Language, People, Numbers

Corpus Linguistics and Society

Edited by
Andrea Gerbig
and Oliver Mason

Language, People, Numbers

LANGUAGE AND COMPUTERS: STUDIES IN PRACTICAL LINGUISTICS

No 64

edited by
Christian Mair
Charles F. Meyer
Nelleke Oostdijk

Language, People, Numbers

Corpus Linguistics and Society

Edited by

Andrea Gerbig and Oliver Mason

Rodopi

Amsterdam - New York, NY 2008

Cover design: Pier Post

Online access is included in print subscriptions:
see www.rodopi.nl

The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISBN: 978-90-420-2350-5

Printed in The Netherlands

On the occasion of Michael W. Stubbs' 60th birthday

Table of Contents

Introduction

Oliver Mason and Andrea Gerbig

The papers in this volume are all concerned with structural aspects of language in relation to its users in a variety of socio-cultural situations. All papers are based on the assumption that only the use of authentic language data can inform us about the role and function of language, its structure and use. The papers give qualitative and mostly also quantitative analyses of their data, always with respect to the specific 'work' the language does for its users.

The contributors to this volume have all been in critical exchange with Michael Stubbs' leading, corpus-based work on theories of language structure and use and its applications to education, cognition and culture, literature, and politics. Such discussion is of course always ongoing; it involves innovative theoretical thinking, looks at the most recent data which have become available through new technology, and puts in perspective and evaluates established theories and findings. The present papers are all original work providing such new insights. They are all written in honour of Michael Stubbs' outstanding contributions to this discussion which are often of a programmatic nature, paving the way for more detailed work, following up his theoretical leads.

Susan Hunston's short contribution evaluates and honours the work and influence of Michael Stubbs in major fields of linguistics, thereby providing an outlook on directions of future research. This is followed by three papers of a more theoretical orientation, starting with an article by *John Sinclair*, who had a profound influence on the field of corpus linguistics and on Michael Stubbs himself. Sinclair discusses how imports from other disciplines have shaped linguistics, and points out a number of pitfalls: sometimes imports do not make sense, or are simply misapplied, so that statistical significance tests for example provide a false sense of security where they should not have been applied in the first place. As always, Sinclair is not afraid to be controversial, casting doubt on many assumptions that are usually taken for granted by many researchers.

Continuing in the realm of theory, *Robert de Beaugrande* attempts to find an answer to how systemic a corpus of English is, investigating a number of systems in the process. He investigates the relationship between text and language, and comes to the conclusion that both of them are in fact systemic.

Wolfgang Teubert then looks at the mental lexicon. Meaning has often been ignored by corpus research, as it is much harder a problem to tackle than lexis or even grammar/phraseology. Teubert argues that there could be a fruitful dialogue between cognitive linguists and corpus linguists, and that Michael Stubbs would be the one person who could facilitate such a dialogue.

Michael Byram discusses issues of language and politics. Commonly, the national identity is based on a common (national) language, but how does it work with a supranational entity such as the European Union? Questions arise here of language policies, both regarding communication and cultural issues. And who draws up those European policies? In his contribution, Byram echoes early work by Michael Stubbs on the National Curriculum.

The collection further continues with a set of papers tracing linguistic progress in the description and investigation of diachronic language data, again in relation to the role such language had at the time it was used and what influences result for our present views on language. The first of these papers, by *Wolfgang Kühlwein*, draws connecting lines from lexicological to intertextual to semiotic research, and demonstrates how these can be applied to a language for which we have something we can only dream of for modern languages: a corpus of all known utterances of the language.

Still within the theme of historical work, *David Reibel* looks into empiricism among early grammarians, working on 'traditional grammar', which is often used by modern corpus linguists as something to distance themselves from. Reibel shows that issues such as what constitutes 'proper English' have been around for a long time.

Andrea Gerbig then shows that both sides of the Saussurean dualism 'synchronic/diachronic' can be studied with a diachronic corpus, touching on issues such as language and the representation of reality/shared cultural knowledge, and language change. Gerbig here exploits the fact that her corpus of travel writing is controlled for topic, and captures the experience of a closely defined sub-group. Gerbig's contribution also provides a link from the historical studies to the phraseological ones.

The following four papers emphasise the phraseological aspect of language, the field in which Michael Stubbs has most recently set new standards in a collection of publications. Function words are routinely being ignored by corpus linguists, on the grounds that they are too frequent or have no meaning. John Sinclair looked at *of* in *Corpus Concordance Collocation* (1991), and *Naomi Hallan* here analyses the uses of *out*, with a special focus on the use by children of different age bands and the differences there are compared to adults. Unsurprisingly, the picture is more complex as one would have expected; this again shows that there is no substitute for looking at real data.

In the following paper, *Bettina Starcke* investigates changing discourse prosodies of phrases based around the same nucleus. Starcke finds that the prosody of a phrase relates to its length (effectively its specificity) because the contexts in which the longer phrase is used are more restricted. The shorter variants are also more often used in a literal sense, whereas longer phrases tend to be non-compositional.

Hans Lindquist then compares varieties of English, British and American, and discovers that there are differences between literal and metaphorical uses of a formulaic expression. Even though originating from America, the particular phrase under investigation (*to stub one's toe*) is now about equally frequent in British English, but is predominantly used literally. Studies such as this can provide useful insights into the development and change of language.

Oliver Mason concludes the phraseology section with a new approach to the description of grammatical structure. Mason uses multi-word units (partly based on n-grams as used by other authors in this volume) to investigate the linear sequence of the sentence (in a way also pursued by Sinclair in the initial paper). The degree of success in the analysis of a particular sentence can be linked to parameters such as creativity and naturalness.

In future, (corpus) linguists will have to deal with language beyond the written and spoken word. New communication technology and multi-media computing allow us to look at other manifestations of language, and at the same time they also have an impact on the further development and evolution of language itself. These aspects of language study are discussed by Bublitz and by Carter and Adolphs.

First, *Wolfram Bublitz* discusses the impact of new communication media on the traditional view of communication as a dyadic process, characterised by the participants speaker/hearer or writer/reader. In current online chats there is no longer a simple relationship between the two roles, and the exchange structure is usually no longer comprised of adjacent pairs of utterances. However, Bublitz concludes that this is not exclusively a feature of computer-mediated discourse alone.

Ronald Carter and *Svenja Adolphs* take the notion of a corpus a step further, including not only actual speech, but also non-verbal gestures. In their paper, they describe the experimental set-up for collecting such a corpus, and the problems and issues that one needs to take into account with multiple streams of different kinds of data. Just as computers enabled the use of electronic corpora initially, now advances in video processing allow us to extend the object of study to include the visual dimension.

Henry Widdowson then covers another area in which Michael Stubbs has pushed forward the boundaries: stylistics. Widdowson argues for a distinction of two different entities when looking at Stubbs' work on Conrad's *Heart of Darkness*, namely the text, with all its words and textual patterns, and the novel, with its characters and plot. The issue at stake is then how elements of these two correlate and can be linked, thus escaping the criticism that stylistics is entirely circular in its nature.

Finally, *Guy Cook* investigates Stubbs himself: where is his place in the hocus pocus/God's truth dualism? Reflecting on a wide range of Stubbs' work and on Stubbs' criticism of his (Cook's) own work, Cook summarises the traits that make Michael Stubbs such an influential scholar.

Contributing Authors

Svenja Adolphs
Associate Professor
School of English Studies
The University of Nottingham, UK

Robert de Beaugrande
Professor of English language
www.beaugrande.com

Wolfram Bublitz
Professor
Anglistik/Amerikanistik
Universität Augsburg, Germany

Michael Byram
Professor of Education
Durham University, UK

Ronald Carter
Professor of Modern English Language
The University of Nottingham, UK

Guy Cook
Professor of Language and Education
The Open University, UK

Bettina Fischer-Starcke
Research Associate
Anglistik
Universität Trier, Germany

Andrea Gerbig
Associate Professor
Anglistik
Universität Bochum, Germany

Naomi Hallan
Researcher
Anglistik
Universität Trier, Germany

Susan Hunston
Professor of English Language
Department of English
The University of Birmingham, UK

Wolfgang Kühlwein
Emeritus Professor
Anglistik
Universität Trier, Germany

Hans Lindquist
Associate Professor
School of Humanities
Växjö Universitet, Sweden

Oliver Mason
Lecturer
Department of English
The University of Birmingham, UK

David Reibel
Emeritus Professor of English Language
Eberhard-Karls-Universität, Tübingen, Germany

John McH. Sinclair †
Emeritus Professor of Modern English Language
The University of Birmingham, UK
The Tuscan Word Center, Siena, Italy

Wolfgang Teubert
Professor of Corpus Linguistics
Department of English
The University of Birmingham, UK

Henry Widdowson
Emeritus Professor of English Linguistics
Universität Wien, Austria

Michael Stubbs – A Select Bibliography

Books

2001: Words and Phrases: Corpus Studies in Lexical Semantics. Oxford: Blackwell.
1996: Text and Corpus Analysis: Computer-assisted Studies of Language and Culture. Oxford: Blackwell.
1986: Educational Linguistics. Oxford: Blackwell.
1983: Discourse Analysis: The Sociolinguistic Analysis of Natural Language. Oxford: Blackwell.
1980: Language and Literacy: The Sociolinguistics of Reading and Writing. London: Routledge & Kegan Paul.
1976: Language, Schools and Classrooms. London: Methuen. (2nd ed. 1983).

Edited Books

1983: (coedited with H. Hillier) *Readings of Language, Schools and Classrooms*. London: Methuen.
1976: (coedited with S. Delamont) *Explorations in Classroom Observation*. London: Wiley.

Articles in Refereed Journals

2005: Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*, 14, 1: 5-24.
2003: [with I. Barth] Using recurrent phrases as text-type discriminators: a quantitative method and some findings. *Functions of Language*. 10, 1: 65-108.
2002: Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*. 7, 2: 215-44.
2001: On inference theories and code theories: corpus evidence for semantic schemas. *Text*, 21/3: 437-65.
2001: Texts, corpora and problems of interpretation. A Response to Widdowson. *Applied Linguistics*, 22, 2: 149-72.
1995: Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language*, 2, 1: 23-55.
1995: Collocations and cultural connotations of common words. *Linguistics and Education*, 7, 4: 379-90.
1994: Grammar, text and ideology: computer-assisted methods in the linguistics of representation. *Applied Linguistics*, 7, 1: 1-25.

1989: The state of English in the English state: reflections on the Cox Report. *Language and Education*, 3, 4: 235-50.

1986: A matter of prolonged fieldwork: notes towards a modal grammar of English. *Applied Linguistics*, 7, 1: 1-25.

1984: (with G. Keck) Koschmieder on speech act theory: a historical note. *Journal of Pragmatics*, 8, 3: 305-10.

1983: Can I have that in writing, please? Some neglected topics in speech act theory. *Journal of Pragmatics*, 7: 479-94.

1982: The sociolinguistics of the English writing system: or why children aren't adults. *Australian Journal of Reading*, 5, 1: 30-36.

1981: Oracy and educational linguistics: the quality (of the theory) of listening. *First Language*, 2:21-30.

Other Main Articles

In press 2007: On texts, corpora and models of language. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert. *Text, Discourse and Corpora*. London: Continuum.

In press 2007: Quantitative data on multi-word sequences in English: the case of the word 'world'. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert. *Text, Discourse and Corpora*. London: Continuum.

2006: Corpus analysis: the state of the art and three types of unanswered questions. In G. Thompson & S. Hunston eds *System and Corpus*. London: Equinox, 15-36.

2004: Language corpora. In A. Davies & C. Elder eds *Handbook of Applied Linguistics*. Oxford: Blackwell. 106-32.

2001: Computer-assisted text and corpus analysis: lexical cohesion and communicative competence. In D. Schiffrin et al. eds *Handbook of Discourse Analysis*. Oxford: Blackwell. 304-20.

1997: Language and the mediation of experience: linguistic representation and cognitive orientation. In F. Coulmas ed. *Handbook of Sociolinguistics*. Oxford: Blackwell. 358-73.

1997: Whorf's children: critical comments on critical discourse analysis. In A. Wray & A. Ryan eds *Evolving Models of Language*. Clevedon: Multilingual Matters. 100-16.

1991: Educational language planning in England and Wales: multi-cultural rhetoric and assimilationist assumptions. In F. Coulmas ed. *A Language Policy for the European Community*. Berlin & NY: de Gruyter. 215-39.

1990: Language in education. In N. E. Collinge ed. *An Encyclopedia of Language*. London & NY: Routledge. 551-89.

1989: (co-author) *English for Ages 5 to 16*. [Cox Report]. Dept of Education and Science and Welsh Office.

1984: Discourse analysis and educational linguistics. In P. Trudgill ed. *Applied Sociolinguistics*. London: Academic Press. 203-43.

Other Articles in Conference Proceeding, Festschriften, Edited Books, Etc.

(in prep): Technology and phraseology: with notes on the history of corpus linguistics. In U. Römer & R. Schulze eds *The Lexis-Grammar Interface*. Amsterdam: Benjamins

2007: On very frequent phraseology in English: structures, distributions and functions. In R. Facchinetti ed. *Corpus Linguistics Twenty-five Years On*. Amsterdam: Ropdopi.

2007: Inferring meaning: text, technology and questions of induction. In A. Mehler & R. Köhler eds *Aspects of Automatic Text Analysis*. Berlin: Springer. [Festschrift for Burghard Rieger] 233-53.

2006: On teaching critical rationalism: reconciling linguistic and literary text analysis. In A. Gerbig & A. Müller-Wood eds *How Globalization Affects the Teaching of English*. Lampeter: Mellen. 15-29.

2006: Exploring *Eveline* with computational methods. In S. Goodman & K. A. O'Halloran eds *The Art of English: Literary Creativity*. Basingstoke: Palgrave Macmillan / Open University. 138-44.

2005: Using language corpora to study pragmatic meaning. In A. Schuth, K. Horner & J. J. Weber eds *Life in Language*. Trier: Wissenschaftlicher Verlag. [Festschrift for Wolfgang Kühlwein] 3-15.

2004: A quantitative approach to collocations. In D. Allerton et al. eds *Phraseological Units: Basic Concepts and their Application*. Basel: Schwabe. 107-19.

2004: Conrad, concordance, collocation: heart of darkness or light at the end of the tunnel? The Third Sinclair Open Lecture. University of Birmingham.

2003: [with A. Pusch] Frequent terms in linguistics: a pilot corpus study for a pedagogical word-list. In C. Tschichold ed. *English Core Linguistics*. Bern: Peter Lang. [Festschrift for David Allerton] 247-67.

2002: Human and inhuman geography: a comparative analysis of two long texts and a corpus. In M. Toolan ed. *Critical Discourse Analysis*. Routledge. 98-130. [Reprint: original publication 1996 in M. Stubbs Text and Corpus Analysis Oxford: Blackwell. Also reprinted in: C. Coffin et al. eds *Applying English Grammar*. London: Arnold.247-74.]

2000: Using very large text collections to study semantic schemas: a research note. In C. Heffer and H. Saunston eds *Words in Context*. Birmingham University. [CD-ROM.]

2000: Text and corpus linguistics. In M. Byram ed. *Encyclopedia of Language Teaching and Learning*. London: Routledge.

2000: Society, education and language: the last 2000 (and the next 20?) years of language teaching. In H. Trappes-Lomax ed. *Continuity and Change in Applied Linguistics*. Clevedon: Multilingual Matters.

1998: German loanwords and cultural stereotypes. *English Today*, 53 (14, 1, Jan 1998): 19-26.

1998: A note on the phraseological tendencies in the core vocabulary of English. *Studia Anglica Posnaniensia*, 23: 399-410.

1998: Judging the facts: an example of applied discourse analysis. In J. Cheshire & P. Trudgill eds *The Sociolinguistics Reader*. Vol 2. London: Arnold. 1998:348-66.
[Reprint: original publication 1991, in C. Uhlig & R. Zimmermann, eds Anglistentag 1990 Marburg: Proceedings. Tübingen: Niemeyer. 312-331.]

1997: Eine Sprache idiomatisch sprechen: Computer, Korpora, Kommunikative Kompetenz und Kultur. K. J. Mattheier ed. *Norm und Variation*. Frankfurt: Peter Lang. 151-67.

1996: The English writing system. In H. Günther & O. Ludwig, eds *Schrift und Schriftlichkeit / Writing and Its Use*. Vol 2. Berlin & NY: de Gruyter. 1441-45.

1995: Corpus evidence for norms of lexical collocation. In G. Cook & B. Seidlhofer, eds *Principle and Practice in Applied Linguistics*. Oxford: Oxford University Press. 245-56. [Festschrift for H. G. Widdowson].

1995: Educational language planning in England and Wales: multi-cultural rhetoric and assimilationist assumptions. In O. García & C. Baker, eds *Policy and Practice in Bilingual Education*. Clevedon: Multilingual Matters.25-39. [Reprint: original publication 1991, in F. Coulmas, ed. *A Language Policy for the European Community*. Berlin & NY: de Gruyter. 215-39.]

1993: (with A. Gerbig) Human and inhuman geography: on the computer-assisted analysis of long texts. In M. Hoey ed. *Data, Description, Discourse*. London: HarperCollins. 64-85. [Festschrift for John Sinclair].

1993: British traditions in text analysis: from Firth to Sinclair. In M. Baker et al. eds *Text and Technology*. Amsterdam: Benjamins. 1-33. [Festschrift for John Sinclair].

1992: English teaching, information technology and critical language awareness. In N. Fairclough ed. *Critical Language Awareness*. London: Longman. 203-22.

1992: Institutional linguistics: language and institutions, linguistics and sociology. In M. Pütz ed. *Thirty Years of Linguistic Evolution*. Amsterdam: Benjamins. 189-211. [Festschrift for René Dirven].

1992: Spelling in society: forms and variants, uses and users. In R. Tracy ed *Who Climbs the Grammar Tree*. Tübingen: Niemeyer. 221-34. [Festschrift for David Reibel].

1991: Knowledge about Language: Grammar, Ignorance and Society. University of London: Institute of Education / Kogan Page. [Professorial Lecture].

1991: Judging the facts: an example of applied discourse analysis. In C. Uhlig & R. Zimmermann eds *Anglistentag 1990 Marburg*. Tübingen: Niemeyer. 312-331.
[Reprinted in J. Cheshire & P. Trudgill eds The Sociolinguistics Reader. Vol 2. London: Arnold. 1998: 348-66].

1989: On ivory towers and the market place: everyday and specialist knowledge in applied linguistics. In C. S. Butler et al. eds *Language and Literature: Theory and Practice.* University of Nottingham. 27-38. [Festschrift for Walter Grauberg].

1987: An educational theory of (written) language. In J. Norrish & T. Bloor eds Written Language. London: CILT. 3.38.

1986: Language development, lexical competence and nuclear vocabulary. In K. Durkin ed. *Language Development in the School Years*. London: Croom Helm. 57-76.

1986: Lexical density: a computational technique and some findings. In M. Coulthard ed. *Talking about Text*. University of Birmingham: English Language Research. 27-42. [Festschrift for David Brazil].

1983: Understanding linguistic diversity: what teachers should know about educational linguistics. In M. Stubbs & H. Hillier eds 1983: 11-35.

1982: Stir until the plot thickens. In R. Carter & D. Burton eds *Literary Text and Language Study*. London: Edward Arnold. 56-85.

