languages? — A case study of French dialects —.” Paper presented at the Near Language Conference held at Limerick (Ireland), on 16-18 June 2005.

Langacker, Ronald W. 1988. “A usage-based model.” In: Brygida. Rudzka-Ostyn (ed.). *Topics in Cognitive Linguistics*. Amsterdam/ Philadelphia: John Benjamins. 127-161.

Martinet, André. 1964. *Elements of general linguistics,* Translated by Elisabeth Palmer, London: Faber and Faber LTD.

Milroy, James and Lesley Milroy. 1985. *Authority in language : investigating standard English,* London/New York: Routledge.

Morin, Yves-Charles. 1989. “Changes in the French vocalic system in the 19th century.” In: Bert Schouten and Pieter van Reenen (eds.) *New Methods in Dialectology,* Dordrecht / Providence: Foris Publications. 185-197.

Poplack, Shana. 1992. “The inherent variability of the French subjunctive.” In: C. Lauefer and T.A. Morgan (eds.) *Theoretical Studies in Romance Linguistics*. Amsterdam: John Benjamins. 235-263.

Poplack, Shana. 2001. “Variability, frequency, and productivity in the irrealis domain of French”, In: Joan Bybee and Paul Hopper (eds.) *Frequency and the Emergence of Linguistic Structure,* Amsterdam: John Benjamins. 405-428.

Rémy-Giraud, S. 1986. “Étude comparée du fonctionnement syntaxique et sémantique des verbes savoir et connaître.” In: S. Rémy-Giraud et M. Le Guern (dirs.) *Sur le verbe*. Lyon : Presses Universitaires de Lyon. 169-306.

Walter, Henriette. 1976. *La dynamique des phonèmes dans le lexique français contemporain*. Paris: France Expansion.

Yarimizu, Kanetaka. Yuji Kawaguchi and Masanori Ichikawa. 2005. “Multivariate Analysis in Dialectology —A Case Study of the Standardization in the Environs of Paris—.” In: Yuji Kawaguchi. Susumu Zaima. Toshihiro. Takagaki. Kohji. Shibano. Mayumi. Usami (eds.) *Linguistic Informatics — State of the Art and the Future —*, Amsterdam: John Benjamins. 99-119.

Viewpoint and Postrheme in Spoken Turkish

Selim YILMAZ

The syntactic structure of the Turkish language is as follows: (*S) O V + pers*. Depending on the context, the subject may be omitted at the beginning of the sentence since the (verbal or nominal) predicate includes a mark of person in itself. In spoken language, however, this syntactic structure tends to include another element known as the postrheme. The postrheme may be identified as an additional element because its use is restricted to spoken language; on the other hand, in written Turkish, the postrheme results in an inverted structure and is sometimes even considered to be grammatically incorrect. Therefore, in contrast to the abovementioned syntactic structure of written Turkish, the syntactic structure of spoken Turkish follows either one of the following two patterns:

1. (VP) O Pred + pers (Postrheme)	2. (Lig) O Pred + pers (Postrheme)

Figure 1.

Every oral utterance is initialized by a ligateur ("linking word") or a marker of modality that is often a viewpoint (VP) marker. This VP has a pronominal nature of the type "ben" (me) and "bence" (in my opinion), which we will examine in detail at a later stage in this paper.

The utterance in (1) is typical of spoken Turkish with "ben" placed at the beginning of the first clause to introduce a rhematic and modalized utterance followed by the adverb "şahsen" (personally) and the predicate "düşünüyor-um" (I think).

(1) *ben* şahsen şöyle düşünüyorum (TY, AHT 30)[1]
me personally so I think
=> (me) personally, I think so

In this study, we attempt to answer the following questions:

a) With regard to enunciation, when does a speaker use a VP and postrheme?

b) What is the nature of the relationship between the VP and the

[1] The initials correspond to the title of the corpus and the speaker. For all other grammatical abbreviations, see the list provided at the end of this paper.

postrheme and can we consider these two elements as the first VP and the final VP, respectively?

c) What are the values and functions of these two elements in an interaction?

With these three questions in mind, we will explain the case of “ben” (me) at the beginning of the utterance as a typical VP marker and deduce the same role for “ben” at the end of the utterance in the position of postrheme. Nevertheless, we can state that the actual VP marker is the one that is placed at the beginning of an utterance and used to introduce a thematic clause. On the other hand, we will attempt to determine the status of “ben” as a postrheme in the end position. Finally, we will investigate whether or not a postrheme is responsible to mark or even highlight the VP of the speaker in a dialogue.

Considering that the Turkish language has a fairly variable syntax, the VP marker may be found at any position in an utterance. Let us consider the example of the first person singular pronoun “ben” as the VP marker in an oral utterance with the verbs “yap-mak” (to do) and “düşün-mek” (to think).[2]

YAPMAK (to do) / DÜŞÜNMEK (to think)		
Yapıyor-*um* Düşüneceğ-*im*	=> I’m doing => I’ll think	assertive and simple rheme
Ben yapıyor-*um* Ben düşüneceğ-*im*	=> (Me) I’m doing => (Me) I’ll think	VP + rheme
Bunu yapan ben-*im* Bunu düşünecek ben-*im*	=> That’s me who is doing => That’s me who will think	focused VP in rheme
Yapıyor-*um* ben Düşüneceğ-*im* ben	=> I’m doing (me) => I’ll think (me)	rheme + postrheme

Figure 2.

The examples of utterances that contain a VP and postrheme will be selected from five corpora of conversations. These conversations are familiar and friendly discussions recorded in a natural situation. The speakers are young Turkish students and teachers. The following are the different corpora along with their principal subject of conversation as well as their duration:

[2] It is necessary to note that the suffix “-mek/-mak” is the infinitive marker of Turkish verbs.

Corpora	Duration	Number of words
1. Corpus TY *Turistik Yerler* (Touristics Places)	29.12 min	4.095
2. Corpus M *Memleket* (Native Land)	26.25 min	3.672
3. Corpus D *Diller* (Languages)	28.18 min	4.361
4. Corpus E *Eğitim/Öğretim* (Education)	28.77 min	3.303
5. Corpus T *Türkiye* (Turkey)	35.42 min	6.308
TOTAL	147.74 min	21.739

Figure 3.

1. Two main categories of VP markers

Starting with the "predicate" as a syntactic reference, we can establish two important categories of VP in any oral utterance: (a) VP prior to the predicate and (b) VP subsequent to the predicate.

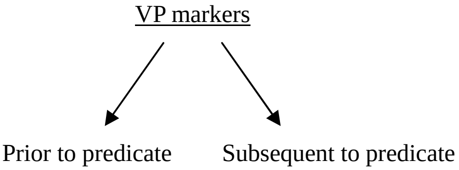


Figure 4.

1.1. VP prior to the predicate: "Ben" and its variants

The VP is an indication of modality that implies the personal thought of the speaker-enunciator. In Turkish, the explicit marker at the semantic level of this subjective modality is "bence," which means "in my opinion." However, when expressing a personal opinion, we can also use the first person singular pronouns "ben" (me), "ben de" (me too), and their variants "benim için" (for me) and "bana göre" (according to me).[3] The choice among these variations lies in the enunciative strategy that is determined by the position of the speaker with regard to the content of the utterance and the interlocutor. In example (2), "ben" introduces an utterance that ends with the person marker "-um" (1st SP) associated with the predicate "düşünüyor-um" (I think).

[3] In *Grammaire de l'intonation*, Morel and Danon-Boileau define the VP for the French language as "*Le point de vue souligne l'identité de l'énonciateur qui sert de caution à ce qui va être dit. On y trouve des expressions autonomes telles que 'moi, à mon avis, pour moi' ou bien encore 'X dit que, selon X...'. (...) L'expression du point de vue est parfois associée à la caractérisation de la valeur de la modalité, repérable dans l'emploi de certains pronoms ('on', 'tu' par exemple) ou conjonctions ('si' et 'quand' en particulier)*" (1998: 40).

(2) kesinlikle *ben* de aynı şeyi düşünüyorum (E, DH32)
certainly me also same thing I think
=> certainly, (me) I also think the same thing.

The explicit VP marker “bence” (in my opinion) comprises a combination of the personal pronoun “ben” and the suffix “-ce.” Here, both the elements fulfill the following functions:

BEN + CE	
BEN	-CE
Identification of the speaker as the enunciator at the moment of speaking => enunciative position (facing the other)	Determination of the VP [the object of the discourse: the experiences of the enunciator] => assumption of the VP by the enunciator

Figure 5.

Although the Turkish language disposes of several markers to express the VP, it is necessary to emphasize that these markers serve to introduce the VP that is established in the entire utterance and not merely in a certain part of it. Based on this fact, we can state that expressing the VP in spoken language is in fact enunciative rather than syntactic. In other words, it would be more adapted to consider this linguistic phenomenon as an enunciative proceeding that is linked to the syntactic structure of the utterance in question, i.e., the utterance in a properly defined context. With regard to this, Hagège, in his book titled *La Structure des Langues* (1982: 100), emphasizes the enunciation and the locutor-line underlying the notion of “ego as the centre of deixis”: “*Il reste qu’une propriété des énoncés linguistiques est d’être ancrés sur la situation d’énonciation. Au centre, celui qui les profère, le locuteur: ego, qu’il se nomme ou non par un “je” explicite, est le point de référence.*”

1.2. VP subsequent to the predicate: the “Postrheme”

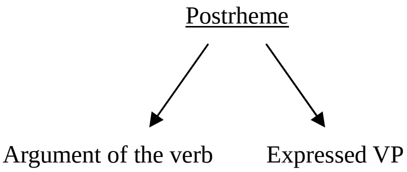
In Turkish, the last element of an utterance is, naturally, the predicate; however, in spoken language, this predicate may be followed by a postrheme, which then becomes the last element of an utterance.[4] The insistence on the terms the “last element” or the “final element” is due to the fact that simply because of their syntactic position, the first and last elements of an utterance naturally possess a particular discursive and enunciative value with regard to

[4] According to Morel and Danon-Boileau, “*Dans le postrhème, l’énonciateur redonne l’élément qui fonde sa prédication. Le repli en plage basse et l’absence de modulation souligne son désir de soustraire à la contestation son dire, dont il considère qu’il est le seul à pouvoir l’énoncer*” (Ibid., p. 30).

the other constituents of the utterance.

In fact, the initial and final positions not only constitute the syntactic boundaries of the utterance but are also a discursive way for the speaker to determine what is essential in his VP in order to draw the listener's attention to it.

Indeed, the postrheme reiterates *a posteriori*—either an argument of the predicate or the expressed VP.[5] Based on this conception and according to its syntactic position after the predicate, this final element is also considered as the "post-predicate" in American linguistics.



Postrheme ---> subjective modality ---> egocentered position

Figure 6.

However, the postrheme may also be defined by what is termed as "dislocation."[6] In their dictionary, Ducrot and Schaeffer (1995: 452) define dislocation as follows: "(...) *C'est le cas pour certaines intonations, et aussi pour certaines structures comme, en français, la dislocation, consistant à détacher un mot, et à le reprendre par un pronom non accentué: un énoncé comme 'Jean, il est venu' ne peut guère avoir pour thème que la personne désignée par le mot Jean.*"

Before analyzing the postrhematic VP categories, we would like to provide an interesting example that presents the value of subjective modality specific to postrhemes that reflect the egocentered position of a speaker. The following is an utterance in which the VP is marked by the postrheme "şahsen," which means "personally." While speaking about cooking specialties in his area, the speaker expresses his personal opinion when he states the following (M, AHT23):

[5] Morel M.-A. (1998), "*Analyse de la structure de l'oral*" (Copy of seminar from DEA/Doctorat), University of Paris III - Sorbonne Nouvelle, Center of French linguistics, EA 1483.

[6] Blanche-Benveniste defines the "postrheme" as a clause dislocated to the right (1997: 68). Moreover, according to her, this constituent also appears in the macro-syntactic classification of a "post-final" element that is situated after the nucleus (1997: 120–121).

(3) onları seviyorum {şahsen}[7]
=> I like them personally/I like these meals personally

Further, the following example represents a general case in which the VP marker is used as a postrheme:

(4) yok hiç hayatımda gitmedim {*ben*} (M, DH4)
no never in my life I have gone me
=> (no) I have never gone there in my life (me)

2. VP in the theme (at the beginning of the utterance)

In the thematic part, a VP expressing subjectivity—or more precisely, the personal judgment of the speaker—is always placed at the beginning of the theme; therefore, it is always placed at the beginning of the utterance. In this case, the central element of the theme or the object of the discourse is the speaker himself. Indeed, the utterance begins with “ben,” which introduces what is to follow—a real experience of the speaker or a personal situation.

Function of the VP: Disjoined lexical support

In this case, the VP marker functions as a “disjoined lexical support”[8] that is dissociated from the rheme by means of intonation and syntax.

(5) ama # *ben* tam olarak hatırlamıyorum (TY, AK6)
but me fully such as I don’t remember
=> but (me) I don’t remember completely

On the intonational curve of this example, we notice that with “ben” in the theme, the intonation increases to 300 Hz immediately after the ligateur “ama,” which has a lower intonation, and then decreases slightly. Such an intonative contour is typical of a theme that generally presents an intonative rise followed by an intonation decrease with regard to the rheme. Therefore, in this representative example, we note that “ben” follows this intonative rule.

[7] *Transcription conventions*: {...} postrheme, # silent pause, eee hesitation, <...> overlapping or covering of voices, (...) non-pronounced segment, and __ (underline) emphatic stress.

[8] This terminology is taken from Morel and Danon-Boileau (1998: 37–41) and is in perfect agreement with a VP that begins the thematic portion of any utterance in spoken Turkish.

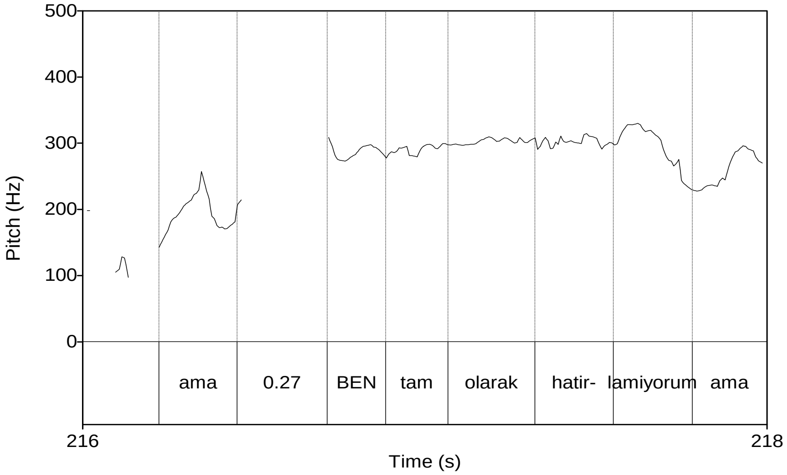


Figure 7.

The following are two examples using “bence” (in my opinion) and “benim” (to me):

(6) *bence* dünyadaki bi(r)çok insan için geçerli {bu olay} (TY, AK34)
in my opinion in the world most of person for true this fact
=> in my opinion, this fact is true for most people in the world

(7) ama *benim* hatırladığım kadarıyla yaklaşık yani elli yıllık filan
but to me that I remember so much around in other word 50 years so
bi(r) geçmişi var {üniversitenin} (TY, AK6)
one past there is the university
=> but for me, as much as I can remember, that is to say, it has a past of about fifty years, or so, the university

3. VP in the rheme (in the middle of the utterance)

The VP may also be marked in the rheme. In this case, the marker that is used most frequently is the personal pronoun “ben,” whose syntactic position is first in the rheme. Even if “ben” introduces the rheme, the predicate is generally accompanied by the suffix of the person, either in its nominal or verbal form. The speaker asserts himself in the discourse by using the structure “ben + rheme,” which may be paraphrased as “there really is me (myself) who....”

Function of the VP: To support the rheme (Support of the theme)

In a case in which the VP introduces the rhematic part of the utterance, the VP marker functions as a support for the predication of the rheme.

(8) e # ama işte # gidilen yerlerde *ben* şunu gözlemledim (TY, AHT36)
but there are visited in places me that I have observed
=> but here we go, I have observed that in the places which are visited

In this utterance (see graph 7), the theme is uttered at a low intonation (under 200 Hz), while the rheme shows two intonative increases to approximately 300 Hz. Moreover, it is clear that “ben” is dissociated intonatively, and there is no segmental or suprasegmental break of tone between this marker and the end of the theme. Based on this fact, we can deduce that “ben” is attached to the theme rather than to the rheme; however, in this case, the context also allows us to determine the particular part of the utterance to which “ben” belongs. The VP marker is indeed in close syntactic and semantic association with the rheme, and this is why we have two intonative peaks that increase to approximately 300 Hz. The intonative value of “ben” under 200 Hz (above 150 Hz) also shows that this constituent performs the particular function of introducing the rheme. The rheme, generally, presents a decreasing intonation until it reaches a low level (below 200 Hz).

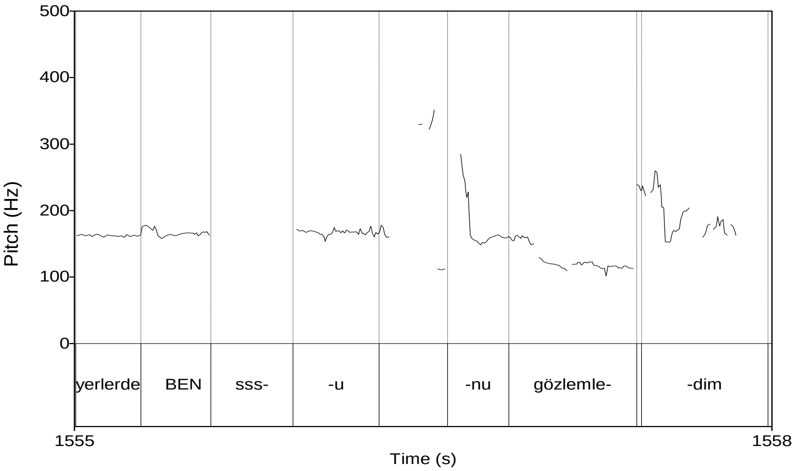


Figure 8.

Let us now consider other examples of utterances that contain the VP markers “benim” (ben + poss), “beni” (ben + acc), “bence” (ben + vp), and “bana” (ben + dat).

(9) yani Rizeli[9] olmanıza rağmen *benim* <dikkatimi çekti> (TY, AHT15)
in other words from Rize that you are although to me my attention has attracted
=> in other words, although you are from Rize, that has attracted my attention

[9] “Rize” is the name of a Turkish city located near the Black Sea.

(10) İstanbul çok kalabalık ## kalabalık olması *beni* çok rahatsız ediyor (T, MG6)
Istanbul more populated that it is to me much bothering it does
=> Istanbul is more populated; the fact that it is more populated upsets me a lot

(11) İstanbul'daki üniversitelerin Anadolu'daki üniversitelere göre
of Istanbul the universities of Anatolia to universities respect to
bence çok daha büyük avantajı var (E, MG33)
in my opinion more great advantage there is
=> in respect to the universities of Anatolia, the universities of Istanbul,
in my opinion, have many advantages

(12) a ama gerçekten bi(r) hafta da olsa *bana* yetti (M, DH30)
but really one week even if it is to me it has sufficed
=> but really, even a week, it has sufficed to me

4. VP in the postrheme (at the end of the utterance)

Function of the VP in the postrheme: Supplementary and/or complementary modality and explicitation marker of the rheme.[10]

In the position of the postrheme, the VP marker acts as a supplementary modality since it occupies the position immediately after the predicate, which is already associated with the marker of person. In this case, the role of the posthreme is to render explicit the reference of the argument that is expressed in the predication of the rheme. This reference is to the speaker himself and is marked with "ben."

4.1. Emphasizing a thematic element (the object of the discourse)

The postrheme is characterized by the fact that it is not separated from the rheme by a pause or any other clue, such as the hesitation token "eee,"[11] and that it never introduces a predication. Generally, the postrheme is short and brief in syntactic terms comprising only one to two words, and its role is to semantically complete the rheme with regard to the spatio-temporal or modal fields. In example (13), the postrheme "İstanbul'un" (of Istanbul) with the genitive "-un" shows that it is in fact a thematic element, i.e., a part of the object of the discourse that the speaker intends to emphasize for his interlocutor.[12]

It must be reiterated that emphasizing a thematic element by means of a postrheme is an operation that is undertaken as a part of the subjective

[10] This is due to the fact that the predicate is also suffixed with the marker of person.

[11] Such hesitation is expressed in different ways in different languages, for instance, it is transcribed as "euh" in French.

[12] With regard to this, Morel and Danon-Boileau remark that "*Le postrhème apparaît largement quand la stratégie de discours l'exige, notamment dès que la discordance se confirme*" (Ibid., p. 30).

modality and in relation with the estimation or argument of the speaker. In this example, the element that must be emphasized according to the speaker is the “economical status of Istanbul”; therefore, this is marked in the postrheme by a nominal construction and the genitive “İstanbul’un” (of Istanbul).

(13) fakat insanlar # daha ziyade ekonomik boyutuyla
but people rather economical status
<................................theme................................>
ilgileniyor {*İstanbul’un*} (Türkiye, MG8)
he interests of Istanbul
<rheme> <postrheme>
=> but people are rather interested in the economical status of Istanbul

4.2. Egocentric (egocentered) position with “ben”

The postrheme as a sign of modality is usually expressed by the VP marker “ben” (me) and its variants, such as “bence” (in my opinion), “benim için” (for me), and “bana göre” (according to me). Subjectivity with regard to the use of the personal pronoun “ben” is the dominant modality in the postrhematic position; this type of postrheme depicts an anaphoric return to what has been said in order to assume the discourse with a strong egocentric position.[13]

V + person (-m) {ben}

Figure 9.

In this predicative formula—where the verb is followed by the postrheme “ben”—there is a need for an assertive rheme that is composed of a more or less modalized predication.

(14) Türkiye’nin gerçekten turizm olarak çok önemli bir yeri olduğunu
of Turkey really tourism such as more important a place that it is
<.......................................theme...>
düşünüyorum {*ben*}
I think me
<......rheme......> <postrheme>
=> (me) I think that Turkey really is in an important position at the touristic level

[13] “*L’énoncé constitué d’un rhème et d’un postrhème est en quelque sorte bouclé sur lui-même: Le rhème exprime une conclusion polémique, une prise de position modale forte, mais l’élément qui la fonde est exprimé dans un postrhème qui se soustrait à la coénonciation, et l’argument qu’il donne devient ainsi irréfutable*” (Ibid., p. 30).

In the above example, the postrheme is the personal pronoun of the first person singular “ben”; using this, the speaker indicates that the matter being discussed is indeed his personal VP. However, this leads us to the following question: Why does the speaker need to use “ben” in the postrheme? In order to answer this question, we need to examine the theme and rheme that precede and, in a way, induce the emphasis on the VP in the postrheme.

Actually, the syntactic structure of the theme and rheme presents a semantic and modal value quite nottable; this explains the use of “ben” at the end of the utterance:

a) The epistemic modality with the adverb “gerçekten” (really) and the appreciative modality qualifier “önemli” (important) in the thematic part

b) The verb “düşünmek” (to think) with a subjective value that directly implies the speaker in the rheme.

The syntactic structure of the most semantized and modalized theme and rheme explains the function of emphasis on and the egocentered position of the postrheme “ben.” In French, the translation of the postrheme “ben” is placed in its usual position at the beginning of the utterance as a VP marker.

The structure “*V + 1SP(-m) ben*” with a VP marker in the postrheme position shows that there are two operations:

a) Identification of the subject of the process expressed

b) Assuming a modal position with “ben.”

In other words, first, by associating the suffix of the first person singular “-m” with the predicate, the speaker highlights that it is he who initiates the action, and second, the speaker expresses a position with respect to his statement by using “ben” and assumes his utterance. In the enunciative field, “ben” loses its status as a personal pronoun when it is used as the postrheme; in fact, it is a VP marker that has a supplementary subjective modality, which is used to complete the assertive rheme. *The speaker uses the marker of modality “ben” as a means to draw the listener’s attention to his personal opinion*.[14]

We can paraphrase the suffix of the person (1st SP) associated with the verbal predicate and the postposed pronoun of the first person singular as follows:

V + 1Ps (-m)	=> I am here saying or doing something
Postrheme “ben”	=> I am assuming what I have said (*anaphoric function*)

Figure 10.

[14] In other words, the syntactic placement of the postrheme at the end of the utterance may ideally be considered as an *enunciative specificity of the oral language*. Its role is to highlight the *modo-enunciative position* of the speaker who thus marks his VP (pertaining to subjective modality with an epistemic or appreciative value).

4.3. Egocentered assumption with "bence"

When the postrheme is the VP marker "bence" (in my opinion) that also is an egocentered position, but rather assuming the VP which is in the foreground. In other words, what is being discussed when the postrheme "bence" is used is the judgment that has just been expressed by the speaker and the assumption of the VP that results from the utterance. In French, "bence" may be translated using expressions such as "that's what I think" and "here is my opinion," depending on the context.

(15) başka ülkeleri gördüğün zaman ülkeni
other the countries that you see when your country
<................................theme...................................>
çok daha fazla geliştirebilirsin {*bence*} (TY, AK34)
much more too you can develop in my opinion
<..................rheme..................> <postrheme>
=> if you see other countries, you will develop your country more (you will improve the development of your country)

As evident from the intonational graph of this utterance, the thematic part of the utterance is higher and more modulated than the rhematic part. While the theme presents a modulation at approximately 350 Hz, the rheme decreases progressively toward 200 Hz at the end on the postrheme "bence," which is the lowest point of the utterance. This graph perfectly reflects the intonative characteristics of any oral utterance and postrheme that are always decreasing or flat at a low level (200 Hz or below).

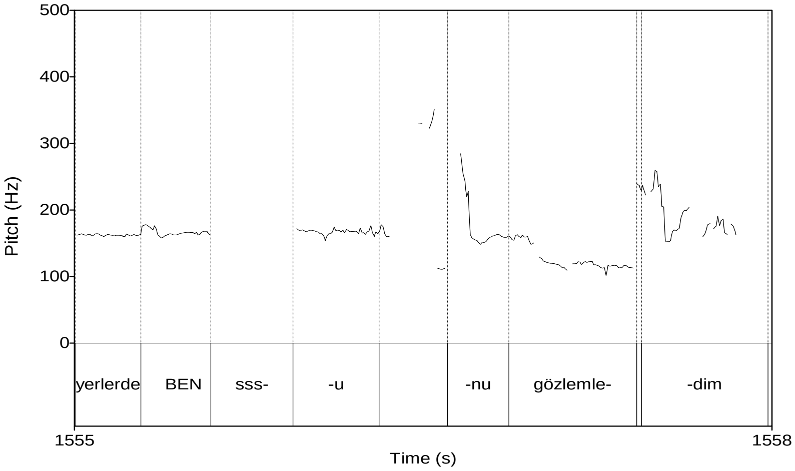


Figure 11.

4.4. VP with an epistemic modality

There are many cases in which the postrheme refers to an epistemic modality with adverbs such as "aslında" (in reality, in fact), "gerçekten" (really), "belki" (perhaps), etc.

Enunciative function (emphasizing the epistemic value of the predicate): In this case, the speaker wants to highlight the degree of truth of his statement, particularly with regard to the predicate.

Let us examine some examples of utterances that render explicit the epistemic value of a speaker's VP. The postrheme with an epistemic modality is often expressed only with one word, whose grammatical nature is that of an adverb; however, it is obvious that natural speech always presents more specific cases, as is exemplified in the second example provided below:

(16) işte Rize'de de vardır {*mutlaka*} (M, AHT23)
there is in Rize also it must be (certainly)
=> here we go, so must it also be in Rize, certainly

(17) yani ben sevemedim {*işin aslı*} (M, AHT28)
in other words me I couldn't love (in reality)

(18) yine aynı şeyleri yaşayabiliyorsunuz bu da bir imkândır {*aslında*} (TY, AHT36)
again same the things you can live that also a possibility (-dir : to be) in reality
=> you will live the same things again, that's also a possibility (in reality)

(19) ve ben benim de hoşuma gidiyor {*doğrusu*} (M, AHT20)
and me my too my pleasure it is going truly
=> and me, even to me, it gave me pleasure, truly

(20) ama ben bunun faydalarını gördüm {*hak(i)katten*} (M, AHT20)
but me to that these profits I have seen (really)
=> but me, I have seen the profits gained by it, really

(21) bizim Ayşe de çok seviyormuş biliyo(r)sunuzdur {*belki*} (M, AH29)
our Ayşe too more she loves you have to know (perhaps)
=> apparently our (friend) Ayşe likes that too, you may know, perhaps

4.5. VP with an appreciative modality

In other contexts, the postrheme may also represent the appreciative modality with qualifiers such as "büyük" (big), "önemli" (important), "ilginç" (interesting), etc. Utterance (22) is interesting as it ends with a postrheme and presents the particularity of having two different modalities. In fact, the postrheme comprises two modalities among which the first—"gerçekten" (really, truly)—has an epistemic value and the second—"enteresan"[15] (interesting)—has an appreciative value.

Enunciative function (emphasizing the speaker's subjective appreciation of the utterance's content): In this case, the speaker highlights the qualitative aspect of his utterance in order to draw the listener's attention to the appreciation that he has just expressed. In examples (22–24), the postrheme emphasizes the appreciative value of the speaker's VP. This type of postrheme is generally composed of two words, the first one being an adverb; the second, an adjective (except in the last example):

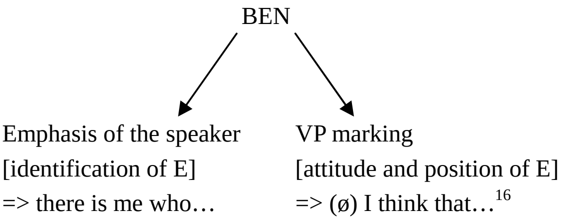
(22) belki çok mükemmel olaylara şahit oldu belki ihanetlere şahit oldu
perhaps many extraordinary to events witness he has been perhaps to treasons witness he has been
{*gerçekten enteresan*} (TY, AHT37)
really interesting
=> perhaps, he has witnessed many extraordinary events, perhaps he has witnessed treasons (really interesting)

(23) *ben* de sevmem <{*çok fazla*}> (M, DH28)
me too I don't like much more
=> I don't like it (so much) either when...

(24) ama *ben* çok ee sevemedim {*onu da maalesef ki*} (M, AHT30)
but me more I couldn't like that also unfortunately
=> but me, I couldn't like more (and that's unfortunate)

Conclusion

The following two operations are undertaken using "ben(ce)" as the VP marker: on the one hand the speaker considered as a subject speaking at the time of enunciation, and on the other hand, the position of the speaker concerning what he states facing the listener. Let us schematize this dual function of "ben."


BEN

Emphasis of the speaker
[identification of E]
=> there is me who…

VP marking
[attitude and position of E]
=> (ø) I think that…[16]

Figure 12.

[15] Word borrowed from French for which the Turkish equivalent is "ilginç" (interesting).

[16] In French, this syntactic structure is rendered more explicit with the use of two successive personal pronouns "moi je...," and it is rendered more explicit in Turkish, firstly, with a facultative personal pronoun and, finally, with a personal suffix "(ben) ... V + pers."

The VP marker “ben(ce)” may be followed by a nominal or verbal element, and the VP will be centered on this element.

BEN	
Ben + N	*Ben + V*
VP on the object of the discourse: consensus (H+)	VP on the action: consensus (H+)
VP on the object of the discourse: discordance (H–)	VP on the action: discordance (H–)

Figure 13.

The following figure recapitulates the values and functions of VP according to its syntactic position:

Two types of relationships of the VP marker BEN(CE)			
1. Syntactic and semantic relationship		2. Enunciative relationship	
Morphosyntax		*İntonation*	
=> morphosyntactic structure of an utterance		=> intonative structure of an utterance	
a) syntactic value	b) semantic value	a) co-enunciation	b) co-locution
=> Relationship of “ben(ce)” with the other constituents of an utterance		=> Relationship between the speaker and the listener (enunciator/co-enunciator/co-locutor)	

Figure 14.

The rheme is associated with assumption and assertion and is therefore endowed with a supplementary modality, such as a subjective modality with “ben” or “bence” in the postrheme, an appreciative modality with “iyi, güzel” (well, good), and an epistemic modality with “gerçekten” (really), “kesinlikle” (certainly), and “belki” (perhaps).

Through an analysis of the different corpora, the following points can be highlighted:

a) The most frequent modality expressed in the postrheme is the subjective modality with the VP marker “ben.”

——*Why?* This can be explained based on the fact that the assertive value of the rheme is compatible with the egocentered value of “ben.” Moreover, it can be explained, in particular, by the fact that the speaker-enunciator resorts to “ben” in the postrheme when he feels the need to express his personal VP in his discourse in order to show the listener his egocentered position with regard to his statement.

——*Why place this marker after the rheme and not before it?* This is because if the rheme is already assertive and sufficiently modalized, the speaker will not need any additional modal marker in order to assert the rheme. On the other hand, the positioning of the postrheme after the predicate is the best way to draw the listener’s attention to “ben,” i.e., to the position of speaker and to his utterance. *This confirms the fact that this type of VP expression is indeed a sort of supplementary and/or complementary modality related to the context.*

b) Above all, one must analyze the syntactic, semantic, and modal contents of the thematic and rhematic parts of an utterance in order to be able to understand the value and function of the postrheme. Thereafter, the semantic and modal relationships that exist between these constituents need to be tested.

The expression of the VP takes its root in the entire utterance; this is evident from the fact that “ben” is generally situated at the beginning of the utterance, a position in which the operations of *thematization* and *modalization* are typically introduced. In fact, these operations are made either at the beginning of the theme and the rheme or at the end of the utterance; “ben” is never dissociated from the preceding and following elements by a pause or any other suprasegmental clue.

Syntactic positions of “ben”		
BEN *at the beginning of the utterance* (supporting the theme)	**ben ø** theme + rheme	No break (or mark) between “ben” and the following theme
BEN *in the middle of the utterance* (introducing the rheme)	theme + **ben ø** rheme	No break (or mark) between “ben” and the following rheme
BEN *at the end of the utterance* (postrheme)	theme + rheme **ø ben**	No break between the rheme and postrheme “ben” that follows

Figure 15.

At the conclusion of our analysis of typical oral utterances through several corpora, we observed that in Turkish, there exists a strict relationship between the VP expressed in the utterance and the postrheme, which is frequently used in oral Turkish. Here, we have a *complementary relationship* that is clarified by the value of the *supplementary modality* of the postrheme that is placed at the end of the utterance in order to complete not only the VP of the speaker-enunciator but also the modal position adopted by him in relation to the listener.

Moreover, this modal operation is like an enunciative process that comprises *bringing into action the VP markers* in three distinct positions. This depends on the discursive strategy of the speaker-enunciator with regard to his utterance and in relation to the listener at the time of speaking.

Position of “Ben(ce)”	Value/Function
Thematic VP	Doubting the VP tendency toward a consensus (H+)
Rhematic VP	Assertive VP + assumption convergence (H+)/discordance(H–)
Postrhematic VP	Conclusive VP + anaphoric value Egocentered position (H–)

Figure 16.

In conclusion, the following statistics present the results (Fig. 17) based on the five corpora (2.5 hours) that display the frequency with which the VP markers are used:

PDV	*in theme*		*in rheme*		*in postrheme*		TOTAL	
Ben	71	42%	29	17%	14	8%	*114*	*67%*
Bence	8	5%	14	8%	8	5%	*30*	*18%*
Benim	21	12%	4	2.3%	3	1.8%	*28*	*16.1%*
TOTAL	*100*	*59%*	*47*	*27.3%*	*25*	*14.8%*	*172*	

Figure 17.

Based on the number and the position of the VP markers used in our corpora, we conclude the following: (a) The VP marker that is most used is “ben” and (b) the most frequent syntactic position for the VP is the thematic part of the utterance. Thus, these conclusions from our corpora provide us with a clear understanding of constructions in oral Turkish, even if these numbers (and the frequency of use of the VP) vary from one corpus to another. However, all that remains with regard to future research is to analyze the VP in other types of corpora in order to enable a generalization of our results, integrating the intonative pattern of these markers in diverse contexts.

Abbreviations

acc: accusative, dat: dative, E: enunciator, H: intonative height, Lig: ligateur, O: object, OD: object of discourse, VP: viewpoint, poss: possessive, pers: person, pred: predicate, SP: singular person, S: subject, V: verb.

Bibliography

Blanche-Benveniste, Cl. 1997. *Approche de la langue parlée en français*. Paris: Ophrys.

Danon-Boileau, L. 1994. “La personne comme indice de modalité” in *La Personne*. *Faits de Langues 3*. Paris: PUF, 159–167.

Ducrot, O. & J.-M. Schaeffer. 1995. *Nouveau dictionnaire encyclopédique des sciences du langage*. Paris: Editions du Seuil.

Guentchéva, Z. et alii. (1994) “Interactions entre le médiatif et la personne” in *La Personne*. *Faits de Langues 3*. Paris: Ophrys, 139-148.

Hagège, Cl. 1982. *La Structure des Langues*. Que sais-je? Paris: PUF.