1982: Written language and society: some particular cases and general observations. In M. Nystrand ed. *What Writers Know*. New York: Academic Press. 31-55.

1981: Scratching the surface: linguistic data in educational research. In C. Adelman ed. *Uttering, Muttering*. London: Grant McIntyre. 114-33.

1981: Motivating analyses of exchange structure. In M. Coulthard & M. Montgomery eds *Studies in Discourse Analysis*. London: Routledge & Kegan Paul. 107-19.

1981: What's the relationship between sociolinguistics and language teaching, please? In H. Eichheim & A. Maley eds *Fremdsprachenunterricht im Spannungsfeld zwischen Gesellschaft, Schule und Wissenschaften*. Munich: Goethe Institut. 27-45.

1980: What is English? Modern English language in the curriculum. *English in Australia*, 51:3-20.

1980: (with M. Berry) The Duke of Wellington's gambit: notes on the English verbal group. *Nottingham Linguistic Circular*, 9,2. 143-62.

1980: The sociolinguistics of literacy. In T. Bessell-Browne et al. eds *Reading into the Eighties.* Adelaide: Australian Reading Association. 99-113.

1979: (with B. Robinson) Analysing classroom language. In M. Stubbs, B. Robinson & S. Twite *Observing Classroom Language*, Block 5, PE232. Milton Keynes: Open University Press. 5-59.

1976: Keeping in touch: some functions of teacher-talk. In M. Stubbs & S. Delamont eds 1976: 151-72.

1975: Teaching and talking: a sociolinguistic approach to classroom interaction. In G. Chanan & S. Delamont eds *Frontiers of Classroom Research.* Slough: NFER. 233-47.

Michael Stubbs: a theoretician of applied linguistics

Susan Hunston

University of Birmingham

My brief for this short paper is to consider Michael Stubbs' influence on the field of linguistics. This is not something to be undertaken lightly, partly because Michael's erudition vastly exceeds my own, and partly because consideration of the role he plays raises questions concerning the relationship of theoretical and applied linguistics, between theory and practice. Michael Stubbs' work has always been located within Applied Linguistics, in the sense that many of his concerns, especially his abiding interest in education (e.g. Stubbs 1976; 1980; 1986b; 1995b), would normally come under the heading of 'applied' research. His contributions to Critical Discourse Analysis and to literary stylistics come into this category too. However, his work forces us to recognise that 'applied' in this sense in no way implies 'theory-impoverished' or 'intellectually inconsequential'. What it does imply is the observation and analysis of naturally-occurring language in its social context, and a demand for a linguistic theory that takes such language as its starting point.

A case in point is one of Stubbs' papers from the 1980s, ' "A matter of prolonged field work": notes towards a modal grammar of English' (Stubbs 1986a). This takes as its data naturally-occurring language from a large number of sources and focuses on those aspects of spoken or written language that express speaker/writer attitude, in particular those that express commitment to or detachment from the truth of a proposition. The paper brings together a number of linguistic topics: attribution, speech act theory, vague language, verb aspect, verb process types, and connectors among them. Stubbs points out that although some acts performed through words (such as 'excommunication') are non-negotiable, illocutionary acts proper are open to hedging, emphasis and verbal prevarications of many kinds. Speakers exploit the resources of language to give their assertions the weight of authority or to remove from themselves any responsibility for the truth-value of propositions. The unifying theme in the paper is that speaker attitude is central to language description, and that consideration of attitudinal and interactional factors can challenge assumptions or solve problems in fields such as speech act theory and syntax. The focus on commitment / detachment and on the averred source of propositions finds echoes in Sinclair (1988), Cooper (1981), and Tadros (1993), and in much subsequent work on written discourse in particular. It is central to considerations of how knowledge is constructed and transmitted (Hunston 1993). The centrality of the attitudinal and the interactional to language theory and description is a theme taken up by numerous writers including Martin and White (2005) and the papers in Hunston and Thompson (2000). What remains typical of Stubbs' approach is the insistence that intuition-dependent theories and naturally-occurring data be placed alongside each other, the latter both informing and challenging the former.

Like John Sinclair, Stubbs was led by his interest in the patterns of lexis and grammar in naturally-occurring discourse to exploit the growing power of computers to analyse large quantities of text. Much of Stubbs' work since 1995 shares some assumptions with Sinclair's approach (e.g. Sinclair 1991, 2004), and he is personally associated with at least four major contributions to the field. The first two follow the practice of placing large amounts of data and particular analytical techniques at the service of traditions that typically draw on smaller amounts and different methods (Critical Discourse Analysis and literary stylistics). The second two offer critical and visionary accounts of some corpus linguistic practices themselves.

Text and Corpus Analysis (Stubbs 1996) was among the first publications to unite the insights of corpus investigation techniques with those of more traditional discourse analysis and with the use of text in investigating cultural practices. In the book, Stubbs presents examples in three sets of contextual parameters. His study of *happy* and *happiness* in two Baden-Powell texts uses concordance lines taken from only those texts. On the other hand, the study undertaken in collaboration with Andrea Gerbig of ergative verbs in two sets of textbooks uses a much larger dataset and a greater degree of statistical processing (see also Stubbs and Gerbig 1993). Finally, the study of cultural keywords and their most frequent collocates uses frequency information from a number of very large reference corpora (see also Stubbs 2001). Mautner (2007: 8) describes such work as uniting qualitative and quantitative methodologies. Put simply, Stubbs notes that frequency is important because it reflects the prevalence in a given cultural context of a particular formulation. For example, he observes (1996: 184) that frequently-occurring formulations such as *child care, care in the community* and *care and resettlement of offenders* reflect (and perpetuate) a cultural context in which diverse groups of individuals are construed as constituencies requiring a common institutional response. Another theme is the conveying of covert evaluative meaning through intertextuality: a given instance has a 'hidden meaning' because of the way a word or phrase in it is used in a number of other texts. Stubbs offers many examples of this, one being the collocations associated with *the streets* that connote danger and menace (Stubbs 2001: 203-206; cf Channell 2000). By using frequency to interrogate intertextuality, Stubbs demonstrates that a concern for numbers does not condemn the researcher to a level of abstraction that precludes interpretation or sensitivity to context.

Many researchers have used methods similar to Stubbs' to further the agenda of Critical Discourse Analysis, to some extent in response to Stubbs' own critique of CDA methods (Stubbs 1997). Mautner (2007), for example, uses the 'collocational profile' of the word *elderly* to argue that the term can be regarded as ageist. Baker (2006) offers a number of studies using different techniques, including a concordance-based study of *refugees* in British newspapers which shows the predominant discourse contexts of the word. Coffin and O'Halloran (2006) take as their starting point a single text – an article from the Sun newspaper about migrants to Britain from Eastern Europe – and explore it alongside a corpus of editions of the same newspaper. They argue that a number of phrases that are evaluatively ambiguous in the context of the single text alone (such as *poverty-stricken former Soviet states* or *desperate for…any job at all*) can be shown to resonate with unpleasant connotations when their phraseologies are examined in the larger corpus. Coffin and O'Halloran draw on the concepts of logogenesis, ontogenesis and phylogenesis to model the relations between a single text and the reader's longer-term experience of texts from the same journalistic source. Partington (2004: 19) gives a name to the union of discourse and corpus techniques: Corpus-Assisted Discourse Studies.

A second area of applied work is literary stylistics. Stubbs' interest in the interaction of language and literature (evident in his sociolinguistic analyses of literary texts in Stubbs 1983) has culminated in his studies of Conrad's *Heart of Darkness* (2004; 2005). In Stubbs 2005 he offers a number of studies, using word frequency, collocation, keyword analysis and word distribution analysis, to corroborate and extend observations about Conrad's novel that have been made by literary critics. For example, critics have observed that much of the novel is vague and imprecise: descriptions are hazy, actions are indeterminate. It is not possible even to know what Kurtz means by his (infamous) cry *The horror, the horror*. Whereas critics have drawn attention to the motif of indistinctness, and to the noticeable frequency of lexical words such as *murky, blurred, darkness*, Stubbs goes further and establishes that grammatical words indicating vague reference (*something, someone, somewhere* and so on) are significantly more frequent in *Heart of Darkness* than in a general Fiction corpus or in written English in general. Critics have also commented, not always positively, on the repetitious nature of Conrad's prose. Stubbs establishes wherein that repetition lies, not just in individual words but in patterns such as 'the noun of (a) neg-prefix adjective noun' (*the darkness of an impenetrable night, the sea of inexorable time, the stillness of an implacable force* and so on). He also comments on spatial-reference phrases such as *to the end of the* and *in the middle of the*, which appear to be frequent in *Heart of Darkness*, but which Conrad in fact uses with the same frequency as is general in English, though with less specific referents. With this work, Stubbs adds to a growing body of research using corpus techniques to study literary works (e.g. Semino and Short 2002; Culpepper et. al. forthcoming; Toolan, forthcoming). What is notable about Stubbs' work is that he takes as his starting point the critical literature on his chosen writer, which he then supplements and sometimes challenges.

To some extent, then, Stubbs acts as an ambassador for corpus linguistics in the wider Applied Linguistics community. What is noticeable, however, is his insistence on the need both to question and to develop methodological assumptions, his refusal to take easy routes of interpretation. These concerns are most apparent in his warnings on the interpretation of statistics (Stubbs 1995a) and in his work on both systematising and diversifying the study of phraseology (Stubbs 2001; 2002; 2006). We find echoes of this caution in Coffin and O'Halloran's (2006) and Mautner's (2007) caveats about the limitations of corpus techniques and the need for triangulation in methodology. And Stubbs' concern to extend methods of examining recurrent phraseology and collocation / colligation is reflected in Mason (this volume); Fletcher (2006); Rayson (2006); Groom (2007) among others.

For me, Michael Stubbs' most profound legacy is probably his theory-oriented writing that integrates corpus investigation techniques, and indeed discourse analytical methods, with the 'bigger questions' in linguistics. Although like many corpus linguists he illustrates his arguments with specific examples – the use in English of the word *proper*, for instance (Stubbs 2001: 156-159), or Conrad's use of *something* (Stubbs 2005), or the way that judges in Britain use phrases such as *you may think* (Stubbs 1996: 113-117) – these carefully-observed phenomena are never ends in themselves, but a step towards a wider vision. In a number of wide-ranging papers (e.g. 1993; 2000; 2006) Stubbs contextualises corpus studies within a number of intellectual traditions, making him one of the leading theorists of corpus linguistics today, as well as one of its most respected practitioners. I shall take his 2006 paper 'Corpus analysis: the state of the art and three types of unanswered questions' as a case in point. The paper, as is typical, recounts a number of sample analyses: the distribution of number of collocates across words, the pragmatic import of the phrase *ripe old age* (which might be contrasted with Mautner's *elderly*!), the relation of collocates and schemata, with *money* and *value* as exemplars. More profoundly, though, he relates such findings to questions about language that have been asked by researchers from very different traditions, theorists whose work is often by-passed as irrelevant by other corpus linguists. He raises Chomsky's distinction between description and explanation (cf Meyer 2002: 2-4), not to dismiss it but to argue for its applicability to observations such as collocation. Returning to an earlier concern (cf Stubbs 1986a) about the discipline of pragmatics and naturally-occurring language, he argues that a corpus-inspired view of the consistency between form and function 'rescues pragmatics from the notion that it is condemned to deal with idiosyncratic, one-off, context-bound interpretations' (Stubbs 2006: 27). Finally, he notes that work such as his own on 'cultural keywords' (e.g. *care* above) can be used to complement the work of philosophers on the construction of social reality (ibid: 30-32). Taking Searle's example of *money* as a socially constructed entity, he demonstrates that a corpus can be interrogated to discover how people talk about money, and therefore how such entities come to be construction and transmitted. In other words, he offers corpus investigation techniques, and the theories about language that have arisen from them, as a way of answering questions about language from outside the corpus field, but he also places a demand upon corpus work to meet the challenges of those questions and not to dismiss them as irrelevant.

Michael Stubbs is a modest and accessible writer, scrupulous always in recognising influences. His own influence on other researchers is huge. To take (somewhat flippant) quantitative data, a survey of the sixteen extant volumes in a series of books on corpus linguistics reveals that he is referenced extensively in no fewer than twelve of them. Interpreting that data more qualitatively, we can see that many of the key ideas, the most apt examples, the most profound questions, are to be found in his writing.

References

Baker P. 2006. *Using Corpora in Discourse Analysis*. London: Continuum.

Channell J. 2000. 'Corpus-based analysis of evaluative lexis'. In Hunston and Thompson (eds.) 39-55.

Coffin C. and K. O'Halloran 2006. 'The role of appraisal and corpora in detecting covert evaluation' *Functions of Language* 13: 77-110.

Cooper M. 1981. 'Aspects of the structure of written academic discourse and implications for the design of reading programmes'. In Hoedt et al. (eds.) *Pragmatics and LSP: proceedings of the 3rd European symposium on LSP, Copenhagen, August 1981*. Copenhagen: Copenhagen School of Economics. 403-433.

Culpepper J., Hoover, D., Louw, B. & Wynne M. forthcoming. *Approaches to Corpus Stylistics*. Routledge.

Fletcher W. 2006. 'Real-Time Identification of MWE Candidates in Data from the BNC and the Web'. BAAL Corpus SIG meeting, Oxford, April 2006.

Groom N. 2007 *Phraseology and epistemology in humanities writing*. Unpublished PhD thesis, University of Birmingham.

Hoey M. (ed.) 1993. *Data, Description, Discourse: paper on the English Language in honour of John McH Sinclair*. London: HarperCollins.

Hunston S. 1993. 'Evaluation and ideology in scientific discourse'. In M. Ghadessy (ed.) *Register Analysis: theory and practice*. London: Pinter. 57-73.

Hunston S. and G. Thompson (eds.) 2000. *Evaluation in Text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press.

Martin J. and P. White 2005. *The Language of Evaluation: appraisal in English*. London: Palgrave.

Mautner G. 2007. 'Mining large corpora for social information: the case of *elderly*' *Language in Society* 36.

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

Partington A. 2004. 'Corpora and discourse, a most congruous beast'. In A. Partington, J. Morley & L. Haarman (eds.) *Corpora and Discourse*. Bern: Peter Lang. 11-20.

Rayson P. 2006. 'Right from the word go: identifying multi-word-expressions for semantic tagging'. BAAL Corpus SIG meeting, Oxford, April 2006.

Semino E. and M. Short 2002. *Corpus Stylistics: speech, writing and thought presentation in a corpus of English writing*. London: Routledge.

Sinclair J. 1988. 'Mirror for a text' *Journal of English and Foreign Languages* 1. 15-44.

Sinclair J. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair J. 1994. *Trust the Text: language, corpus and discourse*. London: Routledge.

Stubbs M. 1976. *Language, Schools and Classrooms*. London: Methuen.

Stubbs M. 1980. *Language and Literacy*. London: Routledge.

Stubbs M. 1983. *Discourse Analysis*. Oxford: Blackwell.

Stubbs M. 1986a. A matter of prolonged field work: notes towards a modal grammar of English. *Applied Linguistics* 7: 1-25.

Stubbs M. 1986b. *Educational Linguistics*. Oxford: Blackwell.

Stubbs M.1993. 'British traditions in text analysis: from Firth to Sinclair'. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.) *Text and Technology: in honour of John Sinclair*. Amsterdam: Benjamins. 1-36.

Stubbs M. 1995a. 'Collocations and semantic profiles: on the cause of the trouble with quantitative studies' *Functions of Language* 2: 23-55.

Stubbs M. 1995b. 'Educational language planning in England and Wales: multicultural rhetoric and assimilationist assumptions'. In O. García & C. Baker (eds.) *Policy and Practice in Bilingual Education*. Clevedon: Multilingual Matters. 25-39.

Stubbs M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs M. 1997. 'Whorf's children: critical comments on critical discourse analysis'. In A. Wray & A. Ryan (eds.) *Evolving Models of Language*. Clevedon: Multilingual Matters. 100-16.

Stubbs M. 2000. 'Society, education and language: the last 2000 (and the next 20?) years of language teaching'. In H. Trappes-Lomax (ed.) *Continuity and Change in Applied Linguistics.* Clevedon: Multilingual Matters. 15-34.

Stubbs M. 2001. *Words and Phrases: Corpus Studies in Lexical Semantics*. Oxford: Blackwell.

Stubbs M. 2002. 'Two quantitative methods of studying phraseology in English' *International Journal of Corpus Linguistics* 7: 215-44.

Stubbs M. 2004. 'Conrad, concordance, collocation: heart of darkness or light at the end of the tunnel?' *The Third Sinclair Open Lecture*. University of Birmingham.

Stubbs M. 2005. 'Conrad in the computer: examples of quantitative stylistic methods' *Language and Literature* 14: 5-24 .

Stubbs M. 2006. 'Corpus analysis: the state of the art and three types of unanswered questions'. In G. Thompson & S. Hunston (eds.) *System and Corpus: exploring connections*. London: Equinox. 15-36.

M. Stubbs & A. Gerbig 1993. 'Human and inhuman geography: on the computer-assisted analysis of long texts'. In M. Hoey (ed.) 64-85.

Tadros A. 1993. 'The pragmatics of text averral and attribution in academic texts'. In Hoey (ed.) 115-138.

Toolan M. forthcoming. *Narrative Progression in the Short Story*. London: Palgrave.

Borrowed ideas

John Sinclair †

Tuscan Word Center, Siena

Abstract

Linguistics fits uneasily in the panoply of academic disciplines, with many links to the soft sciences, more reliance on measurement as the examinable body of language grows in the electronic domain, and with its roots still firmly in the humanities; the unique property of intuitive access to meaning keeps it apart. It is thus in a position to adopt conventions, concepts, terms and working practices from a wide range of sources, and while this is a fundamental strength, there are dangers. This paper suggests that some influences have been uncritically imported, obscuring fundamental properties of language that are not shared by other data types.

Of the four instances discussed the first and third deal with sampling language, and the other two with the textual property of linearity. The first is a cautionary tale about how a general working practice in many data types became almost inviolate in early corpus linguistics although quite irrelevant; the second emphasises that terminology should be chosen with care; the importation of hierarchical terms based on the prefix meta- is in danger of obscuring the key property of the linearity of text. The third point returns to sampling issues and wonders why chance plays such a large role in our statistical analyses of large corpora when we know that the words of a text are chosen and arranged to create meanings. The final mini-study argues that the adoption of conventions of logical notation in formal grammars suppresses an essential sub-property of linearity – the directionality of text – and unnecessarily cuts down the structural options that should be available in description.

1. Introduction

Academic disciplines have a tendency to be insular. They shape their arguments, terminology and experimental methods in order to make the best descriptions of their data, and thus their categories and methods often do not lend themselves to applications to other data, and they can act as deterrents to interdisciplinary work.

However, younger disciplines can often profit by importing models and techniques from established areas of investigation, either from specific subject areas or from general experience. For instance "scientific method" is an imprecise but powerful attitude to the handling of data, experiment and argument that encompasses such valuable notions as objectivity, replicability of procedure and the role of falsification attempts; these notions are not identified with any single subject but are widely adopted in the experimental sciences and beyond.

Some disciplines are so important to others that they are in part shaped by their applications; statistics is one of these; a number of disciplines are in essence applications of statistics to particular bodies of data, and many more disciplines rely heavily on statistical tools. Logic is so basic that every discipline is supposed to take it for granted and apply it without comment; in addition the descriptive techniques of formal and mathematical logic are widely used in an explicit form, with formalisms and conventions imported in bulk; mathematics has similar status. These are not conventions to be trifled with.

Computer science, as a very young discipline, has borrowed most of its procedures and practices and a lot of its arguments from these bedrock disciplines, but with its own status as a central service discipline it is also a remarkably productive source of models, notions and procedures. Binary mathematics existed before computers but has been developed to a point where some of its concepts are becoming part of the general vocabulary of academics and even extend to the public arena. "Eighty gigs" means "lots of storage" in domestic computing.

Linguistics is also in many ways a young discipline, though the study of language has been with us for a long time. In the last half-century it has been boosted by pressure from applications such as language learning, in turn deriving from aspects of globalisation, and it has matured with the help of modern technology. The study of the spoken language was speculative and anecdotal until the invention of cheap recording devices, and the meaningful patterns of both spoken and written text are only just being uncovered through the computational analysis of large corpora.

Linguistics is not a "pure" science because its touchstone is meaning, and meaning is partly determined by the perception of individuals, and accessed via their reports. The intuition, as it is somewhat misleadingly called, is a decision-making mental facility which is non-negotiable, differs from one person to another, and offers no reasons for its decisions; any reasons advanced by an informant are bogus. Intuition has a delphic status in appearing to be quite arbitrary, mysterious and impenetrable, leaving the scholar to sort out how to interpret its "proclamations".

While the intuition maintains an element of subjectivity which is right in the centre of any linguistic argumentation, linguistics as a subject does not sit easily among the humanities because of its heavy reliance on experimental methods and, nowadays, computing. Nor is it more than peripheral among the social sciences. It sits uneasily at a disciplinary crossroads. Contributing further to the unease is the poor standard of applications that has been achieved. In the present state of the world, good-quality applications of linguistics are much sought after and would be highly prized. If a reliable means of deriving meaning from texts, comprehensively and automatically, could be achieved it would constitute the launching pad of a major improvement in the efficiency of social institutions, media services, international understanding and security. Despite very large investment over decades, this goal seems only to recede.

The present position is fairly desperate, and linguists are losing credibility because practitioners of other disciplines seem to achieve better results in the solving of language-oriented tasks, speech recognition being the most notorious of these in recent years. Gradually scholars are beginning to face a most unpalatable prospect – that their models and theories are faulty and are leading them astray and contributing to the routine failure of applications. To remedy this will be a lengthy process, however, because of the large investments and the threats to the careers of thousands of people. Also constructive criticism at this level of profundity requires not just attacks on existing beliefs, but the emergence of more reliable alternatives to put in their place.

This paper is far too limited in scope to tackle such a major problem, and devotes itself to a minor piece of ground-clearing. It makes a claim that some notions and procedures have been imported into linguistics rather uncritically from other disciplines or the wider academic scene, and are perhaps contributing to the inadequacy of models and arguments. These only compound the problem rather than causing it, but the introduction of even a small amount of clarity in a few areas could highlight the larger need for development and accelerate its onset.

We will consider four areas where concepts and/or routines have been imported into linguistics; three we will deal with rather briskly as aperitifs, and the last in more detail. The first and third deal with sampling language, and the other two with the textual property of linearity. One is the question of data sampling conventions for language text, the second raises the matter of the linearity of text and how it can be obscured by unfortunate terminology; the third concerns the relation between linguistic patterns and chance, and the fourth returns to linearity and directionality and considers the use of logical notation to represent structure.

2. Disclaimer

Some of the concepts that will be discussed below are, in their natural habitat, complex and sophisticated, and are the product of advanced research in their parent disciplines. Several would take more than one paper even to outline satisfactorily; to assess them critically would require skills well beyond those of the present author. I would like to make it quite clear that I do not intend to engage with such matters, for which I am ill-equipped, but only to take up the way in which the concepts and routines associated with them are applied within the discipline of linguistics. A philosopher of science, a statistician, a logician will probably wince at the superficiality of the arguments put forward, and my only defence is that the way in which the imported ideas are dealt with here comes from my own experience in linguistics.

Nor should it be assumed that I am trying to protect linguistics from outside influences, to be promoting the view that external imports are undesirable. Far from it – one of the shaping influences on my attitudes to language study has been the realisation that the really mould-breaking ideas have come from outside the subject, and not from developing notions and observations derived from inside. As an example of this, perhaps the most innovative and far-reaching development in linguistic perception in the last fifty years was the philosopher Austin's idea of illocutionary force (Austin 1962). By arguing that sentences did other things than just *mean*, Austin opened up the prospect of structures above the sentence, including interactive constructions. Before Austin, linguists had been embarrassingly short of relevant comments to make on the nature, structure, direction and results of language interaction.

Austin's work was subjected to much criticism from linguists just because it enabled the study of discourse to get going, and thus isolated the many grammarians who still see the sentence as the boundary of the organisation that they seek to describe. The kind of influence that I will draw attention to in this paper, however, is not of this "breakthrough" variety such as Austin, but it is much lower key; it is the kind of model, concept or practice that is adopted with little or no criticism, but just imported as an apparently self-evident concept or a procedure; one that is well established in several disciplines, so why not linguistics as well?

3. Sample size

The issue of the number of words that constitutes a proper sample of text is one that, happily, is now a matter of history. I begin with it because, since it is no longer a burning issue, we can attain a certain objectivity in retrospect, which may help us when we engage with matters which are currently accepted uncritically.

Some forty-five years ago the compilation of corpora began in earnest. There were pioneers even before 1960, in particular Father Busa, and there were by 1970 several models of corpus architecture available. Father Busa was engaged in the huge task of indexing the whole of the works of St Thomas Aquinas (http://www.corpusthomisticum.org/it/index.age), Bernard Quemada had compiled the Trésor de la language Français (Imbs and Quemada, 1988), a team in UK had prepared the first corpus of transcribed spoken language (Krishnamurthy, ed. 2004), and a team at Brown University in USA published and made available a million-word corpus of selected English from American publications of the year 1961 (http://khnt.hit.uib.no/icame/manuals/index.htm). Of these, only the Brown was readily accessible, offered with characteristic American generosity, and it rapidly established itself as the archetypal corpus (Léon 2005). For more than twenty years it was the reference point for anyone wanting to know what a corpus was, and its architecture was still being replicated in corpora in the 1990s, compiled specifically to be compared with Brown and its UK clone, LOB (for all the Brown clones see http://khnt.hit.uib.no/icame/-manuals/index.htm).

One million words, five hundred samples, each of around two thousand words, in a range of sixteen roughly-described genres classified as either informative or imaginative prose; the samples were of a uniform size but the genres differed greatly in the number of samples in each; so for example Learned

Prose had eighty samples, around 160,000 words, while Science Fiction had only six samples, totalling around 12,000 words. Although the many clones varied these proportions quite a lot, the steady element was the 500 samples of 2000 words in each.

All of these figures reflected what was possible in 1963 or so, achieved despite the puny capacity and power of computers, the ghastly problems of data entry, the ponderous and unsuitable programming languages and the difficulty of handling the operating systems. Yet strangely, with the technology advancing at breakneck speed, even ten or fifteen years later the dimensions and character of Brown seemed to restrict people's vision of what could be achieved. When the corpus that became The Bank of English (http://www.collins.co.uk/books.aspx?-group=153) was designed in 1980, aiming initially at a modest five million words and rising to twenty million, it was difficult to persuade potential backers that such dimensions were not ridiculously extravagant. And in the summer of 2006, a billion-word collection was announced by Oxford University Press, and data from a trillion words from the Internet is available from the Linguistic Data Consortium (http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LD-C2006T13).

In this context the 2000-word sample became an untenable limitation, and had to rise, and rise sharply. Unfortunately one of its least valuable features was retained and defended stoutly for some years. This was the idea that all text samples should be of the same size, to facilitate comparisons. It was claimed that a uniform size of sample was an essential feature of sampling technique in all serious science, and a fixture of "scientific method".

This was patently ridiculous. Back in the days of Brown, the sample size had to be small in order to include sufficient variety in a general corpus. Since few published documents are only 2000 words in length, the corpus was made up of fragments. This was felt to be the lesser evil, because if whole documents had been selected there would not be nearly enough variety in a million words – ten smallish books would fill the corpus to the brim. The penalty, however, was serious. One fairly obvious feature of a text is that it is not the same all the way through. In barest outline it has a beginning, middle and end, but it is likely to have a much more elaborate structure than that, and each aspect of its internal structure leads to different phraseology, different vocabulary and different structures.

When the dimensions of corpora increased and it became possible to handle many millions of words, one of the first restrictions to be dropped was the fixed sample size. Corpora like The Bank of English consist of whole documents and transcribed speech encounters, of widely varying sizes. It was possible to get the permission of rights holders for this corpus because there was never any intention to make and distribute copies of the Bank of English, only to provide scholars with access to it via the emerging Internet. Other large corpora have suffered from copyright restrictions, preventing them from including whole documents; so the British National Corpus, for example (http://www.natcorp.ox.-ac.uk/corpus/permletters.html) restricts itself to 40,000 words or ninety per cent of any text, whichever is the smaller figure. There the question arises, which ninety per cent gets in?

It may be difficult nowadays to look back and consider a time when a corpus consisting of texts of varying dimensions was called unscientific and of little use to serious researchers, who apparently were thought to lack the techniques of comparison of datasets of differing sizes. The great variation in size of the genres in the Brown corpus, which perhaps should have been a cause for concern, attracted little comment.

4. Poor taste in terminology

The second concept that has been uncritically imported is the idea of "meta…". I can deal with it briefly here because of recently having published a short paper devoted to it (Sinclair 2005a), to which this is a footnote. See also Ädel (2006: 213-219).

"Meta…" is a prefix which comes from philosophy, and is very old, as shown, for example, by the word *metaphysics*. Originally meaning something beyond or outside the word it prefixes, it nowadays is more specific. To quote Wikipedia, it is "used to indicate a concept which is an abstraction from another concept, used to analyze the latter." "*Meta language*", it instances, "refers to a type of language or system which describes language."

I do not wish to question the use of "metalanguage" in this sense, though there is a risk of misunderstanding that can arise from this usage. The professional terminology may quite reasonably be called metalanguage, but the sentences that are composed using that terminology are sentences in the ordinary language. Here, the prefix *meta-* may be considered appropriate, but *language*, with its implication of being used in communication, is misleading. Only the terminology can be considered meta, not their deployment in sentences.

That is to say, in the sentence "A SIMPLE SENTENCE consists of a single main clause" (Trask 2000: 24), the phrases *simple sentence* and *main clause* can be designated terms in the metalanguage, but the remainder is just ordinary English, and could appear again in something like "A light lunch consists of a single main dish."

This distinction is essential if we are to avoid confusion. The stream of speech is linear and no segments of it assume a superior or abstracted position with reference to the rest of it. Words and phrases whose reference is to aspects of the language system behave just like other words – their relationship with the language system is purely semantic.

Nor do I want to spend time on the use of the term *metadata* to describe the information about a text that is often gathered to aid its classification when it is added into a corpus. It is a very silly and pretentious use of such a term, but innocuous except in one context, to which I will return. To be sure, "metadata" about something is external to it, but it is not a set of abstractions from it; in fact it is the opposite – it is external information that cannot be derived in any direct way from the text, such as the date of birth of the author.[1] But merely to be data about an object but external to it is hardly sufficient reason to be in a "meta" relationship to it, in the way the word is used nowadays. Or else you could consider your passport meta-you.

Where "metadata" can be pernicious is when it is not kept separate from the text to which it is external; indeed I suspect that the origin of this use of the term is the old-fashioned practice of mixing data *about* a text with the text itself, in such strange constructs as "DTDs" (Document Type Definitions) or "headers" that were enthusiastically promoted as enrichments to ordinary text, e.g. in the Text Encoding Initiative (http://www.tei-c.org/). Despite warnings that such insertions were in danger of corrupting the text, because they can never be reliably removed, many corpora of earlier periods are irretrievably damaged. The term chosen, "metadata", should have been sufficient warning in itself, because textual metadata has by definition no place in text; nevertheless it was regularly stuffed in, with predictable consequences.

The term to which I took, and take, exception is *metadiscourse*, which is indistinguishable from ordinary ("object-") language except that its topic is itself and its cotext. The term is particularly associated with comments about the talk that is evolving and of which the comments are part. Other words for the same category are *self-reference* and *reflexivity*, and I will use the former in this paper.

From a linear point of view, the role of discourse self-reference is different according to where the referent lies in relation to the referring item. If it is in text of the future (even in the immediately succeeding text) then the self-reference is a *prospection*, and acts as a preface, *advance label* (Tadros 1985) or something similar, introducing what follows. If the referent is in the immediately preceding text, then the self-reference *encapsulates* it, cancelling its interactive function. Encapsulation is the normal function of most cohesive devices, but it is unlikely that proponents of the "metadiscourse" category would accept all anaphoric and homophoric references as belonging to this category (it would be hard to exclude *and, but* and *or* from a full list of words and phrases whose occurrence entailed the existence of previous states of the text).

In understanding prospection and encapsulation it is not necessary to postulate that some segment of a text is in a conceptually superior position to its surroundings. Spoken text as a physical event is momentary, and so only its meaning-trace is available for reference. The meaning-trace remains relevant for a short time, but its relevance decays unless it is encapsulated. Written text is physically longer-lasting, and the conventions of writing stress that care is taken to make the cohesive references clear and precise; however, it seems that readers do not normally take advantage of the option, always open, to return to an earlier place in the writing to check an encapsulation; they prefer to rely on the immediate text to give them enough to continue looking ahead rather than behind. So in practice the permanence of writing has little effect upon the process of engaging with text.[2]

The key event for "metadiscourse" is where a self-reference cancels a prospection, in an utterance like:

1. Interviewer: How do you intend to achieve this
   Politician: That's a very interesting question

   (Francis & Hunston 1987 p.127)

A question prospects an answer, and since text is linear the expectation is set up that the next utterance will begin an answer. This is a strong prospection, and participants will act on it and interpret the next utterance as some kind of answer unless it is clearly one of the few types of move that challenge, defer or otherwise divert the discourse from its immediate goal. One of the recognised diversions is self-reference, where, instead of answering a question, you talk *about* it. The question becomes part of the subject-matter, its meaning-trace is removed from the linear organisation and its interactive function is thereby cancelled.

In the example above the politician now has the option of ignoring the question altogether, because the prospection no longer applies.

All utterances prospect at the very least that they will be treated as contributions to the linear discourse, requiring attention. So in Ädel's example (op.cit. p 1) the word *that* presumably refers to an attributed statement.

2. I never said that!

Rather than (2) being somehow elevated from the surrounding discourse, a better representation of the textual relations is that the referent of *that* is "demoted" from participation in the ongoing discourse, which of course remains inalienably linear. But it is difficult to see the relations in this way when the technical term suggests the opposite.

For these reasons I believe that the term *metadiscourse* is a barrier to clear thinking on the way in which language refers to itself as it goes along. The alternative that I proposed some time ago is *plane change,*[3] where referents are, by being referred to, moved away from the plane on which the interaction is taking shape. In this representation the linearity of the discourse is not threatened, but no other aspect of the description is disturbed.

5. What to count, what not to count

I was not looking forward to writing this section, because the intricacies of statistical calculations and argument are not an interest of mine, but in corpus study statistics seems to be unavoidable. However a recent paper by Kilgarriff (2005) does all the difficult bits, leaving me with only the need to comment briefly. He starts from the self-evident premise that language text does not occur at random. Despite this, he points out, all the statistical measures we use are based on the possibility of randomness, so if our results cannot be distinguished from random then our only possible conclusion is that our test data were not extensive enough. This unsatisfactory position arises because (a) randomness is not an option and (b) randomness gets less likely as the data set gets more extensive.

Kilgarriff distinguishes four types of association that might pertain, for example, between co-occurring words in a text. They could be "Random, Arbitrary, Motivated or Predictable (R, A, M, P)." He continues (p. 263) "The bulk of linguistic questions concern the distinction between A and M." Unfortunately, the computer has no way of distinguishing A from M because this distinction depends on the meaning, so the whole point of the practice is called into question. Later in the paper (p.264), Kilgarriff tries to make a virtue out of the fact that the primary test – of randomness – tells the linguist nothing of interest. "Making false assumptions is often an ingenious way to proceed."

Making assumptions that *may* be false – even assumptions that are very likely to be false – is an accepted component of scientific method, but making assumptions that are *known* to be false is not likely to be a prelude to acquiring useful information.

Most of Kilgarriff's points are perceptive, however, and could be the basis of a variety of interpretations; in particular they could lead to more drastic conclusions than his own concerning the comparison of corpus patterns with random or chance occurrences. If we follow through his arguments, they could lead to the whole practice of applying chance-based calculations being questioned and revised. After all, the only statistically-relevant fact that is known about a corpus is that its distribution does *not* occur by chance, so why use chance as a criterion of relevance? Whether the occurrence of a pattern beats or does not beat chance predictions tell us nothing about the meaningful units and their relations.

Before suggesting alternatives, I should make it clear that there are a number of active areas of textual research to which the remaining arguments in this section are not relevant. These areas of investigation make full use of conventional statistical techniques and develop their own penetrating routines without focusing at all on the meaning of the texts under study. There lies an important distinction; studies of language variety, of authorship – any study which examines solely the numerical properties of a body of textual material has no interest in the communicative events of which the texts are the physical record. For example one of the earliest studies, Zipf's Law, established a relationship between the number of times a word form occurs and its rank in a wordlist ordered by frequency. What the words mean, and what sort of meaningful relationships they contract with each other, are not directly relevant to their frequency distribution or their conformity to arithmetic laws.

Returning to consideration of the meaning-oriented investigations, we note that chance-based measures are in constant use, notwithstanding the reservations we might have about their relevance. They have for many years given linguists general pointers towards which usage patterns are worth consideration, and no doubt they will continue to do so for some time. However, in so far as they essentially report on whether or not textual patterns fall inside or outside the range indicated by chance occurrence, then they are stop-gap at best, and could be misleading. Their prevalence is probably a result of a misunderstanding about the nature of a corpus as a sample of a language, to which we now turn.

The distinctive character of a corpus as against other collections of texts is that it claims to be reasonably representative of a language or a language variety (Sinclair 2005b). Since language text is a population without limit, and a corpus is necessarily finite at any one point; a corpus, no matter how big, is not guaranteed to exemplify all the patterns of the language in roughly their normal proportions. But since there is no known alternative method for finding them all, we use corpora in full awareness of their possible shortcomings.

We have no reason to believe that two corpora drawn from the same population using the same methods should show the same distribution of phenomena. In fact we have every reason to believe the opposite; we know that the characteristic patterns of a text are dependent on what the text means, and that unless the meanings in the two corpora are replicated, the texts will not exhibit the same patterns.

If we take a shovelful of sand from a sandpit and examine its constituency, and then take another shovelful, we can postulate that the two shovelfuls will share many constituent features, because we have no reason to believe that there are orderings within the sandpit. But in the case of language text we *know* that there are orderings, in fact we know that the reason that the text *exists* is because it is ordered for communication, and so it is meaningless to take shovelfuls of text and expect them to have similar constituency.