Kerbrat-Orecchioni, C. 1999. *L’Enonciation. De la subjectivité dans le langage*. Paris: Collection U-Linguistique. Armand Colin. 4th edition.

Le Querler, N. 1996. *Typologie des modalités*. Presses Universitaires de Caen.

Morel, M.-A. 1994. “Les pronoms en français oral” in *La Personne*. *Faits de Langues 3*. Paris: PUF, 169–173.

Morel, M.-A. & Danon-Boileau, L. 1998. *La Grammaire de l’intonation*. L’exemple du français. Paris: FDL-Ophrys.

Morel, M.-A. 1998. “Analyse de la structure de l’oral” in *Copy of seminar from DEA/Doctorat*. University of Paris III – Sorbonne Nouvelle, Center of French Linguistic, EA 1483.

Rossi, M. 1999. *L’intonation. Le système du français: Description et modélisation*. Paris: Ophrys.

Sarıca, M. (ed). 2005. *Sözlü dil yapısı* (The structure of spoken language). Yeni dilbilim kuramları ışığında. Istanbul: Multilingual.

Uras Yılmaz, A., S. Yılmaz, & M.-A. Morel. (eds). 2004. *Vers une grammaire linguistique du turc*. Istanbul: Multilingual.

Yılmaz, S. 2001. *Le système hypothétique en turc: De la morpho-sytaxe à l’énonciation*. Ph. D. (2000), University of Paris III – Sorbonne Nouvelle, Lille: Septentrion (ANRT).

————. 2002. “Analyse des constituants post-rhématiques en français et en turc” in *L’oral d’ici et d’ailleurs 2002–2003*, Recherches actuelles sur l’oral spontané, EA 1483, Presses Universitaires de Paris III, 2002–2003, 99–108.

————. 2005. “Le pronom personnel comme marqueur de point de vue dans le dialogue oral” (L’exemple de “moi” en français et de “ben(ce)” en turc), in *Dilbilim Günleri / Linguistic Days*, Istanbul University, Faculty of Letters, May 12–13, 2005.

Nonreferential Use of Demonstrative Pronouns in Colloquial Malay

Isamu SHOHO

1. Introduction

We can classify the uses of Malay demonstrative pronouns into three categories based on the type of object they refer to: (1) objects in the real world, (2) what has been said before or conversely, what will be said later, and (3) mental images. The first use is restricted to spoken Malay, whereas the other two appear in both written and spoken Malay. In addition to these uses, we can discern another use of demonstrative pronouns that is found only in spoken Malay—the nonreferential use in which demonstrative pronouns have, wholly or partially, lost their function of indicating an entity. Instead, they have gained a new function of expressing the feelings that accompany speech. To support this contention, we will explore examples in the Corpus of Colloquial Malay (CCM) for which data were collected at the end of 2003 in cooperation with the UKM (Malaysia National University). The main purpose of this paper is to show the type of feeling that can be expressed by Malay demonstrative pronouns in nonreferential use.

2. Reference to objects in the real world

In Malay, we find two demonstrative pronouns, i.e., *ini* (this) and *itu* (that). *Ini* is used when the object or person in question is near the speaker, while *itu* is used when the object or person in question is far from the speaker. In colloquial Malay, *ini* and *itu* are shortened to *ni* and *tu*, respectively. In sentence (1), *ni* is used to refer to the six kittens encircling the speaker since the six kittens are near him.

1) *Kok setakat kucing baik takyah cakap. Buang masa aku aje.*
Alaa…kembar jugak ni tok. Hi, hi, hi!
(I don't want to listen to your story if it's simply about cats. It's just a waste of my time.
Why, another pair of twins here, Uncle. Ha, ha, ha!)
(Ujang 8/15/2004, p. 12)

On the other hand, *tu* in sentence (2) refers to the baby that is a little far from its mother.

2) *Alaaa…leceh betul la! Leceh! Leceh! Tu anak abang jugak la!*

(Come on…It bothers me so much! That child is also yours!)

(Ujang 8/15/2004, p. 16)

In the following sentence as well, we can see the difference between the use of *ni* and *tu*.

3) *Rangka basikal tu, tuk wan nak pakai lagi?*
Tak nak, tuk wan nak buang ni!
(Are you going to use that bicycle frame, Granpa Wan?
No. I'm going to throw it away!)

(Gila-Gila 5/1/1997, p. 6)

In this sentence, *tu* is used after the word *basikal* (bicycle) because the bicycle is located at a distance from the speaker (Wan's grandchild). On the contrary, Granpa Wan uses *ni* to refer to the same bicycle because it is near him.

The following example that is quoted from the CCM can also ascertain the above mentioned distinction between *ni* and *tu*.

4) A6: /// *[a] apa jenis ### jenis jenayah tu?*
B7: *[oh] ni, macam ni ### penyeluk saku, perompak* ///.
(What kind of crime is that?
Oh, are you asking about this article? It is about cases of pickpocketing and burglary.)

(CCM: B2 [30Nov#7; Jenayah; separuh bebas bersemuka])

In sentence (4), speaker A6 asks speaker B7 about the article that he/she is reading at the time. The newspaper carrying the article is at a distance from speaker A6; therefore, A6 uses *tu* when referring to the newspaper. On the other hand, the same object, i.e., the newspaper, is near B7; hence, B7 uses *ni* while referring to it. In this situation, the absolute distance between the speaker and the object in question is irrelevant with reference to the choice of demonstrative pronouns. The relative distance between the object in question and the participants in the conversation is important to determine the demonstrative pronouns that should be used. We have to determine which participant is more near to or far from the object in question. We can imagine a situation in which a newspaper is located only half a meter away from one participant, but its relative distance from the other participant is less than its distance from the former participant, the former chooses *tu* over *ni*. In the situation in which sentence (4) is uttered, the newspaper is relatively near participant B7, irrespective of its absolute distance from participant A6. In this situation, the former participant uses *tu* when referring to the newspaper that is located relatively near the other participant who uses *ni* to refer to the same newspaper.

In the above cases, the objects and persons are concrete in that they have a certain form and are tangible. In addition to these cases, both demonstrative pronouns can be used to refer to intangible things like events, occurrences, or conversations.

5) *Oii! Ini rupanya kerja enko Cubin…aku dah tau.*
(Hey! This may be your work, Cubin…I know you did it.)
(Ujang 8/15/2004, p. 13)

6) *Ha…tu kawan abang dah datang…*
(Look. Your friend is coming…)
(Gila-Gila 5/1/1997, p. 6)

In sentence (5), *ini* refers to a situation that appears to be created by Cubin. In sentence (6), *tu* refers to the fact that the speaker's husband's friend is approaching them.

3. Reference to what has been said before or what will be said later

In addition to these uses, *ini* and *itu* can refer to the meaning in the context. To be more exact, *ini* can refer to what will be said, while *itu* can refer to what has been said as shown in examples (7) and (8).

7) *Tolong la aku Udin. Aku takut ni. Ada orang nak jumpa aku la.*
(Help me, Udin. I'm scared. There's a person who wants to see me.)
(Ujang 8/15/2004, p. 5)

8) *Tapi masalahnya dia bawak geng sorang, tu yang aku takut tu.*
(But the problem is he came with a gangster. That's why I was scared.)
(Ujang 8/15/2004, p. 5)

In sentence (7), *ni* refers, in advance, to what will be said by the speaker, i.e., *Ada orang nak jumpa aku la* (there is someone who wants to see me). In sentence (8), *tu* refers to what the speaker said in the preceding context, i.e., *dia bawak geng sorang* (he brought a gangster along).

We find another use of *tu* in the following sentence wherein the same demonstrative pronoun can refer to what has been said by the addressee, not the speaker.

9) *Eh, macam bagus aje bunyinya tu…tapi kita nak bisnis apa ha?*
(Oh, that sounds interesting…But what business are we going to do?)
(Gila-Gila 5/15/1997, p. 25)

Unlike *tu*, *ni* cannot be used to refer to what has been said by the hearer in the preceding context, and here lies the difference between the two demonstrative pronouns.

(*I*)*ni* and (*i*)*tu* can also be used to refer to what has been understood by the speaker and the hearer, that is, what has attained the status of the common topic of conversation. Sentences (9) and (10) provide such examples.

10) *Kepala hotak engkau Bob, mana boleh buat bisnes macam tu kat hostel, itu kan illegal!*

(You fool, Bob. How can we do such business in the hostel? That's illegal!)

11) *Kita buat bisnes ni masa tengah malam aje, mesti sukseslah!*
(We'll do this business only late at night. I'm sure we'll succeed!)

In sentence (10), *tu* is used to refer to what has been proposed by the hearer, while *itu* is considered to refer to the common topic, i.e., to carry out business in the hostel. In sentence (11), *bisnis ni* (this business) refers to the image of the business that both are focusing on.

In sentence (12), which is cited from the CCM, we can find a division between the referring functions of *ni* and *tu*. In this case, *tu* (in A12) refers to what has been said by the hearer, while *ni* (in B13) refers to what has been selected as the main topic and now dominates the speaker's attention. In the preceding context, in sentence B8, *kes-kes jenayah macam ni* is used. It is clear that *kes-kes jenayah macam ni* and *jenayah seluk saku ni* refer to the same thing; however, the latter (in sentence B13) does not refer back to the former. *Ni* is used in each sentence to refer to the element that has gained the status of the main topic and that now dominates the speaker's consciousness.

12) A12: *Kalau kau tahu, apa jenayah seluk saku tu?*
B13: *Jenayah seluk saku ni biasanya berlaku di tempat-tempat awam, tempat orang ramai, tumpuan orang ramai ///.*
(CCM: B2 [30Nov#7; Jenayah; separuh bebas bersemuka])

13) B8: *Tapi kes-kes jenayah macam ni tak boleh pandang ringan tau.*
(CCM: B2 [30Nov#7; Jenayah; separuh bebas bersemuka])

In sentence 12, *Seluk saku tu* (spoken by A12) is translated as "such cases of pickpocketing as you have said." In sentence 12, *Jenayah seluk saku ni* (spoken by B13) is translated as "such cases of pickpocketing as I have picked up as the main topic."

At this point, we should keep in mind the fact that the speaker who does not pick up the conversation topic continues to use *tu* to refer to the topic introduced by the other party. It is only at some later stage after both parties can be considered to have shared some conversation topic that the speaker is entitled to use *ni*, as shown in sentence (13).

14) A18: *Tapi [tu la] tapi tak boleh pandang ringan juga [a] jenayah ni [dak].*
(Indeed what you've said is right, but we shouldn't make light of this kind of crime.)
(CCM: B2 [30Nov#7; Jenayah; separuh bebas bersemuka])

In sentence (14), speaker A18 begins using *ni*, which implies that pickpocketing has become the common conversation topic between both parties. In other words, this speaker has reached the stage where he/she has shared the common topic with the other person.

4. Reference to a mental image

Both demonstrative pronouns can be used to refer to a mental image stored in the speaker's memory. The only necessary condition for making the use of these images possible is that the speaker should have a certain image in his/her mind that can be referred to by the demonstrative pronouns. It is not always necessary for the hearer to share the image. In case both participants share the image, it is regarded as common knowledge between them, and the conversation begins around this image.

15) *Habis le kita bang. Kalau tunang anak kita tu tahu…tentu disamannya kita!*
Tak apa…klon dia kan ada?
(That's that, Darling. If the fiancée of our daughter comes to know this…he'll sure fine us!
Don't worry…She has her clone, doesn't she?)
(Gila-Gila 5/15/1997, p. 13)

The referent of *tunang anak kita tu* (the fiancée of our daughter) in sentence (15) has not been mentioned in the preceding context. It is mentioned for the first time in this sentence. In this case, reference is made not to what was said before but directly to the mental image of the daughter's fiancée that is stored in the wife's mind. This mental image is shared by her husband. Therefore, between the wife and her husband, there is no doubt about the identity of the fiancée. It can be said that they have common knowledge about the fiancée who is discussed in the conversation. This explains the use of *tu* in this sentence.

In the following sentence, *tu* is used in the same manner as in sentence (15), the only difference being that the personal pronoun, and not the proper noun, is placed before *tu*. In this case as well, *tu* refers to the mental image of a person expressed by the personal pronoun *dia*.

16) *Hebat sangatkah dia tu? Rupanya dia tidak macam Si Joker atau Si Penguin.*
(Is he so great? It seems he's not like those guys, like Joker or Penguin.)
(Gelihati 305 Mei 1997, p. 6)

All the examples we have considered thus far deal with the function of referring to the mental image, which can be considered as having attained the common topic status. However, this type of use is not always based on the firm conviction that the other person and the speaker have common knowledge about a topic.

The following example is one such case. Speaker (B31) in sentence 17 uses *tu* based on the vague assumption that the other person may be aware of the criminal case that occurred in Kuala Lumpur. This does not require the speaker to be confident of the possibility of sharing his/her knowledge.

17) A29: *Ada anak yang kayapun boleh juga ///.*
B30: *Memang [a], penahkan ada satu kes tu.*

B31: *Kes yang berlaku kat KL tu.*

In the following sentence, both persons use *tu* for referring to a specific Saturday in the past. In this case, both persons succeed in recollecting that specific Saturday in the past that was stored in their memory.

18) B17: *Aku call kau tak dapat-dapat hari Sabtu tu kenapa?*
A18: *Hari Sabtu tu, ye ke, tak dapat ek?*
(CCM: G1 [29Nov#2; Bandar dan kampung; bebas bersemuka])

5. Nonreferential use of demonstrative pronouns

In the foregoing chapters, we have classified the Malay demonstrative pronouns into three categories based on the type of entity referred to, i.e., on-the-spot use, contextual use, and conceptual use. With regard to contextual and conceptual uses, we can find examples not only in written Malay but also in colloquial Malay. Examples of on-the-spot use can only be found in colloquial Malay, which can be naturally predicted from its nature. On-the-spot use is valid only when the speaker and the hearer are present in the real world because this type of use refers to objects in the real world. In addition to on-the-spot use, difference between written and spoken Malay can also be found in another kind of reference, i.e., nonreferential use, which will be discussed in this chapter. In nonreferential use, demonstrative pronouns can be considered to have lost the function of referring to objects and to have instead gained the new function of expressing the speaker's feeling or the mental state accompanying his/her speech. In the remainder of this chapter, we will deal with the nonreferential use of demonstrative pronouns and will investigate the type of feelings or mental states expressed by them.

5.1. Making a vague reference to the situation

Tu and *ni* in sentences (19) and (20) given below cannot be said to refer to any specific thing. If we delete these demonstrative pronouns, it does not lead to any meaningful change in the interpretation of these sentences.

19) *Faris, nak ke mana tu?*
(Faris, where are you going?)
(Menggapai Kasih Ayah, p. 3)

20) *Kamu nak ke mana ni?*
(Where are you going?)
(Menggapai Kasih Ayah, p. 3)

Based on this fact, it can be said that these demonstrative pronouns do not constitute an indispensable part of these sentences. They just provide a basis for drawing the conclusion that Faris appears to have gone out. These demonstrative pronouns are not involved in direct reference, but they make a vague reference to the situation that enables Mak Eton (the maid who attends to Faris's family) to infer that Faris is going out. This situation includes the manner in which he dresses himself, his restless attitude, the hasty manner in which he wears his shoes, and the stealthy manner in which he is ready to leave the room. Sentences (19) and (20) appear on the same page and there does not seem to be any difference in their cognitive sense. The only difference lies in the distance from the speaker. Sentence (19) is uttered by the maid when she was in the kitchen and at a time when Faris who was standing in the porch is ready to leave in a stealthy manner. Thus, there is some distance between Eton and Faris. On the contrary, sentence (20) is uttered when Faris is told to sit for lunch and he turns back to stand near Eton. In this situation, he stands face to face with Eton. Interestingly, the use of *ni* and *tu* in this case that makes a vague reference can be said to reflect their use in the real world.

In the following sentence, *ini* appears on two occasions. When it first appears, it refers to the scar on the speaker's foot. Here, the use of the demonstrative pronoun *ini* is the same as that considered in chapter 2. With regard to the second appearance of *ni,* it is clear that the demonstrative pronoun *ni* does not refer to anything specific. It is used to direct the hearer's attention to the present situation, which safely justifies his speech. An attempt to paraphrase *ni* leads to a rough translation as follows: "as you can be assured of my word if you see the present situation."

21) B838: *Dia kata ini nanah ni.*

(He said, "The wound suppurates as you can be assured of my word if you see this terrible condition.")

(CCM: F1 [29Nov#1; Hantu; bebas bersemuka])

It should be noted that this example uses two different forms of the proximate demonstrative pronoun, i.e., the unabridged form of *ini* and the shortened form of *ni*. This is done to show that there are two different uses of this pronoun.

5.2. Expressing irritation, fretfulness, and impatience

In the following sentences, we find another type of nonreferential use where the demonstrative pronouns abort the referring function. These pronouns are used to express accompanied feelings of irritation, fretfulness, and impatience.

22) *Engko nak apa gemok? Aku nak cepat ni.*

(What do you want, fatty? I'm in a hurry.)

(Ujang 8/15/2004, p. 23)

23) *Anak nak susu tu bang, pegi la buat cepat. Saya banyak keja ni.*

(The child needs milk, darling. Go and make some milk quickly. I've got a lot of work to do.)

(Ujang 8/15/2004, p. 13)

24) *Cuba bukak mata tu besar-besar sikit. Nampak tak?*
(Try to open those eyes little more wide. Can you see this time?)

(Gelagat 5/1/1997, p. 47)

In sentence (22), the person who is referred to as *aku* (I) was asked to stop his motorcycle by a gross man. In response, the former remarked irritatingly that he was in a hurry and asked the latter to make way for him. In sentence (23), the wife is occupied with a lot of work and impatiently offers this as the reason why she is unable to take care of her child.

The function of *tu* (*anak nak susu tu bang*) is the same as that in the preceding chapter, i.e., when making a vague reference to the situation. In sentence (24), on being repeatedly asked if he could see anything, the speaker answers in the negative. This made the listener lose her temper, and she impatiently asked him to open his eyes in order to see better. Usually, when referring to another person's eyes, we use *matamu* (your eyes); however, in this situation, the latter uses *tu* instead of *mu* (your), which expresses her feeling of irritation.

In sentence A58 (in 25), *ni* does not function as a demonstrative pronoun but expresses speaker A's fretfulness and aims to urge speaker B to begin narrating a story without wasting any time.

25) B57: *[ooo], aku /// aku /// cerita la sikit kan nak dengar cerita [aaa].*
(I'll tell a story. Do you want to hear it, don't you?)
A58: *[aaa], aku nak dengar ni, cerita la!*
(I'm eager to hear it. Tell it now.)

(CCM: A1 [30Nov#8; Bandar dan kampung; bebas bersemuka])

In sentence (26), the speaker is disgusted with the other person's clamorous and insistent demand regarding his mother's age when she was in love with his father.

26) A609: *Macam mana ni ha…*
(How can I answer correctly?)

(CCM: D2 [30Nov#2; Keluarga; separuh bebas melalui telefon])

5.3. Expressing dissatisfaction and disagreement

As shown in sentence (27), *ni* is used to express the speaker's disagreement with the old man's remark that their conversation topic is not about kittens but about human babies and that it is useless to talk about kittens. Protesting against the old man's condemnation, he contends that the same can be applied to not only human beings but also kittens.

27) *Kok setakat kucing baik takyah cakap. Buang masa aku aje.*
Alaa…kembar jugak ni tok. Hi hi hi!

(I don't want to listen to your story if it's simply about cats. It just wastes my time. Why, another pair of twins here, Uncle. Ha, ha, ha!)

(Ujang 8/15/2004, p. 13)

With regard to sentence (28), speaker A242 asked the other party to come to his village, which is five hours away from Melaka. The latter requested the former to drive them to the village in his chauffeur-driven car. However, the speaker informs them that he does not have a chauffeur because his father did not like to employ one. Instead, his father prefers driving his car himself.

28) A242: *[Hei...]! Sorry [aih…] aku bukan ada pemandu…ayah aku tak suka ada pembantu ni…ada pemandu…ayah aku suka bawa kereta.*

(I'm afraid we don't have a chauffeur because my father doesn't like to employ one. My father would prefer driving his car himself.)

(CCM: D2 [30Nov#2; Keluarga; separuh bebas melalui telefon])

5.4. Expressing disgust

Sentence (29) is uttered by a mother who is displeased with her child who is absorbed in playing games. The child is so engrossed in his games that he is unwilling to help his mother; this attitude displeases his mother, and she utters sentence (28).

29) *Kau ni main je! Pegi kedai kejap!*

(You do nothing but play! Go to the store now!)

(Ujang 8/15/2004, p. 13)

The following example is about childhood memories recalled by the other person who used to play naked in the water ditch. On hearing this, speaker A69 is disgusted with the the attitude of that person, who is sufficiently shameless about describing their nakedness when playing with water. The two appearances of *ni* have the same function of expressing speaker A's disgust at the other person's brazenness.

30) A69: *[haha], oo kau ni melucukan la.*

(Really? Your story is too funny to believe.)

A70: *Apa la kau ni, tak malu betul.*

(I'm ashamed of you! You are so shameless!)

(CCM: A1 [30Nov#8; Bandar dan kampung; bebas bersemuka])

In the following sentence, *ni* is used to express the other person's boldness to wear a skirt even though in Malaysia, an Islamic state, not many girls wear skirts designed in the western style.

31) B542: *Ko nie memang, suke pakai macam tu.*

(You usually like to wear a western style skirt.You are so bold to do it, I think.)

(CCM: E2 [25Nov#4; Syarikat Jepun; bebas bersemuka])

Sentence (32) was jokingly said by speaker A13, who was surprised at the other person's forgetfulness.

32) A13: *Kau ni kita baru kenal, kau dah tak ingat lak aku, alamak [ai…].*
(I'm surprised you are so forgetful. We've just known each other, but now you have forgotten me and cannot recognize me. What a man!)
(CCM: D2 [30Nov#2; Keluarga; separuh bebas melalui telefon])

Sometimes *ni* is used after the second person with function to express intimacy. The following sentence is an example.

33) *Jadi kau ni, Hamdan! Anak Tok Mudim Jamin? Semakin erat tubuh kerdil Hamdan didakapnya.*
(So, you are Hamdan, really? The son of Jamin the circumciser? He embraced the small body of Hamdan more and more tightly.)

In sentence (34), we find *tu* used after *ko* (in the formal style: *kau*). *Tu* expresses speaker A's displeasure at the other person's provocative remark that is intended as a joke about speaker A's inappropriate way of wearing a skirt.

34) B544: *Alah, bukan nampak comel, nampak mengade-ngade.*
A545: *Ko tu, mesti ko jelous, takpe-takpe.*
(Why, you don't look lovely with your skirt, you look saucy.
You rat! You must be jealous of me, but I don't care.)
(CCM: E2 [25Nov#4; Syarikat Jepun; bebas bersemuka])

In contrast with the use of *ni* (in sentence 32), which is intended as a joke, the use of *tu* in sentence (34) is intended for rebuke or condemnation. The following sentence shows another canonical use of *tu*.

35) *Kau tu! Baik-baik sikit cakap.*
(You unashamed fellow! Beware of what you say, I'm warning you!)

This sentence is uttered when the speaker is offended by the other person's words and warns that person to be careful about what he/she says. The speaker intends to rebuke the offending words spoken by the other person and hints at retaliating against it. We can see the same difference between *ni* and *tu* in the following two sentences.

36) *Dia ni, orang cakap tak mahu dengarlah.*
(When I speak, that chap doesn't lend an ear to what I say. That's the trouble with him.)

37) *Dia tu, orang cakap tak mahu dengar langsung.*
(That stuck-up creature pays no regard to what I say.)

In sentence (36), the speaker considers the other person's attitude of paying little attention to what he/she says to be troublesome. However, there is no indication of rebuke for the other person. In contrast to sentence (36), the speaker of sentence (37) rebukes the interlocutor for his bad manners.

5.5. Expressing resistance and defiance

Another use of *ni*—to express resistance or defiance—can be found in sentence (38). The boy is told to fetch a memo pad and a pencil to write down the items required by his mother who instructs him to do so because he is prone to forget the things he is asked to buy. However, he does not want to obey his mother and says that it is troublesome to write down each item to be purchased. He believes that he has a good memory.

38) *Ambik kertas ngan pensel. Tulis, kang lupa plak!*
Lecehlah tulis-tulis ni. Mak cakap je. Orang ingat punya!
(Take paper and a pencil, write it down in case you forget it!
It's troublesome to write this and that. Just you name it, and I'll remember it.)

5.6. Disparaging others or oneself

In sentence (39), the speaker is angry with himself and his friends for being unable to think of exterminating annoying flies with a broom. Disparaging himself, he remarks about how foolish they were not to think of it.

39) *Macam mana la kami tak terfikir yang lidi penyapu boleh membunuh lalat! Bodoh betul kami ni ye?*
(How come we didn't come up with the idea that a broomstick can kill flies! We're really stupid, don't you think?)

Sentence (40) is uttered by speaker B338, who is ashamed of his bad memory that is kiddingly attributed to his senility even though he is not actually senile.

40) B338: *Eh, pandai je umur ayah aku empat puluh lapan, aku ni nyanyuk pulak…tak silap aku umur ayah aku empat puluh enam, mak aku ///.*
(How smart you are! My father is forty-eight years old. Oops, I'm in my dotage. If my memory is correct, my father is forty-six years old and my mother is...)
(CCM: D2 [30Nov#2; Keluarga; separuh bebas melalui telefon])

In contrast to sentence (40), the use of *tu* in sentence (41) disparages the interlocutor's poorness and wretchedness.

41) *Wragh…kah! Kah! Siapa la yang nak kat you…miskin dan hina tu!*
Kah! Kah! Kah!
(Har, har, hee-haw! Who do you think will approach you, you poor and wretched fellow?)
(Gila-Gila 5/15/1997, p. 16)

5.7. Boasting about oneself or expressing confidence

In contrast to the use considered in chapter 5.6., *ini* is also used for boasting about oneself or expressing confidence. Sentence (42) exemplifies this function. This sentence is uttered by an aerobics instructor who is confident of her slim and appealing body. Her pride in her ideal body is expressed in her words *tubuh yang ramping dan menawan macam saya ini* (slim and attractive body like mine).

42) *Saya jamin selepas dua bulan puan-puan akan memiliki tubuh yang ramping dan menawan macam saya ini…bla…bla…bla…bley…*
(I'm quite sure that after two months of training, you will find yourself fit and appealing just like myself.)
(Gelagat 5/1/1997, p. 27)

Sentence (43) is uttered by a girl who is inwardly confident of her feminine charm, which can certainly fascinate her friend's elder brother.

43) *B258: Aku ni tak ada ciri-ciri wanita istimewa abang ko ke…yang dia minat?*
(How about me? Don't you think I have feminine characteristics which attract your elder brother?)
(CCM: D2 [30Nov#2; Keluarga; separuh bebas melalui telefon])

In the following sentence, by using *ni* after *rumah aku*, the sentence is interpreted as if speaker B224 is boasting about having the latest phone model in the house.

44) B224: *[He eleh] pegang telefon pun boleh lenguh la…lain kali guna telefon yang canggih sikit, macam rumah aku ni, aku guna telefon canggih ma…boleh jalan-jalan sambil buat kerja lagi.*
(Only holding the cellular phone makes your hand numb, doesn't it? Better you buy a new model of phone like the one I have in my house. I use the latest phone model with which I can talk on the cellular phone hands free while walking or doing work.)
(CCM: D2 [30Nov#2; Keluarga; separuh bebas melalui telefon])

6. Addition of *ni* and *tu* to obtain a rhyming effect

In spoken Malay, we find many cases in which sentences beginning with *ni* or *tu* end with the same demonstrative pronouns; the sentence begins and ends with *ni*. In the same way, a sentence that begins with *tu* also ends with *tu*. The appearance of the same demonstrative pronoun at both ends of a sentence contributes to a rhyming effect. The following sentences are such examples.

45) *Eh…radio. Wuih…ni model klasik ni!*
(A radio, isn't it? A classic model besides!)
(Gelagat 5/1/1997, p. 32)

46) *Eh…Emy anak sapa kau bawak balik ni?*
Ni yang aku nak cerita ni
(Em, whose child are you carrying?
For that matter, I'll tell you later.)

(Gelagat 5/1/1997, p. 79)

47) A5: *[Aaa…] Ni sapa ni?*
(Er, who is speaking?)
(CCM: D2 [30Nov#2; Keluarga; separuh bebas melalui telefon])

48)=8) *Tapi masalahnya dia bawak geng sorang, tu yang aku takut tu.*
(But the problem is that he came with a gangster. That's why I was scared.)
(Ujang 8/15/2004, p. 5)

In the four examples above, the first of a pair of demonstrative pronouns is used in the usual function, i.e., the function of referring to what has been said (48), what is in the situation (45), what is caught by sensory organs (47), and a matter (46).

Apart from these cases, there are other cases in which both of a pair of demonstrative pronouns are void of a referring function. The following sentence is an example.

In sentence (49), speaker A527 is tired of waiting in vain for a bus for a long time.

49) A527: *Nie dah petang dah nie.*
(It has already become dark, you know.)
(CCM: E2 [25Nov#4; Syarikat Jepun; bebas bersemuka])

In this example, both the appearances of *nie* have nothing to do with the referring function. Sentence (49) is also an example of adding the same word to produce a rhyming effect. Another *dah* is added after *dah petang* to produce an effect of euphony.

7. Concluding remarks

In this chapter, we will consider the relation between the referential use of Malay demonstrative pronouns and their nonreferential counterparts. First, we will compile all the uses we have considered in the preceding chapters and arrange them according to the use of the demonstrative pronouns. The following chart is obtained after arranging the data. In this chart, a cross implies that the pronoun "cannot be used," while a circle implies that it "can be used."

Demonstrative pronouns Function	*ini/ni*	*itu/tu*
1. Outer-lingistic reference	○	○
2. Inner-linguistic reference	○	○
3. Referring to mental images	×	○
4. Making a vague reference to the situation	○	○
5. Expressing irritation, fretfulness, and impatience	○	○
6. Expressing dissatisfaction and disagreement	○	×
7. Expressing disgust	○	○
8. Expressing resistance and defiance	○	×
9. Disparaging (oneself and others)	○	○
10. Boasting about oneself or expressing confidence	○	×
11. Obtaining a rhyming effect	○	○

What attracts our attention with regard to nonreferential use (from 4 to 11) is the high frequency of *ini/ni* as compared with that of *itu/tu*. With regard to the uses of 6, 8, and 10, these functions are only relevant to the first person. Based on this, we can infer that there is a correlation between the first person and the use of *ini/ni*. The same inclination can be observed in the referential use of demonstrative pronouns, as presented in the following pair.

50) *Tangan saya ni sakit sejak tiga hari yang lalu.*
(This arm of mine has been aching since three days ago.)

51) *Kenapa tangan awak tu?*
(What has happened to that arm of yours?)

Now let us consider the use of 4. In this function, both of the demonstrative pronouns can be used and the choice of using either one is based on the distance from the speaker to the hearer, who is a part of the situation. Similarly, in their outer-linguistic use also, the choice is based on the proximity of the target to the speaker.

With reference to the use of 5, this function expresses the speaker’s irritation, fretfulness, and impatience. When the first person pronouns appear, *ini/ni* can be used. On the other hand, with regard to a part of the hearer’s body or belongings, *itu/tu* is used after the word that implies “them.”

With regard to the use of 7, this function is relevant to second and third persons, and both demonstrative pronouns can be used. Its choice is based on the degree of disgust. If we want to express pleasantries, jokes, or intimacy in the sentence, *ini/ni* is preferred. On the other hand, *itu/tu* is preferred if we want to laugh or vent our spleen.

With reference to the use of 9, this function is for disparaging oneself and others. When we want to disparage oneself, *ini/ni* is used after the first person; while disparaging others, *itu/tu* is used after the second and third persons.

Bibliography

Abdullah Hassan 2002. *Tatabahasa Melayu: Morfologi dan Sintaksis untuk Guru dan Pelajar*. Pahang Dmak Muraru: PTS Publications & Distributions Sdn. Bhd.

Nik Safiah Karim (et al.) 1993. *Tatabahasa Dewan* (Edisi Baharu). Selangor Darul Ehsan: Dewan Bahasa dan Pustaka.

3.
Linguistic Informatics

Introduction

Yuji KAWAGUCHI *(Center of Excellence (COE) Program Leader)*

In January and October, 2005, a symposium and national conference were held at the TUFS. Round-table discussions with the COE program promoters were initiated at the national conference, and in the symposium, there was an open debate entitled “Is the Integration of Linguistic Theory and Language Education Possible?[1]” which presents the very objective of the UBLI. Important suggestions are derived from this symposium.

First, it is expected to help specialists of computer science in developing language modules that are easy to manipulate and continue; the modules must stir the learner’s motivation and realize more effective language learning. However, without interaction and feedback from the learners’ side, which are crucial for language learning, language modules would be no more than simple drills. Although it might be possible to obtain interaction and feedback from an e-learning system, it may also be important to understand in what respects computer-assisted language learning (CALL) can contribute toward sophisticating the learner’s language ability and in what respects face-to-face language teaching can be complementary with CALL. For specialists of linguistics, it is expected to analyze different linguistic usages and elicit their pedagogical pertinence.

Second, it is necessary that English and Japanese pedagogues in Japan, who have accumulated teaching experiences and skills for a long time, apply and render their intellectual resources and academic results to teaching and learning Asian and other languages. It seems that specialists of general linguistics and linguistic typology can collaborate with them. With the use of multimedia in e-learning, learners now find it easier to grasp the communicative situations and intentions of living languages. However, in order to sophisticate their communicative ability, tasks and task activities should play a central role in class-room language teaching. It is evident that the theory of enunciation and discourse analysis will provide many valuable ideas to conceive effective task and task activities.

Based on the discussions of the above symposium, we realize three significant orientations of linguistic informatics: (1) Corpus-based analysis of linguistic usages, (2) Typological study of different languages, and (3)

[1] For details on the symposium, see *Working Papers in Linguistic Informatics* 9, published in Japanese, 124–139.

Balancing the weight of e-learning with task-based face-to-face teaching. Both 1 and 2 need to be applied to language education ultimately; however, they should also be counterbalanced, i.e., equilibration between individual language-directed corpus analysis and typological or cross-linguistic analysis. These important directions are discussed in the first three contributions of this chapter.

Based on the presumption that, more than universality, the objective of single-language research is to clarify the actual use of the language in question, Susumu Zaima, in his “German Language Research Methodology Based on Language Use —Language Use, Application and Evaluation—,” insists on the following considerations: (a) the necessity for use-based data analysis, (b) the necessity for applicability and other evaluative perspectives, and (c) concrete examples of envisaged German language research based on these considerations. He claims that we will be able to say that single-language research studies of the German language have departed from the simple “intellectual games” and have arrived at a new type of German language research in the form of a “science” after the results of the abovementioned research have been verified based on the evaluative criteria of usefulness in applications for language education, machine translation, etc., and a framework that evaluates each of these research results has been established. However, he adds that this destination is yet quite far.

In “Developing Grammatical Modules Based on Linguistic Typology”, Makoto Minegishi illustrates the objectives and process of developing web-based teaching materials for the grammar of seventeen languages based on linguistic typology. The teaching materials for grammar consist of two components. One is a set of individual grammar courses of seventeen languages. The other one, the development of which we will focus on, is a cross-linguistic grammar course to provide a bird’s-eye view for grammar in general, based on the data abstracted from the former component.

In “Introducing a Task Activity for Less Proficient Learners —Enhancing the Relationship Among Form, Meaning and Use—”, Hideyuki Takashima et al. assume that in Japan, many learners of English lack in opportunities to use the language outside the classroom. Even in the classroom, English is taught based on a structural syllabus, and the learners focus mainly on forms. As a result, they think that the learners have trouble making use of grammaring (Larsen-Freeman, 2003)—the ability to use grammar structures accurately, meaningfully, and appropriately. Therefore, they believe that it is important to provide learners with the opportunities to use the language while focusing both on both form and meaning. As a means to providing this kind of opportunity, they suggest a language activity called Task Activity (TA). They introduce a TA in detail, using an original example

TA that was carried out in a lower secondary Japanese public school. By referring to the observations made in the experiment, they suggest how TAs can be used for learner evaluation and feedback.

On October 4, 2005, a workshop was held the day before the national conference. This workshop entitled “What is Linguistic Informatics—Contributions of Linguistics, Applied Linguistics and Computer Sciences” had an educational objective: nurturing a new type of language education researcher with expertise in the fields of linguistics, language education and computer science, in short, a typical researcher of linguistic informatics. The researcher is expected to remove the barrier that has traditionally stood between linguists and language educators. Several reports were presented by the postgraduate students of TUFS. We could publish here three interesting contributions, all of which were supervised and revised by the members of the UBLI.

In their “The Relationship between VOT in Initial Voiced Plosives and the Phenomenon of Word-Medial Plosives in Nigata and Shikoku,” Mieko Takada and Nobuo Tomimori, analyze the Voice Onset Time of word-initial voiced plosives in the Shikoku area and northeastern Nigata Prefecture (where there is nasalization of voiced plosives and/or voicing of voiceless plosives in the word-medial position). The results exhibit regional and generational differences of these regions and the patterns of correlations between slightly voiced word-initial plosives and word-medial nasalization and/or voicing obstruents. These results, in addition, are compared with the results from Tohoku to Kanto.