It is not clear to me why we deal in probabilities when analysing corpora. In front of us are not probabilities but actualities, and those should be the focus of our attention. Any actual set of events can be trivially converted into a probability, so that if A and B co-occur *n* times in a corpus, we can assert that there is a probability that they will co-occur approximately *n* times in a similar corpus. To which the obvious response is "So what?" – and we know that it is a poor prediction anyway. The aim of studying language in corpora is to describe and explain the observed phenomena, not to predict what some other corpus might contain. In any case, Kilgarriff establishes firmly that probabilities between corpora are not appropriate, no matter how similar the corpora.[4]

All the familiar measures of significance, the chi-squared, the log-likelihood, the t-score, z-score and the poor MI are predictive;[5] if they were simply descriptive they would not need to invoke probabilities. Corpus linguistics needs its own methods of statistical analysis, which should be purely descriptive and which should quantify linguistic concepts and categories.

The confusion may lie in the word *sample* itself. A corpus is a carefully selected collection of texts, involving a great deal of expert human judgement. A statistical sample is "expected to be selected in such a way as to avoid presenting a biased view of the population" (Wikipedia). These are diametrically opposed concepts; nothing could be more "biased" in its selection methods than a corpus. So perhaps no warning bells sounded when a corpus began to be treated as if it were the kind of sample which is amenable to statistical analysis.

Once the tangentiality of our present measures of the relative importance of text patterns is accepted in the corpus community, it will not be long before alternatives emerge. It is too early to say in detail what is likely to happen, but a concentration on the collocations and other co-occurring features, rather than on their components, is long overdue and is a productive entry tactic.

6. Little boxes

The final section of the paper considers the use of labelled bracketing as a notational device. In the section above on metadiscourse, it was stressed that language text is linear and directional. Whether in speech or writing, only one event is happening at once. Position in a linear string is potentially significant, in that it can contribute to the realisation of meaning. Text is directional in that it does not usually mean the same thing if reversed, if indeed it means anything at all. The meaning does not necessarily survive any changes in the positioning of the units of realisation.

All this almost goes without saying; Saussure ascribes two fundamental qualities to the linguistic sign, its arbitrariness and its linearity, in successive sections of the *Cours* (Saussure 1916). While much has been made of the first quality over the years, very little mention is made of the second although it is at least as important a quality; perhaps the fact that it applies to text – *parole* – rather than to *langue* has relegated it to a secondary status.

Back in the early days of experimentation in computational analysis of language, linguists led by Geoffrey Leech at Lancaster University were the UK pioneers (and are still leaders twenty years later). During a presentation in 1984 Leech made a passing remark that their software worked better if the sentences were input backwards.[6] It was light-hearted enough, but I could not get it out of my head, and it is one of the starting points of the argument of this section. I repeat the question I asked at the time – Why, if this is the case, do we not speak backwards?

For the present discussion, we need only note that Leech had detected some imbalance of signs in the sequencing, such that in one direction the text was easier to divide into meaningful segments. That in turn suggests some imbalance at the boundaries of such segments. Text structure is not symmetrical, although the descriptions tend to depict it as a construction of neatly nesting units. Several years passed before I found a possible explanation for this phenomenon.

7. Labelled bracketing

First of all, let us study an example of the kind of notation that is at issue. Formal grammars are often displayed as tree diagrams or similar networks, for ease of understanding, but these are equivalent to the notational conventions of "labelled bracketing". They conform to constraints like those of Halliday's "taxonomic hierarchy" (1985) and so can be represented by strings of symbols in several

levels of bracketing. Symbols attached to the brackets themselves indicate the structural value of the contents. As an example, let us consider Figure 1, which applies a labelled bracketing to the following sentence.[7]
The boy stood on the burning deck, whence all but he had fled.

[[[[the][boy]] [[stood]]
S MC NP det /det n /n /NP V v /v /V

[[[on] [[the][burning][deck]]]]]
A PP prep /prep NP det /det adj /adj n /n /NP /PP A /MC

[[[whence]] [[all] [[but] [[he]]]]
SC A conj /conj /A NP pron /pron PP prep /prep NP pron /pron /NP /PP /NP

[[[had]] [fled]]]]
V aux *have* /*have* /aux vb /vb /V /SC /S

Figure 1: labelled bracketing

Legend: S= Sentence, MC= Main Clause, NP= Noun Phrase, det= determiner, n= noun, V= verb, A= Adjunct, PP= Prepositional Phrase, prep= preposition, adj= adjective, SC= Subordinate Clause, conj= conjunction, pron= pronoun, aux= auxiliary verb

The structural description above is intended to be non-controversial, the application of a consensus grammar to an often-cited specimen sentence. Notice that between *deck* and *whence*, there are five brackets closed and three opened. I presume that this is intended to represent in some way the structural perceptions of the person originating the sentence and any person who experiences it as a communication.

The question arises, how does the text signal to the reader these eight pieces of structural information? The appearance of the word *deck* tells us that, as a noun, it is capable of carrying the role of headword of the noun phrase that we know started with *the* two words earlier – but it does not settle the matter, because *deck* could just as well be a modifier of another noun, say *tarpaulin, winch* or *hand*, in which case the noun phrase [NP], the prepositional phrase [PP] and the main clause [MC] would not be terminated immediately after *deck.* So the information must largely arrive with the occurrence of *whence*. This word is a fairly reliable structural marker, telling us of the beginning of a subordinate clause of place and therefore by implication the end of the previous one; if a new clause has begun, and is not embedded in another clause, then it follows that the previous clause must be terminated. And if the clause is terminated, then the restrictions of the taxonomic hierarchy mean that all its component structures are also terminated – the PP and its NP in this case.

The important point to note from Fig.1 is that there is no actual signal of the end of the main clause or of its subsidiary structures, but there is a clear signal of the beginning of the subordinate clause, which triggers all the closing brackets.

As to the opening brackets, since *whence* is a conjunction we can open a bracket for a clause and another for an initial adjunct, since it is a subordinator we can open a bracket for a subordinate clause, and of course since it is a word we can open a bracket for a word. The amount of structural information supplied by the choice of *whence* is very substantial.

The point where boundary information is the next most concentrated is between *he* and *had*, where there are four closing brackets and three opening ones. As a pronoun, *he* is almost certain to stand as a NP on its own, and this would explain the closure of the first two brackets. The outer structures, however, are not definitely closed by the occurrence of this word, because it could easily be followed by *and the captain*, thus prolonging the prepositional phrase and therefore the original NP starting with *all.*

The appearance of *had* makes it clear that the original NP and all its components are terminated. The combination *he had* signals fairly definitely that they are separated by a subject-predicate boundary, and thus justifies the last two closing brackets, as well as opening a bracket for the predicator [V]. The assignment of *had* to the status of auxiliary is not at this juncture clear, and is only confirmed by the occurrence of *fled*. The sentence could have continued "all but he had guns."

8. Complete and finished

It is helpful at this point to draw a distinction between structures which are *complete* and whose which are *finished.*[8] A complete structure is one which is well-formed and thus has meaning-potential, in Halliday's terms – it is an abstract concept, not without problems but in many cases specification is straightforward. On the other hand a finished structure is a segment of text which is actually terminated. So in our specimen sentence, *the boy* is both complete and finished as a nominal element, but in another text it could occur as a component of an indefinitely large number of other nominal structures beginning *the boy….*

The general point to be made here is that in understanding text the movement from recognising a structure as complete to appreciating that it is finished is largely a matter of hindsight – settled by the appearance of a word which is incompatible with the evolving structure. The possibilities are:

1. complete and finished
2. complete but unfinished

3. incomplete but finished

There are no examples of possibility (3) in the test sentence, but they are common enough in conversation; here is one:

A: I mean you know its not important its just er
B: What do you mean ……
(Francis & Hunston 1987 p.144)

A case can be made for *I mean* and *you know* to be freestanding discourse units that do not require anything to follow them, but *its just* is clearly an unfinished unit, cut off by the superimposition of B's question (though the tell-tale *er* suggests that A was glad not to have to finish the structure. Sinclair and Mauranen (2006) offer an up-to-date analysis of such phenomena.

Of the thirteen words in our specimen sentence, only two include in their meaning the notion that they are final in the phrase which they wholly or partly realise. These are the conjunction *whence* and the pronoun *he*. That is to say, it is difficult to imagine any way in which the conjunction *whence* could be elaborated; *he* on the other hand can be qualified in limited ways (e.g. *he who…*)[9] but extensions such as these are very rare in normal texts. So the analyst can be fairly confident in placing a boundary after *whence* indicating that it is a full elements of clause structure despite being a single word. This reflects the expectations of readers of the original sentence. Note that this built-in boundary marking applies only to single words and to a very small number of them, some of the subordinating conjunctions perhaps. In the case of the other eleven words it is the occurrence of the following word that establishes the boundary.

We can now revise Fig. 1 by suppressing all boundaries except those that can be predicted before the next word occurs. To avoid clutter I am also removing the word brackets, which are redundant structurally and only serve as positions for the word class to be identified.

```
[  [  [the boy [stood [  [on  [the burning deck
S  MC NP        V      A  PP   NP

 [  [whence]   [ all [ but  [he] [ [ had   fled
SC  A       /A   NP    PP     NP /NP V aux
```

Figure 2: reduced labelled bracketing

There are now fourteen opening brackets and two closing ones; twelve of the structures are not explicitly signalled as being finished; they are simply replaced by another structure. Therefore twelve of the closing brackets in the original analysis must be inferred from hindsight, albeit momentary hindsight. There is thus a large imbalance between the way in which the openings and closings of structures are signalled, and I presume that Leech's team, back in the eighties, became aware of this.

This finding can be supported and partly explained by theory. Chomsky (1957: 23-25) pointed out a long time ago that the set of all well-formed sentences of a language could not be limited. This property arises from the declaration of iterative rules in the grammar; an iterative rule is one which has the same symbol on both sides of the arrow of derivation. This kind of rule can be applied over and over again, generating longer and longer sentences and never reaching an end-point. Just one iterative rule would give the property of limitlessness to the set of sentences generated by a grammar, but in a natural language there is a range of iterative rules which extend sentences in a rich variety of ways.

It follows from this property, also, that there is no such thing as the longest sentence in a language, since any candidate could simply have an iterative rule applied to it, or reapplied if it was present already in the phrase structure. Below the sentence, indefinite extension is not guaranteed, but a moment's reflection will confirm that in all cases where a complete structure contains more than one word the same feature will apply; there are, as we have seen, a few candidates for "terminators" among the conjunctions and perhaps pronouns, but those are operating at the lowest level of structure.

Figure 2 is a fairly acceptable analysis of the sentence, in that it assigns labelled brackets where these are clearly indicated in the text. However, its lack of symmetry in bracketing would cause it to be rejected by any parser of the usual variety; in fact one of the first checks of the parser is to count the number of opening and closing brackets and ensure there is the same number of each. There is a huge discrepancy here. Options for resolving the situation include:

> a) introduce an automatic "bracket equaliser", which regularises the notation. So after *boy* (fig 2) a closing bracket would be added, for example. I expect that this would present few problems, but it transforms the structure into something that is not justified by the facts.
> b) revise the underlying model, the conventions of bracketing, in such a way that structures which do not signal their finishedness within themselves, are simply left open.

Option b) is the closest to the facts, and it allows for the structure to be determined further by punctuational or prosodic features. *The boy*, in the abstract world of completeness, carries forever the potential, as a noun group, for being continued in a number of diverse ways. In the actual sentence, this potential is over-ridden by the appearance of *stood*. Such a model will bring out the dynamics of text, which especially in writing is often neglected.

It is an untidy model compared with the cool symmetry of logical structures, but we are often made aware of how untidy language is when it is in use in communication; so there is no surprise here. It will be interesting to see how such a model will compare with the traditional forms of analysis; despite the lack of any disagreement about the structure of the text, we have seen that there are seriously different ways of representing that structure.

9. Punctuation and prosodic features

The lack of signalling finishedness does not appear to detract much from people's ability to speak and write effectively, and to understand both modes of communication. In writing there has developed a system of punctuation which sometimes helps in boundary assignment, the comma especially, though it is often ambiguous. We are accustomed to using fairly settled conventions of punctuation nowadays, but these were gradually stabilised by the printing industry, are thus of very recent origin, and cannot be considered essential components of the language system.

The three levels of boundary that we have in focus are that of sentence, clause and group/phrase. The tricky boundary is the lowest one, and the system of punctuation while helpful, is not decisive. Also it does not explain why a reader only rarely needs punctuation marks. In understanding a text a reader must make some assignment of boundaries, perhaps subliminally.

Some punctuation marks signal that a structure is finished; so a full stop after *fled* would allow us to close the predicate, the subordinate clause, and the sentence as a whole. It would be quite natural, though not obligatory, to place a comma after *deck,*[10] so that *whence* did not have to carry all the large informative load which is otherwise placed on it. The redundancy of punctuation allows a sharing of the information load.

From the point of view of boundary marking, punctuation marks in English support closings. If the initial capital letter of a sentence is considered part of the punctuation system, then it is the only one that marks an opening. According to present-day practice, punctuation is not permitted between elements that form a syntactic unit below clause level, e.g. between subject and predicator of a clause, or between a preposition and its object. The distinction between defining and non-defining relative clauses follows this practice, in that the syntactic unity formed by the defining clause cannot be interrupted by punctuation.

The full stop, question mark and exclamation mark all indicate clearly the end of a sentence. The colon and semi-colon indicate clearly the finish of all units below the sentence. But the occurrence of a comma, while excluding some options, does not unequivocally terminate all structural units that are open when it occurs. It acts as a resolution of tension between the demands of linearity, which includes some elastic limit on the size of meaningful segments, and the realisations of the complex hierarchies of the grammar. The physical size and length of the realisation of each abstract category is one factor in the balance; the demands of the structure of which it is a part form another; the possibilities for enhancement, extension, elaboration etc. ad infinitum form a third mechanism for concatenating words and phrases together. Balancing this is the need to keep the discourse in bite-sized pieces, so that a listener or reader will not get confused in real time with the complexity of the message. So after five words or so, the pressure will increase to place a comma at a boundary that in other circumstances might not merit such a formal mark.

In the spoken mode, there are patterns of tone contour which again give clues to when a speaker is finishing a structural unit. Brazil's (1997) description gives a clear picture of the harmonious co-ordination of choices that is natural conversation.

10. Completeness

Having got thus far, it is worth a few moments' attention to the details of completeness. It is an intuitive decision, whether an evolving structure is complete or not, and so may not always be clear or logical, but always indicative of something in the structure that is worth consideration. An analyst is too detached to offer a reliable commentary on intuitive matters, but we have to do the best we can. It is a rough-and-ready decision-making process, far removed from the "well-formedness" that formal grammars envisage, but using similar criteria.

Let us look at cases. Recall the first instance of a complete but not finished structure above:

The boy stood

Since STAND is an intransitive verb, this phrase is well-formed at some level of abstraction, but it looks unlikely with these actual realisations. The incidence in The Bank of English of a punctuation mark following *stood* in a thousand instances chosen without bias is 42, less than 5%. As used in this sentence, the word *stood* seems to require some adverbial element to terminate it. So the argument for the structure being complete at this point is weak and not conclusive.

The boy stood on

This is much less likely; although *stood* now has an adverb, our intuitive feel for the actual phrase is that *on* is a preposition, so we await the object; it is incomplete. With a similar verb, *stayed on*, there is clear evidence that the structure is complete, and of course *walked on, drove on* show *on* frequently as an adverb. In the spoken language there would be a stress distinction marking the difference, with the preposition unstressed, and there is a hint of irony in this piece of doggerel in that *on* occurs on a stressed syllable in an iambic rhythm, a trap for inexperienced reciters. There are no plausible instances in the Bank of English of *stood on* finishing an active verb structure.

all

This is complete as an NP, but it is not finished. There are plenty of attestations in the corpus for *all* realising the subject position in a subordinate clause of place, so while *all but…* is also found it is much less frequent.

There are good reasons why a set of notation conventions based on Figure 2 should be developed; the directionality of text is prominent, structures like discontinuity will be better described and there will be much more flexibility in describing the way a sentence can develop. The symmetrical labelled bracketing that we are accustomed to cuts off, quite unnecessarily, many developmental options for the structures. With asymmetric bracketing we are much closer to symbolising the textual signals.

11. Conclusion

The late, brilliant Nick Lafitte, who took up linguistics in flight from his previous career in Econometrics, used to wonder (pers. comms.) why linguistic descriptions, statements about language, were so different in their nature from language text. His interest was aroused by redundancy, which was routinely said to be a major feature of language text. Nick presumed from this that either the structural representations of text would exhibit this feature, or the descriptive categories. But there is no treatment of redundancy at all in descriptions or the theories that lie behind them; it just disappears. Presumably this discrepancy is not just a mistake or oversight, but rather that the ground rules of the theories and descriptions preclude a feature like redundancy, despite its noticeable presence in text. These ground rules are imported from outside, and perhaps they are not quite right for the job; we frequently say that language is unique because its theories are also written in language (see "meta-" above) but we then describe language phenomena without taking advantage of this coincidence.

Something of the same can be said of linearity, which is so obviously a major constructional feature of text that it is usually taken for granted. It is a pity that we lose sight of it, because it could act as a brake against over-indulgence in abstractions and over-complex representations of simple phenomena. In most descriptions linearity all but disappears in favour of multi-layered hierarchies, which do not always seem to be strongly motivated. A recent study (Sinclair and Mauranen op.cit.) seeks to describe language text while maintaining linearity for as long as possible.

My framework for comment on the above issues is the network of relations between text and meaning, because those relations form the apex of language description. I find that it is all too easy to mix up conventions of practice and properties of the data (e.g. sample size), and we must be vigilant in protecting research against vested interests. I find that we sometimes import terminology from other disciplines without sufficient care, especially if it sounds good (e.g. meta-); ill-fitting terminology can certainly distort your thinking. I find that it is not appropriate to measure the salience of patterns of combination in texts by means of predictions concerning the distribution of the component word forms (conventional corpus statistics). New ideas are needed here. And I find that the notation of labelled bracketing, almost universally accepted in formal grammars, is a faulty representation of text structure; like poor terminology, it impedes clear thinking, and has inhibited grammarians from developing notations which better represent the texts.

Notes

1 There is a well-established branch of computational linguistics whose techniques are developed for determining authorship from internal textual evidence; these produce genuine metadata (see JLLC, *passim*). However their results are of course only postulates, and can be discussed and disputed with arguments that are quite different from the way in which arguments about external evidence are conducted.

2 This is a big topic, too big to pursue in this paper. The written form of the language has two distinct functions; one is to be read, in which case I would claim that readers behave quite like speakers in that they do not rely on the possibility of returning to an earlier state of the text. The other function is to make and keep records, where the text must carry all necessary detail and must cohere, avoid ambiguity and the like. While these are distinct functions, a user in reading mode who gets into difficulties can switch to record-keeping mode in order to resolve the problem. In ordinary, everyday reading this does not seem to be a common tactic.

3 An early treatment of this point is to be found in Sinclair (2004 (1982): 51-66).

4 Op.cit.: 268-270. Actually, Kilgarriff does not conduct a straightforward comparison of his two selections from the same large corpus, but associates the words with POS tags, which complicates the issue considerably, and makes it much less likely that the samples will match. He does not discuss his motivation for doing this.

5 I say "poor MI" because, in my limited experience of significance measures, MI is the only one which has to have both its head and tail chopped off before it makes sense. It is not difficult to understand that the tail of a significance measure gets less and less interesting as it goes down, and that a cut-off point is desirable in practical applications; however, to remove the items that are shown to be the *most* significant is bizarre. I put this point to Kilgarriff at TALC in Lancaster, 1994, but he did not address it.