In “On the Semantic Structure of English Spatial Particles Involving Metaphors,” Yasutake Ishii and Kiyoko Sohmiya confirm the polysemous nature of English spatial particles and the significance of the roles played by metaphors and metonymies in the semantic extensions of spatial particles. They argue that the lexical meanings of a spatial particle comprise an image schema, metonymies, and conceptual metaphors. They also propose that verb phrases containing spatial particles are interpreted based on several different interpretation patterns.

Norie Yazu and Yuji Kawaguchi present the paper “Language Policy and Language Choice —A Case Study at Canadian Government Institutions—” The objective of this study is to analyze the language choice made by bilingual public servants in Canada and clarify the factors that govern their choice. The data was collected by a questionnaire survey conducted in ten federal government institutions in the National Capital Region. This study focuses on how bilingual anglophones and francophones actually use English and French, the two official languages of Canada that have equal legal status, in a work environment in which their language use is “planned” by language policy.

German Language Research Methodology Based on Language Use — Language Use, Application and Evaluation —

Susumu ZAIMA

1. Introduction

I believe that all public research—not just language research—is a social action, and consequently, it should have some kind of social significance. Moreover, I think that the purpose of single-language research is to analyze how native speakers of a language form and use sentences in the said language and how these sentences are comprehended. This analysis lays greater emphasis on individuality than on the universality of language.[1]

In this paper, I shall consider the German language and remark on (a) the necessity for analysis based on language use data, (b) the necessity for an evaluative perspective on applicability, and (c) concrete analyses of the German language based on these considerations. As single-language research, the purpose of this paper is also to explore the possibilities of preventing German language research, going no further than simple "intellectual

[1] It will be easier to analyze how native speakers of German comprehend German sentences than how these speakers form German sentences. With regard to this point, Susumu Zaima (2004) states the following:

In general, speakers are induced by stimuli in the real world and have desires to express and communicate. They also form expressions and messages based on a system of rules comprised of categories and recognition formats. These expressions and messages are then transferred to a linguistic form able to be physically perceived, which is based on a fixed system of rules; whereupon the recipient, based on this physically perceivable linguistic form, reconstructs the meaning of the sentences intended by the speaker, and then reconstructs the expressions and messages. Historically, within this flow of language use, analysis was conducted with a focus on the generative aspect (namely, the system of rules for linguistically expressing the content recognized by the speaker). However, if we were to emphasize the focus on substantiveness, then rather than this type of generative aspect—since it is possible to observe and test how recipients comprehend existing linguistic expressions in a more precise manner—we should be focusing on the perceptive aspect (namely the system of rules by which the linguistic expressions are comprehended by the recipient) as a subject of analysis in which verification is more feasible.

For example, the question of how sentences with separable prefix verbs are comprehended by native speakers of German is a subject of profound interest.

games."

2. The Necessity of Analyzing Data on Linguistic Performance

First, I raise the point that if we attempt to clarify precisely how native speakers of German form and use German sentences and how these are comprehended (hereinafter, these actions will be simply referred to as "the language use of native German speakers"), we would require to use extensive language use data based on a corpus.

In single-language research—which analyzes the systems of rules for sentence formation—the language is tested from morphologic, syntactic, and semantic perspectives to extract "properties," "categories," "regularities," and "rules" that are composed of the properties and categories.

For example, in the results of his analysis, Susumu Zaima (1986) states that, as indicated in (1) below, the semantic property known as a "result state" is included in the meaning of transitive verbs which form the "statal passive." Further, he assumes the following rule for the "statal passive": it is formed using transitive verbs that contain the "result state" as a semantic property in their meaning.

(1) (a) abschließen: "lock" (a verb that reflects a qualitative change of state)
Die Tür ist abgeschlossen.
The door is locked.
(b) ausstatten: "equip" (a verb that reflects a change of state due to the attachment of an object)
Der Raum ist mit einer Klimaanlage ausgestattet.
The room is equipped with an air conditioner.
(c) anbinden: "tie up" (a verb that reflects the attachment or detachment of an object)
Der Hund ist angebunden.
The dog is tied up.

In (1), sentence (a) expresses a "change of state" of the patient himself or herself; sentence (b) expresses a "change of state" in which an object is attached to or detached from the patient; and sentence (c) expresses a "change of state" in which the patient has something attached or detached. Based on this, it appears completely natural that the "statal passive," which reflects the "result state," is formed using a transitive verb that contains the "result state" as a semantic property in its meaning.

Furthermore, in the results of his analysis, Tomoaki Seino (1991) states that in the case of verbs that reflect actions taken against the body parts of others, the meaning of verbs that co-occur with the syntactic form "Subject +

Verb + Dative + Prepositional Phrase" is "action-centered" (refer to sentence (a) in (2) below); on the other hand, the meaning of verbs that co-occur with the syntactic form "Subject + Verb + Dative + Accusative" is "results-centered" (refer to sentence (b) in (2) below). In this type of expression, he assumes the rule that if nothing more than "action" is expressed, then the syntactic structure "Subject + Verb + Dative + Prepositional Phrase" is used; however, if information up to the "results" is expressed, then the syntactic structure "Subject + Verb + Dative + Accusative" is used.

(2) (a) Subject + Verb + Dative + Prepositional Phrase ⇔ Action-centered
Die Mutter sieht *dem Kind* **in die Augen**.
The mother looks into the child's eyes.
(b) Subject + Verb + Dative + Accusative ⇔ Results-centered
Die Mutter wäscht *dem Kind* **die Hand**.
The mother washes the child's hands.

In the pattern of expression that reflects the "result state" that a body part subject to an action has undergone, displaying the body part in the accusative form is also consistent with the semantic role of "totality," which the accusative is generally perceived to possess. "Results" only occur once the patient experiences the complete action.

Even if the verb is an "action-centered" verb, if there is a "result state" pattern of expression supplemented by a "result-predicate adjective," then the body part subject to the action will be expressed in the accusative form, without a prepositional phrase.

(3) (a) <Action-centered> Er schlägt ihr **auf die Schulter**.
He hits her shoulder.
(b) <Results-centered> Er schlägt ihr **die Schulter** ***wund***.
He hits her shoulder, and inflicts an injury.
Reference: *Er hat ihr *auf die Schulter* **wund** geschlagen.

In the results of his analysis, Susumu Zaima (1987a) states that in sentence (a) in (4) below, the verb "schütteln" adopts the normal usage of the verb "to *shake*," and the syntactic structure also corresponds with this. In contrast, in sentence (b), the verb "schütteln" is used as a "means" (*through shaking*), and the structural meaning of the sentence "move X from Y" is expressed by the syntactic structure "Subject + Verb + Accusative + Directional Prepositional Phrase."

(4) (a) Er schüttelt den Baum.
He shakes the tree.
(b) Er schüttelt **die Äpfel vom Baum**.
He shakes the apples from the tree.

Confirming that the same type of correspondence exists between sentence (a) and sentence (b) in (5) below, both of which use the verb "klopfen" (to *knock*), he considers the following rule for forming semantic sentence structures: the role of expressing a "means" is assigned to the verb and that of expressing semantic sentence structures is assigned to its syntactic structure.

(5) (a) Er klopft ihr auf die Schulter.
He knocks her shoulder.
(b) Er klopft **den Staub von den Schultern**.
He knocks the dust from her shoulder.

However, when considered in terms of the aforementioned purpose—to clarify the actual language use of native German speakers—conventional research methods like those mentioned above, which do not use extensive language corpus data and which focus on the combination of verbs and "essential constituents," are inadequate due to at least the following three reasons.

First, this research does not include consideration of the semantic content of the constituents of a sentence. For example, both sentences (a) and (b) in (6) below are grammatically correct; however, in terms of the actual language use, sentence (a) is acceptable but sentence (b) is not.

(6) (a) Bei Müllers ist jemand krank.
Someone in the Mueller family is sick
(b) *Bei Müllers ist jemand gesund.
* *Someone in the Mueller family is healthy.*

This demonstrates that the semantic content of the constituents of sentences (in this instance, the semantic content of the adjective) is connected to the acceptability of the sentence in the actual language use. The question of acceptability is one of world knowledge. Accordingly, in order to clarify the actual language use of native German speakers, we must also take into account the semantic content of the constituents of sentences.

Second, the research does not take into consideration "non-essential constituents." For example, according to Eiko Iguchi (1984), in (7) below, sentence (a), which is an impersonal passive sentence containing a motion verb, is not acceptable; however, sentence (b), which has been supplemented by a “non-essential constituent,” is acceptable.

(7) (a) *Nach London wird geflogen.
[Literal translation] *To London is flown.*
(b) Nach London wird nur einmal am Tag geflogen.
[Literal translation] *To London is flown only once a day.*
(There is only one flight a day to London.)

It would appear that the reason why sentence (a) is not acceptable is because as a result of making the sentence passive, there are no longer sufficient constituents for forming a theme-rheme structure. Furthermore, the reason why sentence (b) is acceptable is that as a result of supplementing the sentence with “non-essential constituents,” it is now elementally possible to restructure the theme-rheme structure (refer to Susumu Zaima, 1987b, for details). This example demonstrates the connection of “non-essential constituents” with the acceptability of a sentence. Consequently, in order to clarify the actual language use of native German speakers, we must also take these “non-essential constituents” into consideration. Despite the fact that they are not grammatically necessary, if “non-essential constituents” are used, they gain some kind of expressive role.

Third, the research does not (cannot) take into consideration the “gap” that arises between the data on actual language use and the sentences that are perceived to be possible under the rules which are extracted from sample-type data (including the data from informants, etc.). For example, according to Hiroaki Yamada (2001), with regard to the joining of perception verbs and infinitives, as shown in (8) below, depending on the informant, transitive verbs are able to be used as infinitives. However, according to the data of actual language use, only the instances of intransitive verbs are observed, as shown in (9) below, and those with transitive verbs, such as in (8), are not observed.

(8) Ich sehe seinen Sohn *das Auto waschen.*
I see his son washing the car.
(9) Ich sehe ihn *davonlaufen.*
I see him running away.

Furthermore, with regard to infinitives in the “lassen”-constructions, according to Yasuhiro Fujinawa (2002), in the case of transitive verbs, depending on the informant, it is possible to show the semantic subject in the accusative form, as shown in (10) below. However, there are almost no such instances in the data on language use (an exception is the case of verbs of cognition, as in (11) below). Moreover, in cases where the "actor" is a specific person, there is a tendency to express the semantic subject by using prepositional phrases, as in (12) below.

(10) Er lässt **seinen Sohn** das Auto waschen.
He lets his son wash the car.

(11) Er lässt **die Zuhörerinnen und Zuhörer** diese Situation mit anderen Augen sehen.
He allows the male and female listeners see the situation through different eyes.

(12) Er lässt einen Brief **von seiner Tochter** schreiben.
He makes his daughter write a letter.

These cases demonstrate that there is a certain "gap" between the data on actual language use and the sentences that are perceived to be possible under the "rules" derived from the data obtained through informants. Consequently, in order to clarify the actual language use of native German speakers, we also need to clarify such "gaps" and the factors that give rise to them (for example, the ability to remember and other factors constraining the application of the rules).

In sum, I have stated that in attempting to clarify the actual language use of native German speakers, we must incorporate the semantic content of the constituents of sentences—including the semantic function of "non-essential constituents"—into our analysis. Further, we must incorporate into our analysis both the "gap" between actual language use and the sentences that are perceived to be possible from sample-type language data (including the data from informants, etc.), as well as the constraining factors that give rise to these gaps. The conclusion that can be drawn from these facts is that in order to clarify the actual language use of native German speakers, it is essential that the base of the analysis be data on actual linguistic performance that has already been subject to constraints from various factors associated with linguistic performance, including the concrete semantic content of the constituents of sentences and "non-essential constituents."[2]

[2] Linguistic analysis is conducted on the premise that the language is practiced based on a fixed "system of rules." (If this presumption cannot be made, then there is no point to linguistic analysis.) However, a "system of rules" can only be observed through concrete linguistic phenomena (= language use). If this is the case—moreover, because this is the case—the starting point for linguistic analysis lies in the language use data (= corpus), and even in this sense, a corpus-focused viewpoint is but natural.

3. The Results of Single-language Research, Evaluation and Application

At the outset, I stated that public research as a whole—and not merely language research—is a social action, and consequently, it should have some kind of social significance. Next, I will discuss the limits inherent in research on the German language. Subsequently, for the purpose of making German language research more meaningful for society, I emphasis the need to consider applications for language education, machine translation (automatic translation), etc.

If a certain conclusion (hypothesis) was presented based on certain language data, then conventionally, the logical order leading up to that conclusion would have been made the subject of verification; however, there was no verification conducted on the significance of that type of conclusion (Susumu Zaima, 2002). Heide Wegener (1985) defines the "dative" in the German language as follows:

> Gemeinsames Charakteristikum dieser Varianten (*des Dativs*; author's note) ist das Merkmal des Betroffenseins oder Sich-Betroffenfühlens. Daher gilt BETR als semantische Grundfunktion des Dativs.
>
> (A common feature of these variations of the dative is the *merkmal* that there is a sense of being affected or having been affected. Subsequently, BETR (effect) is recognized as the basic semantic function of the dative.)

The question that first arises pertains to how we should view the significance of the conclusion derived from this type of language data analysis. If we conduct research on the "dative" and find that such a thing actually exists, then the extraction of the very thing itself becomes the subject of the research. Thus, it becomes a question of whether the conclusion—as the result of the analysis—has ultimately become the extraction of the very thing itself. In other words, it becomes a question of whether the conclusion is consistent with the "truth." However, Mariko Hasegawa (1999: 87) indirectly states in the following comments that the direct recognition of the "truth" is impossible:

> Science is a sequence of hypothesis-building. It is the process of selecting the hypothesis closest to the truth by observing and cross-checking opposing hypotheses. Besides this, is there any better method available for acquiring knowledge related to the natural world that surrounds us?

If we extend this concept, then as long as we do not understand what the "truth" is, it is impossible to ever know which of the opposing hypotheses is "closest to the truth," as expressed by Hasegawa. Previously, regarding the topic of research into the German language, it was remarked that “we should be seeking the linguistic truth that can be discovered within the language” (Susumu Zaima, 1987: 4). We need to reconsider the existence of this kind of “linguistic truth.”[3]

Without being limited to Wegener’s research (Heide Wegener, 1985) research—in particular that which is actually being conducted as single-language research—could not be undertaken for discovering what already exists somewhere in a single definite form. Therefore, it should be undertaken for analyzing and organizing a variety of language data based on the respective linguistic knowledge of individual researchers and for creating improved “theoretical constructions.” Ukichiro Nakaya (1958: 19) remarks in the following statements that through the eyes of science, humans have also created laws for natural phenomena.

> In the world of science, we oft use the words “natural phenomena,” “actual natural appearances,” “the laws between these phenomena,” and other things, but these have all been discovered by humans. (...) By “discovered” we actually mean the state of nature found by science. (...) True nature is possibly something quite different, and is *probably* something very different indeed. Nevertheless, since it is humans who are seeing, the same as there is no alternative to seeing through the eyes of humans, when science looks at nature, there is no choice but to look at it through the eyes of science. To put it another way, nature is recognized through various viewpoints presently used in science, namely the forms of scientific thinking ...

It could be assumed that “discovery” through language research is “built up by analyzing and organizing a variety of language data based on the respective linguistic knowledge of individual researchers” (even supposing that research to extract the “truth” is effectively impossible). However, since this research is based on real data, it is conceivable that the more detailed the analysis, the “more truthful” it is. So long as we remain within the framework of linguistic phenomena, we cannot avoid the situation wherein the evaluation of the study findings ends up hinging on the subjective criterion of persuasiveness. In some cases, this will probably be satisfactory; however, if the type of research sought is that which can satisfy social demands more directly, a new framework for research that overcomes such limitations will need to be created. One such possibility is the introduction of “applicability.” In concrete terms, in language research, this would entail setting goals for practical applications for language education or machine

[3] With regard to the topic of whether theoretical “constructions” in generative grammar “really exist,” refer to Satoru Nakai (1999) and Kuniyoshi Sakai (2002).

translation for instance. In this case, the results of the research would not be the subjective criterion of persuasiveness; rather, they would ultimately be evaluated according to the external criterion of practicality. If an evaluative correlation can be established between language research and practical application, then language research might be added to one of the "sciences" whereby the "casualties" can be argued.

Rather than the "truth" of the German language, research in German that seeks applicability will attempt to obtain results that are effective in terms of application (specifically for language education or machine translation); further, more than being a "deep" analysis, the research will be a "broad" analysis. The social significance of creating things (= "grammar"), which can be applied more effectively to language education or machine translation, will become largely evident.[4]

In sum, I have stated that the purpose of extracting the "linguistic truth" in language research is effectively impossible. I have also discussed that as a new form of research that is capable of satisfying social demands more directly, a possible purpose of language research could be the application of the research findings to language education or machine translation, for instance.

4. The Concept of German Language Research and the Associated Analytical Work

Next, we will proceed to the question regarding the nature of research on the German language that is conducted on the premise of practical applications such as language education or machine translation. In general terms, this type of research does not entail a minutely in-depth description of grammatical phenomena. Instead, it involves a broad analysis and description of the basic phenomena concerning the "actual" language use of native German speakers, including the frequency of use, from the perspective of practical applications. Surveying frequency is one of the methodologies that have been made possible only recently, thanks to the advancement of IT technology (using corpus). This type of research should not oppose the conventional type of research wherein the extraction of linguistic rules would be carried out through sample-type data; rather, it should be positioned as one of the two means necessary for clarifying the language use of native German speakers.

Based on the above discussion, I will comment on the analytical work

[4] This does not imply that each and every researcher should conduct his/her research while questioning the social significance of their respective studies. Ideally, researchers who seek the "truth" based on the actual data, and others who emphasize applications, should collaborate to present their results language research to society.

that is currently being undertaken with the aim of producing the “Dictionary of Combination Frequencies of Basic German Verbs” (tentative title). Step 1 is to randomly collect examples from a corpus, and then to survey both the structure of German sentences, including “non-essential constituents,” and the frequency of their usage. With regard to the randomly collected sentences, since this collection will include sentences with high usage frequencies, it will ultimately analyze data that have high levels of importance in language use. The following are the results of a survey of 119 examples of the verb “verbringen” (to *spend*), which were collected from the public corpus of Institut für Deutsche Sprache (Mannheim). (Four main types of sentences are given. Since the semantic classification of the constituents of sentences depends on a subjective judgment, the figures vary depending on the understanding of the criteria.)

(13) (a) Agent + Verb + Object (Time) + Place 50 / 119 examples
... **verbrachten** den letzten Nachmittag im Prager Museum.
…spent my last afternoon in the Prague Museum.

(b) Agent + Verb + Object + Place + Other (+ Other) 40 / 119 examples

b-1 Agent + Verb + Object + Place + 1 Item 27 / 119 examples
... **verbrachte** mit seiner Familie einige Ferientage im Bündnerland.
… spent a few days’ vacation with his family in the Buendnerland region.

b-2 Agent + Verb + Object + Place + 2 or more Items 13 / 119 examples
... **verbringt** wegen eines Rückenleidens täglich eine Stunde oder mehr in der Badewanne ...
... spend an hour or more every day in the bathtub because of back pains…

(c) Agent + Verb + Object + Other 21 / 119 examples
... können Mütter mit ihren Kindern einen gemütlichen Nachmittag **verbringen**.
… mothers can spend the afternoon relaxing with their children.

(d) Agent + Verb + Object + Other + Other, etc. 8 / 119 examples
Dieser Papst **verbringt** täglich allein mehrere Stunden im stillen Gebet.
This Pope spends several hours alone everyday quietly in prayer.

According to this survey, in the case of the verb “verbringen,” the most common sentence structure is <Agent + Verb + Object + Place> (approximately 42 percent), followed by the <Agent + Verb + Object + Place + 1 item> sentence structure (approximately 23 percent). Of the 119 examples, these two structures accounted for 77—an overwhelming number.[5]

The main point that can be confirmed from the surveys conducted to date is the fact that, in general, the number of constituents in one sentence is fairly constant. Even in the case shown above, 71 examples (approximately 60 percent) of the sentences comprised three items including the subject, and no less than 98 examples (approximately 82 percent) of the sentences comprised four items. Frank Mielke (1997) surveys the frequency of combination phrases for the verb “reagieren” (to *react*). He also states that sentences of two or three items, including the subject, account for approximately 70 percent of the 212 sentences that contain the verb “reagieren.” According to him, “Die überwiegende Mehrheit der Sätze enthält lediglich zwei oder drei Satzelemente” (*The majority of the sentences contain at most only two or three sentence constituents.*). These types of surveys on the syntactic frequency of use will not only show the sentence structure that is most prevalent with certain verbs, they will also provide us with empirical data on the standard number of constituents in a single sentence for themes of interest (also related to the ability to remember).

Based on the information from Step 1, Step 2 involves analyzing the usage frequency of the constituents in sentences for certain verbs. Of particular concern will be the “optionality”—in terms of frequency—of the constituents of a sentence that are identified as being “optional essential constituents.” For example, according to the aforementioned perspective of Frank Mielke (1997), the frequency of “causal constituents”—as in sentence (a) in (14) below—is 107 of 212 examples (approximately 50 percent), and the frequency of “manner constituents” as in sentence (b) is 164 of 212 examples (approximately 77 percent).

(14) (a) <Causal constituents> Die Pupillen reagieren *auf Licht*.
The pupils react to light.

[5] There are instances wherein a single phenomenon is expressed using different sentence structures. For example, in the following example, the difference in the sentence structures of sentences (a) and (b), which use the verb “wischen” (to *wipe*), arises due to whether the same phenomenon is perceived as “movement” or as a “change of state.”

(a) Er wischt die Krümel vom Tisch.
He wipes the bread crumbs from the table.

(b) Er wischt den Tisch.
He wipes the table clean.

This type of example considers the following types of analyses: (a) what kind of items are involved in the “situation” in which the verb in question is involved; (b) how these items are combined; and (c) by what morphologic and syntactic structures this combination is expressed. For further details, refer to Akiko Kawashimo (1998).

(b) <Manner constituents> Sie hat *blitzschnell* reagiert und ...
She reacted very quickly, and…

In the case of the verb “reagieren,” the frequency of “causal constituents” of reactions (the existence of which is logically essential) will be considerably lower than that of “manner constituents.”

A certain type of difference is also clearly observed in the frequency of “non-essential constituents.” For example, according to the results of the survey of 60 examples for the verb “warten” (to *wait*), the data for which was collected from the abovementioned corpus, “time constituents” as in sentence (a) in (15) below (24 of 60 examples: 40 percent) have a higher frequency than “place constituents” as in sentence (b) (8 of 60 examples: approximately 13 percent).

(15) (a) <Time constituent> Ich hätte *keine Nacht länger* warten können.
I could not have waited a single night longer.
(b) <Place constituent> *Vor der Tür* wartete das Fluchtauto.
The escape car waited in front of the door.

In learning the reality of the actual language use of native German speakers, it could be stated that the above example indicates that surveying and analyzing “non-essential constituents” is also crucial. Despite the fact that “non-essential constituents” are not grammatically necessary, since they are used, they are perceived to play an intrinsically important role as bearers of information. Obtaining a correct understanding of the role of the “non-essential constituents” demands the adoption of a new perspective.[6]

Step 3 involves analyzing the lexical or semantic categorical combination frequencies for the purpose of clarifying actual language use in more concrete terms. For example, according to Minkyeong Kang (2003), who analyzed the causative alternation verb “brechen” (to *break*), in the case of the transitive verb usage, nominating concrete things as objects, as in sentence (a) in (16) below, is the most infrequent usage (5 of 83 examples) followed by the nomination of body parts, as in sentence (b) (30 of 83

[6] In example (7), I remarked on the relationship between impersonal passive sentences and “non-essential constituents.” If the prepositional phrase with “auf” that is an “optional essential constituent” to the verb “warten” is omitted, then “non-essential constituents” will most likely be added. This suggests that there is a correlation between the omission of “optional essential constituents” and the addition of “non-essential constituents.” Further examination of the usage of “non-essential constituents” is required in the context of the condition for establishing acceptability.

examples); the most frequent case is that of nominating abstract events as objects (48 of 83 examples).

(16) (a) **Die Teeblätter** werden nicht ..., sondern gebrochen. (5 / 83 examples)
The tea leaves are not..., they are broken
(b) ..., seit er sich ... **den linken Knöchel** gebrochen hat. (30 / 83 examples)
..., since he broke his left ankle...
(c) Damit wurde erstmals **das Monopol** ... gebrochen ... (48 / 83 examples)
As a result, the monopoly was broken for the first time.

In the case of intransitive verb usage, the most infrequent usage is the nomination of body parts as the subject, as in sentence (a) in (17) below (3 of 17 examples), followed by nominating concrete things as subjects, as in sentence (b) (6 of 17 examples); the most frequent case is that of the nomination of abstract nouns as subjects, as in sentence (c) (8 of 17 examples). In comparison with transitive verb usage, there is a greater proportion of linking to concrete things in the usage of intransitive verbs.

(17) (a) Links war der Oberschenkel mehrfach gebrochen, ...
On the left, the thigh was also broken in a number of places…
(b) ... Gerichtsrat Walters Deichsel ist ... gebrochen ...
… Court secretary Walter's pole is broken…
(c) Wenn zum Beispiel eine Beziehung auseinander bricht ...
If, for example, a relationship is cut off…

The question of how significant the above percentages are is an issue that should be discussed in the future; however, if we conduct these types of surveys using a corpus that is limited to certain genre (for example, a corpus of expressions used when foreigners travel in Germany), then we are likely to obtain more characteristic frequency results. However, it is true that by using this type of survey, the reality of the phrase combinations will become more concrete. At the same time, we should also be able to observe the precise extent to which the combination of verbs and phrases are flexible or restricted; further, if these combinations are restricted, we should also be able to determine the types of restrictions.[7]

Step 4 involves the analysis of the trends related to how native speakers of German put information together, by conducting a survey on word order in language use. Assuming that there are no special communicative factors, such as emphasis, the aspect that should be considered is the order in which the constituents, which play a role in sentence structure, are usually put together. For example, Itsuko Tokita (2005), who analyzed sentences of transitive verbs that take inanimate datives, demonstrates that this type of verb, as in sentence (b) of (18) below, is divided into a verb group that exhibits word order properties similar to those of the animate dative of sentence (a) and a verb group of reverse word order properties as in sentence (c). In other words, the <dative + accusative> word order is 73 percent for the verb “schenken” (to *give*) in sentence (a), approximately 76 percent in the case of the verb “hinzufügen” (to *add*) in sentence (b), and, in contrast, only approximately 10 percent in the case of the verb “aussetzen” (to *expose*) in sentence (c).

(18) (a) Er **schenkt** ihnen Kleider und Schuhe.
He gives them clothes and shoes.
(b) Er **fügte** der Suppe etwas Salz **hinzu**.
He adds some salt to the soup.
(c) Sie wollte ihn dem Verdacht **aussetzen**.
She tried to pin the suspicion on him.

In other words, from the perspective of word order trends, verb groups

[7] Minkyeong Kang (2005) examined the possibilities of causative alternation phenomena using lexical levels. According to Kang, there are instances, as in (i) below, where causative alternation may be possible with the same noun. Depending on the noun, the use of intransitive verbs may be limited, as in (ii); further, depending on the noun, the use of transitive verbs may be limited, as in (iii). If we take lexical levels into account, certain limitations are observed for causative alternation.

(i) (a) Sie öffnete die Schlafzimmertür, ...
She opened the bedroom door…
(b) ... da öffnet sich die Zimmertür ...
… at that moment, the door in the room opened ...
(ii) (a) hat ... der vierte Angeklagte ... sein monatelanges Schweigen gebrochen.
The fourth defendant broke the several months of silence.
(b) *Sein Schweigen brach.
* *The silence broke.*
(iii) (a) *Jemand/*Etwas reißt den Geduldsfaden.
* *Someone / something wore his patience.*
(b) An der Kasse im Supermarkt reiße Männern ... der Geduldsfaden ...
The patience of men at supermarket cash registers… runs out…

that are semantically similar to the verb “hinzufügen” will display tendencies that conform to the verb “schenken,” and verb groups that are semantically similar to the verb “aussetzen” will display the opposite trend. It could be stated that this fact shows that the order in which native speakers of German put information together varies depending on the verb (specifically, the semantic properties of verbs, and even more specifically, how things are perceived).

Saoko Miyamoto (2004) conducted a survey on word order, covering approximately 130 sentences containing adjectives of manner constituents; the sentences were collected from the first section of Bernhard Schlink’s “Der Vorleser” (1995). According to Miyamoto, if we exclude those constituents that grammatically always appear at the end of sentences (for example, place constituents that have a close connection with a past participle or verb), then the adjectives of manner constituents will, in principle, be placed at the end of the sentence, as shown in (19). Since manner constituents are usually used to carry new information (only being subject to partial negation), this is consistent with the general word order rule that new information is generally placed toward the end of a sentence.

(19) Neben der Tür waren auf der einen Seite die Briketts **ordentlich** geschichtet ...
The briquettes were piled up on one side next to the door…

Furthermore, during sentence formation, it would appear that the newness/oldness of information and the essentiality of the constituents are also closely connected with the meaning of the verb or adjective. According to Yuichi Fukuda (2005), in general, prepositional objects that are connected with adjectives provide information that is already known. Further, of these, most adjectives that heavily reflect rational judgment, such as “interessiert” (*interested*) and “stolz” (*proud*), co-occur with prepositional objects more often, and adjectives that heavily reflect emotion, such as “ärgerlich” (*angry*) and “traurig” (*sad*), are used without prepositional objects more often.

Step 5 involves the analyses of the frequency of language use, by regarding the usage frequency as a reflection of the interests held by native speakers. For example, we can consider the abovementioned survey results of Frank Mielke (1997) according to which, when native speakers of German “reagieren” (to *react*), they have a greater interest in the “manner” aspect than in the “cause” aspect.

Here, I would like to refer to two other similar analyses. Yoshiyuki Muroi (1992) concluded that the combination frequencies for direction constituents vary depending on whether the subject to the verb “fahren” (to *drive*) is a “person” or a “vehicle.” He demonstrated survey results according to which, when the subject is a "person," as in sentence (a) in (20) below, there is a greater frequency of it combining with direction constituents and that when the subject is a "vehicle," as in sentence (b), there is a greater frequency of it combining with "route (place) constituents." From these results, he concludes that, depending on the subject of movement, native speakers of German will take an interest in different ways.

(20) (a) **Ich** fahre mit dir nach Kirchbach.
I will drive with you to Kirchbach.
(b) Vorn fahren **fünf Schmuckwagen** der Genossenschaft.
Five decorated vehicles of the cooperative will drive at the front.

Masahide Kamegaya (1999) conducted a corpus survey on the frequencies with which the movement verb "steigen" (to *get into*/*out*) combines with prepositional phrases and prefixes that express a "source" or a "goal." He demonstrated survey results according to which, when a "goal" is being expressed, unlike in sentence (b) in (21) below, simple verbs expressing the goal are used more often, as in sentence (a). He further stated that if a "source" is being expressed, unlike in sentence (a) in (22) below, complex verbs that do not express the source concretely are used more often, as in sentence (b). From these results, he concluded that when native speakers of German express a "goal," they tend to express it in concrete terms, but when they express a "source," they tend to express it in an abbreviated manner.

(21) (a) Er steigt **in den Zug**.
He rides on the train.
(b) Erst als ich **eingestiegen** war und der Bus anfuhr ...
Only once I had boarded and the bus had departed…
(22) (a) Er steigt **aus dem Zug**.
He gets off the bus.
(b) Er kam mit dem Auto, **stieg aus** ...
He came by car, got out…

Based on the results of the aforementioned analytical work, Step 6 involves analyzing the questions of how the lexis of the German language structures the content of expressions necessary in language use and how semantic content is transmitted to individual words. For example, with regard to the latter question, Minkyeong Kang (2001) collected approximately 320 change-of-state verbs and categorized and analyzed them from the viewpoint of whether they have only causative or non-causative usage or both. According to Kang, in German, change-of-state verbs can be divided into three categories: transitive verbs, such as that (23) below, which contain tools or means within the meaning of the words and consequently possess only causative usage; intransitive verbs, such as that in (24), which reflect events in which human participation is not possible and consequently possess only non-causative usage; and transitive-intransitive verbs (verbs that possess both transitive and intransitive verb usage) and transitive-reflexive verb (verbs that possess both transitive and reflexive usage), such as that in (25), which reflect events wherein human participation is possible and which are unmarked in terms of tools or means.

(23) Er hat die Tür aufgeschlossen.
He unlocked the door.
(24) Die Knospen werden bald aufspringen.
The flower buds should open soon.
(25) (a) <Transitive-intransitive verb> Sie trocknet die Wäsche auf dem Balkon.
She dries the washing on the balcony.
Die Wäsche ist getrocknet.
The washing is dry.
(b) <Transitive-reflexive verb> Er löst die Tablette in Wasser auf.
He dissolves the tablet in water.
Die Tablette löste sich in Wasser auf.
The tablet dissolves in water.

The way in which the content of expressions necessary in language use is structured, and the way in which it is transmitted to individual words is a linguistic phenomenon that is clearly individual in nature.

The analytical study progressed from Step 1 to Step 6 can be regarded as part of the linguistic empirical studies. Step 7 involves the analyses of the “gap” between language use data and the possibilities determined by grammatical rules, such as those observed in the examples relisted in (26) below. Since it is connected to physiological factors related to the brain, the analysis of the “gap” between language use data and the possibilities determined by grammatical rules will require collaborative research with such fields as brain physiology. Although this will be a subject of great interest, we will have to wait to see substantial developments in the future.

(26) → (8) Er sieht seinen Sohn das Auto waschen.
He sees his son washing the car.
(9) Ich sehe ihn davonlaufen.
I see him running away.

(10) Er lässt seinen Sohn das Auto waschen.
He lets his son wash the car.
(12) Er lässt einen Brief **von seiner Tochter** schreiben.
He makes his daughter write a letter.

I mentioned a number of studies currently being undertaken in order to realize the concepts described above. These can be summarized as follows:

<Step 1> Collect data from a corpus and analyze the structure of German sentences and the frequency of their use.

<Step 2> Collect data from a corpus and analyze the usage frequencies of the constituents in sentences for each verb.

<Step 3> Analyze the lexical frequency of constituents in order to clarify actual language use in more concrete terms.

<Step 4> Analyze the trends related to how native speakers of German structurally assemble information, by conducting a survey on word order in language use.

<Step 5> Consider the frequency of language use as a reflection of the interests held by native speakers, and analyze the usage frequency for patterns of expression.

<Step 6> Analyze the questions of how semantic content necessary for language use is structured and how it is transmitted to individual words.

<Step 7> Analyze the "gap" between language use data and the possibilities determined by grammatical rules.

5. Conclusion

Based on the presumption that, more than universality, the objective of single-language research is to clarify the actual use of the language in question, in this paper, I have remarked on the following: (a) the necessity for analysis based on language use data, (b) the necessity for applicability and other evaluative perspectives, and (c) concrete examples of envisaged German language research based on these considerations.

When the results of the abovementioned research have been verified based on the evaluative criteria of usefulness in applications for language education, machine translation, etc., and a framework that evaluates each of these research results has been established, then we will be able to say that single-language research studies of the German language have departed from simple "intellectual games" and have arrived at a new type of German language research in the form of a "science" based on "casualties." However, this destination is yet quite far.

References

Fujinawa, Yasuhiro. 2002. “Corpus ni yoru Futeishidzuki Taikaku Kobun Bunseki: Lassen no moto ni okeru Jirei o taisho ni” (Corpus-based Analysis of Accusative Constructions with Infinitives: Targeting examples of *lassen*), in *Corpus ni yoru Kobun Bunseki no Kanosei*, (Japanische Gesellschaft für Germanistik Studienreihe 009), edited by Inokuchi, Yasushi, 60-75.

Fukuda, Yuichi. 2005. “Keiyoshi to Zenchishikaku-Mokutekigo narabini Kanshi tono Kankei” (The Relationship of Adjectives with Prepositional Objects and Articles), Graduation Thesis, German Major, Faculty of Foreign Studies, Tokyo University of Foreign Studies.

Hasegawa, Mariko. 1999. *Kagaku no Me Kagaku no Kokoro* (Eye of Science, Soul of Science), Iwanami Shoten 623.

Iguchi, Eiko. 1984. “Hininsho-Judou no Youhou” (Usage of the Impersonal Passive), in *DER KEIM* Nr. 8, Doitsugo Kenkyukai, Graduate School, Tokyo University of Foreign Studies, 3-16.

Kamegaya, Masahide. 1999. “Corpus o Mochiita Ninchironteki Hindo Bunseki — *steigen* to *ab*-, *aus*-, *auf*-, *einsteigen* o Chosataisho ni —” (A Cognitive Frequency Analysis Using Corpora — A survey of *steigen* and *ab*-, *aus*-, *auf*-, *einsteigen* —), in *Tokyo University of Foreign Studies Gengo Kenkyu IX*, 197-204.

Kang, Minkyeong. 2001. “Doitsugo ‘Jotaihenka-Doshi’ no Togoteki Imiteki Bunseki” (Syntactic and Semantic Analysis of German ‘Change-of-State Verbs’), in *DER KEIM* Nr.25, Doitsugo Kenkyukai, Graduate School, Tokyo University of Foreign Studies, 5-28.

Kang, Minkyeong. 2003. “Gengo Unyo ni Miru Doitsugo ‘Jotai Henka Doshi’ no Ji-Ta <Kenkyu Noto>” (Intransitive-Transitive German ‘Change-of-State Verbs’ Seen in Language Uage <Research Notes>), in *DER KEIM* Nr.27, Doitsugo Kenkyukai, Graduate School, Tokyo University of Foreign Studies, 47-69.

Kang, Minkyeong. 2005. “Jotai-Henka-Doshi to Shiekikoutai” (Change-of-State Verbs and Causative Alternation), in *Gengo Johogaku Kenkyu Hokoku 7*, edited by COE “Usage-Based Linguistic Informatics”, Tokyo University of Foreign Studies, 399-442.

Kawashimo, Akiko. 1998. “Koi-Doshi no Togo-Kozo ni motozuku Imiteki Tenkai — Doshi Gogi to Togo-Kozo no Kankei —” (Semantic Development Based on the Syntactic Structure of Action Verbs — The relationship between the meaning of verbs and their syntactic structure —), Master’s Thesis, Graduate School of Area and Cultural Studies, Tokyo University of Foreign Studies.

Mielke, Frank. 1997. “Frequenzbasierte Gebrauchsanweisungen für Verben

— Fall-beispiel: *reagieren*,” in *Doitsugo Kyoiku* 2, edited by Japanischer Deutschlehrerverband, Japanische Gesellschaft für Germanistik, 62-72.

Miyamoto, Saoko. 2004. “Doitsugo ni okeru Youtai no Fukushiteki-Keiyoushi no Gojunteki Bunseki” (Analysis of Word Order of German Adverbial Adjectives of Manner), Graduation Thesis, German Major, Faculty of Foreign Studies, Tokyo University of Foreign Studies.

Muroi, Yoshiyuki. 2003. “Multilateral Interpretation of Corpus-based Semantic Analysis — The Case of the German Verb of Movement *fahren* —,” in *Linguistic Informatics I Proceedings of the First International Conference on Linguistic Informatics,* edited by Yuji Kawaguchi and Toshihiro Takagaki, 83-91.

Nakai, Satoru. 1999. *Gengogaku wa Shizen Kagaku ka* (Is Linguistics a Natural Science?), Showa

Nakaya, Uchikiro. 1958. *Kagaku no Hoho* (Methods of Science), Iwanami Shoten, G50.

Sakai, Kuniyoshi. 2002. *Gengo no Nokagaku*, (The Neuroscience of Language), Chuokoron Shinsha, 1647.

Seino, Tomoaki. 1991. “Shintai-Hyogen niokeru Yonkaku-Mokutekigo no Kino” (The Function of Accusative Objects in Body Expressions), in *Kumamoto Journal of Culture and Humanities* No. 35, 138-152.

Tokita, Itsuko. 2005. “Tadoushibun no Museibutsu Sankaku to Chuuiki Gojun (1)” (Inanimate Datives in Transitive Sentences and Word Order in the Middle Place (1)), in *Gengo Johogaku Kenkyu Hokoku 7*, edited by COE “Usage-Based Linguistic Informatics”, Tokyo University of Foreign Studies, 383-398.

Werner, Heide. 1985. “Der Dativ im heutigen Deutsch,” Tübingen, Narr (Studien zur deutschen Grammatik 28).

Yamada, Hiroaki. 2001. “‘Chikaku Doushi + AcI’-Kobun no Imiteki Tokucho — Chikaku-Doushi Hyogen no Hindo Bunseki o Toshite —” (Semantic Characteristics of ‘Verbs of Perception’ — through frequency analysis of expressions containing verbs of perception), Master’s Thesis, Graduate School of Area and Cultural Studies, Tokyo University of Foreign Studies.

Zaima, Susumu. 1986. “Doitsugo no ‘Jotai Judo’” (‘Statal Passive’ in German), in *Tokyo University of Foreign Studies Gogaku Kenkyusho Goken Shiryo* 5, 1-29.

Zaima, Susumu. 1987. “Doitsugo Kenkyu no Ichihoko” (A Direction of German Language Studies), in *Doitsu Bungaku* Nr. 79, Japanische Gesellschaft fur Germanistik , 23-34.

Zaima, Susumu. 1987a. “Verbbedeutung und syntaktische Struktur,” *Die Deutsche Sprache*, Heft 1, 35-45.

Zaima, Susumu. 1987b. “Einige Überlegungen zum „Generativen Mechanismus“ für die Generierung der deutschen Sätze,” in *Doitsugo no‘Togoteki Imiteki Seisei Mechanism’*. (Grants-in-Aid for Scientific Research FY 1995 – FY 1997 (Basic Research (C) (2)), Kenkyu Seika Hokokusho), 107-117.

Zaima, Susumu. 2002. “Doitsugo Kenkyu no Hohoron” (German Language Research Methodology), in *Atarashii Doitsugo Bunpo Kochiku no Kokoromi* (Studienreihe 011, Japanische Gesellschaft fur Germanistik Studienreihe 011), Edited by Inokuchi, Yasushi, 60-68.

Zaima, Susumu. 2004. “Doitsugobun no Keisei no Kisoku-Taikei” (System of Rules for the Formation of German Sentences), Grants-in-Aid for Scientific Research FY 2001 – FY 2003 (Basic Research (C) (2)), Kenkyu Seika Hokokusho, 1-234.

Developing Grammatical Modules Based on Linguistic Typology

Makoto MINEGISHI

Introduction

The objective of this paper is to introduce the purpose, process and structure for developing web-based educational materials for the grammar of 17 languages, based on linguistic typology.

Since 2001, the Tokyo University of Foreign Studies has been developing a system for web-based foreign language educational materials, under the auspices of the COE Program of Japan’s Ministry of Education, Culture, Sports, Science and Technology. The main feature of these educational materials is their modularity. In other words, four independent modules including pronunciation, dialogue, vocabulary and grammar have been integrated into one set of educational materials for each language. Development of the education materials for the 17 languages is nearly complete, and they can be accessed at our website.

Designing a collection of language educational materials in a modular way such as this is currently considered to be an unusual method. This is because many claim that the communicative approach is more effective and that foreign languages should be studied in a comprehensive manner, although it is difficult to test these assertions. However, the disadvantages created by this sort of modularity can be overcome by the use of hyperlinks. By connecting modules with hyperlinks, the student is able to jump from one module to the next at any time. As a result, the student, while studying the principles of grammar in the grammar module, can easily find practical examples of how these principles are applied by referring to samples within the dialogue module.

There is an additional advantage in having the grammar module independent of other modules: this enables the student to compare the grammar of his or her chosen language, the target of study, with the grammar of another language. In this way, the student of world languages is able to acquire an overview that may help him or her to better grasp the meaning of grammar.

The grammatical educational materials of the system above (to be called the G Module) are made up of two components. The first component, which is the principle one, is an integrated collection of the grammar courses for the 17 individual languages. Each course basically follows traditional teaching methods (when they exist). The second component, which is smaller, is the cross-linguistic grammar course, which presents a common framework for understanding grammar in general, and is based on data from the first component.

The section below defines the objective, development process, and structure of the two components of the G Module. We will pay special attention to the cross-linguistic grammar course because cross-linguistic educational materials such as these are quite uncommon in the world of the Internet.

Individual grammar courses in the G Module

The main part of the G Module is comprised of the 17 independent grammar courses. It is to be assumed that the students who study foreign languages using the G Module are primarily Japanese university students, as well as working adults and high school students. However, the English and Japanese language materials are exceptions to this rule: English is generally studied by Japanese children from a young age, and Japanese is usually learned by non-Japanese people as a foreign language.

Generally speaking, Japanese adults have experienced learning English through their primary and secondary education. Their desire to learn a language is based on the practical application of language skills or because of a certain hobby, but it is presumed that they do not have a general knowledge of linguistics.

Students may choose one language from the G Module and learn the grammar of that language step-by-step. This approach is very efficient in the study of a certain language because it can be used in combination with the modules for pronunciation, dialogue, and vocabulary.

Design and features of the individual language courses

Based on the knowledge and teaching experience of experts, the contents of the teaching materials for the G module have been sequenced so that the necessary grammar items (grammatical features) can be selected from beginner to intermediate level, enabling systematic study. Grammatical terminology also uses the traditional technical terms of each language. The student's program of study is not completed simply through educational materials on the Internet. The student is also expected to use other educational materials and dictionaries for reference, so the terminology used in the course material must be consistent with traditional educational materials.

This means that the web-based educational materials for each separate language in the G module have been transplanted from printed materials that follow traditional grammatical education methods; what has been added are advantages offered by the Internet such as interactivity and multimedia using sound and graphics.

Typological features of the educational materials of 17 languages

Table 1 shows the individual language grammar course prepared for each language along with the typological features of that language such as the region where it is spoken, genealogy, morphological type, word order and topic prominency.

One feature of the G module is that 10 of the 17 languages are from regions in Asia east of Turkey. Also, all the languages in the G Module are classified as Dependent Marking Languages, as defined by Nichols (1986), and no Head Marking Language is included.

Table 1. Languages with Typological Features

Language	**Spoken Area**	**Genealogy**	**Morphological Type**	**Word Order**	**Topic Prominency**
Portuguese	W. Europe	Indo-European	Inflectional	VO	
Spanish	W. Europe	Indo-European	Inflectional	VO	
French	W. Europe	Indo-European	Inflectional	VO	
English	W. Europe	Indo-European	Inflectional	VO	
German	C. Europe	Indo-European	Inflectional	VO/OV	+
Russian	E. Europe	Indo-European	Inflectional	-	+
Arabic	Middle East	Afro-Asiatic	Inflectional	VO	
Turkish	Middle East	Turkic	Agglutinative	OV	
Mongolian	E. Asia	Mongolian	Agglutinative	OV	
Korean	E. Asia	Unknown	Agglutinative	OV	+
Japanese	E. Asia	Unknown	Agglutinative	OV	+
Chinese	E. Asia	Sino-Tibetan	Isolating	VO	+
Vietnamese	M. SE. Asia	Austroasiatic	Isolating	VO	+
Laotian	M. SE. Asia	Tai-Kadai	Isolating	VO	+
Cambodian	M. SE. Asia	Austroasiatic	Isolating	VO	+
Indonesian	I. SE. Asia	Austronesian	Isolating	-	+
Tagalog	I. SE. Asia	Austronesian	Isolating	-	+

Abbreviation: W. = Western, E. = Eastern, C. = Central, SE. Southeastern, M. = Mainland, I. = Insular

The trends revealed by Table 1 are described below.

From a language genealogy perspective, the bias is Indo-European (6, Portuguese, Spanish, French, English, German, Russian), followed by Austronesian (2, Tagalog, Indonesian), Austroasiatic (2, Cambodian, Vietnamese), Afroasiatic (1, Arabic), Turkic (1, Turkish), Sino-Tibetan (1), Tai-Kadai (1, Laotian), Mongolian (1, Mongolian), Unknown (Korean, Japanese).

From a geographical distribution perspective, there is a regional bias toward the languages of the Eurasian continent among the G Module languages. Six of the languages are European, two are Middle Eastern, four are East Asian, three are mainland Southeast Asian, and two are insular Southeast Asian. On the other hand, the languages of the North and South American continent, Oceania, New Guinea, and Africa have not been included.

In terms of the traditional morphological types, European languages tend to be inflectional, for instance the Eastern European languages, represented by Russian, with its proximity to Asia. German, a language of Central Europe, is regarded as inflectional in that it has case marking articles in its noun phrases, as compared to other Western European languages.

The languages stretching from Turkish in the west of the Asian continent to Japanese in the east are agglutinative languages.

The languages in Southeast Asia from China are isolating languages. The isolating languages from China to mainland Southeast Asia do not have derivational affixes, as opposed to the isolating languages of insular Southeast Asia, which do not have inflection but feature various derivational affixes.

When focusing on verbs and objects in word order typology, the languages of Europe, as a rule, feature a VO word order. Languages from the Middle East to East Asia have an OV word order, and languages from China in East Asia to mainland Southeast Asia have a VO word order.

Word order is related to topic prominency, the next typological feature. In languages that have a topic-comment sentence structure with the topic placed at the beginning of the sentence, even if VO is the basic word order of the language, OV word order is possible, placing O as the topic at the beginning of the sentence. In addition, in case topic prominency is higher, as in Indonesian and Tagalog, that language’s word order category in Table 1 is shown as a ‘–’ (minus sign) because a VO and OV distinction does not make sense.

Table 1 also shows German with both VO and OV word order. The order is decided upon according to the sentence structure as well as topic prominency. German is regarded as the transition between the VO common in Western Europe and the – (minus) of Eastern Europe (Russian). Likewise, Chinese is considered as the transition between the East Asian OV and the Southeast Asian VO word order.

The "Topic Prominency" column indicates languages which have the means to express the "about what" of the sentence through morphosyntactic means. All languages have some means of indicating what the sentence is about in their dialogical structure, and those means are considered to display topic prominency if they give it priority over, for example, the syntactic form such as word order.

In many languages, topic prominency is manifested through topic-comment word order and placing the topic at the beginning of the sentence.

In Table 1, among the languages that have topic prominency (+), Russian expresses the semantic role of the sentence's initial topic using the morphological means of noun inflection. In German, the semantic role of the noun which is the topic is indicated using the morphological means of an article, and in Japanese and Korean it is shown by the morphological means of an affix or particle attached to the topic noun, which is the same for the languages of insular Southeast Asia.

Languages from East Asia to mainland Southeast Asia are topic prominent, and manifest topic prominency only through the syntactic means of placing the topic at the beginning of the sentence, and special morphological means are generally not used. In that case, the semantic role of the topic is decided by the semantic relationship of the noun and verb.

As explained above, Table 1 shows that the typological features including morphological types of words, word order, and topic prominency nearly correlate with the geographical distribution of the languages of Europe, East Asia, and Southeast Asia, respectively.

The table also reveals that Russian, Arabic and Chinese are transitional types located on the boundary of each of their typological distributions.

Individual language courses as an introduction to the cross-linguistic grammar course

The morphosyntactic typological features of each language explained above are viewed as reflected in traditional or standard educational methods of grammar education as a foreign language. For example, educational materials for Russian, which is of the highly inflectional type, cannot be formed without considering the systematic arrangement of inflections that take place in verbs and nouns. On the other hand, in the grammar of an isolating language such as Cambodian, what comes first is the study of basic word order and structure through the combination of the verb and noun, and the choice of topic in dialogue based on pragmatic information. Therefore, choosing a language to study from among the individual courses means choosing a certain educational method that corresponds to that language, which should lead students to compare their own native language with the target language in terms of typological similarities and differences.

In general, the two points raised below are involved in efficiency when the student undertakes grammar studies in the individual language course.

First, if the grammatical type of the target language is the same as that of the learner’s native language, the learning process should become efficient due to the help of analogy with the grammatical system of the student's native language.

The second point is relevant not only to grammar but also to pronunciation, vocabulary, and the writing system of the target language. The proximity of the spoken area of the learner’s native language to that of the target language should make language learning more effective.

For learners whose native language is Japanese, it is likely that the languages in Central Asia, as well as Turkish, Mongolian and Korean in East Asia, will be easier to learn, as they share typological features, as shown in Table 1. On the other hand, learners may find difficulty in studying grammatical systems of languages that are different, morphosyntactically and typologically, such as the inflectional languages of Europe and the isolating languages of Southeast Asia.

Furthermore, when it comes to learners who speak Japanese who wish to study Chinese, it is important to note that, as both languages are spoken in East Asia, they share many cultural features such as writing systems, cultural background, and cultural vocabularies. As a result, Chinese should be easy for Japanese speakers to learn, even though these two languages are different morphosyntactically and typologically.

Through the study of the individual language grammar course, finding that the relative difficulty or ease can be attributed to the typological and regional features of the language being studied may lead students to another sphere of interest beyond learning the grammatical system of the individual language — he or she may also become aware of the general typological features of languages. In other words, being aware of the similarities and differences among languages leads the student to the topics of linguistic typology and the geographical distribution of language. Learning a foreign language can, therefore, become a good opportunity for the student to view that language in the broader perspective of languages in general, i.e., to survey languages from the typological and geographical points of view, to question how a certain grammatical function may be realized in languages of different types, and eventually, to ask the ultimate question, “What is grammar, and what is language itself?” — the answer to which is an important objective of the study of general linguistics.

Cross-linguistic grammar course in the G Module

In order to respond to the interest in linguistic typology, and moreover to the budding interest in general linguistics that students develop through undertaking the individual language grammar course, the G module provides a cross-linguistic grammar course based on the typological knowledge, which is a unique feature added to the language education materials.

The following explains (1) the objective, (2) the process of developing the cross-linguistic grammar course, and (3) the structure of the education materials.

Anticipated students

Currently, the cross-linguistic education materials are primarily for Japanese native speakers, and then for speakers who understand Japanese, who have chosen one of the individual language courses listed above and are currently studying or have completed their studies.

Japanese people, no matter what their capabilities may be, are supposed to have studied English for a minimum of three years during their compulsory education. Therefore, it would not be an exaggeration to say that English ranks first and foremost in the study of foreign languages in Japan. As a result, when Japanese students choose a language included in the individual language courses of the G Module, it means that they are learning a third language, following English.

According to Table 1, Japanese is an agglutinative language that has OV word order and topic prominency. In contrast, English is an inflectional language with VO word order and does not feature topic prominency. The areas in which the two languages are spoken are remote, as well.

Therefore, for the typological features and geographic distribution as listed in Table 1, Japanese students would probably think of the target language they choose in the G Module on a certain point of a line that assumes Japanese and English at opposite ends of the spectrum. This thought would itself be the first step to a language typology and an appreciation of general linguistics, since the learner has begun to consider the typological and geographical attributes of the target language.

The cross-linguistic grammatical course assumes students who, through the experience of studying an individual language, have discovered an interest in the general typological features of languages, their geographical distribution and, further, the cross-linguistic features of relations in form and meaning in languages.

Objective of Developing the Cross-linguistic Educational Materials

The objective of developing the cross-linguistic educational materials is to provide those who study language from the point of view of linguistic typology with answers to the following questions concerning grammar itself, beyond the rules of specific individual languages.

In other words, the question of “what grammar is” shall be clarified by answering queries such as: (a) what sort of grammatical categories can be seen in the specific grammatical features of the languages; (b) what sort of linguistic forms are used to realize grammar; and (c) how the linguistic form which is used to express the grammatical category co-occurs with lexical items or a certain lexical category. Put simply, the goal is to provide elementary educational materials for language typology as well as general linguistics study, in response to the interest the student has developed in the course of individual language study. Concerning the three questions above, the following is what the cross-linguistic educational materials should provide.

Contents of the cross-linguistic educational materials

(a) *What sort of grammatical categories best describe the specific grammatical features of the different languages?*

In developing cross-linguistic educational materials, the first requirement is an abstraction and generalization of what sort of grammatical categories best describe the grammatical features of various languages.

From the perspective of de Saussure cited below, it is impossible to predict which would appear as the morphosyntactic features among the meaning and function of words.

> *The interpenetration of morphology, syntax and lexicology is explained by the fact that all synchronic features are ultimately of the same kind. No boundary between them can be laid down in advance. Only the distinction earlier drawn between syntagmatic relations and associative relations suggests a classification which is indispensable, and which fulfills the requirements for any grammatical systematization.* (de Saussure: Harris 1983: 134)

Therefore, it is impossible to presuppose what sort of grammatical category should exist in every language. For example, it is well known that grammatical categories such as gender and number, though very common among languages of the Indo-European language family, do not universally exist. Concepts such as case, transitivity and passivization, which are often taken up as topics in typology, are also not universal to all languages.

However, on the other hand, grammatical categories in existing languages do not exist in a purely arbitrary state, completely chaotic and with no regard for principle.

The first reason for this is that language rests on the basis of cognitive systems possessed by all human beings. Thus, the grammatical concepts of a given language must not be such that they can be understood only by the group of people who speak that particular language, but rather must, by definition, be intelligible to all people.

The second reason is that natural languages are limited by human physiological constraints, in particular regarding the transmission of sound. The linearity of linguistic signs, stated as de Saussure’s second principle, is a universal constraint to all languages. Human languages, consequently, have developed devices that mark the cohesion of sequential language forms (such as phrases) to overcome the constraints of sound. Examples include the expression of agreement in the gender and number of an article, a numeral, adjectives, and a noun within a noun phrase.

In addition, the cross-linguistic course should not only deal with the general conditions of human language as described above, but also elucidate how grammar should be, as explained below.

First of all, there are specific grammatical categories in the individual languages, forming several clusters according to the spoken area, linguistic genealogy and language types.

The distribution of grammatical categories in general agrees with that of language genealogy or the spoken area. For example, number, as a grammatical category, prevails among the languages of the Indo-European family, although, depending on the language, it may appear as a dichotomy, singular or plural, or in tripartite opposition, singular, plural, or dual. On the other hand, number, as a grammatical category, generally exists neither in East Asian nor Southeast Asian languages, but in many of the languages of these regions, classifiers, often used together with the numeral, are closely connected with numerical concepts. Certain languages express with the classifier not only a number, but also the specific shape of the countable items and the difference in whether they exist individually or as an abstraction. The grammatical feature of individuality through this sort of classifier mostly handles the function of definite/indefinite articles in the Modern Indo-European language family.

The above example shows that the human cognitive basis such as “cognition of number,” “distinction of individual versus abstract” and “cognition of visual shape/form” is reflected in one of the three grammatical categories of number, definitiveness or individuality, depending on the language. We can assume that these three categories form a cluster, as they are realized in the nominal affix, article or classifier for many languages. The reason for the clustering of grammatical categories may be the mutual connection of the three cognitive functions mentioned above.

Secondly, certain grammatical categories have various formal realizations according to the language. For example as explained above in the grammatical category of number, for languages in the European region from Indo-European roots, the number takes shape in a noun and adjective affix or article, whereas in many languages in East Asia or Southeast Asia, it takes shape in the combination of the numeral and classifier. Here as well, we can see geographical distribution, and it is important also to look for involvement of typological features, the former language group being inflectional, and the latter being agglutinative or isolating.

As shown above, the grammatical categories of various languages form clusters, and they are distributed according to the spoken areas, language genealogy or language types. In the cross-linguistic course, the first objective is to clarify what sort of relation exists between specific grammatical categories in languages and the spoken area of the language or language types.

(b) In what form is the grammar realized?

In order to present a bird's eye view of the grammaticality of languages through these education materials, we focus on what sort of linguistic forms the linguistic functions, which are shared among languages, are realized in. This can be called a sort of functional approach.

Specifically, this means describing a certain linguistic function as it is realized in a particular linguistic form, or presenting a syntagma of linguistic forms in languages. Taking as an example the function concerning the general idea of number explained above, the idea of number is expressed in the syntagma with a noun as its head (i.e., noun phrase) in English, by means of the definite or indefinite article expressing the definiteness/indefiniteness of the referent, and by numerals and plural affixes showing the number of the referent.

de Saussure's two relations between forms

As described above, de Saussure remarks that linguistic signs are united together in either syntagmatic or associative relations. The former relation means the formation of syntagma by the sequence of linguistic signs. Among the latter associative relations is included the paradigmatic relation, whose elements are limited in number and mutually exclusive, such as the singular and plural forms of a noun.

Generally, grammatical notions are realized either in the form of syntagmatic expansion or as in the selective paradigm. The two sorts of relations among linguistic signs can also be seen between independent word or free forms, and bound forms such as stems and suffixes that make up words. Bloomfield (1933) states that syntax describes the former, and morphology describes the latter. According to the traditional grammarians, the notion of grammar consists of both syntax and morphology.

The classic descriptions of the grammar of Latin and Sanskrit are mainly concerned with the paradigmatic relation in morphology. On the other hand, modern language theory is focused on syntactic relation, such as describing the structures formed with words.

The cross-linguistic course mainly describes how nominal and verbal syntagma are formed as chains of forms. We focus on the morphological aspect of languages because common rules are often found in the morphology, especially in the word formation, of languages, whereas in the syntactic aspect, the structure and meaning of a syntactic construction must be specified for each language except for the basic principle of syntagma formation, as explained below.

Sapir's algebraic formula

Concerning the basic principle of syntagma formation, Sapir (1921) states that, when the two linguistic forms "A" and "B" unite, the simplest way to express a grammatical idea is to form a sequence "AB" through simple juxtaposition, where uppercase "A" and "B" each stand for radical elements such as the root and stem. For example, in English "They come," is a structure at the syntactic level that follows this principle. Formation of a sentence following this sort of juxtaposition is seen in many languages around the world.

Sapir (1921:25) further gives an example "sing-er" as a structure in the morphological level, and expresses it by the "algebraic formula" of "A+b." Here the capital "A," as in the earlier example, stands for a radical element such as the root or stem of a word, and the lower case "b" expresses a subsidiary grammatical increment. In other words, "A" is a lexical item which carries a lexical meaning, and "b" is a bound form which carries a grammatical function, which is typically realized in an affixation such as a prefix, suffix, infix, circumfix, and so on.

Following the algebraic formula of Sapir and the syntagmatic/associative relations of de Saussure, we examine how a grammatical notion is realized in languages.

The grammatical description in terms of the classic paradigm applies to the above "A+b" where "A" stands for a noun (or a verb) and "b" its inflectional suffix, respectively. In this case, "b" as the bound form of definite numbers b_1, b_2, b_3, ... b_k appears in a mutually exclusive, paradigmatic relation. Likewise, the case marking of nouns in Russian, the declension of articles in German, and verb inflections in other

Indo-European languages and Arabic, which are all expressed in the same formula “A+b” standing for a paradigmatic relation, are therefore regarded as equal in function in the G module.

As for the similar case system, Japanese marks cases by adding a relatively independent word, a postposition, to a noun, which could be expressed in Sapir’s formula of “A+B”, where “B” stands for the postposition. “B” has the property of an independent word which appears “after A.” In other words, it appears as an element in syntagmatic relation with the other element “A,” rather than being an element of the mutually exclusive paradigm.

It must be noted that, as mentioned above, the number of elements b_1 to b_k in a paradigmatic relation are limited to a certain number, whereas the number of lexical elements in a syntagmatic relation is potentially indefinite, such as “A+B+C+D, ... etc.” Especially if “B,” “C,” “D,” etc., are of a highly lexical property, they are regarded as grammaticalized forms of lexical items.

Among the languages of the cross-linguistic course, in the grammars of the isolating languages such as Chinese and the mainland Southeast Asian languages, or of the agglutinative languages such as Japanese, grammaticalized lexical items appear in syntagmatic relation, whose number in total is indefinite.

The outline of cross-linguistic grammar is a line in which one end of the spectrum hosts inflectional languages like Russian, in which grammar appears paradigmatically, and the other end hosts isolating and agglutinative languages, in which grammar appears syntagmatically. Between these poles are the languages that feature some grammatical system of compromise. This viewpoint is close to the classic typology proposed by von Humboldt.

For example, although English has lost most of its inflectional properties, its grammar is largely characterized by paradigmatic expressions, with one part combined with syntagmatic expressions. English tense is accomplished in a paradigmatic relation consisting of the non-past tense, past tense of verbs, and future tense of modal auxiliary “will”; on the other hand, apart from this paradigm, the future may be expressed as a syntagmatic extension, “to be going to do,” that indicates the immediate future.

In this context, depending on the language, if there are more inflective qualities in a language, its grammar is expressed as paradigm, and if there are more independent qualities in each grammatical form (specifically in agglutinative languages and isolating languages), its grammar is expressed as syntagma. The example of the “be going to” of the English tense system is an example of the existence of a system of compromise which incorporates a grammatical system, where independent vocabulary such as “going” is grammaticalized.

The cross-linguistic course displays the way grammar takes shape in association with language genealogy, typology and geographical distribution.

(c) In what way does it co-occur with lexical categories?

The above example of the English tense system, the grammaticalization of "going," with its high degree of independence as vocabulary, indicates that the distinction between "B" and "b" in Sapir's formula is not always discrete, but rather, continuous.

We consider there to be various levels to the realization of grammatical categories, related to the degrees of freedom of their appearances, ranging from lexical elements that have a high level of freedom to affixes that are dependent upon other linguistic forms. An affix that in itself has a low level of freedom of appearance subcategorizes the free forms with which the affix co-occurs in parts of speech, or word classes. In other words, for many languages a group of words formally expressing a particular grammatical category forms the basis for classifying the parts of speech for that language.

The cross-linguistic course shows how the phrases with the two word classes, the noun and the verb, as their head, are enlarged and form syntagma.

One reason we focus on nouns and verbs is that the number of word classes differs depending on the language. According to Sapir and Bloomfield, the distinction of noun and verb is believed to be common for almost all languages, whereas the existence of other parts of speech depends on the language.

Another reason, related to the functionalist viewpoint, is that, because of the prevalence of a function that describes "about what" (reference) and "being in what sort of situation/situational change" (predication), the noun describing the former and the verb the latter, respectively, are believed to be fundamental to all languages.

Based on this line of thought, the cross-linguistic educational materials describe how nouns and verbs form phrases with the other grammatical constituents in many languages.

It then goes on to describe the way more than two forms are morphosyntactically united to form an even larger grammatical structure.

Process of Educational Material Development

As mentioned above, in the development of cross-linguistic educational materials, it is not realistic as a methodology to establish grammatical categories via a top-down approach prior to explaining how those are actually realized as concrete grammatical features in the individual languages.

In the initial stages of developing the cross-linguistic grammatical education materials, firstly we worked to specify the grammatical features in each individual language course for representative languages of morphologically different types, i.e. Spanish, Russian, Chinese and Japanese, in order to abstract common grammatical categories. Then based on the data, we examined how the grammatical categories corresponded to the grammatical features which appeared in the rest of the individual language courses.

Through this process, we obtained data on what sort of grammatical categories can be abstracted from the grammatical features that exist in the languages of the world from a typological point of view.

In the educational materials we developed, each of the main grammatical categories is described with a "definition" and further "details," if any. These grammatical categories are also hyperlinked to actual examples of the grammatical features through which they are realized in each language. In the list of hyperlinks, respective links are displayed as a button with the name of the language when a particular grammatical category is realized as a form in that language, and when it does not exist, that language name appears on a gray background. By glancing at the language names, either highlighted or grayed out, students are able to see whether a particular grammatical category exists in a target language, as well as what sort of regional distribution that category has. Moreover, by clicking to select one or more languages, students can confirm exactly how the grammatical category is formally realized in each of those languages.

Abstracting and organizing the cross-linguistic language functions realized in each language enables the students to make a cross-linguistic survey of many languages. For example, in many languages, expressions concerning perception and emotion take exceptional forms, such as particular sentence patterns, affix/particle marking, and so forth, depending on the language.

Structure of the Cross-linguistic Educational Materials

The actual cross-linguistic educational materials are structured as described below:

The cross-linguistic course consists of 20 lessons. The steps for each lesson are from 1 at the least and 9 at the most. Further additions to or restructuring of these lessons and steps may be undertaken during the further development of the educational materials in the future.

Table 2 displays the titles of the 20 cross-linguistic lessons and the number of steps in each.

Table 2. Structure of Cross-linguistic Course

Lesson No.	**Title**	**Number of Steps**
Lesson 01	Language and grammar	5
Lesson 02	Basic concept of grammar	9
Lesson 03	Grammar with a focus on nouns	9
Lesson 04	Grammar with a focus on verbs	3
Lesson 05	Verbs and time	6
Lesson 06	Important expressions and verbs	8
Lesson 07	Events and subjectivity (Part 1)	5
Lesson 08	Notice to participants	5
Lesson 09	Events and subjectivity (Part 2)	3
Lesson 10	Important functional expressions	3
Lesson 11	Expressing emotion and thinking	3
Lesson 12	Pronouns and indications	3
Lesson 13	Expressing questions and indefiniteness/concession	4
Lesson 14	Expressing negations	2
Lesson 15	Expressing admiration/emphasis	2
Lesson 16	Phrase expansion	3
Lesson 17	Modifier and adjective expression	5
Lesson 18	Connecting phrases/verses	2
Lesson 19	Word class conversion	5
Lesson 20	Word order and meaning	1

An outline of each of the above lessons is given below:

Lessons 1 and 2 provide a basic introductory framework for understanding linguistics and grammar.

Lesson 1 explains the significance of the birds' eye view of languages, the objective of using the cross-linguistic educational materials, and rudimentary definitions of "sentence" and "grammar."

Lesson 2 introduces the basic grammatical concepts including "grammar," "parts of speech," "nouns and verbs," "phrases," "functional words," "basic word order," and so forth.

Lesson 3 discusses the way a phrase is structured, with a noun as its head. From a functional viewpoint, the noun has a role of explaining "about what" in sentences.

In order to carry out the reference function, certain "limitations" are placed on nouns. In Lesson 3, these limitations, such as "definite/indefinite," "collective/specific," "spatial distance," "owner," "quantity/order," and "specific form," are displayed in hyperlinks which indicate the actual noun phrase constituents of each language.

Also, nouns in some languages possess purely formal gender and class. In addition, some languages have a case system to show semantic relations between nouns and other linguistic forms, in particular predicate verbs. Gender, class and case are combined in some languages to function as the agreement that shows cohesiveness within noun phrases. Lesson 3 gives an explanation of these formal features.

Lessons 4 and 5 focus on how a phrase is structured, with a verb as its head. Verbs have the function of predication, or describing “what sort of situation, or situational changes” regarding the referent noun. For this purpose, verbs are limited in several ways. In Lesson 4, “limitations of subject” and “limitations by means of other participants” are discussed. Since some languages have developed rich systems for tense, aspect and so on, “limitations through temporal relation between utterance and event” for verbs are described separately in Lesson 5.

Lesson 6 gives an overview of important expressions. Expressions including basic verbs tend to have exceptional sentence constructions. In Lesson 6, various expressions are introduced, especially concerning “presentation,” “ownership,” “existence and presence,” “identification assessment,” “natural phenomena,” “movement” and “time and seasons.”

Lessons 7 and 9 cover expressions of “events and the speakers’ subjectivity related to them.” Depending on the language, the psychological attitude of the speaker is asserted mainly through verbal inflections and sentence final particles.

In Lesson 7, “indicative expressions,” “conditional expressions,” “imperative expressions,” and “hearsay/speculative and indirect expressions,” concerning mood and modality are introduced.

Lesson 8 shows how the participant of an event is focused on and expressed in a sentence. This is often expressed as “voice” through verbal inflection systems and auxiliary verbs. Lesson 8 covers “active voice,” “passive voice,” “middle voice” and “causative expression.”

Lesson 9 is similar to lesson 7 with the expressions for “subjectivity of the speaker related to events,” but it contains expressions with auxiliary verbs and combinations of several verbs (verb serialization).

Lesson 10 takes up expressions regarding “ability/possibility,” “hope and demand,” and “duty/necessity,” which are among the important functional expressions using modal auxiliary and subsidiary verbs.

Lesson 11 contains expressions that are often exceptional cases concerning sentence construction and affix/case marking in many languages, such as “emotion and thinking,” “perception and knowledge/understanding” and “the condition of emotion/feeling/body.” These are not only exceptions to morphosyntactic patterns, but also exceptions in that they are limited to certain predicate verbs.

Lesson 12 deals with deictic indications by means of pronouns. Pronouns in themselves are usually free forms, but in some cases, they can be clitics added to a verb. The latter pronouns are considered to be a formal expansion of verbs.

Lessons 13 to 15 introduce important functional expressions including “questions,” “indefiniteness and concessions,” “negations” and “exclamation/emphasis.”

As explained above, while Lessons 1 through 15 deal with how nominal (or pronominal) and verbal structures are given limitations from the functional viewpoint, Lesson 16 and subsequent lessons describe how the structures are given formal expansions.

Lesson 16 deals with adpositions: prepositions and postpositions. Cases when the adpositional phrase co-occurs with a predicate and assumes a specific semantic role are also described.

Lesson 17 mainly takes up adjectives that act as modifiers to limit quality/volume/manner of the referent noun. In addition, we show the case of nominal expansion with added clauses.

Lesson 18 explains “assumptive and conditional clauses” through conjunctions connecting noun phrases, verb phrases and clauses.

Lesson 19 describes conversion of word classes, or parts of speech, regarding languages which explicitly classify word classes by means of word forms. The lesson includes “nominalization of verbs,” “infinitives” and “participles.”

While the lessons so far are concerned mainly with the formation and expansion of phrases, focusing on nouns and verbs from a morphological and functional viewpoint, lesson 20 gives expressions that have their meaning by means of word order, or syntactic combination of words.

Conclusion

This paper reports on grammar educational materials that are currently under development, with emphasis on the objective, development process and structure of cross-linguistic materials.