6 At the ICAME conference, May 1984, in Windermere, hosted by Lancaster.

7 The conventions are that below each opening bracket is a symbol designating the structural value of what is inside the bracketed segment;

below each closing bracket is the appropriate symbol prefaced by a diagonal slash. S=sentence, MC=main clause, NP=noun phrase, det=determiner, n=noun, V=verbal element, v=verb, A=adjunct, PP=prepositional phrase, prep=preposition, adj=adjective, SC=subordinate clause, conj=conjunction, pron=pronoun, aux=auxiliary verb.

8 This contrast cuts across the hallowed distinction between langue and parole, competence and performance. It first came to my notice as the origin of textual effects in the analysis of Wordsworth's poetry (Sinclair 1972). In being exploited for stylistic effect, I used the terms *arrest* (structure incomplete, new structure initiated) and *extension* (structure complete, more material added without initiation of a new structure). Very recently, in working on linearity in grammar, I have returned to the distinction because it plays an important role in *chunking* (see Sinclair and Mauranen 2006).

9 The Bank of English offers 1755 instances of "he who" or "he whom", of which quite a number show the pronoun followed by a defining relative clause. But these are characteristic of certain marked text types, from the domain of religion, or faked antiquity, gnomic utterances or just a particular pomposity. In cleft structures like "It was he who answered." a conventional analysis would have "he" as the complement on its own.

10 Metrists might reasonably argue that the placement of a line-end after *deck* is equivalent to an auxiliary punctuation mark.

References

Ädel, A. 2006 *Metadiscourse in L1 and L2 English.* Amsterdam and Philadelphia, John Benjamins.

Austin, J. 1962 *How to do things with words*; ed. J.Urmson, Oxford, Clarendon Press.

Brazil, D. 1997 *The Communicative Value of Intonation in English.* [2nd Edition]. Cambridge: Cambridge University Press.

Chomsky, N. 1957 *Syntactic structures*, The Hague, Mouton.

Francis, G. and S. Hunston 1987 "Analysing everyday conversation" in Coulthard, R. (ed.) *Discussing Discourse*, Birmingham, ELR Monograph no. 14.

Halliday, M. 1985 *Introduction to functional grammar*, London, Edward Arnold.

Imbs, P. and B. Quemada 1988 , *Trésor de la langue française*, Paris, Gallimard.

Kilgarriff, A. 2005 "Language is never, ever, ever random" in *Corpus Linguistics and Linguistic Theory* 1-2, 263 – 275.

Krishnamurthy, R. ed. 2004 (1970) *English Collocational Studies* by J. Sinclair, S. Jones and R Daley; London, Continuum.

Léon, J. 2005. "Claimed and unclaimed sources of Corpus Linguistics", *Henry Sweet Society Bulletin*. N°44. pp. 36-50.

Saussure, F. 1916 *Cours de linguistique générale* compiled by C. Bally and A. Sechehaye, Paris: Payot.

Sinclair, J. 1972 'Lines about Lines', in B. Kachru, B. Stahlke, F. W. Herbert (eds) *Current Trends on Stylistics*, Edmonton, Linguistic Research Inc. 251-61, reprinted in R. Carter (ed.) *Language and Literature*, London: AlIen & Unwin 1982, 163-76.

Sinclair, J. 2004 "Planes of discourse" in *Trust the Text*, London, Routledge, 51-66 (reprinted from S. N. A. Rizvil (ed.) *The Two-fold Voice: Essays in honour of Ramesh Mohan*, Pitambar Publishing Co., India, 1982).

Sinclair, J. 2005a "Language as a string of beads: Discourse and the M-word" in E. Tognini-Bonelli and G. Del Lungo Camiciotti (eds.) *Strategies in Academic Discourse***,** Amsterdam and Philadelphia, John Benjamins; Studies in Corpus Linguistics 19, 163-8.

Sinclair, J. 2005b "Corpus and Text – Basic Principles" in M. Wynne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice* Oxford, Oxbow books. Also available in electronic form: http://ahds.ac.uk/-linguistic-corpora/http://www.ota.ox.ac.uk/documents/linguistic-corpora/.

Sinclair, J. and A. Mauranen 2006 *Linear unit grammar – integrating speech and writing*, Amsterdam and Philadelphia, John Benjamins.

Tadros, A. 1985 *Prediction in Text*; Birmingham, ELR Monograph no.10.

Trask, R. 2000 *The Penguin Dictionary of English Grammar*, London, Penguin Books.

Bank of English, The http://www.collins.co.uk/books.aspx?group=153.

Brown corpus http://khnt.hit.uib.no/icame/manuals/index.htm.

Index Thomisticum http://www.corpusthomisticum.org/it/index.age.

LOB corpus http://khnt.hit.uib.no/icame/manuals/index.htm.

How 'systemic' is a large corpus of English?

Robert de Beaugrande

www.beaugrande.com

Abstract

That both the text and the discourse are 'systemic' entities, is, I believe, a principle of general consensus within Systemic Functional Linguistics, although in a sense related yet distinct from the 'systemic status of a 'language'' plausibly following the distinction between 'actual' versus 'virtual' (or 'potential', to use a less overloaded term). Now that very large corpora of authentic text and discourse are readily accessible, we can take up the question of whether such a corpus may in turn be 'systemic' in a sense mediating between the poles of this distinction: both 'intersystemic' and 'intertextual' at once. The present investigation adduces newly extracted corpus data to answer this question in the affirmative, notably by demonstrating how these data project and confirm 'systemic' tendencies exerting pressures that can modify, expand, or alter the language system itself. Indeed, such demonstrations might pass unnoticed if these very factors did not guide the methods of search and retrieval I deployed.

1. Language and text as 'system' and 'systemic'

Among the foremost achievements of 'systemic functional linguistics' (SFL) has been to expound an alternative view of the relation between *language* and *text* (e.g. Halliday 1992; Martin 1992); intentionally interrelated texts can be said to constitute a *discourse*, the most common of course being a conversation.

A 'language' is a *potential system*; a 'text' is an *actual system*. Thus, a complex process of *actualisation* is implicated in the production or reception of any text. By operations of selection and combination, a set of *intrasystemic* choices become a set of *intratextual* choices, and the relation between these two sets is *intersystemic*. Along the way, the *systemic function* of any given expression can be reset: adjusted, specified, weighted, colligated, collocated, and so on.

Prior to SLF, the trend in language studies and linguistics often was to take the process of actualisation for granted and begin the 'investigation' with a handful of samples invented by the investigator, doubling, one might say as actualiser. Since the language was assumed to be uniform (heterogeneous) across an entire language community, the identity of the actualiser was judged immaterial for the 'analysis'. This expedient logic was turned back to front and made circular: those aspects or features of language were "investigated" which were judged the most uniform, the rest being airily relegated to pragmatics, stylistics, rhetoric, sociolinguistics, or whatever seemed most opportune for the evasion. The "speaker" was "idealised" and "homogenous" into the blandest possible human, who never said anything worth hearing.

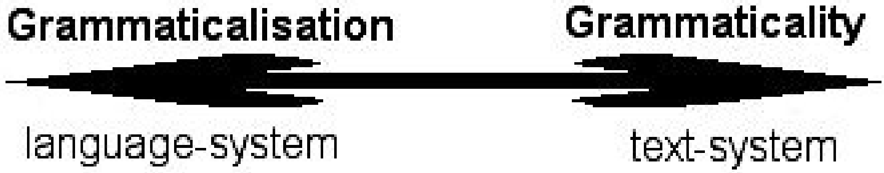
I shall argue that the gap between language and text might be most effectively relaxed by a *very large corpus of texts*, whose ability to widen the coverage of the actualisation moving from potential to actual, and the representative status of the population of actualisers, increases and diversifies with size and balance. Admittedly, no corpus of any size can lead us to a complete display, but we have repeatedly learned that as our corpora grow, the picture becomes sharper and more facetted, often unveiling curious surprises: nouns whose plural does not match the meaning of the singular, verbs whose passive does not follow from the active, and so on. In essence, these insights contribute to the deconstruction of the premature hypothesis of uniformity and its centrality I have cited. They demonstrate the vital interaction between grammar and lexicon well beyond how they are conventionally conceived.

However, to my knowledge, the complex process of actualisation has not been thoroughly analysed as such — and not just its initiation as a system property (e.g. speech enunciation) and its termination as a text property (e.g. sentence analysis). The outstanding question remains how and how far participants in communication — whose knowledge of the potential system certainly *cannot* be uniform — perform the process with sufficient (though hardly complete) uniformity to arrive at the condition reassuringly known as 'understanding'. The answer I have repeatedly suggested in my books is that the process is so devised as to favour a temporary 'tuning' which can not only occur during the overall process of actualisation, but must also be prefigured and supported in the design strategies of both language and text. To revise the familiar terms, the 'process' aims at an interactive and convergent constellation of outcomes in the 'product', especially by the multifarious ways in which some choices or sets of choices raise or lower the probability and suitability of other choices (cf. Halliday 1992).

2. Intersystemic processes between language and discourse

Provisionally, then, I shall propose ten *interactive processes of actualisation* and note some data where dynamic pressure and evolution seem to be still operative, taken from my own English Prose Corpus (EPC, 100 million words of 'classic' texts), the British National Corpus (BNC, 100 million words of contemporary texts), and the Internet accessed via AltaVista (uncountable words in on-line texts, annotated with $^{\text{WWW}}$). These processes have been implicated in the educational goal of **standardization** of English, mandated in the Tories' cloud-cuckoo-land **National Curriculum English** (1988) to quite disparate degrees, if at all (Beaugrande 2004), which is like a road map grandly displaying only the goals but not the roads for going there.

Among the most capital errors of the "generative" approach and its epigones has been to situate **Grammaticality** as prefigured in the system, whereas in fact it can only be the textual product of **Grammaticalisation**.



'Grammaticality' thus can only be accessed via texts, however bland and banal, after the process has been achieved. I have never seen or heard a convincing example of 'grammaticality' or 'ungrammaticality' in the potential system. The flurry of 'sentences' conjured up to demonstrate the ostensible borderline were simply instances where grammaticalisation has been either trivialised [1] or vandalised [2].

[1] John knew what Mary was doing (Annie Tremblay)$^{\mathrm{WWW}}$

[2] *Which room is there in a very strange beast with enormous antlers and five arms? (Joseph Sabbagh and Lotus Goldberg)$^{\mathrm{WWW}}$

The jammy sun-drenched students at the University of Manoa in Hawaii were not supposed to draw the immanent inference that Mary is up to something naughty or sneaky (or worse) [1]. Still less were the frost-bitten students at MIT and McGill supposed to notice that the freak anatomy of the 'beast' smuggles in non-grammatical unacceptability [2].

The ongoing dynamics of grammaticalization can be documented by the at times wilful conversion and confabulation among word classes:

[3] The principle of inverse irreversibility. An inquest into scientific methodology, from the Popperian hypotheticodeductive perspective, with Kuhnian paradigmatic nonreconstructionism to Feyerabendian counterinductivism (New Scientist)

[4] Cyclosporin A, a cyclic undecapeptide, is a potent immunosuppressant that binds to a peptidyl-prolyl cis-trans isomerase, cyclophilin.The cyclosporin A/cyclophilin complex inhibits the calcium- and calmodulin-dependent phosphatase, calcineurin. (Nature)

Although the marked items are lexical items in an ordinary sense, the only way I can see to understand them is to run a grammatical back-analysis into the constituents.

A quite different clue of dynamics is the rise of items whose grammatical functions are not accountable in grammar-books. The adverb 'so' is placed and given stress as an intensifier in positions it would not usually occupy:

[5] Well I'm just so going to bed now (Xanga)WWW

[6] And I'm just so going guy crazy and I want one (MySpace)WWW

One wonders how you 'so go to bed' – taking a ferocious running jump, perhaps. But how 'so going guy crazy' differs from 'going so guy crazy' remains for me a mystic rite of teenagerhood.

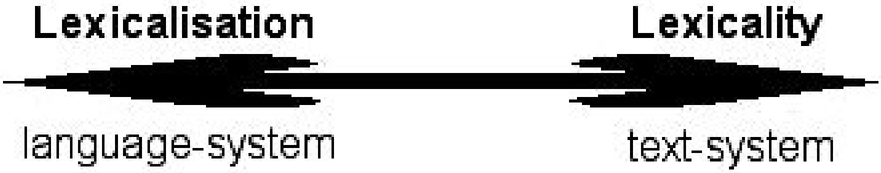
Another grammatically displaced item recently emerging is the negative 'not' postponed after an (ironically flavoured) positive statement, viz.:

[7] why do these old men discover they can get new careers lumbering ape-like through rock history. One awaits the geriatric Techno of the next century with interest. Not. (New Musical Express)

[8] James Baker III and the seven dwarfs of the "Iraq Study Group" have come up with some simply brilliant recommendations. Not. (Columbus Free Press)

Yet the risk of being shouted down as a gormless berk before you get to utter your Negative may not be a trivial one.

Though it is equally essential, the process of **Lexicalisation**, which engenders the **Lexicality** of the text-system, has been relatively neglected, presumably because it would not fit well either into the tinker-toy formalisms of trees and networks of "formal grammar" or into the cut-and-dried "lesson plans" of the "standardisers".



Yet even the most freeze-dried "grammarians" could not deny that their frail constructions are like hollow eggshells without Lexicalisation to bring them to life.

Another disquieting factor is that Lexicalisation is far less stable and more dynamic than Grammaticalisation. Discourses about computer construction and high-tech race-cars are rife with lexical items hardly one would have understood some years back and perhaps few enough would today:

[9] The nForce 680i SLI board is eVGA's flagship motherboard with support for the Core 2 Duo and Quad that uses our new chipset (Morry Teitelman)WWW

[10] Enhance the under-hood appearance of your V [V-8-engine block]. A lightweight carbon-fiber tower-to-tower brace replaces the standard steel brace. In addition, the radiator and engine cover features an attractive carbon-fiber appearance. *(Cadillac)*^{WWW}

The dynamics are also readily manifest in the spreading of non-technical expressions. One amusing source can be regional varieties of English with a phonetically ironic side-effect, such as Irish English 'gackawacka' for a tiresome fool [11] or African-American English (or Ebonic) 'badonkadonk' for a curvaceous female behind [12], reputedly cloned from 'honky-tonk', a music to which such an, erm, asset can be paraded to eye-catching advantage

[11] How could she admit how silly Sarah always made her feel? "Oh, let me write her a reply, oh do!" Aislin said, "We'll pay back this gackawacka for all her stupid blather." (*Maureen Monahan*)^{WWW}

[12] How J-Lo [Jennifer Lopez] squeezed her badonkadonk into this Zum Zum [trendy fabric] wonder is damn impressive (*Carpet Burn*)^{WWW}

University student discourse is renowned for droll contributions like 'borassic' for broke [13], and 'trollied' for drunk beyond control [14].

[13] Saz is borassic and will be until she gets a job after graduation. She needs money for essentials like a Kylie programme on Saturday. (*Specimen Days*)^{WWW}

[14] Due to being so trollied on his birthday in Copenhagen, Axel asked folks how to get to Amsterdam Centraal (not only wrong city... wrong country!!!). (*Team Plastique*) ^{WWW}

So intricate is the interaction between Grammar and Lexicon that it would often be more apt to use the systemic term **Lexicogrammaticalisation** that engenders the **Lexicogrammaticality** of the text-system. But consistent use would be cumbersome, and we might be content to deploy the other terms as 'short cuts', keeping in mind we are doing so (Halliday 1994).

The vibrant dynamics of the Lexicogrammar are best shown in its creation of **Colligations** (habitual grammatical combinations), as in [15], and **Collocations** (habitual lexical combinations), as in [16].

[15] Author of Bush Biography Commits Suicide. If you believe that, I've got some ocean front property in Arizona I'd like to sell (*Wake Up America*)^{WWW}

[16] Sometimes I come here by myself, just to relax, think things over. By yourself?" Kelly narrowed her eyes with mock suspicion. 'The eligible bachelor on the pull?' (*Stone Cold*)

These combinations lack the fixity of traditional "idioms", which all too easily mutate into clichés; and some elements are more variable than others. The jibe of clueless credulity in [15] is fairly stable in the conditional "if"-clause, but the variety of worthless or jocular scams, usually land or bridges, may vary considerably, even to "property on the moon".

Complex systems generate a certain chance margin, where Collocations may seem to abruptly and unwantedly revert to their basic lexicality:

[17] Your face can open doors (*BBC Online*)^{www}
[18] Jury is still out on composting toilets (*Salem Statesman-Journal*)

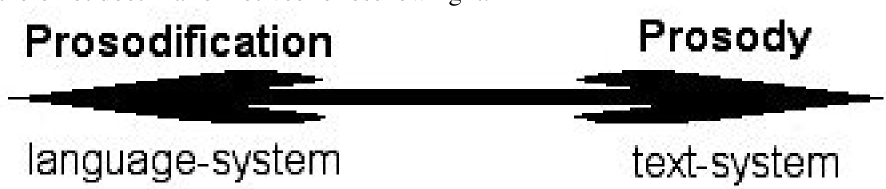
As factors within the doughty project of the standardisation of English as a foreign language, Grammaticality is plainly more cultivated than Lexicality in the discourse of learners; probably their routine dissociation during "instruction" is responsible. Some data from my students at the United Arab Emirates University seemed Grammatically regular but Lexically bizarre, as when I was struggling to convey the Medial Transitivity in English whilst they struggled in turn to model it on the Passive as the familiar alternative to the Active in Arabic:

[19] *my sample text*: Mrs Bennet fidgeted about in her chair, got up, and sat down again. (*Pride and Prejudice*)
student responses:
The chair was fidgeted up and down by Mrs Bennet.
Getting up and sitting down got fidgeted in her chair.
Mrs Bennet's chair was fidgeted about, was got up, and was sat down again.

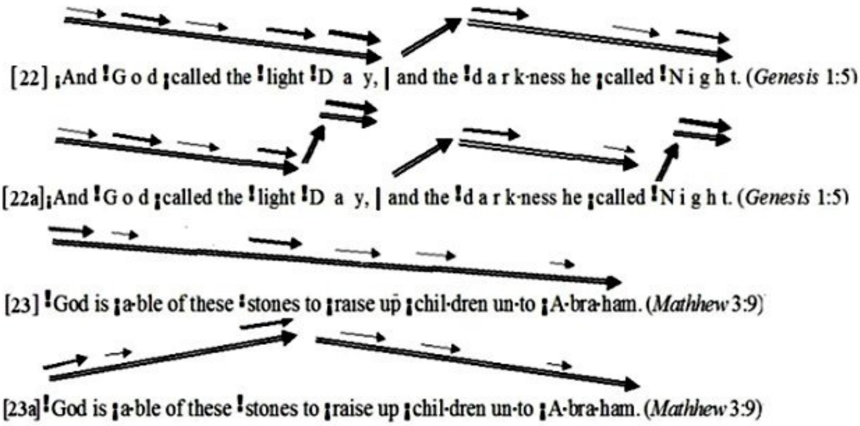
By contrast, other Arab student data seemed both Grammatically and Lexically deviant:

[20] Miss Raymond looks smelly [smiley] face but speaks in pride ways. She collects her hair in the back. Her teeth look when she talks, and she owns angry tone. She is a liar person who lied to disappear her ignorant.
[21] If anyone dressed by the name footman he will be shame that they don't even want to wear their clothes. In the US was not respect and tricker man and swindle person.

Prosodification is a vital process for engendering the **Prosody** of the Text-System, whose neglect – aside for the pronunciation of individual sounds and words -- at nearly all levels of education and research would be astonishing, were there not doctrinaire motives for eschewing it.



It is far harder than Grammar and Lexicon, however construed, to standardise by 'rules' or judge for 'correctness' or 'rightness', much less fix in a 'grade'. Depending on personal interpretation, either [22-22a] or [23-23a] would be quite acceptable to me. (Hollow arrows show pitch contour; thick filled arrows show strong stress; thin filled arrows show weak stress; upright bars indicate a pause [cf. Beaugrande 2004].)



[22] ¡And !God ¡called the !light !Day, | and the !dark-ness he ¡called !Night. (*Genesis* 1:5)

[22a] ¡And !God ¡called the !light !Day, | and the !dark-ness he ¡called !Night. (*Genesis* 1:5)

[23] !God is ¡a-ble of these !stones to ¡raise up ¡chil-dren un-to ¡A-bra-ham. (*Mathhew* 3:9)

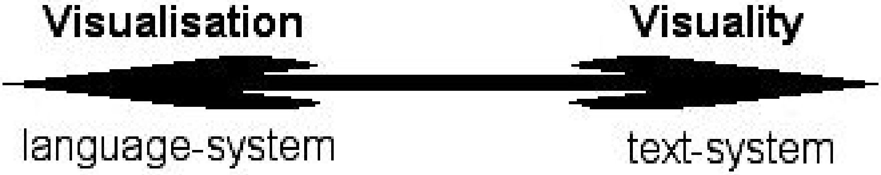
[23a] !God is ¡a-ble of these !stones to ¡raise up ¡chil-dren un-to ¡A-bra-ham. (*Mathhew* 3:9)

Some sequences allowed by the Lexicogrammar seem Prosodically unappealing [24-25] (Slovene student data):

[24] The sea floor is in closer to the shore solid.
[25] On the slope of Cape Roenk, typical sub-Mediterranean species, despite the fact that it has northern position and that the substratum is less flyshy, live.

This is because Slovene tends to complete the Subject-Predicate connection at the end of an Utterance.

The least studied and taught is the process of **Visualisation** that engenders the **Visuality** of the Text-System.



Written language is after all designed to be looked at, from the ornate parchments of the high middle ages and the renaissance to the flashy websites of today. Internet browsers also allow for the easy transmission of photographs that complement and expand the significance of the written text, as in this report from *Der Spiegel* (English edition, June 2006) on the 'gigantic orb' placed slap in front of the famous Brandenburg Gate to commemorate the World Football Cup hosted by Germany:

> [26] The massive sphere glows an eerie blue at night and its gaping maw seems to swallow up people as they march inside. But what looks like a scene from a cheap sci-fi flick is something more nefarious than aliens enslaving mankind.

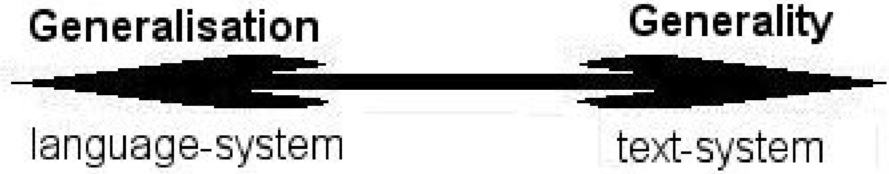
Unfortunately, the organisers went over the top by launching a barrage of fireworks whose smoke enveloped the square within seconds and reduced visibility to nothing, offering a test of what Berlin was probably like when the Red Army came for a visit in 1945.

The very symbols of modernity and unity are discursively metamorphised by this Visuality into an atavistic and sinister ambience with harsh historical overtones.

These, then, are the four systemic processes of Actualisation I hold to be indispensable for the creation of Texts. The fact that they have been so unevenly explored or addressed in projects of standardisation suggests why some significant issues have barely been raised; and why so many non-natives well versed in Grammar and Lexicon somehow still do not sound or write like native speakers.

Moreover, further processes (or meta-processes) are implicated which apply to the organisation and evolution of those expounded so far and which, to my knowledge, are largely missing from both research and pedagogy, because they are generically inimical to the static notions at the centre of their enterprises.

The process of **Generalising** engenders the **Generality** in or among Text-Systems, that is, the extent or reach of regularities such as a large corpus can reveal.



On the whole, once this process takes up some domain of a system, it seems set to run its full course, sometimes altering whole patterns or 'paradigms', as if speakers were subject to its will rather than the reverse. Grammatically, the formation of English Plural with '-(e)s' not merely managed to displace older formations with stem vowel shift or with '-(e)n', but continues to emit spontaneous new Plurals, sometimes more than one, viz.:

[27] Use the criterions of deployment or a readiness exercise […] and sustainment based on functional inspection criterions (*Air Force*) $^{\text{www}}$
[28] Student was weak in two criterias of the exemplary performance for this objective. (*Baylor University*) $^{\text{www}}$
[29] Use of datums in product definition is required in order to specify part features that are used as a basis for functional relationship with other features. (*Candoris*) $^{\text{www}}$
[30] I am having a problem to get all datas into one table. I must get datas from 2 other tables. (*databasejournal*) $^{\text{www}}$

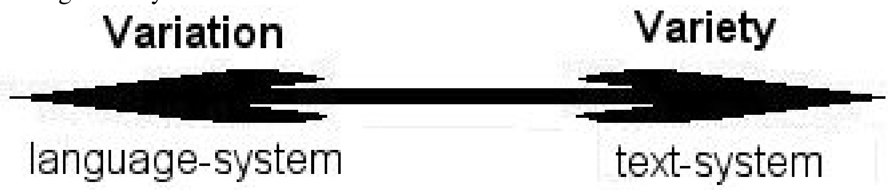
By contrast, the formation of the Past Tense of Verbs with the Ending '-d' or '-ed' or '-t' did not generalise quite so thoroughly and so has left behind some 'non-standard' detritus which rouses the ire of schoolteachers:

[31] Mostly, we jes clumb up on the shed top, inna shade of a tree, and passed the time (*Zeke & the Hoss-Puppy)* $^{\text{www}}$
[32] That iijit [idiot] Frenchman got tryin some fool trick walking a timber stick and got upsot into the wet. (*Man from Glengarry)* $^{\text{www}}$
[33] All of us 'fans' ranned outside and we saw him running to the bathroom! (*Xanga)* $^{\text{www}}$
[34] He was teaching his boy Melvin how to play some baseball so he stolt this baseball bat off the churchhouse softball team. (*Digging Postholes)* $^{\text{www}}$

Teachers fail to appreciate that the alternates they call 'wrong' or 'bad English' are consistent with the system of a regional English and thus resist extirpation. Such holds especially for the universal Negative 'ain't', e.g., for 'isn't', 'haven't', and 'didn't' [35-36].

[35] She <u>ain't</u> exactly my girlfriend, but we spend loads of time together. I <u>ain't</u> asked her if she's my girlfriend (*Billy Bayswater*)
[36] The police shoot them three fellas, but they <u>ain't</u> get Alfred. (*Seeing in the Dark*)

A converse systemic process is **Variation**, which engenders the **Variety** in or among Text-Systems.



It operates especially in domains of instability and complexity of a degree not typical of English in general, such as the subsystem of Pro-Nouns, which alone sustain formal distinctions in 'Case' and 'Gender'. The flock of variations I encountered suggests speakers being aware that forms differ but finding them tiresome to 'use correctly' (cf. Beaugrande 2007). For example, I found Possessive Pro-Nouns occurring in the Subject or Object Forms, e.g.:

[37] Me mum and me dad are separated, like, and me dad reads *The Sun*. (*NME*)

[38] I remember this song 'Shaddup You Face' [by Joe Dolce], spawning the annoying catch-phase as a response to almost anything. (*Fast-Rewind)* ^{www}

[39] Dude, he face is alright, if you just glance at him you can tell right away it's Sheva [soccer star Andriy Shevchenko]. (*Soccer Gaming Forums*)^{www}

[40] In the illustration she hair looks dark. I think they did a perfect job (*mugglenet*) ^{www}

[41] I'm not fond of Carlisle. We took us caravan up that way a few years ago (*BNC data*)

[42] And we wont stop till we have 'em puttin' they feet in they mouths (*Rapsearch)* ^{www}

Reflexives, which are doubly coded for number, turned up a veritable zoo of Variations, e.g.:

[43] I like to think of myselves as a catalyst for innovation (*Ecademy)* ^{www}

[44] new members introduce yourselfs here with a bit of info on yourselfs (*invisionfree)* ^{www}

[45] Stewart found hisself with his back to goal, layed it neatly back to Edwards who ABSOLUTELY SMASHED it into the far top corner of the net (*Birmingham City)*^{www}

[46] she was also surprised to hear it; she had never thought of herselves as strong (*The Valkyrie)*^{www}

[47] As the security forces transform and rid itsselves of the baggages of the apartheid past, they will be able to sufficiently fight crime (Thabo Mbeki in *The Mail and Guardian*)^{www}

[48] We see ourself as the biggest club in Britain, with a stadium to match (Mark Hateley of Queen's Park Rangers in *Today*)^{www}

[49] we want to profile ourselfs as 'The Best Grower of Thomson Seedless Grapes In The Country'. (*NCubeExports)*^{www}

[50] Willie Calder wants to know if anyone recognises theirself in this Class Photo from the Gravesend Sea School, 1957 (*Merchant Navy Memories)*^{www}

[51] They have the third biggest city to theirselfs and are the only team within a 90 mile radius (*Soccer 24/7)*^{www}

Creative Variations in the Lexicogrammar were already briefly mentioned. Threats of punishment for some mistake or neglect can be expressed as whimsically misappropriating some piece of the hearer's anatomy:

[52] By the third day I expect third-years to work alone, and if you slip up, gal, I'll have your guts for garters! (*Hospital Circlers*)

[53] That's a nice bit of double-barrelled lying. Quick. Out with it, or I'll have your skin for a cigar case. (*First of Midnight*)

[54] A secretary whips away the remote. 'Keep that handy', I warn him, 'or I'll have your head for a hat-rack.' (*The Dyke & the Dybbuk*)

[55] Matt put in a warning. 'Just let Bill hear you say you're the hostess and he'll have your ears for horse blinkers.' (*Wilder's Wilderness*)

The 'smash and grab raid' as the most primitive robbery from British shops, as in [56], is varied for occasions that are sometimes more similar, e.g., a police raid' with 'sledgehammers' [57], and sometimes less so, e.g., lively sports [58] or pigging out [59].

[56] The thief [...] smashed a hole in the shop window using a hammer and grabbed about £3,000-worth of gold jewellery before making off on foot. (*East Anglian Daily Times*)

[57] A major heroin dealing ring was believed smashed today after police made a series of raids in Liverpool. [...] 55 officers, some carrying sledgehammers, launched their 'smash and grab' raids on homes in the Everton and Kirkdale areas. (*Liverpool Daily Post*)

[58] At the County Ground, it was daylight robbery; a smash and grab raid by Charlton. They had 3 attacks and scored 2 goals. (*television news*)^{BNC}

[59] At British Petroleum's annual meeting last year, there were protests about a 'smash and grab' raid by one group who scoffed too many sandwiches. (*Daily Telegraph*)

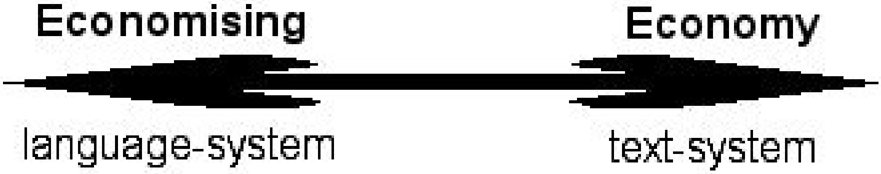
Perhaps the most dynamic Variation of our times is occurring in Prosodification, namely the wholesale spread of so-called *Estuary English* outwards from London, South East England, and the 'estuary' of the river Thames:

The pronunciation of British English is changing quite rapidly. Estuary English may now and for the foreseeable future, be the strongest native influence upon RP [Received Pronunciation]. For large and influential sections of the young, the new model for general imitation may already be 'Estuary English', which may become the RP of the future. (Rosewarne 1984)

Significant numbers of young people see Estuary English as modern, up-front, high on 'street cred', and ideal for image-conscious trendsetters. Others regard it as projecting an approachable, informal, and flexible image. (Coggle 1993)

I can't say if the trend reflects any deliberate defiance against the image of the 'upperclass twit', but it would be far safer than speaking Received Pronunciation in urban centres like Soweto or Kingston Jamaica.

The systemic process of **Economising** engenders the **Economy** in a Text-System where much can be said with quite modest resources.



Among the most economical patterns in the lexicogrammar, which has nonetheless been snubbed by most grammar books, are Non-Clauses lacking Subject or Predicate (Beaugrande 2007), e.g.:

[60] John Major is now being exposed for what some of us always warned that he was. A fake. A flake. A wimp. A phoney. (*Daily Mirror*)
[61] The wedding would come off right enough but the reception would linger on night after night. Yeah. Aye. Singing. Drinking. Oh indeed. (*Oral History Project*)

At most, grammar-books treat them as 'elliptical' versions of full clauses, which for mysterious reasons have not been uttered. (Ignorance? Laziness? Bad manners? Laryngitis?) In order to avoid recognising Non-Clauses, *The Comprehensive Grammar of the English Language* (Quirk, Greenbaum, Leech, and Svartvik 1985: 889f) cobbled together a Rube-Goldberg conversion device with no less than seven mechanisms of "ellipsis" with steadily diminishing certification — "strict", "standard", "situational", "structural", "weak", "virtual", and "quasi-".

Lexically, spoken English manifests a trendy tendency to shorten lexical items down to single syllable [62-67], even if the result yields the same item standing for different sources [65-67].

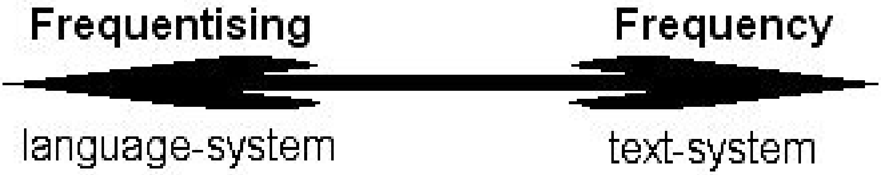
[62] Okay will you excuse me, I'll be back in a mo [moment] (*Shropshire County Council*)$^{\text{BNC}}$
[63] I got kissed lots these hols [holidays], how about that? (*The Prince*)
[64] Poor old Johnnie Ray. They tagged him the Nabob Of Sob, the Prince Of Wails and the Cry Guy. […] The girls thought he was brill. [brilliant] (*New Musical Express*)
[65] It was the bottle that first got Frank into diffs [difficulties]. He goes on binges. (*Diamond Waterfall)*
[66] I used a code compare program to see the diffs [differences] between the two files. (*osCOMMERCE*)$^{\text{WWW}}$
[67] Fitting higher gear diffs [differentials] will improve the top speed of the slow revving engine. (*Know Your Land Rover*)

Phrases too get clipped:

[68] Iris washed her hands and filled the kettle. 'Might as well have a cuppa while you're waiting.' (*Finishing Touch*) [cup of tea]

[69] When we lost 5-0 at Liverpool a couple of weeks ago we all got together and had a heart-to-heart. We sorted a few things out (Andy Thorn of Crystal Palace in *Today*) [heart-to-heart talk]

The process of **Frequentising** (despite the brittle term) merely engenders the **Frequencies** in or among Text-Systems.



Frequency data are problematic if not paradoxical in a recurring sense for data-based research on language and texts: the more straightforward the extraction of data, the less so is its interpretation. Frequent data should signal Generality, but may merely be trivial for items like 'of' and 'in' that are so multi-functional.

Queried for the basic Verb forms for the jolly old 'five senses', the BNC returned 'see' at 115,100 occurrences, 'hear' at 13,079, 'touch' at 2431, 'smell' at 1108, and 'taste' at 672. The disparities are striking, but cannot be uncritically adduced as evidence that sight is many times over the most important sense. Numerous attestations of 'see' are barely related to vision, such as 'understand' [70], 'consider' [71], or 'unmask' [72]. If you 'don't see' something, you can mean it doesn't or won't happen [73]. And so on.

[70] 'You see — 'he began. 'I do see. I have seen for two years. I see why you have come here, penniless.' (*Longest Journey*)

[71] Stanley sees Blanche as a threat to his marriage and his affection for Stella (*School essay*)^{BNC}

[72] For the first time she saw through his mask of precocious intelligence (*Middle Kingdom*)

[73] He's a nice boy, but I don't see him making it to the top. (*City of Gold*)

Some frequencies may appear so general as to be unaccountable. I once made an extensive tally of male and female Pro-Nouns in two corpora installed in concordance programmes and differing quite markedly in sources, text sizes, discourse producers, and dates. My EPC was then at 392 complete 'classic' texts from 253 writers since Shakespeare; the BNC had 4214 partially incomplete contemporary texts, including a laudable contingent of transcribed spontaneous conversations.

Using the subcorpora of 'literary' works in the EPC, back then with 23,936,418 words in 303 texts, I queried the Female and Male Subject Pronouns 'she' and 'he'. I found the proportion of 184,961 for 'she' and 331,595 for 'he' — 55.77% Female to Male. I then ran the same queries on the BNC and found 'she' at 352,872 and 'he' at 640,736 — 56.26%. The close fit was astounding. A bit unnerved, I calculated the totals for *all* Female Pronouns and all Male Pronouns. In the EPC, I got 389,007 Female and 691,387 Male to yield – yes, 55.77%; in the BNC I got 658,965 Female and 1,204,215 Male to yield – yes again, 54.72%.

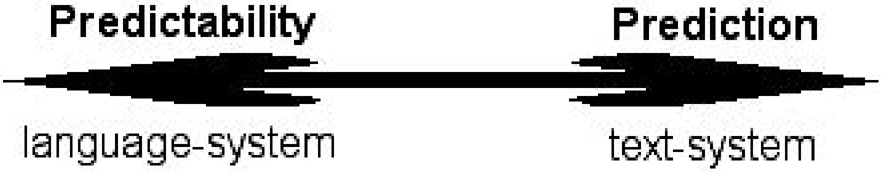
Feeling my workstation fading over into a peak in Darien, I did the same searches on all the comedies of Shakespeare and got 2263 Female and 4090 Male — 55.33%. A strange attractor?

I experimented with various selections to see if the uncanny consistency could be disturbed. In a mini-corpus of three Jane Austen novels, Males were only 76% as frequent as Females. But when I added just three novels by Dickens, two by P.G. Wodehouse, two by Virginia Woolf, plus one each by Fielding, Thackeray, and Oscar Wilde, the figures came out 34,978 Female and 64,133 Male — a ratio of 54.53%.

On my next computation, I calculated the ratio for three female authors programmatically writing about women and their concerns: Margaret Oliphant, Mary Wollstonecraft, and Charlotte Perkins Gilman (just under 300,000 words together). There, the Female Pronouns (5817) were 232% as often as Male (2602). But I only needed to add in the voluminous, mild-mannered Bullfinch and Pepys and I got figures of 8,985 and 16,600, with the Females smack dab at 54%.