Linguistic typology, which is the study of the diversity and universality of language, might seem somewhat distant from the practical field of foreign language education. In the development of these educational materials, we attempt to provide foundational materials for linguistic typology and general linguistics obtained by generalizing the content of grammatical study of each language from the viewpoint of linguistic typology.

With developing the materials, we abstracted the grammar for the many languages included in the educational materials, and through that process, we have acquired a method of generalizing grammatical categories and their regional distribution; this in itself is a heuristic method based on trial and error. In this sense, the cross-linguistic educational materials described, rather than being drawn from established educational content, should be considered as still being under development, or as a mid-term report on the typological research being conducted. In the future, further improvements of the cross-linguistic grammar materials are expected through integration of other language education materials.

References

Bloomfield, Leonard: Language. London: George Allen and Unwin. 1933. Rev. Edition 1935.

Nichols, J. 1986. “Head-marking and Dependent-marking Grammar.” Language 62.56-119.

Sapir, Edward: Language. New York: Harcourt, Brace and World. 1921.

de Saussure, Ferdinand: Course in General Linguistics. Translated by Harris, Roy. London: Gerald Duckworth and Co. Ltd. 1983.

Introducing a Task Activity for Less Proficient Learners — Enhancing the Relationship among Form, Meaning and Use —

Hide TAKASHIMA, Chihiro INOUE, Chiaki YAMANE, Natsuko UZAWA,
Mayo NAGATA, Takayuki SADAHIRO and Yukiko SHIMAMURA

0. Introduction

In recent years, task-based language teaching (TBLT) has received wide recognition for its practicality and effectiveness (e.g., Nunan 1989; Long and Crookes 1992; Willis 1996; Skehan 1996; Bygate, Skehan, and Swain 2001; Ellis 2003). This has resulted in a shift from a form-focused to a more meaning-focused method of instruction in language teaching (Richards and Rodgers 2001). As this shift suggests, it is essential to teach learners not only grammar, but also the meaning of the target structure by encouraging them to *use* the language. However, it is very difficult to apply TBLT in English as a foreign language (EFL) countries like Japan[1].

This difficulty is attributed to two factors. First, Japan is an EFL country. As opposed to the situation in an English as a second language (ESL) country like the Philippines, English is not a necessity in the daily lives of the Japanese; this leads to the lack of input and output as well as of interactions, all of which makes it difficult to accomplish Tasks[2] outside the classroom (Takashima 2005). Since learners face no immediate need to use English — aside from learning it to pass the written examinations for entrance into upper schools — their motivation for studying and mastering the language is not strong enough, especially with regard to speaking.

[1] In one study, tasks were successfully introduced experimentally at the senior high school level (the technical college level) in Japan. Sugiura (2006) investigated the applicability of two different kinds of activities — tasks and task activities — as a means of improving the participants' understanding of the present perfect and the past tenses. She concludes that in order for tasks to work effectively in improving the learners' usage of particular structures, focused activities, such as Task Activities, should be introduced first. Thus, it is necessary for learners to successfully engage in these task activities before attempting regular tasks.

[2] Capitalized nouns such as Tasks and Task Activities take on a pedagogical meaning in his paper in that they are all language activities in the classroom which are specifically designed to enhance students' grammatical accuracy as well as their appropriate use of structural forms.

Furthermore, as learners advance from junior to senior high school, an increasing number of English classes use the Grammar-Translation Method because it encourages success on the entrance examinations. Even classes that seemingly impose oral practice often do not advance beyond reading aloud, so opportunities for meaningful interaction, if any arise, are rare; this results in a patterned practice. In these situations, learners will naturally be unable to acquire both the form and the function of the target structure, both of which elements are necessary to complete a task.

Second, teachers have to complete a structure-based textbook screened by the Ministry of Education in accordance with *the Course of Study,* the compiled national guideline specifying the content to be taught, within a limited number of lessons. This number of lessons consists of three classroom hours per week in junior high schools and approximately five in senior high schools. As a result, although the learners are able to complete the Drills and Exercises that appear at the end of each unit, they cannot use these structures freely when they appear in other units or appropriately combine them with the other structures addressed in class[3]. In other words, the learners are incapable of utilizing Grammaring, which is described by Larsen-Freeman (2003) as the ability to use grammatical structures accurately, meaningfully, and appropriately. Therefore, it is vital to provide opportunities for the learners to realize that no matter how subtle the differences, different forms convey different meanings.

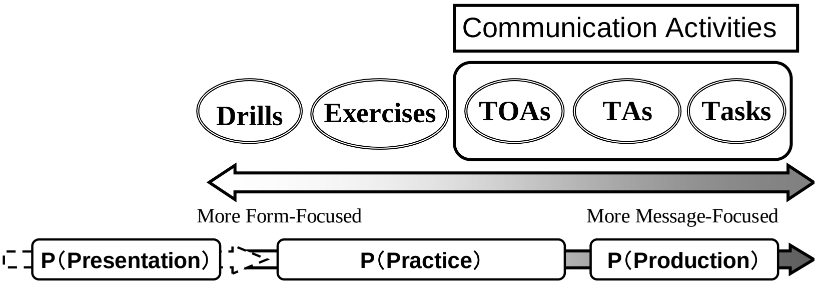
What then can we do to ensure that the learners can communicate in a more meaning-focused way? We believe the answer lies in a Task Activity (TA). A TA is a simulated pair/group activity intended to force learners to exchange information in English using the structure(s) targeted in the lesson. In other words, the Tasks are simulations of real-life activities, which are usually managed in terms of tasks. A TA is an activity specifically designed to bridge the gap between Tasks and explicit, structure-based instruction. In this paper, we propose that TAs be activities that are necessary to attain or restructure the grammatical knowledge gleaned from basic Drills and Exercises. The following sections discuss the characteristics of TAs by providing a sample TA.

[3] This is especially true, for instance, of the distinctions between the uses of the present perfect and the past tense forms, or conditional and subjunctive sentences. In Japanese, which is an aspectual language, “I’ve finished my homework” and “I finished my homework” are translated in a similar manner; this is similar to the translations of “If I win the lottery, I will buy a big house” and “If I won the lottery, I would buy a big house.”

1. Characteristics of a Task Activity

How is a TA different from a Task? Let us first examine the definition of a “Task.” According to Skehan (1996: 20), “tasks are activities which have meaning as their primary focus. Success in tasks is evaluated in terms of achievement of an outcome, and tasks generally bear some resemblance to real-life language.”

Tasks are divided into two types — unfocused and focused — according to the degree to which they focus on meaning and/or form (Ellis 2003). Thus, unfocused ones are focused only on message. When Takashima (2005: 50) categorizes the major language activities in the classroom into five main types, he places Drills, which are mostly form-focused activities at the farthest left end of the continuum and Tasks at the farthest right. Here, both unfocused and focused forms are combined in the term Tasks. (See Figure 1.)



*from Takashima (2005)

Figure 1. The five main types of classroom language activities and their relationship with the 3Ps

Among the major language activities, five of them, though their boundaries are blurred, can be placed on the continuum according to the degree, again, to which the activities involve focus on form or meaning. Three of them, Drills, Exercises and Task-Oriented Activities (TOAs), roughly correspond to the practice level of the traditional 3P approach, while the remaining two TAs and Tasks, to the production level. We now provide a detailed explanation of the five activities.

Drills, which are located on the far left of the continuum, strictly emphasize the learning of forms and are represented by pattern practices and repetitions. Exercises, such as gap-filling and translation problems, lie to the immediate right of Drills and place more importance on meaning. Examples of a Drill and an Exercise are provided below.

Example 1. Drills

Teacher: I get up at six every morning.
(Use *Kim*.)
Learners: Kim gets up at six every morning.
Teacher: I have breakfast at seven.
(Use *she*.)
Learners: She has breakfast at seven.
…

Example 2. Exercises

Learners Questions	1	2	3
What time do you get up every morning?			
What time do you leave home?			
What time do you arrive at school?			

In the example Drill, learners are required to listen to the teacher's cue and replace the subject of the sentence. The learners must then conjugate the form of the verb in accordance with the new subject. In the above Exercise, the teacher prompts the students to question three of their peers and record their answers on the grid. Both Drills and Exercises have similar purposes for the students — to form associations and memorize the structures presented in class. However, the other three types of activities, TOAs, TAs, and Tasks, focus more on meaning and the production of language.

Of these three activities, Tasks, located at the far right of the continuum, most closely approximate real communication. As part of a Task, learners are required to complete a problem-solving activity, for example, comparing two tours and deciding which one to join through discussion. The focus is solely on the conveyance of meaning, and learners speak freely using pieces of their knowledge. However, one limitation of Tasks is that learners with low proficiency or fluency may face difficulties and be unable to express their thoughts. There is a large gap between the proficiency level necessary for performing Tasks and that of most Japanese students, who usually do not advance beyond the level of practice Drills and Exercises.

Compared to Drills and Exercises, TOAs are activities that place greater emphasis on meaning. In principle, TOAs are problem-solving activities that focus on one target structure and provide learners with a model dialogue and the necessary vocabulary; as a result, they force learners to focus mainly on form. Although the use of TOAs has long been the goal of communicative activities in Japan's English education, there is a sizeable gap between the level of English proficiency required to complete TOAs and that required to complete Tasks, which do not provide model dialogues or require the creative use of language. In order to bridge this gap between TOAs and Tasks, we propose the use of TAs, which provide several steps for the learners to follow, enabling them to gradually attain the desired level of communication.

This form-focused activity was created in order to compensate for the lack of a more meaning-focused activity while simultaneously allowing learners to compare more than two difficult, related structures and differentiate their uses in the context of a discourse. TAs are meaning-focused in the sense that they provide neither model dialogues nor instruction on how to use a particular structure in the activity. Instead, they merely offer instructions to follow in order to complete the given task; thus, they are called Task Activities. On the other hand, TAs are form-focused in that they are designed to require learners to use particular opposing structures (such as the present and the past tense forms), or grammatical items within the same category (such as modal auxiliaries) to complete the task or goal. Since TAs are a kind of Task, some of the particular structures on which the activity is intended to focus may occasionally not be used; in these cases, feedback from the teacher is necessary (Loschky and Bley-Vroman 1993).

In a typical TA, learners are given a situation, for example, “You are a customer at a travel agency. You only have $40, but you want to see as many sights as possible,” and instructions such as “(1) Introduce yourself to the travel agent and (2) Find out the characteristics of each tour.” The steps are designed to require learners to accurately use two or more similar structures (such as interrogatives: “how much,” “how many,” “what time,” etc.) in order to complete the given task; however, the learners are not explicitly told which structures to use. This helps them to realize that the use of inaccurate forms results in inaccurate meaning, which, in turn, hinders the completion of the TA. This comparison and contrast of structures allows learners to focus on both form and function and provides them with an opportunity to enhance both accuracy and fluency.

The characteristics of TOAs, TAs, and Tasks are similar and can be compared from nine different perspectives, as is seen in Chart 1. With the exception of perspectives (e) through (h), all three have similar characteristics. This is due to the fact that they are not independent of one another but part of the same continuum. As the activities move from TOAs to TAs to Tasks, they more closely approximate real communication and place greater emphasis on meaning.

Chart 1. Comparison of the main characteristics of TOAs, TAs, and Tasks

The nine perspectives	TOA	TA	Task
(a) Emphasis placed on the conveyance of meaning	○	◎	◎
(b) Emphasis placed on the completion of a goal and its process	○	◎	◎
(c) Transfer and exchange of information	◎	◎	◎
(d) Information gap	◎	◎	◎
(e) Instructions with steps	◎	◎	×
(f) Model dialogue and specification of a target structure	◎	×	×
(g) Comparison of two or more structures	×	◎	○
(h) Negotiation of meaning	△	◎	◎
(i) Emphasis placed on interest and motivation	○	◎	○

(◎ and ○ denote necessary conditions. ◎ is indispensable, ○ is desirable, △ is dispensable, and × is unnecessary.)

(Based on Takashima 2005)

We believe that TAs, which have the potential to enable learners to achieve the level of language proficiency necessary to complete authentic tasks, are the most efficient type of language activity; thus, they should be actively adopted and practiced in EFL environments.

The main characteristics of a TA can be organized according to the following six points:

1. Focuses on meaning
2. Uses the target language as the ultimate aim of task completion
3. Requires negotiation of meaning for task completion
4. Involves a comparison of two or more structures
5. Presents an information gap between the learners
6. Contains content and activities of interest to the learners

It is worthwhile to note that the "comparison of two or more structures" plays a crucial role in this activity because it also induces cognitive comparison, that is, the comparison of what learners know and do not know (Ellis 1995). In the next section, we will provide an example of an actual TA and explain how the six characteristics mentioned above are fulfilled.

2. Example of a TA and Its Validity

In this section, we will narrow our focus to an example TA that we developed ourselves (see Appendix for the Japanese version and its English translation). This TA is based on a role-play activity; therefore, it involves two activity sheets — Sheet A for one learner playing the part of *Bandolf the Wizard,* and Sheet B for another playing *Frado the Brave.* This particular TA targets beginners who have not mastered to-infinitives. The target structures are the three usages of to-infinitives.

In Japan, to-infinitives are classified into three types according to their function and are addressed separately. The following are the three types and examples of them:

1. To-infinitives used as noun phrases (I want *to eat.*)
2. To-infinitives used as adjective phrases (I want something *to eat.*)
3. To-infinitives used as adverb phrases (He went *to eat* lunch with her.)

However, because of the similarities among the three types, learners are often confused and find it difficult to use them distinctively. Our TA is designed to provide learners with opportunities to accurately use the three types of to-infinitives while focusing on their respective meanings.

Next, we will explain how the example TA in this paper meets the six characteristics mentioned in the previous section.

First, this TA can be called meaning-focused because it requires one set of learners to choose three items they want, explain why they want them, and persuade the others to buy them. For example, a learner playing the role of Bandolf may suggest, “Let’s buy a magic carpet”; explain, “I want the carpet because I want to fly”; and proceed to argue, “We can travel far on the carpet. It is also cheap.” Although learners are required to use to-infinitives to convey meaning, they remain focused on meaning.

Second, “task completion” is achieved when the final decision on what to buy is made within the stipulated amount of time (10 minutes). Therefore, it can be said that for the learners, the ultimate purpose is to accomplish the task.

Third, “negotiation of meaning” is necessary for task completion. Both sides wish to purchase several items, and each individual has certain conditions that he/she cannot relinquish. For example, Frado might wish to buy items that will benefit both him/her and Bandolf. There is also a limitation on the amount of money, and the pair must negotiate to arrive at a decision with which they are both satisfied. In addition, the presence of low-proficiency learners often causes misunderstandings between the pair, which leads to the need for negotiation of meaning.

Fourth, the example TA is designed to require learners to distinguish the uses of “two or more structures,” in this case, three types of to-infinitives, although this is not explicitly stated in the instructions. Learners need to compare functions, establish accurate and appropriate uses according to the situation at hand, and make any necessary modifications. Examples of several possible utterances that include to-infinitives are shown in Chart 2. It is worthwhile to notice that both roles require learners to use to-infinitives to complete the task.

Chart 2. Possible utterances that include to-infinitives

Sheet Step	Sheet A (Bandolf)	Sheet B (Frado)
STEP 1	・I'll be glad to go.	・Please go to the store to buy me something.
STEP 2	・I'd like to buy three things.	
STEP 3	・I really want to buy the magic lamp.	・I want the red drink because I want to get well.
STEP 4	・Do you want anything to eat?	・I want something hot to eat and cold to drink.

Fifth, an "information gap" exists between the paired learners. Neither Bandolf nor Frado knows the characteristics and cost of the items that the other desires or the conditions that the other cannot relinquish (for example, it is stated in the instructions that Bandolf considers the magic lamp to be a necessity). Bandolf must also determine the kind of food and drink that Frado wants him/her to buy.

Finally, the content of the example TAs can draw the learners' attention, thus meeting the sixth characteristic. Role-play makes it easier to motivate less-proficient learners to speak. In addition, considering the popularity of role-playing video games, known among Japanese children as RPGs, the enjoyable and attractive nature of a role-play-based TA enhances its validity.

3. Evaluation by the Use of TAs

TAs, which elicit the usage of the target language, are beneficial not only for learning, but also for assessment purposes, as Ellis similarly suggests with regard to task-based assessment (2003: 279–80).

The aforementioned example TA was, in reality, implemented in a lower secondary public school in Kochi prefecture; based on the utterances monitored in the experiment, it is possible to suggest that TAs also help to assess the learners' accuracy in using the target language, particularly with regard to the structure(s) that the TA is intended to elicit. In our example TA, the target structure comprised the use of to-infinitives, and both correct uses and errors were observed. Correct uses include utterances such as "I want to ..." and "... anything to drink." Erroneous uses are comprised of those such as "... something to eat hot."

In addition to accuracy, fluency can also be assessed through TAs. Fluency can be divided into two types: linguistic and communicative fluency. Linguistic fluency can be assessed according to the utterances themselves, based on the number of meaningful utterances that the learners make during a fixed period of time; this fluency is evaluated in terms of the words per minute and/or type-token ratio of the utterances.

The other type of fluency, communicative fluency, can be assessed based on the completion of the TAs and behavior observed while the TAs are in progress. Communicative fluency includes linguistic fluency as one of its determinants but also embraces other elements — verbal communicative markers as well as non-verbal communicative skills. For instance, while our TA was being conducted, one learner repeated the same word(s), possibly to suggest that she could not recollect the appropriate vocabulary item. Another learner used hand gestures to indicate shifts in the conversation. It is obvious that these elements facilitate communication between the participants in the TAs and accelerate their completion; therefore, TAs can be used to assess communicative fluency.

A few additional suggestions are worth mentioning. One is the washback effect caused by TA-based assessment. In the response sheets of the example TA, one learner commented that she could not use an expression during the TA but remembered it afterward. It is often the case in experimental learning that the items one remembers well are those learned through mistakes. Since TAs can provide learners with opportunities not only to successfully use the target language, but also to fail to do so, they can trigger learning from mistakes and thus exhibit the washback effect as assessment instruments. It is also worthwhile to note the applicability of this kind of assessment for both summative and formative purposes (for a detailed discussion of summative and formative assessment, see Ellis 2003: 283–87).

4. Discussion and Conclusion: Further Possibilities for the Use of TAs

Our example TA was based on a role-play activity. It is neither a guided practice using functional dialogues nor a complicated version of a TOA; rather, it is a simulated task that drives learners to focus on the usage of particular contrasted structures within a natural, if sometimes unavoidably contrived, context. During the activity, learners were required to stop and consider which structure was appropriate in a certain context and which should be used over the other(s), doing so within the context of a more or less natural discourse that might occur outside the classroom.

We believe that TAs based on role-play activities have several benefits.

First, Japanese learners will consider the content of the TA interesting because role-playing video games are popular among Japanese youth, as already mentioned. If we create TAs based on a consistent story in the format of an RPG, learners will be interested in their content and focus on meaning when attempting to complete them. The learners will also be interested in the characters of the story and look forward to further work with TAs. In addition, since learners are used to RPGs, they will easily be able to understand the situation of the TAs. In this way, a TA series with a consistent story similar to that of an RPG can benefit the learners. However, it is important to note that TAs are not restricted to an RPG format. TAs are simulated pair activities and are also applicable to other standard pair activities (see Takashima 2000; 2005 for a variety of TAs).

Whether or not they are based on role-play, TAs are very important. They not only bridge the gap between Tasks and other activities, such as Drills and Exercises, but also verify whether the learners have understood how to convey messages based on the grammatical rules that they have already studied. When the learners fail to use the structures of the TA correctly, the teacher's role, among other things, is to help them recognize the discrepancy between their level and the target level. Feedback is crucial for the learners as well as for the teachers themselves because teachers can evaluate their own instruction according to how well the students are able to use the structures.

Although TAs were originally proposed for use with EFL in Japan, it should also be possible to apply them to ESL settings. Since TAs supply the learners with more opportunities to use English, they would be very effective for learners who speak English only in classrooms or for learners with low proficiency.

References

Bygate, M., P. Skehan, and M. Swain. (Eds.). 2001. *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. London: Longman.

Ellis, R. 1995. “Interpretation Tasks for Grammar Teaching”. *TESOL Q*, 29, pp. 87–105.

————. 2003. *Task-Based Language Learning and Teaching*. Oxford: Oxford University Press.

Larsen-Freeman, D. 2003. *Teaching Language: From Grammar to Grammaring*. Boston, MA: Thomson/Heinle.

Long, M. and G. Crookes. 1992. “Three Approaches to Task-Based Syllabus Design.” *TESOL Quarterly*. 26 (1), 27-56.

Loschky, L. and R. Bley-Vroman. 1993. “Grammar and Task-Based

Methodology”. In Crookes, G. and S. Gass (eds.), *Tasks and Language Learning: Integrating Theory & Practice*. Clevedon, UK: Multilingual Matters. pp. 123–167.

Nunan, D. 1989. *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press.

Richards, J. C. and T. S. Rodgers. 2001. *Approaches and Methods in Language Teaching* (Second Edition). Cambridge: Cambridge University Press.

Skehan, P. 1996. “Second Language Acquisition Research and Task-based Instruction”. In Willis, J. and D. Willis (eds.), *Challenge and Change in Language Teaching*. Oxford: Heinemann.

Sugiura, R. 2006. “Grammar Instruction Through Task Activities and Tasks in the EFL Context”. *Annual Review of English Language Education in Japan (ARELE),* 17, pp. 101–110.

Takashima, H. 2000. *Structure-Based Tasks and Grammar Instruction for Practical Communicative Competence*. Tokyo: Taishukan.

————. 2005. *Task Activities and Tasks for Form-Focused Instruction and Assesment*. Tokyo: Taishukan.

Willis, J. 1997. *A Framework for Task-Based Learning*. London: Longman.

Appendix. Task Activity

魔法使いのお買い物 -Sheet A-

あなたは勇者と旅する魔法使い・バウルです。勇者・ソファーが足に怪我したため、しばらく町に滞在することになりました。買いたいものがあるのですが、２人で使うお金を管理しているのは勇者です。相談に行って、何を買うか決めましょう。
※ □がついた番号のところは、あなたから会話を始めます。

1. ソファーの部屋へ行くと向こうから話しかけてきました。ソファーの話に答えましょう。

2. これはチャンス！ 買い物に行きたいと言ってみましょう。さらに、下の３つのものすべてと、それを買いたい理由と値段を伝えましょう。

買いたいもの

速く走るために欲しい！	空を飛ぶために買いたい！	もっと強くなるために買いたい！
魔法のくつ	魔法のカーペット	魔法のランプ
値段：1200G	値段：3000G	値段：2200G

3. あなたは、魔法のランプは絶対に買いたいと思っています。ソファーの話を聞いて下にメモをとり、何を買うか決めましょう。

☆ 何を買うことに決まりましたか？

【買うもの： 】

4. ソファーは話をして疲れたようです。何か食べる物、飲む物、遊ぶための物は欲しくないか、聞いてみましょう。あれば快く引き受けて、買い物に出かけましょう！ ・・・さあ、買い物へ！！

魔法使いのお買い物 -Sheet B-

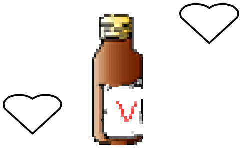
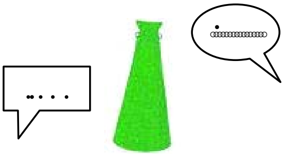
あなたは魔法使いと旅する勇者・ソファーです。足に怪我をしてしばらく町に滞在することになりました。動けないので買い物に行けません。魔法使い・バウルに買い物を頼みましょう。
※ □がついた番号のところは、あなたから会話を始めます。

1. バウルがやってきました。店へ行って買い物をして欲しいと頼みましょう。

2. バウルがあなたに相談があるようです。話をよく聞いて、下にメモをとりましょう。

3. あなたは２人のために、たくさんのものを買いたいと思っています。バウルに下の３つのものすべてと、それを買いたい理由と値段を伝えて、買うものを決めましょう。２人で使えるお金は現在 6000G で、あなたが管理しています。

買いたいもの

元気になるために欲しい！	動物と話をするために買いたい！	魔法の勉強をするために欲しい！
赤のドリンク	緑のメガフォン	黄の本
		
値段：1 本 800G	値段：1200G	値段：3500G

☆ 何を買うことに決まりましたか？

【買ってきてもらうもの： 】

4. 一生懸命話をしたら、小腹がすいてきました。何かぴりっと辛いものが欲しいし、冷たいものも飲みたいし、たいくつなのでゲームもしたいところですが・・・

・・・あとは魔法使いにまかせて、ゆっくり休みましょう！

Do an Errand for Frado —Sheet A— (translated version)

You are Bandolf, a wizard traveling with Frado the Brave.
Frado got his/her leg injured, so you are going to stay in this town for some time. You want to buy something, but Frado has all the money. Go to Frado's room to talk and decide what to buy.
* You start the conversation in the squared number.

1. When you enter Frado's room, Frado starts talking. Respond to Frado.

2. Now's a good chance to tell Frado that you want to go shopping! Tell Frado what three things you are thinking about, the reasons why, and their prices.

What You Want to Buy

To run fast! Magic Shoes	To fly in the sky! Magic Carpet	To be stronger! Magic Lamp
Price: 1200G	Price: 3000G	Price: 2200G

3. You want to buy the Magic Lamp by all means. Listen to Frado, take notes, and decide what to buy.

☆ What did you decide to buy?

<You decided to buy: >

4. Frado seems a bit tired. Ask if Frado wants something to eat, drink, or play with. If Frado does want anything, add it to your list.

・・・Off to your shopping!

Do an Errand for Frado —Sheet B— (translated version)

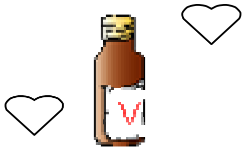
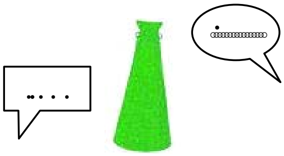
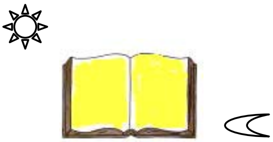
> You are Frado the Brave, traveling with the wizard, Bandolf.
> You got your leg injured, so you are going to stay in this town for some time. As you cannot move, you cannot go shopping. Ask Bandolf to go shopping for you.
> * You start the conversation in the squared number.

$\boxed{1.}$ Bandolf comes into your room. Ask Bandolf to go shopping.

2. Bandolf seems to want to talk about something.
Listen carefully and take notes below.

$\boxed{3.}$ You want to buy what you both can use and as many things as possible. Tell Bandolf the following three items, the reason you want to buy them, and their prices. You have all the money, 6000G.

What You Want to Buy

To be fine!	To talk with animals!	To study Magic!
Red Drink	Green Megaphone	Yellow Book
		
Price: 800G per 1	Price: 1200G	Price: 3500G

☆ What did you decide to buy?

<You decided to buy: ____________________________________ >

4. You feel a little hungry. You want something hot to eat and cold to drink. Also, you want to play a game since you are bored. So, tell Bandolf what you want.

・・・Now, leave the shopping to Bandolf and have a good rest!

The Relationship between VOT in Initial Voiced Plosives and the Phenomenon of Word-Medial Plosives in Nigata and Shikoku

Mieko TAKADA and Nobuo TOMIMORI

1. Introduction

This study analyzes the voice onset time (VOT) in word-initial voiced plosives in the Shikoku area and northeastern Nigata prefecture in order to reveal regional and generational differences in these regions. Further, correlations between taking positive values in word-initial voiced plosives. (that is, those that are “slightly voiced” in Catford 1988) and nasalization or voicing in word-medial obstruents are examined and compared with the results of previous studies from Tohoku to Kanto.

2. Previous studies

In phonetics, plosive consonants that have a contrast between voiced and voiceless have been studied physiologically, articulatorily, acoustically, and perceptually in many languages because numerous world’s languages contain this contrast. However, few studies examine internal variations within each language.

VOT is one of the acoustic characteristics that are usually defined as the time interval between voice onset and the release of occlusion (Lisker & Abramson 1964). This acoustic characteristic has been widely accepted as one of the basic features to distinguish between voiced and voiceless consonants (Lisker & Abramson 1964, Shimizu 1996, etc.). Lisker and Abramson (1964) reported that the VOT of voiceless plosives has a positive value, while that of voiced plosives usually has a negative value in many languages.

In Japanese, previous studies have supported the tendency of the VOT value shown by many other languages (Kobayashi 1981, Shimizu 1993, 1996, etc.). However, other studies have simultaneously reported that the VOT value of Japanese voiced plosives varies widely and may even assume positive values (Homma 1980, Sugitoo 1996).

Voiced plosives without prevoicing are also observed in English and German and are referred to as “half-voiced” (Hattori 1984, Kamei et al. 1996), or “slightly voiced” (Catford 1988). In this study, this type of voicing will be referred to “slightly voiced,” as opposed to “fully voiced,” which describes the sounds with prevoicing. Although the existence of slightly voiced in Japanese has been observed in previous studies, the social factors and the actual conditions of the phenomena have not been reported.

In sociolinguistics, especially in Japanese dialectology, plosive consonants that have voiced and voiceless contrasts are a major topic of study because they show regional and generational variation. However, simple auditory impression without a detailed acoustic analysis was used as the observational method in most studies.

Needless to say, this observational method applies only to phenomena that are recognizable through the auditory mechanism. In Japanese, audible variations of consonants with voicing contrasts are found in the word-middle position but not in the word-initial position; some regions have nasalization of the voiced plosives /b, d, g/ and the fricative /z/ in the word-medial position, and the voicing of the voiceless plosives /t, k/ in word-medial position. Most studies have addressed the word-medial phenomenon, but few studies have addressed the word-initial phenomenon.

Takada (2004 and 2006) studied the word-initial phenomenon from an acoustic perspective and revealed the existence of generational and regional differences. Specifically, Takada (2004) focused on the relationship between the slightly voiced word initial /d/ and the year of the speaker’s birth in the Kanto area. The results revealed a strong correlation between the year of birth and voicing.

Takada (2006) dealt with regional and generational variations in the VOT in the word-initial voiced plosives /b, d, g/ from the Tohoku to Kanto areas. The data of the study were obtained from the speech data of Inoue and his collaborators in 1986–89 (hereafter, “Inoue data”).[1] According to the result of the study, there exists a regional difference between the Tohoku and Kanto areas, the boundary of which runs through Tochigi and Ibaraki prefectures, which, from the northwest to the southeast. These prefectures are conventionally rabeled Kitakanto area. Moreover, in the Tohoku area, there are no generational differences, and the VOT is standardized as a “positive value,” which indicates that voiced plosives are pronounced as “slightly voiced.” On the other hand, there are generational differences in the Kanto area. For speakers born in the early 1900s, the VOT was standardized as a “negative value,” which implies that voiced plosives are pronounced as “fully voiced.” In contrast, for speakers born around the 1970s, the VOT

[1] *Nihongo onsei no chiikisa, sedaisa no oninron-teki, onkyougaku-teki bunseki (Shouwa 61-63-nendo Monbu-sho kagakukenkyu-hi hojokin sogo (A))* “Phonological and Acoustic Analysis of Geographical/Generational Differences of Japanese Sounds (1988 Grant-in-Aid for Scientific Research (A))”) Inoue (1989).

values vary, that are often positive. The result from the Kanto area corresponded with those of Takada (2004). The Kitakanto area is not only located geographically between the Tohoku and Kanto areas but also showed initial voiced plosive values in the midrange of these two regions.

Furthermore, the results of an extensive analysis of the Kitakanto data indicate a possible boundary between the Tohoku and Kanto areas, which corresponds to the phonetic/phonological boundary indicated in many previous studies (Kindaichi 1954, Tatara 1959, Kato & Inoue 1970, etc.).

The results also indicate that slightly voiced word-initial plosives behave in the same way as the phenomenon in Tohoku. Tohoku dialect has some characteristic phonetic/phonological phenomena. For example, according to Kato(1975), (1) phonetic or phonological contrasts between the vowels /i/ and /e/ or /i/ and /u/, (2) the sounds of hiatus /ai/ or /ae/, (3) the nasalization of *Dakuon* (the voiced obstruents /b, d, g, z/) in the word-medial (V$\underline{\text{C}}$V) position, and (4) the voicing of *Seion* (the voiceless obstruents /p, t, k, s/) in word-medial (V$\underline{\text{C}}$V) position, and so on. In particular, the nasalization of *Dakuon* in the word-medial position and the voicing of *Seion* in the word-medial position share a point in common with slightly voiced in word-initial voiced plosives—all of them are phenomenon that relate to the voiced characteristics of consonants. Also, a kind of co-occurrence was revealed between nasalization and voicing in the medial position (the former enables the occurrence of the latter) (Inoue 1971, Kato 1975, etc.). Thus, it is possible to expect a type of co-occurrence between the word-initial phenomenon and the word-medial phenomena.

3. Purpose of this study

Based on the results of the aforementioned previous studies, this study examines whether or not a co-occurrence exists between slightly voiced in word-initial voiced plosives and nasalization in word-medial voiced obstruents and/or voicing in word-medial voiceless obstruents.

In this study, the term word-medial nasalization refers to the following three phenomena of the word-medial voiced obstruents /b d g z/:

1a nasalization of the vowel preceding the obstruents
(e.g., /ageru/ “give” → [ãgeru])
1b prenasalization of the obstruents themselves
(e.g., /ageru/ “give” → [a^{ŋ}geru])
1c nasalization of the obstruents themselves. Note that this applies only to /g/.
(e.g., /ageru/ “give” → [aŋeru])

The term word-medial voicing refers to the following phenomenon of the word medial voiceless obstruents /t k/:

2a voicing of the obstruents themselves
(e.g., /akeru/ “open” → [ageru])

The possible patterns of their co-occurrence are listed as (1) to (4).

1) Word-initial slightly voiced occurs, but neither word-medial nasalization nor word-medial voicing occurs.
2) Word-initial slightly voiced co-occurs with both word-medial nasalization and word-medial voicing.
3) Word-initial slightly voiced co-occurs with word-medial nasalization only.
4) Word-initial slightly voiced co-occurs with word-medial voicing only.

In order to discuss these patterns of the co-occurrence, this study will analyze data from regions in which word-medial nasalization or word-medial voicing are reported to exist. The specific regions are the Shikoku area, where the dialect has word-medial nasalization, and northeastern Nigata prefecture, where the dialect has both word-medial nasalization and voicing. Slightly voiced word-initial voiced plosives have not been previously studied in either of these regions.

4. Methodology

4.1. Data

The data used in this study are taken from Inoue data.[2] The data were collected from approximately 700 speakers between 1986 and 1988 (see Inoue 1989). We report the results of only 39 of the 700 speakers. At that time, the 39 speakers comprised high school students and people of their grandparents’ generation, who were born and raised in their hometowns. Data consisted of surveys and tape recordings facilitated by high school teachers.

Fifteen words that have the voiced plosives /b, d, g/ in the initial position were analyzed in this study; these are shown in Table 1.

[2] The same data set was used in Takada (2005).

Table 1. Analyzed words

Word-initial Consonants	Words
/b/	/biwa/'Japanese lute', /bero/'tongue', /baka/'silly', /boro/'rag', /buta/'pig'
/d/	/deguti/'exit', /dango/'dumpling', /doku/'poison', /daikon/'radish', /doozoo/'bronze statue'
/g/	/gin/'silver', /geta/'Japanese wooden clogs', /ga/'moth', /go/'five', /gunkan/'naval ship'

4.2. Regions and speakers

Figure 1 depicts the geographical distribution of word-medial nasalization and voicing as reported by Kato (1975: translated by Takada). The area encircled in black is the region analyzed in this study: northeastern Nigata prefecture (hereafter, "Nigata") and the area of Shikoku that centers around Kochi prefecture (hereafter, "Shikoku"). Tables 2 and 3 list the speaker characteristics separated by region.

The two ellipses drawn with solid broken lines show research areas of this study: i) northeastern Nigata prefecture and ii) relevant regions in Shikoku area.

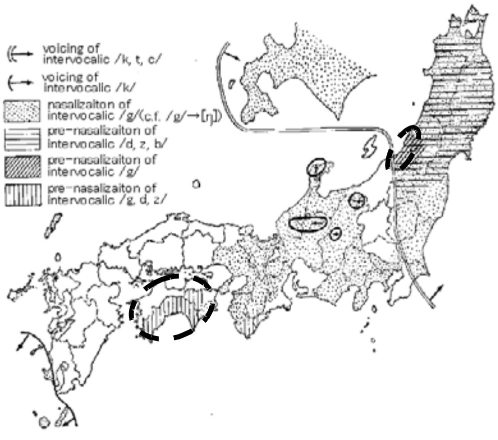


Figure 1. Distribution of word-medial nasalizaition and voicing of Japanese. (from Kato (1975) Figure4)

* Dashed line is added by Takada.

* In Kato(1975), the condition of nasalization was "after vowel."

Table 2. Speakers in Nigata.