I confess myself wholly at a loss to explain such precise generality among frequencies; and the mystery is all the deeper, given the sizes and diversity of the data sets. Any assistance would be highly appreciated.

The systemic process of **Predictability** engenders **Prediction** in or among Text-Systems.



It could plausibly be regarded as the human correlate of *probability*, a factor that information theory dryly proposed to extract directly out of "transition" frequencies, which tear to shreds any conception of 'context'. In discourse, probabilities become active only in so far as they guide predictions.

The Collocation 'I'll see you' is predictably followed by an Adverbial of Time, the most frequent in the BNC being 'later' (79 occurrences), e.g. [74], 'tomorrow' (34), 'again' (17), 'then' (15), in the morning' (15), 'tonight' (7), and 'next week' (7). As such, it can serve as a convenient salutation when parting [74]. Unpredictable continuations can serve as defiance [75] or threat [76].

[74] 'I'll be off, then.' I said: 'Goodbye.' 'Yes', he said. 'I'll see you later'. (*Wasp Factory*)

[75] 'I will trouble you to hand over that purse of gold you had saved to pay for my head.' 'I'll see you hanged first!' raged the Bishop. (*Robin Hood*)

[76] 'Before I let you foul Walter's memory, I'll see you in hell!' she yelled. (*Posthumous Papers*)

If you encounter the Verb 'remanded', you can fairly well predict the next words will be 'in custody' — 273 attestations out of 405 in the BNC, e.g. [77]. 'On bail' lagged behind at 42 attestations, e.g. [78], which says a lot about British justice, whose prisons even Tony Blair admits are 'full to bursting point' (BBC Online). If you read someone 'was remanded to Crumlin Road' [79], you assume (even if you've never been to Belfast) he was packed off to a jail there and not ordered just to loiter on the same street or camp out there 'until June 11'.

[77] Horace Notice, a former British and Commonwealth heavyweight boxing champion, was one of four men remanded in custody last night for a week accused of rioting at an acid house party. (*Independent*)

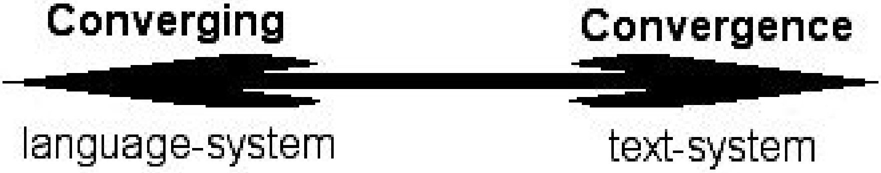
[78] Three workers from McDonalds restaurant appeared in court yesterday charged with making a hoax bomb call to rivals Burger King. They were remanded on bail until March 11. (*Northern Echo*)

[79] Joseph Walsh Wilkinson was remanded to Crumlin Road until June 11. (*Belfast Telegraph*)

By contrast, 'commanded' is less predictable. If the Definite Article is included in the query, I still find no specific next words with special frequency, but there is a definite pattern of military Collocates: 'army, troops, force, militia, regiment, battalion, corps, division, squadron, cavalry'.

'Commanded by' correlates with a predictable metalworks of military brass: 'General, Admiral, Marshal, Colonel, Major, Lieutenant, Captain, Constable' along with their Byzantine combinations (like 'Lieutenant-General'), plus a gallery of titled nobles and notables and, in exactly one case, a 'king' (but only that frilly showoff Charles I).

Finally, the systemic process of **Converging** engenders **Convergence** in or among Text-Systems and concerns the process whereby the diversity of uses and meanings of individual words and expressions come together in a sharable use and meaning for a context.



The impact of Convergence is perhaps most tangible when it effortlessly constructs meanings that are by no means the summing up of 'literal' meanings, viz.:

[80] Pancakes to sell for grave flags (*University Herald*)

[81] Internal memos on tampon introduced (*Washington Post*)

[82] Insecticide sprayed on judge's oral ruling (*Spokane Chronicle*)

[83] Congress votes for running trains over union workers *([Lafayette] Courier and Journal*)

In exchange, unduly effortless Convergence can seem trivial and patronising:

[84] War dims hope for peace (*Columbus Dispatch*)
[85] Cold Wave Linked to Temperatures (*Daily Sun*)
[86] Study says dead patients usually not saved (*Miami Herald*)
[87] Plane Too Close To Ground, Crash Probe Told (*San Antonio Express-News*)

How Convergence can in fact be achieved is only gradually being explained, and the best experimental evidence indicates that it is not done in a way common sense would suggest (Kintsch 1988, 1989).

3. The dynamics of intersystemic processes in Actualisation

In sum, I would propose a model of four properly textual processes, plus six more multi-purpose processes that can also apply, say, to the perception of visual scenes or the audition of symphonic music. I believe that at least these processes are indispensable for the Actualisation that mediates between language-system and text-system. Being dynamic by nature, they tend to evolve, despite the traction of institutional standardisation.

If Walter Kintsch's 'construction-integration model' (cited above) has 'psychological reality', the process of actualisation during reading normally runs between 5 and 500 milliseconds, and such speed leaves us largely dependent on indirect evidence. Whereas Walter's lab work sifts traces in super-fast operations of perception, memory, and response, I have here undertaken to sift traces in longer-term corpus evidence which might plausibly evince an account for certain classes of phenomena that stand out from the ordinary.

Back in the 1980s, Walter ruefully quipped he had arrived at the conclusion that 'reading is too difficult to be done'; we can just cling to the reality that it *is* done, and mostly well enough too. I in turn can adduce the reality that Actualisation is also done; the difficulty lies not in the complex of processes, but in the patchwork of piecemeal models that often mix determinable facts with wishful thinking and thin air.

If, as I have suggested, significant and essential processes have gone mainly unexplored because they don't figure well in such models, then we may be seeing SFL and corpus research nearing a space of convergence for questions which can only be properly tackled with very large sets of authentic data.

And so I come back to the question posed in my title. I am confident both language and text are not merely systemic, but are mutually designed to sustain systemic actualisation from processes to products. As the corpora continue to grow, our insights will be deepened and broadened in both directions: toward language and toward text.

If we once imagined a 'dream' of 'lexis as most delicate grammar' (Hasan 1987), we might now imagine a 'dream' of 'text as most delicate corpus'. (Even dreams inspire.) Actualisation must be intersystemic; a corpus must be intertextual; and I submit that our most auspicious pathways for exploring these vastly rolling wordscapes will be with parallel expeditions in lively contact.

References

Beaugrande, R. de 1991. *Linguistic Theory: The Discourse of Fundamental Works*. London: Longman.

Beaugrande, R. de 1997. *New Foundations for a Science of Text and Discourse*. Greenwich, CT: Ablex.

Beaugrande, R. de 1998. Performative speech acts in linguistic theory: The rationality of Noam Chomsky. *Journal of Pragmatics* 29, 765-803.

Beaugrande, R. de July 2004. *A New Introduction to the Study of Text and Discourse*. Published on the Internet.

Beaugrande, R. de January 2007. *A Friendly Grammar of English*. Published on the Internet.

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

Coggle, P. 1993. *Do you speak Estuary? The New Standard English*. London: Bloomsbury Publishing.

Halliday, M.A.K. 2004. Language as system and language as instance: The corpus as a theoretical construct. In J. Svartvik (ed.), *Directions in Corpus Linguistics* Berlin: Mouton, 1992, 61-77. Also in *Collected Works of M. A. K. Halliday*, (ed. J. J. Webster), vol. 6: *Computational and Quantitative Studies*. New York: Continuum International Publishing.

Hasan, R. 1987. The grammarian's dream: Lexis as most delicate grammar. In M. Halliday and R. Fawcett (eds.), *New Developments in Systemic Linguistics*. London: Pinter, 184-211.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A 'construction-integration model'. *Psychological Review* 95/2, 163-82.

Kintsch, W. 1989. The representation of knowledge and the use of knowledge in discourse comprehension. In R. Dietrich & C. Graumann (eds.), *Language Processing in Social Context*. Amsterdam: North Holland, 185-209.

Martin, J. R. 1992. *English text: systems and structure.* Amsterdam: Benjamins.

Rosewarne, D. 1984. Estuary English. *Times Educational Supplement*, 19, October 1984.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik, 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Weinberg, G. M. 2001. *An Introduction to General Systems Thinking*. New York: Dorset House.

Some notes on the concept of cognitive linguistics

Wolfgang Teubert

University of Brimingham

Abstract

In this contribution, I argue that the cognitive sciences are troubled by some internal contradictions that seem to me difficult to resolve. Cognitive linguistics is the part of the cognitive sciences dealing with language, and the philosophy of mind provides its theoretical underpinnings. Its goal is to describe the language system as a mechanism that processes thoughts into utterances and utterances into thoughts. While thoughts involve intentionality, the processing mechanism is thought to operate without our awareness. But what do we actually know about this mechanism? Is there really a language of thought, and how innate and how universal would it be? What do we know about the mind as the locus where cognition is processed? How dependable is the computational model of the mind? Do the various factions of cognitive linguistics offer scientific evidence or just possible models of how the mind, if there is one, might work? In the end, cognitive linguistics cannot account for meaning. We do not have access to our own or anyone else's mental concepts. Meaning and knowledge, on the other hand, is public; it is what is exchanged, negotiated, and shared in the discourse. Whatever cognitive linguists may be able to find out about our mental representations, to the extent that it is effable it can never be more than a duplication of what we find in the discourse.

1. What is meaning?

Many linguists have such a respect for meaning that they are careful to avoid the issue wherever possible. Traditionally, language study has a strong focus on grammar. Grammar is, if one keeps sorting the elements that make up language long enough, a land of apparent law and order, in which every part and parcel finds, in the end, its pigeonhole. The meanings of words, however, behave disorderly. Words are ambiguous and fuzzy. This is why it always has seemed prudent to leave them to the poor cousins of linguists, to the lexicographers. However, whenever the makers of our dictionaries try to make sense of them it is the linguists who habitually criticise them for all the inconsistencies abounding in even the best dictionaries.

For two schools of linguistics, this is picture is true no longer. Corpus linguistics and cognitive linguistics share the fascination for meaning, though not for much else. Yet they look for meaning in different places. For corpus linguists, the meaning of a lexical item can only be studied in real language data, in the texts in which they occur, in the contexts in which they are embedded. Here miraculously all ambiguity and fuzziness seems to fade. If we read a text we rarely have the problem of not knowing what the words are supposed to mean. Only if we look at these strings of alphabetic characters, with a space in front of them and behind them, in isolation, we start to wonder whether *bank* means 'river edge' or 'financial institution'.

Michael Stubbs' publications have made corpus linguistics popular almost all over the world. *Tout le monde* now uses corpus evidence. By itself, however, working with corpus data does not make one a corpus linguist. More and more linguists, including cognitive linguists, may underpin their investigations with examples discovered in corpora. But corpus linguistics is more. As Mike Stubbs has demonstrated time and again, most convincingly, I believe, in his magisterial *Words and Phrases* (2001), corpus linguistics opens a new perspective on language. It replaces the traditional practice of analysing and categorising linguistic phenomena in the sterile conditions of a post-mortem autopsy, by interpreting these phenomena *in vivo*, in their contexts and with their implicit and explicit links to other discourse events. For this task corpus linguists use corpora, principled samples of the discourse, and computers. They correlate the statistical analysis of lexical correspondences over the corpus to their semantic relevance. Corpus linguistics understands language in terms of open choice and co-selection, as John Sinclair has pointed out repeatedly. The discourse, this entirety of texts that have been and are constantly being exchanged between the members of a discourse community, is a network of intertextual references. Whenever something is said, it is said as a reaction to what has been said in other, previous texts. Only rarely we say something new. Instead, we re-use the phrases and expressions that we find. Now and then we may add our own touch. What a phrase, an expression means is how it has been used and paraphrased in its previous history. Meaning, therefore, is in the discourse. Language has to be viewed as a social phenomenon. That these ideas are now increasingly accepted is largely to the credit of Mike Stubbs. He has been untiring in his efforts to develop a consistent theoretical framework that will let us make sense of the findings the methodology of corpus research is supplying, always sticking to his characteristic lucid, straightforward and unpretentious style that is so typical of this great scholar.

For cognitive linguists, meaning is not in the discourse; rather, it is in people's heads. They view traditional linguists, the philologists for instance, as sitting in a dark cave looking in one direction only, toward the back wall of the cave. Behind them is an open fire providing light, and between the fire and where they are sitting there is a catwalk on which the mental concepts move, casting lexical shadows on the wall. Shadows are all the traditional linguists see. As long as they stay in this position they take the shadowy words for the stuff meaning consists of. If they only turned around, they would be confronted with something more real, with mental concepts or cognitive representations. Regardless whether these things are metaphysical Platonic forms, i.e. what Plato calls *eidos* or *idea*; or whether they are only models of the 'real' things, accepting them in lieu of words would be a step in the right direction, a step taken by cognitive linguists. Once it is taken, they can start working on the last obstacle to truth, the division between brain sciences and mind sciences. Eventually, cognitive linguistics promises, we will grasp what is really out there in the uncharted territories in our heads.

In this contribution, I want to investigate the key element of cognitive linguistics, the mental/cognitive concept/representation as the embodiment of meaning. To the extent there is a common philosophical and theoretical basis of what is now called cognitive linguistics, we find it in the cognitive sciences, as they developed after the demise of behaviourism, and in the philosophy of mind. Many of my arguments refer to these foundations rather than to specific contemporary schools of cognitive linguistics, of which there are many. From my outsider's perspective, there seems to be a tendency in several of these schools towards modelling the linguistic faculty of the mind rather than to demonstrate the reality of this faculty. Because cognitive linguistics is far from being a uniform discipline, it will always be possible for one school or another to maintain that my charges do not apply to them. I am more concerned with the basic ideas driving cognitive linguistics than with individual schools, namely that we have to look for meaning in people's heads. My goal is to show that this is a hopeless enterprise.

Some cognitive linguists draw the main dividing line between one-level semantics and two-level semantics. Those who subscribe to one-level semantics, for instance Jerry Fodor and Ray Jackendoff, deny that there is a systematic difference between the meanings of words (the meaning of the expression) and the respective mental representations. Two-level semantics, on the other hand, holds that mental representations are richer and more specific, and certainly not isomorphic with the meaning expressed in natural language utterances. Stephen Levinson seems to make a distinction between two-level semanticians like himself, on the one hand, and 'Cognitive Linguists' on the other hand, and he applies this label apparently only to one-level semanticians. (Levinson 1996, 24) There is a multitude of views as to the nature of the mental concept, and there are different terminologies. Looking at these texts from the perspective of corpus linguistics, I find myself unable to distinguish between concepts and representations, and between mental and cognitive. Mental/cognitive concepts and/or representations all seem to refer to more or less the same idea.

Still much what I present here has been put forward in irreconcilably different accounts. Dan Sperber and Deirdre Wilson identify three views as to the number of concepts: There could be less than there are words (the compositional view, argued by, among others, Anna Wierzbicka and also Stephen Levinson), there could roughly the same number (a view to which Jerry Fodor and Stephen Pinker seem to subscribe), and there could be infinitely more concepts than words (as Sperber and Wilson themselves believe). (Sperber/Wilson 1998, 186f.) The different camps of cognitive linguistics are not co-extensive with the academic disciplines of the scholars. We find linguists, cognitive scientists, computer scientists, neurologists, philosophers of mind and analytic philosophers.

My main claim is that we know too little about what is going on in our heads to make strong claims about the connection between thinking and language. Thinking is a mental activity that takes place in the brain. Thinking, for me at

least, involves consciousness, and consciousness involves intentionality. Both features are highly enigmatic. It is doubtful if they can be related to brain functions. Then there is the enigmatic nature of thoughts. We do not know the content of a thought unless it is expressed in language. Is it possible at all to distinguish between a thought and its expression? Cognitive linguists, like all linguists, have to start with linguistic expressions. But they contend (and who could possible hold against it?) that any linguistic expression must be caused by a thought, and that its intended destination, in the hearer's head, is again a thought. Can we learn to travel that route? Can we put our fingers on a thought just before it becomes the expression of a thought? What do we know about the thought caused by an expression? How can we trace a thought? How can we describe the content of a thought without using language? If there is more to a thought than what can be verbalised in the expression of it, how can we access this extralinguistic content? To me it seems there is no easy way to resolve these matters. The mental/conceptual concepts/representations I find in cognitive accounts are, at their very best, a duplication (and sometimes even a triplication) of what we can find out about the meaning of a lexical item, about the meaning of what has been said. Indeed, what more could we hope to find in people's heads?
Cognitivists are "*interested ultimately*", in the words of Ray Jackendoff, "*in the manner in which language is embodied in the human brain*." They are not interested in language as "an abstract phenomenon or a social artifact". There "might be properties that language has because of the social context into which it is embedded". But there are also "some important properties ... that can be effectively studied without taking account of social factors". Is this "mental stance" a stance concerning the brain or the mind? While Jackendoff calls language a "mental organ", his ultimate goal seems to locate it in the brain. "Language, vision, proprioception, and motor control" all are "instantiated" by neurons of basically similar design. Thus, in the long run, we should study language as a "physical organ". (Jackendoff 1997, 2ff.; Jackendoff's emphasis) Should we assume, then, that mental concepts or cognitive representations have a physical reality? And will our quest for meaning have come to a conclusion once we have identified the neurons and their nature that correlate with a linguistic expression?

We use language firstly to interact with people and secondly to give our linguistic and non-linguistic interactions a meaning. Our societies are much more complex than those of our closest relatives, the non-human primates. This complexity requires not only a division of labour but also distributed knowledge. We must negotiate who carries out which task, and we must provide the necessary knowledge. Language enables us to trade content. For corpus linguists, language is, first of all, a social phenomenon. Language is public. Language is observable. Outside of language there is no symbolic content that could possibly be conveyed. Whatever we think, we have to express it in language so that others can share it. How we turn our thoughts into language, how we understand what other people say is a matter of speculation. But what something that has been said means is not a matter of speculation but of negotiation. When someone has said something, we, whether we are the addressees of this statement or whether we learnt about it some other way, can discuss its meaning. We may not necessarily agree. It can be a discussion without a conclusion. But beyond our various interpretations of this statement, concordant or discordant as they may be, there is nothing to be found out about its meaning. To know which neural processes and which hormonal outpourings led to the statement, and which neural processes and hormonal outpourings it triggered in the people who were addressed does not lead us to understand the statement any better. Meaning is, just as language, public. It belongs to the sphere of social interaction, not to the realm of mental processes. This is the stance of corpus linguistics, as I see it. Is there any reason why we should take up the cognitive approach to find out about meaning?