Prefecture	Generation	No.of speakers	City or county	Gender	Year of birth	Age
Nigata	Grandparents	N-o1	Toyosaka city	Female	1919	69
		N-o2	Kitakanbara county	Male	1921	66
		N-o3	Kitakanbara county	Male	1921	65
		N-o4	Higashikanbara county	Female	1928	60
		N-o5	Kitakanbara county	Female	1931	56
		N-o6	Shibata city	Male	1940	47
	Grandchildren	N-y1	Kitakanbara county	Male	1968	18
		N-y2	Kitakanbara county	Male	1968	18
		N-y3	Kitakanbara county	Female	1970	18
		N-y4	Shibata city	Female	1970	17
		N-y5	Toyosaka city	Female	1971	16

Table 3. Speakers in Shikoku.

Prefecture	Generation	No.of speakers	City or county	Gender	Year of birth	Age
Kagawa	Grandparents	Ka-o1	Mitoyo county	Female	1912	76
		Ka-o2	Mitoyo county	Female	1919	69
		Ka-o3	Mitoyo county	Male	1920	67
	Grandchildren	Ka-y1	Mitoyo county	Female	1970	17
		Ka-y2	Mitoyo county	Male	1970	17
		Ka-y3	Mitoyo county	Male	1971	17
Tokushima	Grandparents	T-o1	Naka county	Female	1910	78
		T-o2	Naka county	Male	1913	75
	Grandchildren	T-y1	Naka county	Male	1969	18
		T-y2	Naka county	Male	1969	17
Ehime	Grandparents	E-o1	Onsen county	Female	1911	78
		E-o2	Onsen county	Male	1919	69
		E-o3	Kitauwa county	Female	1923	64
	Grandchildren	E-y1	Iyo city	Female	1970	17
		E-y2	Onsen county	Female	1970	17
		E-y3	Onsen county	Male	1971	16
Kochi	Grandparents	Ko-o1	Hata county	Male	1901	85
		Ko-o2	Nakamura city	Male	1902	84
		Ko-o3	Hata county	Female	1912	75
		Ko-o4	Sukumo city	Female	1912	74
		Ko-o5	Nakamura city	Female	1912	74
		Ko-o6	Nakamura city	Female	1915	72
		Ko-o7	Hata county	Female	1916	71
		Ko-o8	Nakamura city	Male	1916	70
	Grandchildren	Ko-y1	Nakamura city	Male	1969	17
		Ko-y2	Nakamura city	Female	1969	17
		Ko-y3	Nakamura city	Male	1969	17
		Ko-y4	Hata county	Female	1969	17

This section provides an overview of each dialect. Nigata prefecture is known to be a prominent boundary between east and west Japanese dialects; the dialectal variety of northeastern Nigata prefecture belongs to Tohoku dialect, unlike other Nigata varieties. In particular, the phonetic/phonological boundary lies near the Agano River, and the speech sounds in the northeastern region indicate Tohoku dialectal characteristics (Kato 1958, Iwai 1983). The same is true for the phenomena of word-medial nasalization and voicing. Therefore, in this region, the VOT in word-initial voiced plosives is expected to be similar to that in the Tohoku or Kitakanto areas analyzed by Takada (2006).

Shikoku contains two distinct dialects: (1) the Asanyo dialect, which contains the dialects spoken in Ehime, Kagawa, and Tokushima prefectures, and (2) the Tosa dialect, which contains the dialects spoken in Kochi prefecture (Doi 1975, Doi & Hamada 1985). Word-medial nasalization has been found mainly in the Tosa dialect and also in parts of the Asanyo dialects. Word-medial voicing was not found in either dialect. Thus, the analysis of these dialects might reveal a co-occurrence between word-initial slightly voiced and word-medial nasalization, without the effects of word-medial voicing.

Before the acoustic analysis of the VOT of initial plosives, the current data set (39 speakers) was subjected to preliminary auditory and acoustic observations to determine whether the conditions of nasalization and voicing occurred in the sample. The results are tabulated in Table 4.

According to the results, in the grandparents' generation, nasalization is relatively common in both Nigata and Shikoku, while voicing in Nigata varies among individuals. On the other hand, in the grandchildren's generation, nasalization and voicing are observed in neither Nigata nor Shikoku.

Table 4. Word-medial Nasalization and Voicing in Nigata and Shikoku.

	Nasalization			Voicing		
	Previous	This Data		Previous	This Data	
Nigata	○	GP	○	○	GP	△
		GC	×		GC	×
Shikoku	○	GP	○	×	GP	×
		CH	×		GC	×

○=often found ×=not found △=sometimes found

GP=Grandparents' generation GC=Grandchildren's generation.

4.3. Methodology for the acoustic analysis

The analog data in cassette tapes are transformed into digital data (48 KHz sampling rate, 16 bit quantization rate), and are analyzed using the acoustic software *SIL Speech Analyzer* (version 2.4 test 3.6). The VOT is judged by measuring the length between burst and voice onset as represented by the wave form and spectrogram. Figures 2 and 3 illustrate a typical example of the wave form and spectrogram of word-initial voiced plosives. Figure 2 is an example of a negative VOT value, and Figure 3 is an example of a positive VOT value.

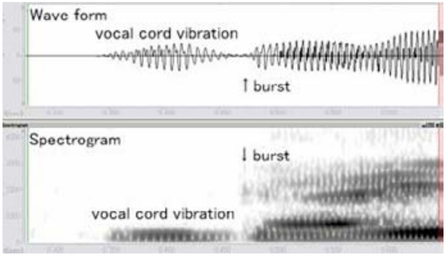


Figure 2. Example of wave form and spectrogram: sound taking negative VOT value. (Ko-o4’s /d/, VOT = -128)

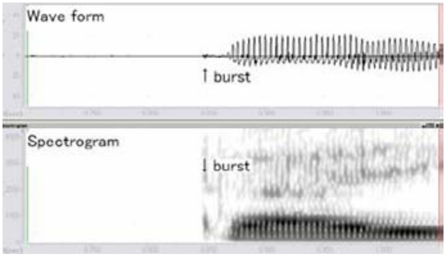


Figure 3. Example of wave form and spectrogram: sound taking positive VOT value. (Ko-y2’s /d/, VOT = +13)

5. Results and discussions

Figures 4 and 5 show the relationship between the VOT values of word-initial voiced plosives and the speaker’s year of birth by means of scatterplots for Nigata and Shikoku, respectively. The VOT values are represented on the x-axis; the speaker’s year of birth, the y-axis. No differences were observed within the region in either Nigata or Shikoku.

The main difference between the results for Nigata and those for Shikoku is reflected in the speakers born in the early 1900s (grandparents’ generation). In particular, the amount of the positive VOT value data concerning slightly voiced pronunciations differs between Nigata and Shikoku. In Nigata, the grandparents’ generation shows more data with positive values than is reflected in Shikoku. This is also true in both Nigata and Shikoku for speakers born around the 1920s, who enabled a direct comparison of regions. In other words, while the word-initial voiced plosives are often expressed as slightly voiced in Nigata, they are standardized to fully voiced in Shikoku.

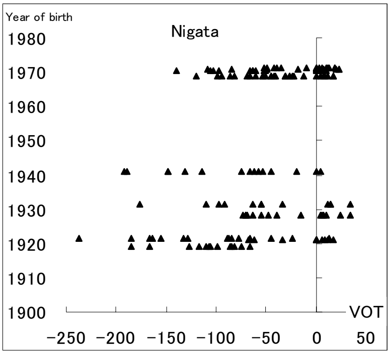


Figure 4. Scatterplot by VOT (m.s.) and year of birth for Nigata.

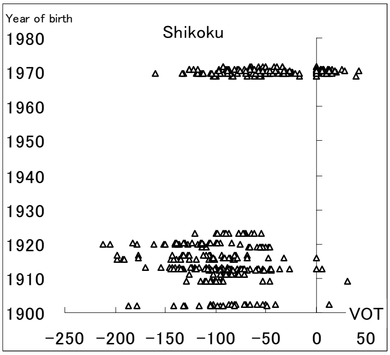


Figure 5. Scatterplot by VOT (m.s.) and year of birth for Shikoku.

On the other hand, no significant difference is shown between Nigata and Shikoku for speakers born around the 1970s. On average, speakers in both regions show four data samples with positive values. The remainder of the data shows negative VOT values, with only a few beyond –150. This tendency is actually a common phenomenon for the young generation in all of the regions analyzed thus far. At the time of study, speakers between the ages of 16 and 18 showed slightly voiced word-initial voiced plosives (slightly voiced) for an average of one third (or more) of their speech production.

Figures 6 and 7 are abstractions of Figures 4 and 5. They indicate the ratios of the data included for each VOT value according to circle size (the larger the ratio, the larger the circle). VOT values were grouped in 50 ms sizes. The data are divided into two generational categories: grandparents' generation and grandchildren's generation. The x-axis indicates the central VOT value in each level, and the y-axis indicates the relative position of the mean of the speakers' years of birth. These figures more clearly demonstrate the tendencies of each generation. In both regions, the grandchildren's generation shows a high ratio of VOT with positive values (that is, slightly voiced), the ratio of VOT with negative values is very low when the absolute value is 100 ms or greater. Nigata, as compared with Shikoku, registers a higher ratio of VOT with positive values. With regard to the grandparents' generations in Nigata, the ratio of VOT is distributed evenly across all values, which range from –200 to 50. However, in Shikoku, VOT values span between –50 and –150, which indicates that their word-initial voiced plosives are standardized to fully voiced and are accompanied by more or less equal lengths of prevoicing.

A further comparison is made between the current results and those reported in Takada (2006). The comparison is shown in its abstracted form in Figures 8, 9, and 10. The regions analyzed in Takada (2006) was three regions—Tohoku, Kitakanto, and Kanto.Tohoku comprises six prefectures: Aomori, Akita, Iwate, Yamagata, Miyagi, and Fukushima. Kitakanto contains two prefectures—Tochigi and Ibaraki—while Kanto contains four prefectures—Gumma, Saitama, Kanagawa, and Chiba.

Kitakanto (Figure 9) has proportions that are the very similar to those of Nigata (Figure 6) because in the grandparents' generation, the ratio of VOT comprises a large range and is evenly distributed from negative to positive. This indicates that the pronunciation of word-initial voiced plosives is not standardized. In addition, in the grandchildren's generation, the proportion of VOT is skewed in favor of the positive values in both regions. However, these two regions vary in that Kitakanto reflects a much higher ratio of positive values than Nigata. In this respect, Kitakanto is similar to Tohoku and Nigata is similar to Kanto.

On the other hand, Shikoku (Figure 7) and Kanto (Figure 10) appear to share similar characteristics both in the grandparents' and grandchildren's generations. These similarities suggest that the generational differences in these regions have similar origins.

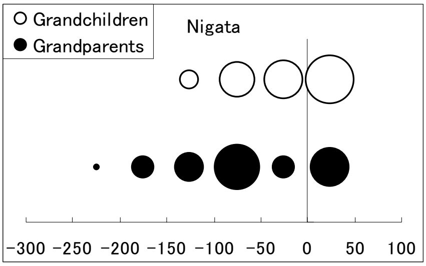


Figure 6. Propotion of pronunciation by VOT group for Nigata (size of circle=the ratio of pronounciation.)

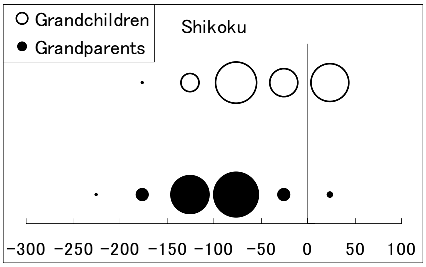


Figure 7. Propotion of pronunciation by VOT group for Shikoku (size of circle=the ratio of pronounciation.)

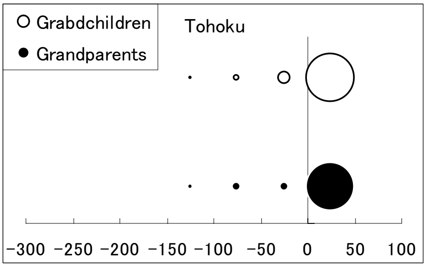


Figure 8. Propotion of pronunciation by VOT group for Tohoku (size of circle=the ratio of pronounciation.)

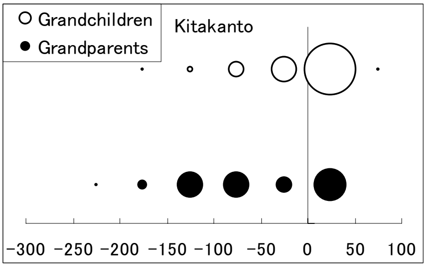


Figure 9. Propotion of pronunciation by VOT group for Kitakanto (size of circle=the ratio of pronounciation.)

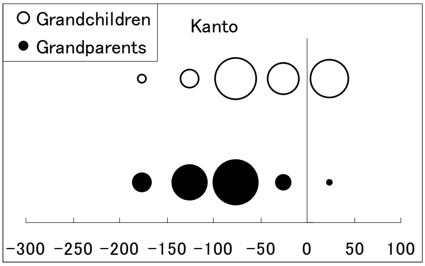


Figure 10. Propotion of pronunciation by VOT group for Kanto (size of circle=the ratio of pronounciation.)

The results reveal the presence of a generational difference, which is indicated by the regional differences that are observed only in the grandparents' generation. The grandchildren's generation does not have such noticeable regional differences. These results support Inoue's claim that with regard to the grandchildren's generation at that time, Japanese had been gradually standardizing throughout Japan. The dialectal features of sounds were disappearing when the data were collected (Inoue 1989; 3). Therefore, this generation's sound patterns are not necessarily based on traditional phonological structure, and it is not valid to analyze the phenomenon of co-occurrence based on the results of the grandchildren's generation. To determine this relationship, it is necessary to consider only the results of the grandparents' generation.

Regarding the Nigata dialect, which has been reported to have both word-medial nasalization and voicing, the grandparents' generation also showed word-initial slightly voiced plosives. This observation agrees with pattern (2) listed in section 3. Regarding the Shikoku dialect, which has been reported to have only word-medial nasalization (but not vocing) in previous studies, the grandparents' generation did not show word-initial slightly voiced plosives. This feature distinguishes the Shikoku dialect from the Tohoku dialect. In the latter, word-initial slightly voiced plosives co-occur with word-medial nasalization, but in the former, do not. From this it is concluded that word-medial nasalization do not imply word-initial slightly voiced plosives. It should be pointed out here that the similarity between Nigata and Kitakanto, illustrated in figures 6 and 9, suggests that these regions may share an areal feature involving "slightly voiced", which can be revealed only by the VOT measurements of the present investigation.

6. Conclusion

This study analyzed VOT in speech data, which were recorded in

Nigata and Shikoku between 1986 and 1988, and described and discuss the co-occurrence between word-initial slightly voiced and word-medial nasalization and/or voicing. This section develops the discussion concerning the patterns of the co-occurrence, in terms of Japanese dialect typology, and points out certain implicational relationship among the three features, i.e. “word-initial slightly voiced”, “word-medial nasalization”, and “word-medial voicing”.

By using the abbreviations, WIS (for word-initial slightly voiced), WMN (for word-medial nasalization), and WMV (for word-medial voicing), I below recapitulate the four patterns of the co-occurrence between WIS and WMN and/or WMV listed in Section 3.

(1) WIS occurs, but neither WMN nor WMV occurs.
(2) WIS co-occurs with both WMN and WMV.
(3) WIS co-occurs with WMN only.
(4) WIS co-occurs with WMV only.

As already mentioned in Section 2 and 5, Pattern (2) is attested in the Tohoku dialect. Regarding Pattern (3) it is unattested in Shikoku analysis (demonstrated in the preceding section). This means that it is not possible to hypothesize that WMN implies WIS (in other words, when there is WMN in a dialect, it does not automatically means that there is WIS in the same dialect). Regarding Patterns (1) and (4), I have not investigated relevant dialects of grandparents’ generation (except dialects of grandchildren’s generation, which shows pattern (1)), which will be researched in my future study.

In addition, the regions that have been analyzed thus far can be divided into the following three types: (1) Regions in which there is no generational difference and every generation pronounces word-initial voiced plosives as slightly voiced(WIS)—the Tohoku type, (2) Regions in which there is a generational difference—the grandparents’ generation produces fully voiced word-initial plosives accompanied by pre-voicing, but the grandchildren’s generation produces both the fully voiced and slightly voiced(WIS) word-initial plosives—the Kanto and Shikoku type, and (3) the border region (mid range of the above two regions) in which there are generational differences—the grandparents’ generation produces both fully voiced and slightly voiced(WIS) word-initial plosives, and grandchildren’s generation produces slightly voiced(WIS) plosives in a higher rate—the Kitakanto area and northeastern Nigata prefecture type. Type (3) can also be referred to as the type that is observed at the boundary region of the Tohoku and west Japanese dialects.

Future studies must reveal the inducement of the generational differences between grandparents and grandchildren, as well as that of the regional differences in the VOT values between Tohoku, including Nigata, and the other areas. To achieve the former, new data or additional old data must be analyzed in order to consider whether generational differences are a diachronic change or an aging process of the language.

To achieve the latter, non-Tohoku dialectal regions in which there is word-medial voicing must be analyzed in order to confirm or deny the existence of co-occurrence between word-initial slightly voiced(WIS) and word-medial voicing(WMV). It is also necessary to analyze all of Inoue's data or new data to reveal the nature of the VOT values of word-initial voiced plosives throughout Japan. Few regions in West Japan have been addressed previously. These analyses will illustrate the relationship between the word-initial phenomena and the phonological structure for each dialect.

Acknowledgment

I would like to thank the speakers of this study. Thanks are also due to Professor Inoue Fumio for offering his data and advice.

References

Catford, J.C. 1988. *A Practical Introduction to Phonetics*. New York: Oxford University Press.

Doi, S. 1975. "Shikoku no Hogen (Shikoku Dialects)". *Hogen to Hyojungo: Nihongo Hogen-gaku Gaisetsu (Dialects and Standard Japanese: Survey of Japan Dialectology).* Oishi, H. and Uemura, Y. (eds.) Tokyo: Chikuma. (in Japanese).

Doi, S. and Hamada, K. 1985. *Kochi-ken Hogen Jiten (Dictionary of Kochi Prefecture's Dialect).* Kochi: Kochi City Foundation for Cultural Activity. (in Japanese).

Hattori, S. 1978. *Onseigaku (Phonetics).* Tokyo: Iwanami. (in Japanese).

Homma, Y. 1980. "Voice Onset Time in Japanese Stops". *The Bulletin of the Phonetic Society of Japan* 163, pp.7–9.

Inoue, F. 1971. "Ga-gyo Shiin no Bunpu to Rekishi (Distribution and History of the Consonant /g/)". *Kokugogaku 86 (Study of Japanese 86),* pp.28–43. (in Japanese).

——— 1989. *Nihongo Onse no Chiki-sa, Sedai-sa no On'inron-teki, Onkyogaku-teki Bunseki (Phonological and Acoustic Analysis of Geographical/Generational Differences of Japanese Sounds).* [Showa 63 nendo Monbu-sho kagakukenkyu-hi hojokin sogo (A) Kenkyuseika hokokusho (Report of Grant-in-Aid for Scientific Research (A) 1988)]. (in Japanese).

Inoue, F., Shinozaki, K., Kobayashi, T. and Onishi, T. (eds.) 1995. *Kanto Hogen-ko1 (Reports of Kanto Dialects 1)*. [Nihon-retto Hogen Sosho5 (Library of the Japanese Dialects 5)]. Tokyo: Yumani. (in Japanese).

——— (eds.) 1996. *Hokuriku Hogen-ko1 (Reports of Hokuriku Dialects 1).* [Nihon-retto Hogen Sosho11 (Library of the Japanese Dialects 11)]. Tokyo: Yumani. (in Japanese).

——— (eds.) 1997. *Shikoku Hogen-ko1 (Reports of Shikoku Dialects 1).* [Nihon-retto Hogen Sosho21 (Library of the Japanese Dialects 21)]. Tokyo: Yumani. (in Japanese).

Iwai, R. 1983. "Chubu Hogen no Gaisetsu (Overview of Chubu Dialects)". *Koza Hogen-gaku 6 (Lectures on Dialectology 6)*. Iitoyo, K., Hino, S. and Sato, R. (eds.) Tokyo: Kokushokankokai. pp.5–29. (in Japanese).

Kamei, T. Kono, R. and Chino, E. 1996. "Han-yuse-on (Half-voiced)". *Sanseido Gengogaku Daijiten 6 (Sanseido Encyclopaedia of Linguistics 6).* Tokyo: Sanseido. pp.1095–1096. (in Japanese).

Kato, M. 1958. "Nigata-ken ni Okeru Tohoku Hogen-teki On'in to Echigo Hogen-teki On'in no Kyokai Chitai (The Phonological Boundary Region of Tohoku and Echigo in Nigata Prefecture)". *Kokugogaku 34 (Study of Japanese 34).* (included in Inoue, F. et al. 1996, pp.63–76.) (in Japanese).

——— 1975. "Hogen no Onse to Akusento (Dialectal Sounds and Accents)". *Hogen to Hyojungo: Nihongo Hogen-gaku Gaisetsu (Dialects and Standard Japanese: Survey of Japan Dialectology).* Oishi, H. and Uemura, Y. (eds.) Tokyo: Chikuma. pp.77–109. (in Japanese).

Kato, M. and Inoue, F. (1970) "Tonegawa Ryuiki no On'in (Phonological System of the Tone River Basin)". *Jinrui Kagaku 22 (Human Science 22).* (included in Inoue, F. et al. 1995, pp.96–125.) (in Japanese).

Kindaichi, H. 1954. "On'in (Phonology)". *Nihon Hogen-gaku (Japan Dialectology).* Tokyo: Yoshikawa Kobunkan. pp.87–176. (in Japanese).

Lisker, L. and Abramson, A.S. 1964. "A Cross-language Study of Voicing in Initial Stops: Acoustical Measurements". *Word 20*, pp.384–422.

Shimizu, K. 1996. *A Cross-language Study of Voicing Contrasts of Stop Consonants in Asian Languages*. Tokyo: Seibido.

——— 1993. "Hesa Shiin no Onse-teki Tokucho: Yuse-se/Muse-se no Gengo-kan Hikaku ni Tsuite (Phonetic Characteristics of Stop Consonants: a Cross-linguistic Study on Voiced-voiceless Categories)". *Journal of Asian and African Studies 45*, pp.163–175. (in Japanese).

Sugitoo, M. 1996. "Chugokugo Washa ni Yoru Nihongo no Muse-shiin/Yuse-shiin to, Yukion/Mukion (The Speech Sounds of Voiceless and Voiced Consonants of Japanese and Aspirated and Non-aspirated Consonants by Chinese Speakers)". *Sounds of Japanese.* Osaka: Izumi. pp.264–285.

(in Japanese).

Takada, M. 2004. “Nihongo no Goto no Yuse Shike Haretsu-on /d/ ni Okeru +VOT-ka to Sedai-sa (+VOT Tendency in the Initial Voiced Alveolar Plosive /d/ in Japanese and the Speakers’ Age)”. *Journal of Phonetic Society of Japan 8-3,* pp.57–66. (in Japanese).

——— 2006. “Goto Yuse Haretsu-on ni Okeru VOT no Chiki-sa to Sedai-sa: Tohoku kara Kanto no Bunseki (Regional and Generational Variations of VOT in Initial Voiced Plosives: in the Tohoku-Kanto Area)”. *Nihongo no Kenkyu 2-2 (Study in the Japanese Language 2-2),* pp.34–44. (in Japanese).

Tatara, S. 1959. “Tochigi-ken Hogen no On’in (Phonological System of Tochigi Prefecture Dialects)”. *Utsunomiya-daigaku Gakuge-gakubu Kenkyu Ronsyu 9 (Bulletin of Utsunomiya University 9).* (included in Inoue, F. et al. 1995, pp. 83–95.) (in Japanese).

On the Semantic Structure of English Spatial Particles Involving Metaphors[*]

Yasutake ISHII and Kiyoko SOHMIYA

1. Semantic structure of spatial particles modeled in terms of metaphors

1.1. Metaphors

In this paper, metaphor is defined as follows:

> A metaphor is a linguistic expression that refers to something that belongs to a domain distinct from the one to which the expression's basic, essential, and literal senses primarily belong. This reference is made on the basis of some kind of similarity that exists between the two things or domains and is established based on encyclopedic[1], contextual, or experiential knowledge that is shared within the same language community.

This term is used to refer not only to certain concrete linguistic expressions but also to the pattern of thinking that functions as the background for their production.

One of the most important arguments made by Lakoff and Johnson (1980), who emphasized the importance of metaphors in human language, is the significance of conceptual metaphors, i.e., correspondences of one concept to another. Conceptual metaphors are established through the combination of some basic spatial concepts called "image schema(ta)", such as CONTAINER and UP-DOWN, and their corresponding things/structures in the real world. The following are some examples of conceptual metaphors and their English realizations:

(1) PERSONAL RELATIONSHIPS ARE CONTAINERS[2]
She is *trapped in* a marriage she can't *get out of.*

(2) MORE IS UP; LESS IS DOWN
a. The crime rate keeps *rising*.
b. The stock has *fallen* again.
(Lakoff 1987:271–283)

[*] This paper was written by Yasutake Ishii under the supervision of Kiyoko Sohmiya. Ishii is, however, solely responsible for all the remaining errors and shortcomings therein.

[1] The conceptual metaphors discussed below are uttered and interpreted using this kind of shared knowledge.

[2] This example is based on Lakoff's (1987:272) statement that "[p]ersonal relationships are also understood in terms of containers: one can be *trapped in a marriage* and *get out of it*".

(3) AN ARGUMENT IS A CONTAINER
  a. That argument *has holes in it.*
  b. You won't *find* that idea *in* his argument.
  (Lakoff and Johnson 1980:92)

(4) HAPPY IS UP; SAD IS DOWN
  a. I'm feeling *up*.
  b. I'm feeling *down*.
  (Lakoff and Johnson 1980:15)

1.2. Semantic structure of spatial particles

The English Spatial metaphors that are addressed in this paper often contain spatial particles[3], as is seen above in (1), (3), and (4), with at least one conceptual metaphor in the background in most cases.

The meanings of spatial particles, which are not usually conceived as metaphors but are actually uttered and interpreted on the basis of conceptual metaphors, are referred to as "metaphors internalized in the lexical meanings" (henceforth, "lexicalized metaphors"). The semantic structure of spatial particles is modeled in Figure 1[4].

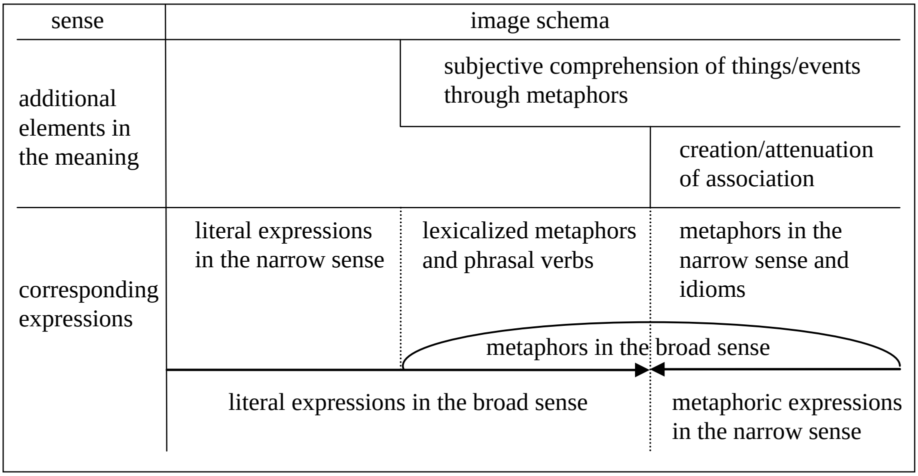

sense
image schema
additional elements in the meaning
subjective comprehension of things/events through metaphors
creation/attenuation of association
corresponding expressions
literal expressions in the narrow sense
lexicalized metaphors and phrasal verbs
metaphors in the narrow sense and idioms
metaphors in the broad sense
literal expressions in the broad sense
metaphoric expressions in the narrow sense

Figure 1. Semantic structure of spatial particles

A brief explanation of this model is given below in order to clarify the argument of this paper (see Ishii 2005, 2006 for detailed discussions of this

[3] In this paper, "spatial particles" refer to prepositions and adverbial particles of the same forms, such as *over* in *turn over* and *up* in *give up*.

[4] Figure 1 is a simplified version of the model presented in Ishii (2005, 2006).

model).

- An image schema, an abstract image of space extracted from the speakers' experience, constitutes the semantic core of spatial particles as their conceptual "sense". Literal expressions in the narrow sense are those that can be uttered and interpreted using only the knowledge of image schemata.
- A lexicalized metaphor is an expression that necessitates an understanding of not only the image schema but also the conceptual metaphor(s) shared within the same language community. Lexicalized metaphors have a literal characteristic in the sense that they are not usually regarded as metaphors (thus, they fall under the division "literal expressions in the broad sense"); however, they also have a figurative characteristic in the sense that they are based on conceptual metaphors (thus, they fall under the division "metaphors in the broad sense").
- A metaphor in the narrow sense is an expression in which speakers create a new relationship between different domains that is inherent in metaphors. An idiom, on the other hand, is an expression in which conceptual metaphor(s) play a less important role and are no longer shared within the same community.

The left side of the figure represents the more physical, concrete, literal, and simple characteristics while the right side represents the more abstract, figurative, and complex characteristics.

The following examples represent the above classification:

(5) The card is *in* the envelope. (a literal expression in the narrow sense)

(6) She is *in* white today. (a lexicalized metaphor)

(7) God is *in* the details. (a metaphor in the narrow sense)

We place the strongest emphasis on the level of lexicalized metaphors. Image schemata abstracted from the images of referents combine with the conceptual metaphors that are important to the language community, thus enabling the figurative meanings of spatial particles to be incorporated as their lexical meanings as well—the lexicalized metaphors embody the conventionalized figurative meanings of spatial particles. We argue that lexicalized metaphors play a significant role in language in general, functioning as a bridge between concrete references and abstract language uses.

1.3. Problems in the model

The semantic structure of spatial particles modeled in Figure 1 (henceforth, "the original model") contains the following two problems:

1. It does not presuppose the polysemy of spatial particles.
2. It does not facilitate the decision of which sense constitutes the image schema.

Let us begin by discussing problem 1. If we define polysemy as "having more than one sense independent of the context, not making us establish a context-dependent meaning from a single sense using the available contextual information", spatial particles can be said to be polysemous, as seen in the following examples (Tyler and Evans 2003:136–141):

(8) turn *up* the volume

(9) be dressed *up*

(10) finish *up* the work

In (8), *up* expresses an increase in volume; in (9), 'in a special or formal way'; and in (10), perfection or completion. However, our original model of the semantic structure of spatial particles was indifferent to the relationship among these meanings because this model focused primarily on the importance of the level of lexicalized metaphors and regarded the uses of *up* mentioned above in (8)–(10) in the same light, as realizations of lexicalized metaphors with conceptual metaphors incorporated.

With regard to problem 2, our original model assumed that those expressions that can be interpreted only through image schemata are literal expressions in the narrow sense. To put it conversely, all literal expressions in the narrow sense that include a particular spatial particle are assumed to represent a single image schema. However, there are various kinds of literal expressions in the narrow sense that do not include conceptual metaphors, as seen in the following examples (Tyler and Evans 2003:183–198):

(11) He stayed *in* for the evening.

(12) The train is finally *in*.

(13) The walls of the sandcastle fell *in*.

Although the landmarks are not explicitly specified in these three examples, (11) expresses that the subject (*he*) remained where he was usually found, such as his home or his hotel room; (12) conveys that the train has arrived at the station; and (13) communicates that the walls of the sandcastle collapsed inward. Since conceptual metaphors are not used in these interpretations, the above three instances of *in* were categorized as literal expressions in the narrow sense under our original model. Then, if all these examples are understood only through an image schema, what image schema of *in* will be shared among them? In the case of (11), *in* designates a point in a bounded space. In (12) and (13), *in* not only designates a point in a bounded space, as was the case in (11), but also refers to a process of entering the space from outside (12) or to the directionality to the inside of a bounded space (13). It is possible to state that the image schema of *in* is highly abstract and embraces all of the above abstract images; however, it would be an overextension to interpret concrete sentences such as those of (11)–(13) using only this kind of highly abstract image. It would be natural to assume the existence of an element that connects image schemata and literal expressions, rather than to equate image schemata with the meanings found in literal expressions in the narrow sense.

In Section 2, we will overview Tyler and Evans's (2003) polysemy model, which presents helpful suggestions to solve these problems, and in Section 3, we will revise our original model by adopting their arguments.

2. Polysemy of spatial particles and Tyler and Evans's model

In this section, we will overview Tyler and Evans's (2003) "principled polysemy model" in order to confirm the polysemy of spatial particles. This model is shown to help solve the problems of our original model of the semantic structure of spatial particles.

Tyler and Evans's view of spatial particles is a cognitive linguistic approach, which can be summarized as follows:

- Spatial particles comprise a polysemous structure consisting of more than one distinct sense; each distinct sense has a "configurational element" in space and a "functional element" associated with the configuration. The spatial configuration of *in*, for example, is an abstract one that has a bounded space as its landmark and a trajector inside the landmark. Its functional element is that of "containment", which is often seen in this type of spatial configuration[5].
- Semantic extensions are achieved through the fixation of the experiential link between the functional elements and the contexts of utterances. This process of fixation is called the "pragmatic strengthening"[6] of functional elements. These extensions result in the formation of polysemous networks of spatial particles.
- The distinct sense of a spatial particle that fulfills some conditions (described below) is its "primary sense", and the mental representation of the primary sense is called a "proto-scene".

Tyler and Evans admit that with regard to the relationship among the distinct senses and the primary sense, it is impossible to produce one feasible model with which everyone will be in agreement. They set out to create their

[5] Tyler and Evans consider that word meanings do not exist within the words themselves, but rather are just a clue to the speakers/listeners' knowledge; thus, the important aspect of understanding meanings is inference—making full use of context and encyclopedic knowledge. They argue that the function of inferences enables the spatial particles, which belong to the closed class, to express infinite spatial relationships. They also argue that spatial relationships (often expressed with spatial particles in English) do not exist objectively in the world but are intrinsically conceptual (Tyler and Evans 2003:50–51) because meanings themselves are intrinsically conceptual in the sense that all experiences are embodied, i.e., interpreted through human cognitive systems.

[6] This term is found in Traugott's (1988) title and was advocated with reference to diachronic semantic changes. For a more detailed explanation of this concept, see Traugott (1989).

principled polysemy model, which explicitly shows the criteria for identifying the distinct senses and the primary sense.

They propose the following two criteria for determining distinct senses (Tyler and Evans 2003:42–45):

A particular use of a spatial particle represents a distinct sense if it

1. has a non-spatial meaning or a configuration that is different from that of the proto-scene, and
2. has an example in which the sense cannot be derived from the context.

Let us begin by reviewing the first criterion. We can detect configurational consistency in the trajectors (*the helicopter* and *the hummingbird*) because they are positioned above the landmarks (*the ocean* and *the flower*) in (14) and (15):

(14) The helicopter hovered *over* the ocean.

(15) The hummingbird hovered *over* the flower.

In contrast, there is no such consistency in (16) and (17):

(16) Joan nailed a board *over* the hole in the ceiling.

(17) Joan nailed a board *over* the hole in the wall.

Therefore, it is possible in (16) and (17) that *over* is used in a distinct sense, to mean COVERING, a meaning that is not possible in (14) and (15). With regard to the second criterion, while we can derive the sense of COVERING from the context in (18), this sense cannot be derived from one's knowledge of reality and the configuration in (16) above:

(18) The tablecloth is *over* the table.

In other words, in (18), it is possible to infer that the table is covered with the tablecloth and cannot be seen as a result of the configuration of the trajector (*the tablecloth*) being above the landmark (*the table*), which is also the case in (14) and (15), and the knowledge that a tablecloth is normally larger than a table. However, this type of reasoning is impossible in (16) and (17); thus, they cannot be interpreted correctly if we are not aware that *over* has the sense of COVERING. Therefore, *over* is identified to have a distinct sense of COVERING fulfilling the two aforementioned criteria.

A primary sense should fulfill as many of the following conditions as possible (Tyler and Evans 2003:45–50):

1. It should be the earliest attested meaning.
2. It should include a spatial configuration found in as many distinct senses as possible.
3. It should be incorporable in compounds or phrasal verbs.
4. It should be the sense associated for a contrasting pair of

spatial particles[7].

5. It should allow the determination of a context that functions as a bridge to an immediately relevant sense[8].

Tyler and Evans actually depict semantic network structures for many English spatial particles in accordance with the above criteria, although they admit that their semantic networks could be revised. Let us consider the depiction of *over* shown in Figure 2 as an example (Tyler and Evans 2003:80–106).

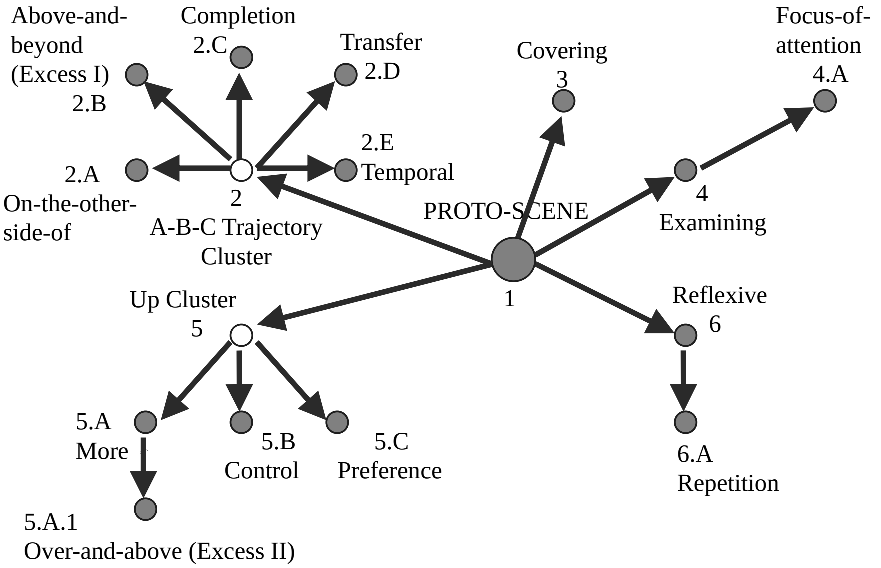


Figure 2. The semantic network of *over*[9]

[7] The primary sense of *over*, for example, is the meaning of the word in contrast to *under*.

[8] For instance, the situation in (18) functions as a bridge between the primary sense of *over* and the distinct sense of COVERING.

[9] Examples of each distinct sense and brief explanations of each cluster (indicated by unshaded circles) are as follows: (1) The picture is *over* the mantel., (2) a complex conceptualization of a trajectory that surmounts a high construct and goes to the other side, (2.A) Arlington is *over* the Potomac River from Georgetown., (2.B) The arrow flew *over* the target and landed in the woods., (2.C) The cat's jump is *over*., (2.D) Sally turned the key to the office *over* to the janitor., (2.E) The festival will take place *over* the weekend., (3) The tablecloth is *over* the table., (4) The mechanic looked *over* the train's undercarriage., (4.A) The little boy cried *over* his broken toy., (5) a complex conceptualization of vertical elevation, (5.A) Jerome found *over* forty kinds of shells on the beach., (5.A.1) The heavy rains caused the river to flow *over* its banks., (5.B) She has a strange power *over* me., (5.C) I would prefer tea *over* coffee., (6) The fence fell *over*., and (6.A) After the false start, they started the race *over*.

It is worthwhile to mention that there is a criticism of Lakoff's network model of the senses of *over* (in which its use in *The plane flew over.* is the central sense; cf. Lakoff (1987:419)) in the background of this model. Tyler and Evans criticize that Lakoff's model involves an excessive amount of arbitrariness, which results from a laxity in the methodological constraints necessary to build the model, that it dismisses non-linguistic functions, and that the semantic distinctions are made too minutely.

Tyler and Evans's discussion provides a reasonable explanation for the semantic extension of spatial particles and seems to conform to the present writers' argument that the image schemata obtained through abstraction from expressions referring to physical space form the semantic core of spatial particles, and that figurative senses are obtained through an additional step of conceptualization of the cores and their combination with conceptual metaphors.

However, the present writers are critical of Tyler and Evans's discussion for the following reasons:

1. Their model does not place sufficient emphasis on the centrality of the primary sense that designates a part of concrete objective space. They argue that both configurational and functional elements are variable in each sense and that the distinct senses in the network are linked only by the extensional relationships between the nodes (distinct senses).
2. Both configurational and functional elements can be null; for example, Tyler and Evans argue that the primary sense of *down* has a functional element of NEGATIVE VALUE, such as invisibility and vulnerability (Tyler and Evans 2003:142); however, it would be difficult to consider the existence of this functional element in the case of expressing a physical downward movement without a value judgment or an emotional implication. It is also possible to consider a case in which a configurational element is lacking, for example, the case of temporal expressions and highly figurative distinct senses. Therefore, the present writers argue that not every distinct sense, including the primary one, necessarily requires both configurational and functional elements.
3. Their argument—that extended senses are obtained through the pragmatic strengthening of configurational and functional elements—seems to be excessively generalized. They fail to emphasize the role played by conceptual metaphors when spatial particles are subjected to semantic extensions.

In Section 3.3, we will observe that these problems can be solved in our revised model.

Although some of the polysemous network models of spatial particles proposed by Tyler and Evans (2003) are used for discussion in this paper, the present writers do not unconditionally accept the primary and other distinct senses of each spatial particle in their models. As the authors themselves admit, their models have room for revision; there are criticisms of their models, such as that of Kunihiro (2005:315–316), who criticizes the absence of a sense of PASS THROUGH for *over*. The present writers use some of Tyler and Evans's network models of spatial particles only to demonstrate that the revised model presented in this paper conforms to the polysemous characteristic of spatial particles. Other polysemy models can be used for this purpose, provided they follow the criteria specified by Tyler and Evans.

3. Revision of the model of the semantic structure of spatial particles

In this section, we will revise the original model we presented in Section 1.2 by adopting the arguments made by Tyler and Evans. We will show that the revised model can not only solve the problems involved in the original model but also provide more comprehensive explanations for the limitations of Tyler and Evans's arguments.

3.1. Revising the original model

3.1.1. Modifications in the revision and their motivations

As a result of the following two modifications made to our original model, the model will be able to

1. fully recognize the polysemy of spatial particles, and
2. presuppose a possible semantic extension based on metonymies, which can exist between image schemata and conceptual metaphors.

The motivation for the first modification is that our original model cannot provide a complete explanation of the relationship among senses because it maintained only that the level of lexicalized metaphors is obtained through the combination of an image schema and conceptual metaphors. However, we can identify senses that are independent of the context in spatial particles; further, they play a significant role in forming verb phrases that include spatial particles, such as phrasal verbs, which will be discussed in Section 4. For these two reasons, we have fully recognized the polysemy of spatial particles, which was not explicitly specified in our original model, in the sense that we can identify multiple "established senses"[10] that are

[10] What Tyler and Evans refer to as "distinct sense" and what the present writers refer to as "established sense" can be regarded as the same. The present writers prefer the latter term because "distinct" gives the impression that there is not a substantial amount of linkage among senses, whereas "established" gives the impression that the senses that have been extended from the primary one are established through the process of conventionalization.

independent of context and related to another. We also admit a further extension from an extended sense (such as from 4 to 4.A in Figure 2), by which we can easily explain the fact that some senses do not remind us of the image schema synchronically—i.e., it is possible that the association of the image schema found in the primary sense weakens through several steps of semantic extensions.

The motivation for the second modification is that our original model, which assumes that literal expressions in the narrow sense can be interpreted only through image schemata, cannot identify image schema when we suppose that the polysemous network structures of spatial particles are composed of several interrelated established senses (see the discussion under problem 2 in Section 1.3.). In order to solve this problem, the present writers have decided to place metonymies (discussed in detail in Section 3.1.2.) between the levels of image schema and conceptual metaphors. By thus positioning the metonymies, it becomes possible to distinguish the primary sense, which can be interpreted with only the image schema, from other literal meanings, which do not necessitate conceptual metaphors but use metonymies in addition to the image schema.

3.1.2. Metonymies

A metonymy is a form of figures of speech in which one event or concept refers to another on the basis of their contiguity or relationship[11]. The following are examples of metonymies (Langacker 1999:198–199).

(19) The coach is going to put some fresh legs in the game. (A part (*leg*) refers to the whole (*player*).)

(20) That car doesn’t know where he’s going. (The whole (*car*) refers to its part (*driver*).)

(21) She bought Lakoff and Johnson, used and in paper, for just $1.50. (The names of the authors (*Lakoff and Johnson*) refer to their book.)

Let us briefly examine Langacker’s (1999:198–200) work for a cognitive linguistic understanding of metonymies. A metonymy refers to a “reference point” (Langacker 1993) that is cognitively salient within a single

[11] Since a part-whole relationship (meronymy) has been traditionally called “synecdoche”, many previous studies (such as Ullmann (1962:219), Lakoff and Johnson (1980:36), and Gibbs (1994:322), who refers to Lanham (1969)) explain that a part-whole relationship is called synecdoche as well as metonymy. At present, however, the term “synecdoche” is often restricted to those uses where a more comprehensive genus refers to a less comprehensive species or vice versa, and a part-whole relationship is often treated as a type of metonymy (as in Langacker (1999), for example; Sugai (2002:159) explains the confusion of the terms found in the literature). This is because unlike synecdoches, which are based on a higher-lower relationship of genus and species, a part-whole relationship does not share a part of the semantic elements; this lack of shared semantic elements is also seen in metonymies, which are based on contiguity.

domain[12], makes the hearer choose the intended object, which is related to the reference point, and prompts him or her to mentally access the object[13]. In short, a metonymy is an expression that refers to an object that is prominent or easy to refer to in linguistic expression but actually designates an object that is relatively less prominent or more difficult to refer to. Metonymy can be considered an efficient linguistic method of reconciling two incompatible elements: necessity of properly directing the hearer's attention to the intended object and the human tendency to speak and think about the most cognitively salient objects.

It is widely acknowledged that metonymies contribute to semantic changes and produce polysemy (Waldron 1979:186–200; Kunihiro 1982: 125–127); Taylor (2003) emphasizes the importance of metonymies as functioning as a basis for metaphors.

3.1.3. Revision of the original model

Figure 3 presents the revised model of the semantic structure of spatial particles based on the discussion in Section 3.1.1. Also shown at the bottom of the figure are the levels to which the distinct senses of *over* illustrated in Figure 2 correspond. (However, we do not intend to show the positions of each distinct sense or the distances between them except for the boundaries between the levels in Figure 3.) Figure 3 demonstrates the existence of a primary sense that can be interpreted through only the image schema without metonymy (the PROTO-SCENE [1] in Figure 3) in the first place, from which extensions are made and each established sense is obtained. The established sense of COVERING [3], for example, can be regarded as having been established after repeated focus on the semantic element of COVERING that is a part of the event represented in (18). Therefore, this provides an example of the establishment of a metonymic use (the "whole" term designates its part) as a sense of a word. To take another example, the A-B-C TRAJECTORY cluster [2] designates a trajectory of something passing through the position represented as the landmark of the PROTO-SCENE [1] in the opposite direction; thus, this metonymic use, in which a part (transit point) designates the whole (trajectory), can be deemed to have been established as the configuration of the cluster.

Figure 4 represents the extent to which the senses obtained from the primary sense and the metonymies extend; this representation is superimposed on Tyler and Evans's network model of *over* (Figure 2).

[12] Langacker calls a set of potential referents a "dominion".

[13] This characteristic of metonymies distinguishes them from metaphors, which are based on the similarity between two objects belonging to different domains.

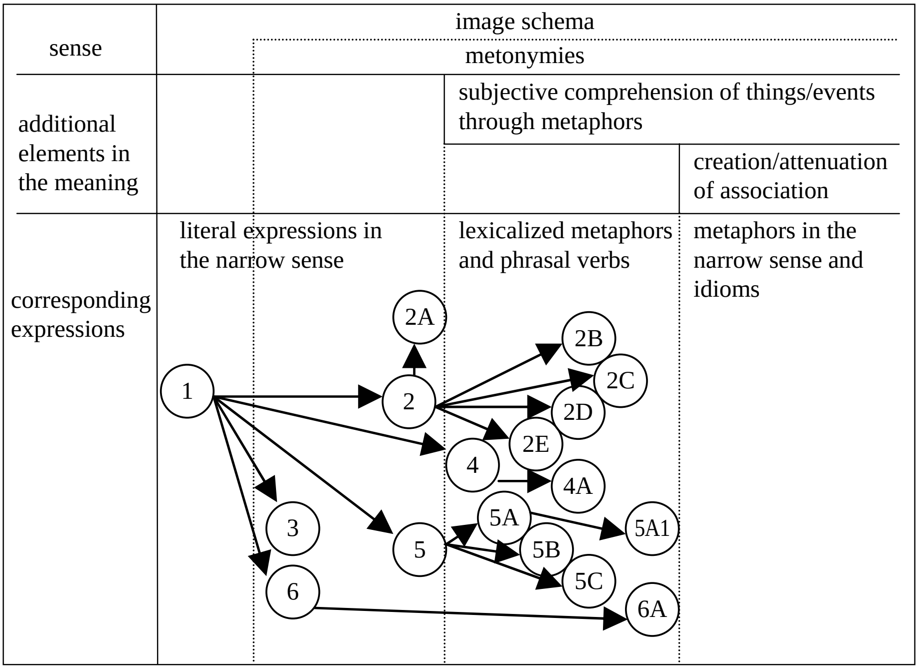

sense
image schema
metonymies
additional elements in the meaning
subjective comprehension of things/events through metaphors
creation/attenuation of association
corresponding expressions
literal expressions in the narrow sense
lexicalized metaphors and phrasal verbs
metaphors in the narrow sense and idioms
1
2
2A
2B
2C
2D
2E
3
4
4A
5
5A
5A1
5B
5C
6
6A

Figure 3. Semantic structure of spatial particles (revised)

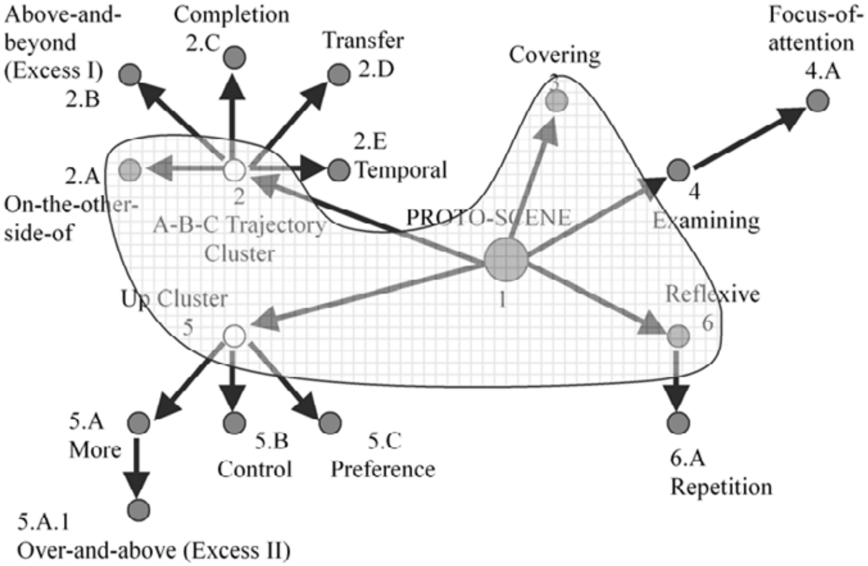


Figure 4. Tyler and Evans's network model of *over* and its relationship with metonymy

The shaded section in Figure 4 shows the metonymic extension from the primary sense [1] to other distinct senses. This figure also demonstrates that those extensions that continue beyond the shaded portion are based on conceptual metaphors. Tyler and Evans (2003:32–36) subdivide conceptual metaphors into “experiential correlation” and “perceptual resemblance”, which combines more than one concept; however, they do not provide more detailed explanations. It is not clear whether the referents of these two terms are identical to the conceptual metaphors, where the former are fractionized and renamed forms of the latter, or whether the former include some new concepts.

3.2. Generalization of the polysemous semantic structure of spatial particles

Tyler and Evans explain that since the clusters ([2] and [5] in their model of *over*) include a single configuration and several functional elements, the extended senses are obtained through the clusters. The present writers generalize the polysemous structure of the senses in spatial particles, including the clusters that have characteristics that are slightly different from those of the distinct senses, as follows:

> The polysemous semantic structure of spatial particles: An image schema that designates a concrete space constitutes the primary sense of spatial particles, whence spatial senses other than the primary sense are obtained through metonymies. This interpretation of the spatial configuration of the primary sense or a sense obtained in this manner is associated with a function that accompanies the configuration. This functional element combines with one or more conceptual metaphors to produce a sense involving the metaphor(s). However, the metaphor(s) involved are internalized into the lexical meaning of the particle, usually causing the senses that incorporate conceptual metaphors to be regarded as literal rather than metaphoric.

3.3. How our revised model addresses the problems in Tyler and Evans’s model

The present writers’ original model considers the criticisms of Tyler and Evans’s model discussed in Section 2.

We first criticized their incomplete emphasis on the centrality of the primary sense. Our revised model underscores that the primary sense of spatial particles is an image schema designating a space, which simultaneously serves as the foundation for semantic extensions and all the established senses.

Our second criticism was that it seems possible for configurational and functional elements to be null. Thus, in our revised model, it is assumed that the established senses used in literal expressions in the narrow sense roughly correspond to Tyler and Evans's spatial configurations, while those used in lexicalized metaphors roughly correspond to their functional elements. The associated functional elements in the former seem to be associated in extensions to the established senses used in lexicalized metaphors rather than to be a part of the established senses themselves. For example, the spatial configuration of the A-B-C TRAJECTORY cluster [2] includes possibilities of interpretation (functional elements) that cause the four senses to be categorized under the class of lexicalized metaphors [2.B–2.E]. However, it is difficult to think that the cluster involves all these functions; the independency of 2.C has been achieved through the production and fixation of the interpretation (functional element) that "being on the other side now is the result of the COMPLETION of the movement to the other side", and this functional element seems to function as a bridge between the A-B-C TRAJECTORY cluster [2] and the sense of COMPLETION [2.C] rather than as a part of an established sense. Therefore, the present writers argue that not all distinct senses, including the primary one, inevitably necessitate both configurational and functional elements.

Our third criticism concerns the excessively generalized nature of their argument that semantic extensions are the result of pragmatic strengthening. In our revised model, we argue more definitely that metonymies are used for semantic extensions within the scope of literal expressions in the narrow sense, while conceptual metaphors are used for those across to the class of lexicalized metaphors.

4. An interpretation model for the V-PP construction including phrasal verbs

In this section, we argue that various verb phrases, which appear to have the same syntactic structure of "verb + spatial particle (+ complement noun)", actually include several interpretative levels, and that this fact is closely related to the semantic structures of spatial particles.

4.1. How verb phrases are addressed in our revised model of the semantic structures of spatial particles

Before discussing the interpretation model of verb phrases, let us observe how verb phrases that include spatial particles (henceforth, "V-PP constructions") are treated in our revised model. Our model mentions the interchangeability of the collocates (verbs and complement nouns used with the spatial particles) in each class of expressions. Although this is omitted in Figures 1 and 3 (cf. Ishii (2005, 2006)), it is briefly explained below. Literal expressions in the narrow sense can syntactically collocate with other words with quite a high degree of freedom, provided that they designate existences and movements in physical space. On the other hand, the interchangeability of the collocates of lexicalized metaphors is restricted by the conceptual metaphors that are shared within the same language community. Phrasal verbs, which consist of verbs and spatial particles and whose meaning cannot be obtained by merely totaling their meanings, are classified as lexicalized metaphors[14]; for example, the entire meaning of the phrasal verb "find out" consists of a combination of the meanings of each component word—*find* ('to discover') and *out* ('from inside to outside')—and a conceptual metaphor that can be identified as VISIBLE IS OUT[15]. This is why the interchangeability of the collocates of phrasal verbs is even lower. With regard to idioms, their background relationships play a less important role and are no longer shared within the same community, possibly because they have been forgotten through history or for some other reason; only the expressions have survived as formulaic ones, which gives them the lowest interchangeability of the collocates. On the other hand, with the exception of idioms, metaphors in the narrow sense are uttered on the basis of the speakers' creation of relationships through their creativity, thus raising the interchangeability of their collocates.

Phrasal verbs and idioms are both subject to a high degree of conventionalization and have fairly fixed expressions. Phrasal verbs are those expressions that have been fixed at the level of lexicalized metaphors, for example, "pick up" ('to select', for instance); idioms are those expressions that have been fixed at the level of metaphoric expressions in the narrow sense, for example, "(be) up in the air" ('undecided', for instance). The fixation of an expression implies the attainment of an overall lexical characteristic.

4.2. Patterns of interpreting verb phrases

Having observed how our revised model addresses verb phrases, we now examine the patterns of interpreting V-PP constructions. We can state that V-PP constructions, which include phrasal verbs, comprise three major linguistic elements: verbs (V), spatial particles (P), and conceptual

[14] Hampe (2000) also asserts that the meanings of phrasal verbs are created from the literal senses of the verbs and spatial particles and (in many cases, several) conceptual metaphors. For example, she argues that the phrase "face up to (a problem)" is understood via conceptual metaphors such as PROBLEMS ARE OBSTACLES, CLOSE IS UP, ACTIVE IS UP, and PURPOSEFUL ACTION IS MOTION TOWARDS A GOAL.

[15] Since there is no defined set of conceptual metaphors with which many researchers agree, in many cases, different people may imagine different conceptual metaphors in the background of the same sentence based on their viewpoint.

metaphors (CM). There is also a non-linguistic element of context, in a broad sense, which includes encyclopedic knowledge of subjects and complements. Thus, the following three types of interpretations of V-PP constructions can be conceived:

1. V + P: No conceptual metaphor is used.
2. V + [P + CM]: An independent established sense of a spatial particle based on conceptual metaphors ([P + CM]) is combined with a verb.
3. [V + P + CM]: Particular conceptual metaphors are incorporated into the V-PP construction as a whole.

The first pattern (V + P) assumes that the particle is interpreted through only its image schema without a conceptual metaphor (but possibly with a metonymic extension). By applying this scheme to our revised model, we can state that literal expressions in the narrow sense are interpreted according to this pattern. This is because the expressions whose meanings do not include conceptual metaphors are necessarily literal expressions in the narrow sense. Examples interpreted according to this pattern are as follows[16]:

(22) He was born *in* Edinburgh, ... (B1D 390, BNC)

(23) . . . she moved *down* the hill to live in more conventional quarters, ... (AC7 1782, BNC)

(24) Philip looked *at* her. (ABX 1317, BNC)

When the V-PP constructions in (22)–(24) are interpreted, each spatial particle is used as a literal expression in the narrow sense, and no conceptual metaphor is used.

The second pattern (V + [P + CM]) can be used to interpret many V-PP constructions, including those known in the present writers' model as lexicalized metaphors. Examples interpreted according to this pattern are as follows:

(25) It must be borne *in* mind that. . . . (APD 83, BNC)

(26) Rabbits scurried *at* our approach. (F9H 1837, BNC)

(27) "That depends *on* what you mean by hope." (A0F 1880, BNC)

In (25)–(27), each particle in the V-PP constructions is used as a lexicalized metaphor; in (25), *in* is used to refer to abstract space, which is based on the conceptual metaphor MIND IS A CONTAINER; in (26), *at* is used to denote reason, which is derived from the conceptual metaphor CAUSATION IS A SEQUENCE IN TIME; and in (27), *on* is used to mean dependency, which is rooted in the conceptual metaphor INFLUENCE IS PHYSICAL SUPPORT.

The third pattern ([V + P + CM]) is used to interpret many phrasal verbs.

[16] Italics are added by the present writers. The same is true for all the other examples from the BNC.

This is because the meanings of phrasal verbs are not constructed by the combination of the verb, spatial particle (and conceptual metaphors), and context, but are semantically independent in that they can to a certain extent be understood without context. The following examples are interpreted according to this pattern:

(28) . . ., *fill in* the coupon on page 216 and. . . . (G2V 501, BNC)

(29) . . . see if the rules relating to unfair play can be *tightened up*. (CH3 5263, BNC)

(30) . . . the only animals capable of using tools to *get at* their food. (HV9 48, BNC)

In (28), *in* is used to denote abstract space, which is based on the conceptual metaphors SPACE ON PAPER IS A CONTAINER and WRITTEN CHARACTERS ARE ENTITIES; in (29), *up* is used to refer to firmness, which is founded on the conceptual metaphor FIRM IS UP; and in (30), *at* is used to mean a destination point, which is derived from the conceptual metaphor ARRIVING AT A PLACE IS OBTAINING THINGS THERE. As can be seen from these examples, many phrasal verbs incorporate a combination of [P + CM] in the unit of [V + P + CM]; in other words, spatial particles and conceptual metaphors are not selected separately but rather a verb merges with a particular established sense of a spatial particle. In contrast, with regard to idioms, the incorporated conceptual metaphors are often synchronically unclear; this renders their pattern of interpretation different from that of [V + [P + CM]] for phrasal verbs.

4.3. Previous literature and data that support our interpretation model of verb phrases

In this subsection, we will examine some of the discussions and data that support the interpretation model of phrasal verbs presented in Section 4.2. First, let us refer to Bolinger’s (1971:112–115) work on phrasal verbs. He classified the combinations of verbs and particles into the following categories:

1. First-level stereotypes: simple combinations of verbs and particles with literal meanings
2. Second-level stereotypes: phrasal verbs
   - 2a. First-level metaphors: phrasal verbs in which only particles are figurative (such as “load up” as compared to “go up”)
   - 2b. Second-level metaphors: phrasal verbs in which all parts are figurative as a whole (such as “make up a face” and “rub out an adversary” as compared to “make up a bed” and “rub out a mistake”, respectively)
3. Third-level stereotypes: idioms such as “put on the dog”

We can say that Bolinger’s argument shows a high degree of conformity to the treatment of V-PP constructions in our model; Bolinger's first-level stereotypes are embraced in our literal expressions in the narrow sense, his first-level metaphors in our lexicalized metaphors, his second-level metaphors in our phrasal verbs, and his third-level stereotypes in our metaphoric expressions in the narrow sense. Bolinger's argument mentions neither the polysemy of particles nor how their semantic extensions are created, but instead makes these classifications only in terms of whether the senses of the verbs and particles are literal or figurative and provides an inadequate discussion of the syntactic collocates of V-PP constructions and the senses of the particles behind the interpretations; nevertheless, his classification can be deemed appropriate.

The existence of these patterns can be demonstrated through examples which contain difficulty in interpreting an ambiguous phrase as a result of several possible patterns from which to choose, or which contain misunderstandings caused by a failure to choose the correct pattern. The former difficulty is clearly expressed in language in (31) and (32), and even selecting the correct interpretation pattern can produce several possibilities of choosing a sense or a misunderstanding, which is exemplified in (33):

(31) Whether the employers were in breach of the statute depended on whether the machine *was* "*in motion.*" In the primary or literal sense of the words it was, but since the machine was not working under power and was only in temporary motion for necessary adjustment, the House of Lords chose to give the words the secondary meaning of "mechanical propulsion." Since the machine was not being mechanically propelled it was not in motion. (FRA 97–99, BNC)

(32) One prisoner put it like this: "I waken up every morning with this pain. It's terrible. It's not that I want to die, but I just want to *get out of my mind*". He means this literally.

It is this wanting to get out of one's mind that creates such a strong demand for drugs in prisons. (G0T 899–903, BNC)

(33) Straight to the point

A FRENCH driver took it literally when he asked for directions and was told to "*go straight over the roundabout*".

He got stuck in the middle in Wolverhampton and had to be rescued. (CH6 717–719, BNC)

The context for (31) is as follows: "This Act [the Factories Act] requires dangerous parts of machines to be constantly fenced while they are in motion. A workman adjusting a machine removed the fence and turned the machine by hand in order to do the job. Unfortunately he crushed his finger." (FRA 94–96, BNC). The problem concerns the manner of interpreting the phrase "(be) in motion" in the provision. Although the "literal sense"[17] involves the interpretation of 'working' based on the second pattern (V + [P + CM]), they interpret the phrase more specifically as 'mechanically working under power'. In other words, in this scene, they interpret the sentence "The machine is in motion." based on the third pattern ([V + P + CM]) (although "is in motion" is not considered a phrasal verb). On the contrary, in (32), the writer states that the phrasal verb "get out of my mind" should be interpreted based on the second pattern (V + [P + CM]) as 'to lose his awareness of reality', not based on the third pattern ([V + P + CM]) as its normal figurative meaning 'to become crazy'. With regard to (33), although the correct pattern (V + P), has been selected, the driver errs in selecting the correct sense of *over*; whereas he should have avoided and gone around the obstacle in the rotary, he collided with it.

Data obtained from psycholinguistic experiments on young children suggest that there are several interpretation patterns for the V-PP construction. We will examine a psycholinguistic experiment by Friederici (1983) on children's reactions to function words in German[18]. She first classified prepositions into the following three categories:

1. Lexical prepositions: "carrying the semantic information of a spatial relation but also some structural information as the head of a prepositional phrase" (Friederici 1983:721); in the present writers' model, these correspond to literal expressions in the narrow sense.
2. Obligatory prepositions: "fulfilling a structural subcategorization requirement" (Friederici 1983:721); in the present writers' model, some of these correspond to phrasal verbs[19] and idioms in that they are fixed.
3. Verb particles: carrying semantic information of themselves or contributing to the change of the meaning of the preceding verb (Friederici 1983:721); in the present writers' model, these are embraced in lexicalized metaphors.

Following this, she then elucidates that acquisition proceeds in the order of

[17] This usage of *in,* which is regarded as "literal" in this example, is classified as a lexicalized metaphor in the present writers' model; thus, the expression is classified as a literal expression in the broad sense.

[18] Friederici's experiments are aimed at German function words; therefore, her argument cannot be directly applied to English spatial particles. However, she does not discuss individual German particles but rather addresses closed-class words in general. Thus, it will be valid to assume that her data support our discussion.

[19] Note that in our model, the particles in phrasal verbs are deemed to have lexical meanings; thus, they do not fully correspond.

lexical prepositions, verb particles, and obligatory prepositions. This roughly corresponds to the order of literal expressions in the narrow sense, lexicalized metaphors, and phrasal verbs and idioms established in the present writers' framework. Therefore, her data can be considered to support the model we present in Section 4.2, which argues that each of these is understood based on different interpretation patterns of verb phrases.

Learners show a tendency to interpret V-PP constructions with the first simplest pattern (V + P), or the second (V + [P + CM]) for some highly frequent independent senses, although the temporal expressions subject to the second pattern and the highly frequent phrasal verbs subject to the third pattern ([V + P + CM]) are learned through the process of "item learning" (Ellis 1999:474). This tendency is not limited to learners' interpretations —learners have a strong tendency to remember and use the senses of particles that can be interpreted with the V + P pattern, even in the context of production (as is seen in the result of an experiment on learners' writing (Ishii 2005:516–517, in press)).

The above data can be regarded as providing valid arguments for the claim that the V-PP construction has several possible patterns of interpretation that are closely related to the semantic structure of spatial particles.

5. Closing remarks and future studies

We have focused on the polysemous and figurative natures of English spatial particles and argued that the lexical meanings of spatial particles consist of an image schema, metonymies, and conceptual metaphors, and that the latter two elements function as the foundation of semantic extensions of spatial particles.

We have progressed to assert that the syntactic structure of a verb and a spatial particle (and a complement), which appear to be the same, include several possible patterns of interpretation and that these interpretation patterns are closely related to the semantic structure of spatial particles.

In order to have a greater quantity of objective evidence for our discussion, we must first consider whether we can collect usage data that will clearly indicate that metonymies play a significant role in semantic extensions.

We would also like to consider whether the knowledge obtained through this kind of research can be applied to language education. Since lexicalized metaphors of spatial particles and phrasal verbs trouble many learners of English, it will be of great importance to find appropriate ways to present these concepts to them.

References

Bolinger, D. 1971. *The Phrasal Verb in English*. Cambridge, MA: Harvard University Press.

Ellis, R. 1999. “Item versus System Learning: Explaining Free Variation”. *Applied Linguistics* 20. 460–480.

Friederici, A. D. 1983. “Children’s Sensitivity to Function Words during Sentence Comprehension”. *Linguistics* 21. 717–739.

Gibbs, R. W. 1994. *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge and New York: Cambridge University Press.

Hampe, B. 2000. “Facing up to the Meaning of ‘face up to’: A Cognitive Semantico-pragmatic Analysis of an English Verb-particle Construction”. In A. Foolen and F. V. D. Leek (eds), *Constructions in Cognitive Linguistics: Selected Papers from the Fifth International Cognitive Linguistics Conference, Amsterdam, 1997*, Amsterdam: John Benjamins. 81–101.

Ishii, Y. 2005. “Eigo no Fuhenkashi no Imi to Goiteki Imi toshite Naizaika Sareta Metafaa [The Semantics of English Particles and the Metaphors Internalized into Their Lexical Meanings]”. In Y. Tsuruga, T. Takagaki and K. Urata (eds), *Gengo Jouhougaku Kenkyuu Houkoku 7: Koopasu Gengogaku ni okeru Goi to Bunpou* [*Working Papers in Linguistic Informatics 7: Lexicon and Grammar in Corpus Linguistics*], The 21st Century COE Program “Usage-Based Linguistic Informatics”, Graduate School of Area and Culture Studies, Tokyo University of Foreign Studies. 497–520.

Ishii, Y. 2006. “Metafaa no Kanten kara Mita Eigo no Fuhenkashi ni Mirareru Imi no Kaisousei [On Three Metaphoric Levels of English Particles]”. In Y. Kawaguchi, I. Kameyama, N. Tomimori and T. Takagaki (eds), *Gengo Jouhougaku Kenkyuu Houkoku 9: Shimpojiumu, Kouenkai, Kenkyuu Houkoku* [*Working Papers in Linguistic Informatics 9: Symposium, Lecture, Report*], The 21st Century COE Program “Usage-Based Linguistic Informatics”, Graduate School of Area and Culture Studies, Tokyo University of Foreign Studies. 186–217.

Ishii, Y. in press. “Metaphors in English Verb-particle Combinations and Learners’ Difficulty with Phrasal Verbs”. In M. Murata, K. Minamide, Y. Tono and S. Ishikawa (eds), *English Lexicography in Japan*, Tokyo: Taishukan Shoten.

Kunihiro, T. 1982. *Imiron No Houhou* [*The Methodology of Semantics*]. Tokyo: Taishukan Shoten.

Kunihiro, T. (translation supervision and annotation). *Eigo Zenchishi no Imiron* [*The Semantics of English Prepositions*]. By A. Tyler and V.

Evans. Tokyo: Kenkyusha, 2005. Trans. by T. Kimura.

Lakoff, G. 1987. *Woman, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago and London: The University of Chicago Press.

Lakoff, G. and M. Johnson. 1980. *Metaphors We Live By*. Chicago and London: The University of Chicago Press.

Langacker, R. W. 1993. “Reference-point Constructions”. In *Cognitive Linguistics* 4. 1–38.

Langacker, R. W. 1999. *Grammar and Conceptualization*. Berlin and New York: Mouton de Gruyter.

Lanham, R. 1969. *A Handlist of Rhetorical Terms*. Berkeley: University of California Press.

Sugai, K. “Synecdoche”. In Y. Tsuji (ed), *Ninchigengogaku Kiiwaado shu*u [*An Encyclopedic Dictionary of Cognitive Linguistics*], Tokyo: Kenkyusha, 2002. 158–159.

Taylor, J. R. 2003. *Linguistic Categorization,* Third Edition. Oxford: Oxford University Press.

Traugott, E. 1988. “Pragmatic Strengthening and Grammaticalization”. In S. Axmaker, A. Jaisser and H. Singmaster (eds), *General Session and Parasession on Grammaticalization: Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistic Society*. 406–416.

Traugott, E. 1989. “On the Rise of Epistemic Meanings in English: An Example of Subjectification in Semantic Change”. *Language* 65. 31–55.

Tyler, A. and V. Evans. 2003. *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning and Cognition*. Cambridge: Cambridge University Press.

Ullman, S. 1962. *Semantics: An Introduction to the Science of Meaning*. Oxford: Basil Blackwell & Mott Ltd.

Waldron, R. A. 1979. *Sense and Sense Development*, Second Edition. London: Andre Deutsch.

Corpus

BNC: the Second Edition of the British National Corpus.

Language Policy and Language Choice — A Case Study at Canadian Government Institutions —

Norie YAZU and Yuji KAWAGUCHI

1. Introduction

Since the enactment of the "Official Languages Act" in 1969, Canada has developed its own official languages policy. Although this Act stipulates that the Government of Canada is committed to "fostering the full recognition and use of both English and French in Canadian Society" (Official Languages Act, 1988, Part VII, 41(b)), Canada's official languages policy is founded on the notion of "institutional bilingualism," in which both official languages may be used by federal government institutions when various services are provided to the public but are not imposed on the public in their private communications. The Act only regulates the language use of the employees of approximately 180 institutions, including the federal departments, crown corporations, and other organizations that are subject to the Act. These employees are required to provide service to the public in both official languages in offices where demand is significant, or where the nature of the office requires it. Their language use is also regulated with regard to internal communication, which is the focus of this paper.

The Canadian federal government has issued a clear policy regarding the "language of work" among federal public servants. In regions designated as bilingual for the purposes of language of work[1], federal public servants may use the official language of their choice. Above all, members of the executive group, who are required to perform supervisory duties, must comply with the proficiency standards of the second official language established by the policy[2]. A "bilingual bonus" is paid to federal public servants who are assessed as being bilingual.

This paper aims to analyze the language choices made by federal public servants in a bilingual workplace by focusing on the comparison between the

[1] These designated regions are the National Capital Region, northern and eastern Ontario, New Brunswick, Montreal, and some parts of the Eastern Townships, of Gaspé, and of the Outaouais in Quebec.

[2] Their second official language abilities must be evaluated as "advanced" (level C) in reading, at least 'intermediate' (level B) in writing, and 'advanced' (level C) in speaking.

language choices made by bilingual Anglophones and bilingual Francophones. Data were collected from a questionnaire survey conducted by Yazu in 10 federal government institutions in the National Capital Region of Canada. In this study, we intend to observe the ways in which language choices are made in work environments that are regulated by a language policy. More precisely, our study focuses on how the two languages, which have asymmetric strength in society and are given equal status by law, are actually used by the people subject to the language policy. In the case of Canada, English, compared to French, is the overwhelmingly dominant language in terms of both demolinguistics and social interactions; this is true not only in federal government institutions but also in Canadian society as a whole[3], with the exception of the province of Quebec, where francophones are the majority.

2. Research in institutions

In the past, a number of linguistic researches have been carried out in institutions. In Canada, the research of one prominent scholar, Heller (1979, 1982) deserves attention. In Heller's research, fieldwork was conducted in an English-speaking hospital in Montreal for the purpose of analyzing the language choices of bilingual workers and patients.

With regard to countries other than Canada, the Wellington Language Project of New Zealand, which began in 1996, analyzed the ways in which people actually communicate at work. Many hours of natural conversation were tape-recorded in organizations such as government institutions, commercial shops, factories, and hospitals. In general, volunteers recorded a range of their daily work interactions over a period of two to three weeks; some kept a recorder and microphone at their desks and others carried the equipment around with them (Holmes 2003). In Australia, Béal (1994), in her study that analyzed cross-cultural communication problems in the workplace, interviewed employees working in a French firm operating in Australia and tape-recorded several hours of the employees' office conversations.

The research whose survey method and study objectives are the closest to those of our present study is the one carried out by Quell (1997, 1998) at the European Commission, in which language choice in multilingual work settings was analyzed by a questionnaire survey. Quell, who was then a

[3] If we examine the entire Canadian population, approximately 68% are Anglophones (English-speaking) and 23% are Francophones (French-speaking). Most Anglophones speak only English, and those who are bilingual in English and French constitute only 17.7% of Canada's total population, most of whom are Francophones (Statistics Canada 2002).

trainee at the European Commission, distributed questionnaires to as many trainees as possible; 274 trainees completed and returned the questionnaires. Although Quell was not detached from the institution he investigated, his study is outstanding in that a fairly large-scale survey was conducted by a single researcher in an international mega-institution, where research for personal academic purposes is usually not feasible.

Formal written permission is usually required from the head of the institution before any action can take place, resulting in a complex and time-consuming process. Our research, which was conducted in Canadian government institutions, was said to be the first to be undertaken by outsiders for personal academic purposes. Although the process was time-consuming, written permission was obtained owing to the assistance of the Foreign Ministry of Japan, which funded our research, and the generosity of the Canadian government; thus, our research was conducted quite smoothly with the cooperation of many Canadian federal public servants. In the following sections, we will describe the survey method and analyze the main results of our survey.

3. Survey method employed in the study

3.1. A previous study

A study that analyzed the language choices of bilingual federal public servants in Canada, a subject in which we take an interest, was conducted in 1993, as a study commissioned by the Commissioner of Official Languages, the ombudsman for the implementation of Canada’s official languages policy. This study, whose results were published as a government report entitled *Negotiating Language Choice in the Federal Civil Service* (Hay Management Consultants 1993), was not intended to be used for academic purposes but, nevertheless, has important implications for our research.

In this study, 52 bilingual federal public servants—who were working in Ottawa and Montreal and who had attained at least an intermediate level of proficiency in the second official language (level B in reading, writing, and speaking)—filled out a questionnaire and participated in group discussions. Due to the small number of respondents and the nature of the questions posed, significant results regarding detailed language use in the workplace did not emerge from the questionnaire survey. However, the group discussions, which were based on the participants’ daily experiences, yielded a large amount of information about their language use in various workplace situations and their perceptions of the factors that influence their language choices. This information helped us create the questions in our survey.

This 1993 government study indicated that in Ottawa, French was generally underused in most situations at work as a result of the disparity in the second official language (OL2) proficiency of Anglophones and Francophones. The highlight of this study was the identification of the ten factors that seem to influence the language choices of bilingual public servants[4], the most important of which were described as follows:

> Linguistic ability itself is a key determinant of language choice, but also a major determinant of the decision to continue a conversation in one's second official language. Regardless of facility in the second language, we heard evidence that most people will tend to revert to their first official language when the subtleties and nuances of the second language become complex. The overriding determinant is functionality...............*situational functionality* appears to be the single most influential. People will choose the path of least resistance to getting the job done. In Ottawa, this means functioning in English, which is the language of the majority. (Hay Management Consultants, 1993 pp.24–5)

Thus, this study states that "linguistic ability" and, more importantly, "situational functionality" (in other words, "efficiency of communication") are the two primary determinants of the language choices of bilingual public servants. In Ottawa (the National Capital Region), although the overall tendency of language choice leans towards English, the majority language, there are such situations in which French, the minority language, is used. For example, this is the case when Francophones declare certain "minority rights" or Anglophones are motivated by organizational "rewards," which imply career advancement.

The objective of our present study is to quantitatively examine the qualitative findings of the 1993 government study. The questions in our questionnaire were designed to examine the detailed language choices of bilingual public servants in various workplace situations, which we term "subdomains" in our study, following the concept used by Chien (2002), which extended Fishman's (1965) "domain" concept[5]. We identified 17 subdomains according to the attributes of the interlocutor, topic of conversation, specific settings, etc., as are shown on the list of questions in

[4] These ten factors were presented in the following three categorizations: individual characteristics (1. psychological processes, 2. cultural identity, 3. linguistic ability, and 4. language acquisition experience); situational characteristics (5. situational identity, 6. situational functionality, 7. work unit integration and teamwork, and 8. situational power); and external characteristics (9. organizational rewards and sanctions and 10. societal rewards and sanctions).

[5] Chien, in her study of language choices in Taiwan, for example, divided her "public places" domain into what she called "subdomains" such as "talking to superiors," "talking to classmates," "at a government (municipal) office," "at a hospital," "at a store (small business)," and "at a department store" (Chien 2002).

the subsequent section. This method enables us to quantitatively observe, for example, the subdomains in which English is used the most by Anglophones and by Francophones or the subdomains in which Anglophones tend to use French.

3.2. Our survey method

The first procedure was to obtain written permission from the Secretary of the Treasury Board of Canada[6]. Referring to the Treasury Board's database on "Anglophone and Francophone Participation by Institution (PCIS, as of March 31, 2001)," we identified several institutions as possibilities for our research. The selection of these institutions was based on the size (number of employees) and the proportions of Anglophones and Francophones in each institution. The average of the latter was 69.0% for the Anglophones and 31.0% for the Francophones (Treasury Board 2002a). We inferred, as was also indicated in the 1993 government study, that in institutions where Anglophones were over-represented, English was extremely dominant, whereas in those where Francophones were over-represented, English was not as dominant. Therefore, in order to represent language use in all institutions, our selection comprised institutions in which Anglophones were over-represented (a > f in table 1), those in which Francophones were over-represented (a < f), and also those in which the proportion of Anglophones and Francophones was fairly balanced (a = f). However, because some institutions declined to cooperate with our research, we were ultimately able to carry out our research in the 10 institutions shown in table 1, some of which had not been originally targeted:

Table 1. Institutions in which our research was conducted

Anglophones over-represented a > f	Anglophones and Francophones fairly balanced a = f	Francophones over-represented a < f
-Fisheries and Ocean -Health Canada -Western Economic and Diversification -Atlantic Canada Opportunities Agency	-Human Resources Development Canada -Public Works and Government Services -Industry Canada	-Heritage Canada -Office of the Commissioner of Official Languages -Tax Court

a: Anglophone, f: Francophone

[6] At the time our research was conducted, the Official Languages Act stipulated that the Treasury Board of Canada was the agency responsible for the general direction and coordination of the policies and programs of the Government of Canada relating to service to the public, language of work, and equitable participation in the Public Service of Canada. Currently, these responsibilities are assumed by the Public Service Human Resources Management Agency of Canada, a newly created agency resulting from the Cabinet shuffle in 2004.

The questionnaires were distributed from the end of October to mid-November of 2002. Anglophones received an English version and Francophones received a French version. Of the 320 copies distributed, 265 were completed and returned, constituting a high return rate of 82.8%, among which 253 were valid—113 from the Anglophones and 140 from the Francophones.

3.3. The second official language proficiency of the respondents

It must be noted that the situation of the OL2 proficiency of federal public servants (OL2: French for Anglophones and English for Francophones)[7] in the National Capital Region somewhat reflects that of the Canadian society as a whole, in which Anglophones form the majority and are mostly monolingual whereas many Francophones tend to be bilingual. A government study reports that even in "bilingual regions for the purpose of language of work," including the National Capital Region, only 51% of Anglophone public servants can speak French, fluently or quite fluently whereas 91% of the Francophone public servants can speak English fluently or quite fluently (Treasury Board 2002).

Since the objective of our research was to examine the language choices of bilingual public servants, the respondents of our questionnaire had to be bilingual. To ensure this, under the supervision of the official languages coordinator in each institution, we identified the public servants whose OL2 speaking abilities were at least "intermediate (level B)" and conducted a random sampling before distributing the questionnaires. However, even after having collected them, we observed a disparity in OL2 proficiency between the Anglophones and Francophones, as is shown in figure 1.

[7] In Canadian government institutions, the *first official language* (OL1) is defined as "the official language with which an employee has a primary personal identification" (that is, the official language in which a person is generally more proficient). Therefore, *the second official language* (OL2) is the official language in which a person is generally less proficient. Further, an *Anglophone* is defined as "any person, of whatever ethnic origin or mother tongue, whose first official language is English" whereas a *Francophone* is defined as "any person, of whatever ethnic origin or mother tongue, whose first official language is French" (Treasury Board, Policy on Official Languages, Chapter 6-1-Glossary). However, it should be noted that for general use in society, an Anglophone is perceived to be a person who ordinarily speaks English and a Francophone, a person who ordinarily speaks French.

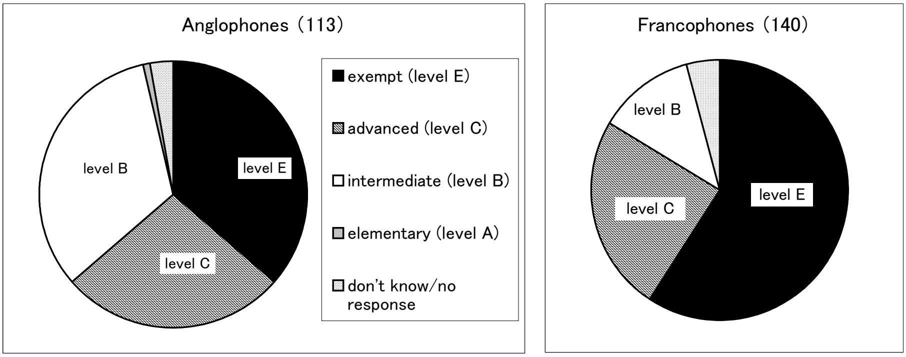


Figure 1. OL2 proficiency of the respondents

Although our respondents are highly bilingual in comparison to the average OL2 proficiency of the public servants as a whole, we can see that our francophone respondents are more proficient in their OL2 than their anglophone counterparts. 83.6% of the Francophones and 63.7% of the Anglophones had achieved the advanced (C level) and the most advanced "exempt" (level E) levels. 59.3% of the Francophones and 36.3% of the Anglophones had achieved level E.

3.4. Questionnaire questions

The questions in our questionnaire, as shown below, pertained to language choices in 17 different workplace situations, which we refer to as subdomains (see footnote 5). We also asked additional questions, such as those regarding the frequency of and reasons for code-switching and the perception of passive bilingualism, which are not focused in this paper.

Questionnaire questions

<when initiating a conversation>

subdomain 1: do not know if the interlocutor is an Anglophone or a Francophone

subdomain 2: interlocutor is the speaker's supervisor

subdomain 3: knowing that the interlocutor is a bilingual colleague (other than the supervisor) whose OL1 is different from one's own

<when answering>

subdomain 4: having initiated a conversation with a bilingual colleague whose OL1 is different from the speaker's in the speaker's OL1, the interlocutor responds in the speaker's OL2

subdomain 5: having initiated a conversation with a bilingual colleague whose OL1 is different from the speaker's in the speaker's OL2, the interlocutor responds in the speaker's OL1 subdomain 6: when the conversation has been initiated by a bilingual colleague whose OL1 is different from the speaker's in the speaker's OL2

subdomain 7: when the conversation has been initiated by a bilingual colleague whose OL1 is different from the speaker's in the speaker's OL1

<topic>

subdomain 8: when speaking with a bilingual colleague whose OL1 is different from the speaker's about work-related topics, including technical terms

subdomain 9: when speaking with a bilingual colleague whose OL1 is different from the speaker's about matters not related to work

<when having difficulty expressing oneself in OL2>

subdomain 10: when speaking in OL2, face difficulty expressing a certain word, phrase, or idea in OL2

<when speaking with a bilingual colleague whose OL1 is the same as that of the speaker>

subdomain 11: when speaking with a bilingual colleague whose OL1 is the same as the speaker's

<when writing e-mails>

subdomain 12: when writing an informal e-mail message to a bilingual colleague whose OL1 is different from the sender's

subdomain 13: when writing an informal e-mail message to a bilingual colleague whose OL1 is the same as the sender's

subdomain 14: when writing a formal e-mail message to a group of colleagues within one's work unit

subdomain 15: when writing a formal e-mail message intended for distribution beyond one's work unit

<when attending meetings>

subdomain 16: when the meeting is attended by more Anglophones than Francophones

subdomain 17: when the meeting is attended by more Francophones than Anglophones

The above subdomains include a variety of situations, including those in which the interlocutor is a bilingual whose first official language (OL1) is the same as or different from that of the speaker, conversations with a supervisor or an unknown person, as well as communications in meetings and e-mails. The nine underlined subdomains denote the situations in which the interlocutor is a bilingual whose OL1 is different from that of the speaker—communication between a bilingual Anglophone and a bilingual Francophone. These will be distinguished from the other subdomains in the calculation of the "overall index" described in the next section.

3.5. Calculation of the overall index

Instead of analyzing the results of each question in the questionnaire, in this paper, we examine the respondents' general language choices in the workplace, using an index that we termed the "overall index". To calculate the overall index, the answers to each question must be quantified. The answer in which we observed the strongest degree of choosing English ("only in English") is allotted a score of 2, the highest positive score, and the answer in which we observed the strongest degree of choosing French ("only in French") is allotted a score of –2, the lowest negative score. Table 2 takes subdomain 1 as an example:

Table 2. Scores allotted in the calculation of the "overall index"

subdomain 1: *"When you are not sure whether the colleague you are about to speak to is an Anglophone or a Francophone, how do you initiate a conversation?"*

answer	score
only in English	2
mostly in English	1
in English or French	0
mostly in French	–1
only in French	–2

Based on this scale, two types of 'overall index' are calculated. One is the "overall index by individuals," which is calculated by totaling the scores of all the subdomains for each respondent to examine the general tendency of his or her language choice. Respondents with a positive total score tend to use more English than French, while those with a negative total score tend to use more French than English. The higher the score, the more English tends to be used, while the lower the score, the more French tends to be used. Two examples are provided below:

Table 3. Examples of the calculation of the "overall index by individuals"

	respondent no. 1	respondent no. 2
subdomain 1	2	–2
subdomain 2	1	0
subdomain 3	2	–1
	:	
subdomain 17	2	–1
total	25	–11

Table 3 shows that respondent no. 1, whose total score was 25, has quite a strong overall tendency to use English, and respondent no. 2, whose total score was –11, tends to use French more than English.

We term the other type of overall index the "overall index by subdomains"; this is calculated by totaling the scores of all the respondents for each subdomain to identify the respondents' overall language choices in each subdomain. Thus, we can observe, for example, the subdomains in which English or French tend to be used predominantly or those in which the use of English and French is quite balanced.

4. Survey results

4.1. Overall index by individuals

Our survey compares two types of overall index by individuals. One is the index that is calculated by totaling the scores of all the subdomains. These subdomains, which include many situations, if not all, that a public servant is likely to experience at work, more or less represent the overall language situation of the respondents' work environment. The other type is the index calculated by totaling the scores of the nine subdomains underlined in the list of questions, which represent the language choices when the interlocutor and the questionnaire respondent are bilinguals with different OL1s. In other words, the latter index focuses on communication between a bilingual Anglophone and a bilingual Francophone.

Figure 2 compares these two types of overall index by individuals based on language groups (Anglophones/Francophones) and the proportion of Anglophones and Francophones in each institution (a > f: Anglophones over-represented, a = f: Anglophones and Francophones balanced, a < f: Francophones over-represented); the first index described above is indicated as "all subdomains"; the second, "between bilinguals."

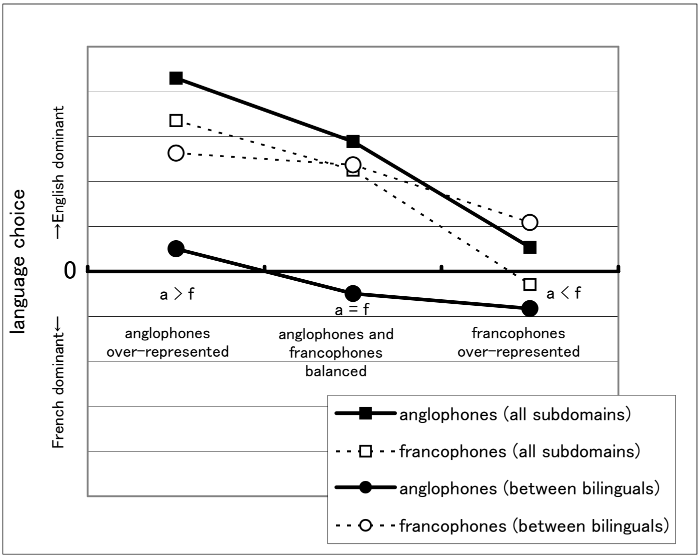


Figure 2. Comparison of the overall indices by individuals

The indices of all subdomains (■ and □) clearly show that, as expected, the language choices of both Anglophones and Francophones are affected by the proportion of Anglophones and Francophones in the workplace. It is inferred from figure 2 that in institutions where Anglophones are over-represented, the work environment is strongly dominated by the use of English. On the other hand, in institutions where Francophones are over-represented, the indices for both Anglophones and Francophones approach the score of 0, which implies that the overall use of English and French is fairly balanced.

What is the most noteworthy in figure 2 is that the "between bilinguals" index for Anglophones (●) differs greatly from that for Francophones (○). This implies that in communications between bilingual Anglophones and bilingual Francophones, Francophones are strongly inclined to choose English, although this tendency is observed to be weak in institutions in which Francophones are over-represented. On the other hand, when bilingual Anglophones communicate with bilingual Francophones, they are not inclined to choose English as much as the Francophones. Rather, except in institutions where Anglophones are over-represented, Anglophones tend to choose French slightly more often than English.

The aforementioned 1993 government study stated that the use of French was insufficient in the workplace but did not elucidate what or who triggers this situation. As depicted in figure 2, our study suggests that it is the Francophones rather than the Anglophones who contribute to the predominance of English in communications between bilingual Anglophones and bilingual Francophones. Bilingual Anglophones are observed to endeavor to choose French, especially in institutions where Francophones are over-represented. In the next section, we will examine which language tends to be used in which subdomain, with special attention to the subdomains in which English tends to be used predominantly by Francophones and those in which French tends to be used predominantly by Anglophones.

4.2. Overall index by subdomains

For our analysis of the "overall index by subdomains," we calculate the indices for Anglophones and Francophones separately and make them comparable. Figure 3 compares the overall language choices of Anglophones and Francophones in each subdomain[8]. We categorize the subdomains in which both Anglophones and Francophones register positive scores as "English dominant," those in which Anglophones register negative scores and Francophones register positive scores as "OL2 dominant," and those in which Anglophones register positive scores and Francophones register negative scores as "OL1 dominant." There were no subdomains in which both Anglophones and Francophones registered negative scores ("French dominant").

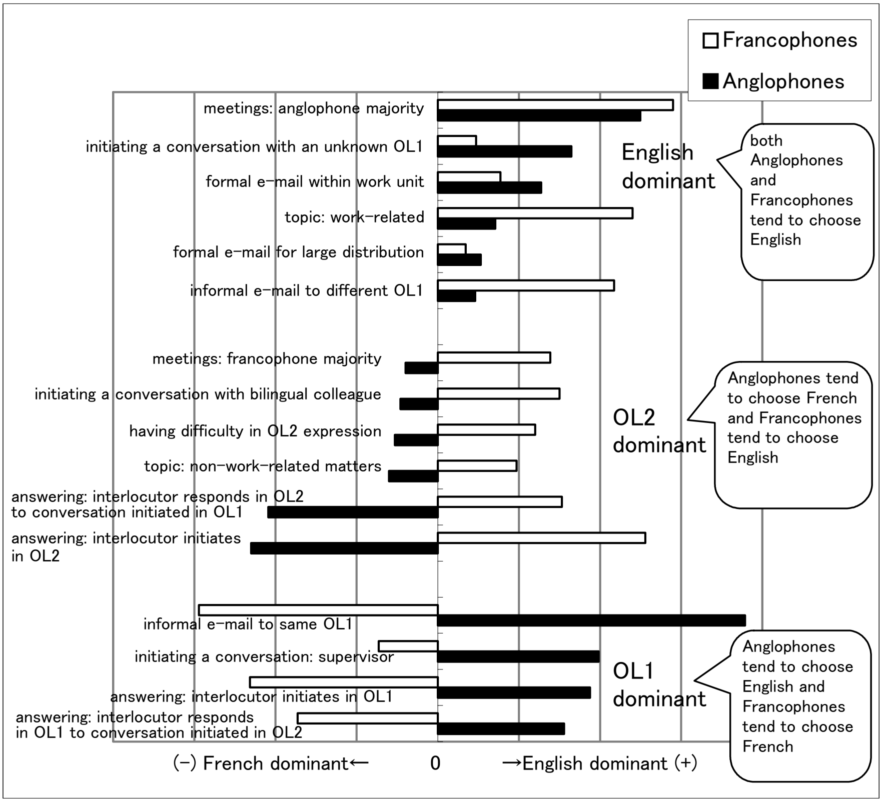


Figure 3. Overall index by subdomains (Anglophones, Francophones)

[8] In figure 3, subdomain 11 ("speaking with a bilingual colleague whose OL1 is the same as that of the speaker") is excluded because the question for subdomain 11 was not asked in the same way as the other questions. However, if we had forcedly included subdomain 11 in figure 3, it would have been included in the "OL1 dominant" category. Our data for subdomain 11 show that Anglophones speak with Anglophones primarily in English and Francophones speak with Francophones primarily in French, which is natural; however, what is noteworthy is that 1 out of 10 Francophones "usually" or "often" speaks with a Francophone in English.

Figure 3 shows that the subdomain in which English is most often chosen by both Anglophones and Francophones is “meetings attended by more Anglophones than Francophones.” This could imply that since some Anglophones are monolingual or have insufficient abilities in French, English tends to be used in meetings. If we examine the English dominant subdomains, we can detect the subdomains in which only Francophones have a strong tendency to choose English. Further, from the OL2 dominant category, we can perceive the subdomains in which Anglophones tend to choose French more than English. In what follows, we will try to ascertain which factors determine the choice between English and French.

In order to pursue this objective, we distinguish between the subdomains that are and are not affected by the language choice or the attributes of the interlocutor and exclude the former from our present analysis. Included in the former are the following six subdomains from the lower portion of figure 3: the four subdomains of “answering,” “writing an informal e-mail message to a bilingual colleague whose OL1 is the same as that of the sender,” and “initiating a conversation with a supervisor.” In these six subdomains, it is natural for the respondents to comply with the language choice or OL1 of the interlocutor. It is in the other subdomains that the respondents make free language choices without constraints from the interlocutor.

With regard to the Francophones, the English dominant subdomains in which only Francophones have a strong tendency to choose English are “writing an informal e-mail message to a bilingual colleague whose OL1 is different from that of the sender” and “when speaking about work-related topics including technical terms” (although this tendency is weak in institutions in which Francophones are over-represented). Including the subdomain “meetings attended by more Anglophones than Francophones,” in which both Anglophones and Francophones have a strong tendency to choose English, these three subdomains have one feature in common—efficiency of communication badly deteriorates when French is used.

With regard to the Anglophones, “when speaking about work-related topics including technical terms” and “writing an informal e-mail message to a bilingual colleague whose OL1 is different from that of the sender” are the two subdomains in which the Anglophones’ tendency to choose English is not as strong as that of the Francophones. In the following four subdomains, Anglophones tend to choose French slightly more often than English (although this tendency is weak in institutions where Francophones are over-represented): “when starting a conversation knowing that the interlocutor is a bilingual colleague whose OL1 is different from one’s own,” "meetings attended by more Francophones than Anglophones," "when having difficulty expressing oneself in OL2," and "when speaking about matters not related to work."

Among these, the first two are the subdomains in which efficiency is not very important. Concerning "meetings attended by more Francophones," it was reported by some of the respondents of our questionnaire and also indicated in the interviews that because Francophones are the minority, this situation is unusual in most institutions, except for the few in which Francophones are over-represented. In our analysis of "when having difficulty expressing oneself in OL2," we discovered that many of our Anglophone respondents do not switch easily to English but persist in French.

Briefly, the subdomains in which English is predominantly used are those in which efficiency is important, and it is the Francophones who contribute to this situation. Anglophones endeavor to use French, especially in subdomains in which efficiency is not very important.

5. General Conclusion

The aforementioned 1993 government study concluded that efficiency of communication is the most influential factor that explains the language choices of bilingual federal public servants. Our study confirms this point with the proviso that it is the Francophones more than the Anglophones that tend to use English to accelerate efficiency when communicating with each other.

The communication strategies of bilinguals are governed by the "linguistic competence principle," which is described by Hamers & Blanc (2000) as follows:

> The code selected in the interaction is that in which the sum of the individual communicative competences of the interlocutors is maximum. Code selection or choice is defined here as the speaker's decision, in a given communicative interactional situation, to use one code rather than another.

If we adapt this principle to our study of communication between bilingual Anglophones and Francophones, English—in which the sum of the individual communicative competences of the interlocutors is higher—is naturally selected. However, the application of the "linguistic competence principle" may be counteracted by social factors such as language policy, and this is what we also observe in our study. Canada's official languages policy strives to create a work environment conducive to the balanced use of both English and French as languages of work. In this environment, a struggle between English and French is constantly observed.

In order to describe this situation, we suggest the following four types of mutual language choices that can occur during communications between bilingual Anglophones and bilingual Francophones: “accommodation[9] to English,” in which both Anglophones and Francophones speak English; “accommodation to French,” in which both Anglophones and Francophones speak French; “passive bilingualism,” in which Anglophones speak English and Francophones speak French; and “hyperaccommodation,” in which Anglophones speak French and Francophones speak English. These are incorporated in figure 4, which depicts the dynamics of language policy and language choice.

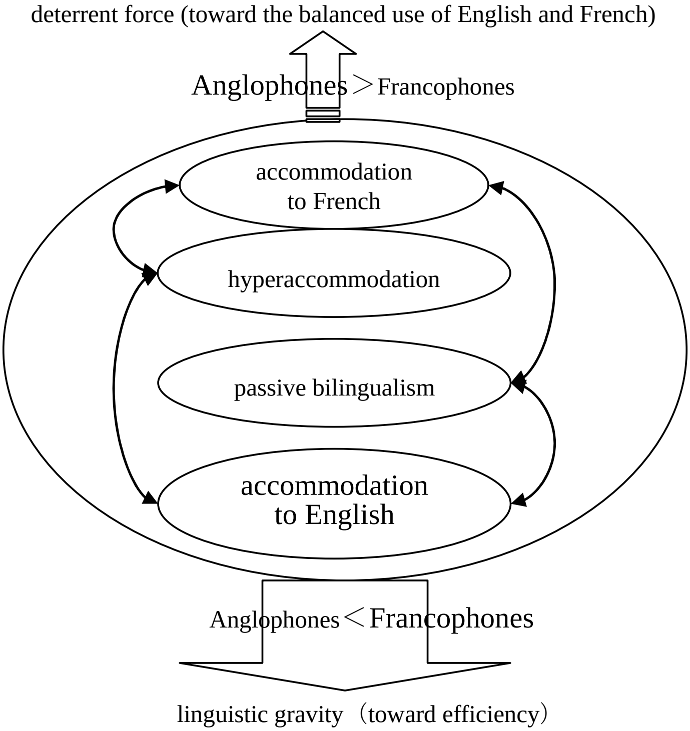


Figure 4. Dynamics of language policy and language choice

It is observed that in actual communications between bilinguals, the first

[9] In this paper, the term “accommodation” is derived from the “accommodation theory” developed by Giles et al. (1973). Two types belong to this theory, “convergence” and “divergence”; for the purpose of our study, the meaning of the term “accommodation” is limited to the former, and the subject of “accommodation” is limited to language choice.

mutual language choice is usually not fixed but fairly frequently converts to other mutual language choices. For example, accommodation to English may in one conversation convert first to passive bilingualism, then accommodation to French, then hyperaccommodation. Our data, which show a high frequency of code-switching for both Anglophones and Francophones[10], support this observation.

Within this interactional situation, an attractive force toward efficiency, which we term "linguistic gravity," progresses toward accommodation to English, which is the most efficient method of communication. In an environment governed by language policy, this linguistic gravity is countered by a deterrent force that progresses toward creating a balanced use of English and French. This deterrent force is strengthened by the Anglophones who recognize the incentives for career advancement.

Although there appears to be ambiguity regarding the definition of "a balanced use of English and French," it can roughly be described as a situation in which accommodation to English does not predominate most of the time. Passive bilingualism was once suggested in a government report (Treasury Board 2002b), but this has never been overtly encouraged. Passive bilingualism can easily shift to accommodation to English as a result of the Francophones' strong inclination toward efficiency and their high frequency of code-switching, as is shown in our study.

Considering that linguistic gravity is always predominant over the deterrent force produced by language policy, overtly encouraging Anglophones to continue with their efforts and Francophones to actively choose French might be the key to the realization of a work environment conducive to the balanced use of English and French.

References

Béal, Christine. 1994. "Keeping the peace: a cross-cultural comparison of questions and requests in Australian English and French" In *Multilingua*, Vol.13.

Chien, Yuehchen. 2002. "Language contact in Taiwan" In *The Japanese Journal of Language in Society*, Vol.4, No.2.

Fishman, Joshua A. 1965. "Who speaks what language to whom and when?" In *La Linguistique*, Vol. 2.

Giles, Howard, Donald M. Taylor & Richard Bourhis. 1973. "Towards a

[10] Our data on code-switching show that 81.6% of Anglophone respondents and 70.7% of Francophone respondents report switching from OL2 to OL1 "sometimes" or more frequently. The figures for code-switching from OL1 to OL2 are 84.0% and 91.4%, respectively. What is noteworthy is that 60.0% of Francophones report switching from French to English "often," "usually," or "always."

theory of interpersonal accommodation through language: some Canadian data” In *Language and Society*, Vol.2, No. 2. October.

Giles, Howard & Philip Smith. 1979. “Accommodation theory: optimal levels of convergence” In Giles, H & R. N. St. Clair (eds.) *Language and Social Psychology*, Oxford: Blackwell.

Government of Canada. 1999. *Official Languages Act*, Ottawa: Minister of Public Works and Government Services Canada.

Hamers, Josiane & Michel H. A. Blanc. 2000. *Bilinguality and Bilingualism*, Cambridge: Cambridge University Press.

Hay Management Consultants. 1993. *Final Report of Pilot Focus Group Results on Negotiating Language Choice in the Federal Civil Service*, Toronto: Hay Management Consultants.

Heller, Monica. 1979. “Bonjour, hello?: Négociation de choix de langue à Montréal” In Pierette Thibault (ed.) *Le Français Parlé: Études Sociolinguistiques*, Edmonton: Linguistic Research.

Heller, Monica. 1982. “Negotiations of Language Choice in Montreal” In Gumperz, J. (ed.) *Language and Social Identity*, Cambridge: Cambridge University Press.

Quell, Carston. 1997. “Language choice in multilingual institutions: A case study at the European Commission with particular reference to the role of English, French, and German as working languages”, In *Multilingua* Vol.16, No.1, Berlin: Mouton de Gruyer.

Quell, Carston. 1998. “Requirements, dynamics and realities of language use in the EU: a case study of the European Commission” In Kibbee, Douglas A. (ed.) *Language Legislation and Linguistic Rights*, Amsterdam: John Benjamins Publishing Company.

Statistics Canada. 2002. *Profile of languages in Canada: English, French and many others, 2001 Census: analysis series*, Ottawa: Ministry of Industry.

Treasury Board of Canada. 2002a. *Annual Report on Official Languages 2001–2002.*

Treasury Board of Canada. 2002b. *Attitudes Towards the Use of Both Official Languages within the Public Service of Canada*, unpublished report.

Yazu, Norie. 2005. *The Official Languages Policy in Canada: Language Choice of Bilingual Public Servants*. Ph. D. dissertation, Tokyo University of Foreign Studies.

Index of Proper Nouns

Names

Index of Subjects

Contributors

Yuji KAWAGUCHI	Faculty of Foreign Studies, Tokyo University of Foreign Studies
Claire BLANCHE-BENVENISTE	Ecole Pratique des Hautes Etudes, Department of French Linguistics, University of Aix-Marseille I
Susan M. CONRAD	Department of Applied Linguistics, Portland State University
Massimo MONEGLIA	Italian Department, University of Florence
Emanuela CRESTI	Italian Department, University of Florence
Susumu ZAIMA	Faculty of Foreign Studies, Tokyo University of Foreign Studies
Toshihiro TAKAGAKI	Faculty of Foreign Studies, Tokyo University of Foreign Studies
José DEULOFEU	Department of French Linguistics, University of Aix-Marseille I
Antonio MORENO-SANDOVAL	Faculty of Philosophy, Autonomous University of Madrid
José Maria GUIRAO-MIRAS	School of Informatics, University of Granada
Maria Fernanda Gorjão Bacelar do Oliveira NASCIMENTO	Linguistics Center, Lisbon University
José Bettencourt GONÇALVES	Linguistics Center, Lisbon University
Maria Amália Pereira MENDES	Linguistics Center, Lisbon University
Sandra Cristina dos Santos ANTUNES	Ph.D. candidate, Linguistics Center, Lisbon University

Selim YILMAZ	Faculty of Arts and Science, Marmara University
Isamu SHOHO	Faculty of Foreign Studies, Tokyo University of Foreign Studies
Makoto MINEGISHI	Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies
Hide TAKASHIMA	Faculty of Foreign Studies, Tokyo University of Foreign Studies
Chihiro INOUE	Ph.D. Program, Graduate School, Tokyo University of Foreign Studies
Chiaki YAMANE	Otsuma Tama Senior and Junior High School
Natsuko UZAWA	Obirin Junior and Senior High School
Mayo NAGATA	M.A. Program, Graduate School, Tokyo University of Foreign Studies
Takayuki SADAHIRO	Japan Business Systems, Inc.
Yukiko SHIMAMURA	Fujimi Junior High School
Mieko TAKADA	Ph.D. candidate, Faculty of Foreign Studies, Tokyo University of Foreign Studies
Nobuo TOMIMORI	Faculty of Foreign Studies, Tokyo University of Foreign Studies
Yasutake ISHII	Ph.D. candidate, Faculty of Foreign Studies, Tokyo University of Foreign Studies
Kiyoko SOHMIYA	Faculty of Foreign Studies, Tokyo University of Foreign Studies
Norie YAZU	Lecturer, Kanda University of International Studies

In the series *Usage-Based Linguistic Informatics* the following titles have been published thus far or are scheduled for publication:

5 **KAWAGUCHI, Yuji, Susumu ZAIMA and Toshihiro TAKAGAKI (eds.):** Spoken Language Corpus and Linguistic Informatics. 2006. vi, 434 pp.

4 **YOSHITOMI, Asako, Tae UMINO and Masashi NEGISHI (eds.):** Readings in Second Language Pedagogy and Second Language Acquisition. In Japanese Context. 2006. vi, 274 pp.

3 **KAWAGUCHI, Yuji, Ivan FÓNAGY and Tsunekazu MORIGUCHI (eds.):** Prosody and Syntax. Cross-linguistic perspectives. 2006. vi, 384 pp.

2 **TAKAGAKI, Toshihiro, Susumu ZAIMA, Yoichiro TSURUGA, Francisco MORENO FERNÁNDEZ and Yuji KAWAGUCHI (eds.):** Corpus-Based Approaches to Sentence Structures. 2005. vi, 317 pp.

1 **KAWAGUCHI, Yuji, Susumu ZAIMA, Toshihiro TAKAGAKI, Kohji SHIBANO and Mayumi USAMI (eds.):** Linguistic Informatics – State of the Art and the Future. The first international conference on Linguistic Informatics. 2005. viii, 363 pp.