Gisela Harras, who prefers two-level semantics distinguishing between 'semantic content' and cognitive or mental concepts, gives an example to tell us why language does not tell the full story. She compares the two sentences 'Open the bottle' und 'Open the washing machine' and concludes that a verb such as *open* can convey a sheer unlimited amount of concepts. The semantics of the verb, she maintains, does not tell us what *open* means when applied to different things. The bottle may have a screw top or a cork. The plumber opens the washing machine at a different spot from the normal user. When we, as the addressees, have to interpret the different utterances containing the verb *open*, we have to go beyond its semantic content, Harras says. The hearer has to apply cognitive mechanisms that will tell her or him which of the many concepts expressed by the verb *open* the speaker has intended. For Dan Sperber and Deirdre Wilson, to whom Harras is indebted for her example, these concepts seem to be individual rather than universal: "A concept, as we understand the term, is an enduring elementary mental structure...It is arguable that each of us has ineffable concepts [e.g. of a special kind of pain] – perhaps a great many of them… For the time being, we will restrict ourselves to effable concepts: concepts that can be part of the content of communicable thought…[T]here are a great many stable and effable mental concepts that do not map onto words." (Sperber/Wilson 1998, 189) How do we have to understand these strange mental concepts which are effable, but do not map onto words?

Let us assume the bottle referred to in the utterance 'Open the bottle!' is a standard bottle of wine of 1998. Then the cognitive representation of the verb *open* would be the concept of 'uncorking'. If we accept this, the question must be allowed how this mental concept differs systematically from the natural language expression. The concept 'uncork' is not only effable, it happens to map miraculously onto the word *uncork*. But how do I move on from the rather general verb *open* to the much more specific concept 'uncork'? Do I have to rely on cognitive mechanisms that make me understand the speaker intended 'uncork' when she or he said *open*? Corpus linguists can easily do without such constructs. They would find citations in their corpus which paraphrase the meaning of *open* in the case of wine bottles, i.e. which tell us what open means here:

> How to *open* a bottle: It's actually pretty simple to *open* a bottle of wine. These steps are for a double-action, or wing, corkscrew, which has two arms (or wings) that help lever the cork out of the bottle. (http://www.ehow.com/how_1715_open-wine-bottle.html)

Meaning is in the discourse. All we have learnt about the meanings of lexical items we have learnt from other people's contributions to the discourse. This is how we learn how to use words when we grow up. Our carers or, a bit later, our peers, explain what they mean. This is how new lexical items are introduced: The text introducing them has to explain them, has to paraphrase them. If I have not been told what *opening a bottle* means, no cognitive mechanism will help me to understand the speaker's intentions. However, if I know the meaning of *open a bottle of wine*, I do not need such a mechanism.

Why do we not readily accept such a simple solution? The reason, I think, is that we attribute too much importance to the meaning of a single word in isolation. We grow up in the belief that words are the core elements of language, and that their meanings are registered in dictionaries. Indeed, there we are confronted with the apparently incontrovertible evidence that words, frequent words in particular words like *open*, are fuzzy, polysemous or ambiguous. A central part of the endeavours of cognitive linguistics are directed at disambiguation, at the resolution of ambiguity. For in contrast to words, our thoughts, we feel, are unambiguous. But is language really more ambiguous than thinking? When we are confronted with a text of common length, do we feel it is ambiguous? Once we give up the belief that words are the core elements, the ambiguity starts disappearing. All we have to do is to replace the linguistic construct 'word' by the linguistic construct 'unit of meaning', defined as a node word plus all the words in its immediate context that make it unambiguous. Instead of wondering about the polysemy of *open*, we now have to deal with a unit of meaning, a lexical item *open* a bottle of wine,* which is fairly monosemous. If our speaker tells us 'Last night I opened a bottle of wine.', we still cannot be absolutely sure that she uncorked it. She might have unscrewed the top, if she is the kind of person to have wine with a screw top. Unless she tells us, we will never know. No cognitive mechanism will provide that information. Once we throw the notion of the word as unit of meaning overboard, ambiguity recedes. Forty years ago, when corpus research was taking off, John Sinclair argued for the principle of collocation which gives us complex lexical items larger than the single word but with only one meaning. (Krishnamurthy 2004, 10) Corpus linguistics has nothing to tell us about how thoughts are turned into linguistic expressions, and how linguistic expressions are turned into non-linguistic thoughts. But corpus linguistics can deal with meaning to the extent that meaning is public and negotiable. How this is done I have sketched in "My Version of Corpus Linguistics" (Teubert 2005).

In this contribution, I will have a closer look at two essential aspects of cognitive linguistics. The first aspect deals with the model of the mind that we find in the cognitive sciences. The second aspect concerns the innateness and universality of the mental lexicon. I will conclude by asking if there is any common ground on which corpus linguists and cognitive linguists can work together.

2. The model of the mind in the cognitive sciences

There is a rather disquieting aspect of cognitive linguistics. It concerns the foundation of the cognitive sciences in general. In the fifties of the last century, cognitivism has replaced behaviourism with the promise to keep up the claim to scientificity established by behaviourism while at the same time pledging to demonstrate and explain directly the working of the mind, without having to resort to ambiguous stimulus-response situations. But then it is not so easy to look into people's heads. A model was needed to spell out how the mind is working.

The computer had just been invented, and its prospects seemed boundless. In the long run, people believed, it was only a matter of size to compete with human reasoning. More memory, more operations per second, more complex programs, and computers would emulate if not surpass human thinking. As they were believed to be, in principle, functionally equal to the human mind, they were seen as the perfect model of cognition, offering the additional advantage of blocking out the erratic impact of emotions. That was the hour of birth of the computational theory of the mind, a theory quickly becoming a doctrine stating that the working of the mind can be understood and described in analogy to the working of the computer. Ray Jackendoff was influential in relating this model to the study of language, for example in his book of 1987 *Consciousness and the Computational Mind*.

This model appealed to those within the human sciences who wanted to look their colleagues in the natural sciences straight into their eyes. The cognitive sciences enabled them to locate the study of the psyche, the human mind, within the sciences. What had been an object of mostly philosophical investigation, with all the arbitrariness of interpretation, could now be based on solid fact. This was a welcome opportunity for linguistics to bid farewell to the increasingly embarrassing lodgings within the *Geisteswissenschaften* and to jump on the bandwagon of scientific and technological modernity. It was there, and not in the humanities, where the big research funds were on offer.

The promise was that thanks to the computational model the mystery of the mind could now be finally solved. Further rewarding consequences soon began to show. Thanks to the new paradigm, other disciplines began to take the cognitive sciences seriously. Therefore it came as no big surprise that when computer scientists set out to invent artificial intelligence, they now, in turn looked to the cognitive sciences as their inspiration. The cognitivists naturally were happy to sell back the blueprints they had copied a few years earlier from the computer engineers, together with some new annotations, to the emerging artificial intelligence community. The new goal was to transfer the now established view of how humans perform mental operations and solve problems of all kinds to computer scientists eager to develop 'intelligent' machines. This cross-fertilisation continues to the present day, as does the ignorance about the circularity involved. So when the computer scientists developed the concepts of parallel processing and connectionism, these new ideas were immediately taken over by the cognitivists. Thus Pierro Scaruffi tells us: "[A] connectionist structure such as our brain works in a non-sequential way: many "nodes" of the network can be triggered at the same time by another node. The result of the computation is a product of the parallel processing of many streams of information." (Scaruffi 2003)

The new developments were greeted by many active in the philosophy of mind. They seemed to vindicate the computational theory of mind. "Connectionism", says Daniel Dennett, "is a fairly recent development in A[rtificial] I[ntelligence] that promises to move cognitive modelling closer to neural modelling, since the elements that are *its* bricks are nodes in parallel networks that are connected up in ways that look *rather* like neural networks in brains." (Dennett 1993, 269, Dennett's emphasis)

The belief that parallel processing and neural networks are all that it takes to make our computers truly intelligent was permeating the whole western(ised) society. This was when everybody was talking about a new generation of computers believed to be able to emulate human intelligence. The new computers could learn (program themselves) how to do things instead of just carrying out the programmer's instructions. They could be trained to reason based on common sense, thus enabling them to pass reliable judgment on matters too complex to be thought through by humans. These new 'virtual machines' would employ an 'architecture' of processors operating in a parallel, interactive way. The relationship between the initial and the resulting states of such a machine would not be determined by pre-programmed commands, but would develop themselves on the bases of huge amounts of training data. This is the idea of connectionism. Once the computer has 'learned' which initial stages lead to which resulting states, this knowledge can be applied to any new initial state displaying the same properties, and the computer will deliver the correct final state.

Eric Pederson and Jan Nuyts, the editors of *Language and Conceptualization*, would not disagree:

> Thus, while 'classical' cognitive theories would consider representations to be virtual 'objects' of some type, manipulated by a 'machinery' of procedures or rules which are somehow implemented in the human brain, connectionist and parallel distributed processing theories consider representations to be simply the resultant characteristics of peculiar states of the 'conceptual system' distributed across the neural networks of the brain… In the latter view, if the notions of knowledge and representations are to be used at all, any characterizations of them beyond the vague ones given above are no longer acceptable as descriptions of actual cognitive mechanisms creating human behaviour. (Nuyts/Pederson 1997, 1f.)

So can we be sure that the brain/mind is isomorphic with a (connectionist) computer? Perhaps not quite, Ray Jackendoff thinks, but that does not do harm to this model. He accepts that "eventually the neuronal basis of mental functioning is a necessary part of a complete theory". But neuroscience today is "far from being able to tackle the question of how mental grammar is neurally instantiated". Therefore "the formal/computational approach is among the best tools we have for understanding the brain at the level of functioning relevant to language, and over the years it has proven a pragmatically useful perspective". (Jackendoff 1997, 9)

For the John Searle of the late eighties it became exactly this equation of computers and human minds which he sees the foundational error of the cognitive sciences:

> If one looks at the books and articles supporting Cognitivism one finds certain common assumptions, often unstated, but nonetheless pervasive.
> First, it is often assumed that the only alternative to the view that the brain is a digital computer is some form of dualism. The idea is that unless you believe in the existence of immortal Cartesian souls, you must believe that the brain is a computer. Indeed, it often seems to be assumed that the question whether the brain is a physical mechanism determining our mental states and whether the brain is a digital computer are the same question. Rhetorically speaking, the idea is to bully the reader into thinking that unless he accepts the idea that the brain is some kind of computer, he is committed to some weird antiscientific views. Recently the field has opened up a bit to allow that the brain might not be an old fashioned von Neumann style digital computer, but rather a more sophisticated kind of parallel processing computational equipment. Still, to deny that the brain is computational is to risk losing your membership in the scientific community. (Searle 1990)

Yet does it really work, this model used both by the cognitive and the artificial intelligence community? Some misgivings are allowed. In spite of billions of dollars invested in artificial intelligence and machine translation systems, results are far from satisfactory. For forty years we have been hearing that success is 'just around the corner'. We are still not there. The epochal endeavour to teach computers 'common sense', the megalomaniac CYC project of a comprehensive ontology of knowledge, never came close to the originally envisaged results (www.opencyc.org/), nor did EUROTRA, the European project for automatic translations from and into all the EU languages, which cost the taxpayers hundreds of millions of euros (http://www-sk.let.uu.nl/stt/eurotra.html). Could it be that the whole approach was faulty?

3. Are concepts inherited and universal?

The second question: Are the mental concepts of cognitive linguistics universal as it has been claimed? For Jerry Fodor there is no doubt that the 'language of thought' (what comes to be called mentalese by Stephen Pinker) is really universal, and that it is in this language of thought, and not in natural languages, that meaning resides: "English *has* no semantics. Learning English isn't learning a theory about what its sentences mean, it's learning how to associate its sentences with the corresponding thoughts." (Fodor 1998, 9; Fodor's emphasis) How literal should we take this claim? In Stephen Pinker's *The Language Instinct* we read:

> People do not think in English or Chinese or Apache; they think in a language of thought. This language of thought probably looks a bit like all these languages; presumably it has symbols for concepts, and arrangements of symbols. … [C]ompared with any given language, mentalese must be richer in some ways and simpler in others. It must be richer, for example, in that several concepts must correspond to a given English word like stool or stud. … On the other hand, mentalese must be simpler than spoken languages; conversation-specific words and constructions (like a and the) are absent, and information about pronouncing words, or even ordering them, is unnecessary. (Pinker 94, pp 81-2)

This quote shows nicely how cognitive linguistics multiplies entities without need, and thus violates Ockham's razor. Natural languages have words (*stool* or *stud*), mentalese has symbols representing the various meanings of *stool* or *stud*, and these symbols stand for concepts. Does that mean that first we have to translate a natural language sentence into mentalese, and then link this mentalese sentence to its cognitive representation, and then we have understood what the sentence means? This would amount to a triplication of our semantic apparatus. More interesting in our current context is that these mentalese symbols correspond *grosso modo* to the meanings of natural language words. Indeed there has not been a lack of endeavours over the last decade to build multilingual conceptual 'ontologies' on the basis of this hypothesis. There has been, for instance, the large project financed by the European Commission and the member states involved with the title EuroWordNet (www.illc.uva.nl/EuroWordNet/) which uses a slimmed-down version of the Princeton WordNet, the largest American English language online dictionary/database (www.wordnet.princeton.edu/). It relates the senses (formerly called 'synsets') given in WordNet to senses of words in a large number of European languages. There is further offspring in form of 'localisations' of WordNet for some of these and some other languages, such as GermaNet (www.sfs.nphil.uni-tuebingen.de/lsd/Intro.html). What is called an ontology is, of course, in reality nothing more than a taxonomy. It is not a reality-inherent classification of whatever we find in reality. Neither is it a language-inherent classification of all the senses (or concepts) that we find in language. Rather it is a classification of English words and their senses as they have been found fit for lexicographic purposes. There you look at single words in isolation and not at words embedded in a text. However, numerous attempts to have people or computers assign the 'proper' sense as found in a given dictionary to a word embedded in the text have yielded rather disappointing results. How else could it be? The entry for the noun *fire* has eight senses in WordNet; in the New Oxford Dictionary of English (1998) it has only two senses, and in the Collins Dictionary (fourth edition, 1998) it has 13 senses (excluding idioms). Assigning senses to words in isolation is, to a large extent, arbitrary, once we leave aside obvious cases such as the two meanings of the commonly invoked example *bank*, a homonym conflating two etymons. There is little evidence that the word senses we find listed in dictionaries relate to some natural language 'reality', and there is even less evidence that we find these senses, as they are listed in dictionaries of one language, in dictionaries of a different language.

For once we proceed from a monolingual to a multilingual perspective, we are bound to recognise quickly that as long as we look at words in isolation they hardly seem to map onto each other from one language to the next. Whatever there might be, in terms of mental concepts in our heads, there is hardly any evidence to call it genetically inherited, innate or universal. Take the German words *Kummer*, *Trauer* and *Gram*. Bilingual dictionaries tell us that their English equivalents are *grief*, *sorrow* und *mourning*. There is no one-to-one relationship, however. The English *sorrow* translates into *Kummer*, when a young girl is left by her lover, it translates into *Gram*, if an old man cannot accept his fate, and it translated into *Trauer* if someone post puberty suffers the loss of someone dear and seemingly irreplaceable. German-English dictionaries invariably offer *sorrow* and *grief* as the equivalents of *Trauer*, but never provide a distinction between the two equivalents. But there must be a distinction. In eight out of ten sentences featuring either *grief* or *sorrow*, native speakers insist that one cannot replace the other. Again we are confronted with the fact that there is no way to deal with the meaning of single words in isolation. It is the context, the situation and an infinity of peripheral conditions which have to me matched. *Kummer* in Thomas Mann's novels is different from *Kummer* as we find it in tabloids. Words out of context, in isolation, are, to a large extent, empty, waiting to acquire a specific meaning from the wider and the narrow context, in particular from the collocates they co-occur with. How should we imagine the mental concepts corresponding to single words? Would they not have to be similarly indeterminate? Or do they have a specified meaning, regardless of the context in which they occur, just as H2O is always the same substance wherever it occurs? Are we expected find the same mental concepts in the heads of all people, regardless which language they speak? How do they get there? Are we born with them?

For Noam Chomsky, it was a well-advised decision to exclude, over decades, the semantic component from his investigation into the workings of the language faculty. To him we are indebted for the apophthegm that a visiting scientist from Mars would conclude that, aside from their mutually incomprehensible vocabularies, all earthlings speak the same language. (Chomsky 2000, 118) That the different vocabularies keep me from understanding people speaking another language is, for the Chomsky of the *Aspects of a Theory of Syntax*, but a surface phenomenon. Just as all languages share the same syntactic deep structure, Chomsky the mentalist seems to believe, they also share a common pool of mental concepts. Whenever he discusses the question of the universality of mental concepts, he invokes the exclusive domain of the innate language organ, with far-reaching consequences. In his article "Language and Interpretation" (first published in 1992) he takes up, in modified form, Jerry Fodor's claim, that concepts are holistic and cannot be decomposed into more basic primeval, concepts, and he agrees with Fodor that all concepts are somehow already present in the human language faculty. "There is, it seems rather clear, a rich conceptual structure determined by the initial state of the language faculty (perhaps drawing from the resources of other genetically determined faculties of mind), waiting to be awakened by experience." (Chomsky 2000 [1992], 64) In another contribution "Language as a Natural Object", reprinted in the same volume, he explains why this must be the case:

> The linkage of concept and sound can be acquired [by children] on minimal evidence… However, the possible sounds are narrowly constrained, and the concepts may be virtually fixed. It is hard to imagine otherwise, given the rate of lexical acquisition, which is about a word an hour from ages two to eight, with lexical items typically acquired on a single exposure, in highly ambiguous circumstances, but understood in delicate and extraordinary complexity that goes vastly beyond what is recorded in the most comprehensive dictionary, which, like the most comprehensive grammar, merely give hints that suffice for people who basically know the answers, largely innately. (Chomsky 2000 [1994], 120)

This is, to say the least, highly speculative. If children around eight years of age really had a fully working vocabulary of 26280 words (12 words x 365 days x 6 years) we should wonder why they do not put it to better use. Be this as it may, though, the underlying question is what it means when Chomsky says that children typically acquire a lexical item in a single exposure which could be only explained by the fact that they "basically know the answers, largely innate." What does the sentence "the concepts may be virtually fixed" mean? On the same page, we also find "There is reason to believe that the computational system [of the mind] is invariant, virtually." Is it just one of those typical Chomsky-sentences whose main purpose seems to add rhetorical fervour to his argumentation but which should not be taken too seriously in terms of their content, such as this sentence, also on page 120: "But there is evidence that the languages [English, German, Latin, Greek, Sanskrit and Chinese] have basically the same inflectional systems, differing only in the way formal elements are accessed by the part of the computational procedure that provides instructions to articulatory and perceptual organs." While Chomsky may not have been aware that the first five languages he mentions belong to the same family, he certainly knew that there are no data confirming what he calls evidence in the case of Chinese. However, his employment of hedging adverbs such as *basically* and *virtually* make every argument invulnerable. "virtually all walls," we could say, "can be painted in black or in white; black and white are basically the same thing."

Unsurprisingly, Hilary Putnam strongly disagrees when it comes to the question of mental concepts: "Contrary to a doctrine that has been with us since the seventeenth century, meanings just aren't in the head." (Putnam 1981, 19; Putnam's emphasis) This is what Hilary Putnam has been saying consistently over many decades. If meanings are not in the head, then the idea of innate mental concepts does not make sense: "A Chomskyan theory of the semantic level will say that there are 'semantic representations' in the mind-brain; that these are innate and universal; and that all our concepts are decomposable into such semantic representations. This is the theory I hope to destroy." (Putnam 1998, 5)

This is how Putnam describes this theory:

> Mentalists who follow Fodor's lead are committed to the idea that there is an innate stock of semantic representations in terms of which all our concepts can be explicitly defined. … *How could such concepts as* carburetor *be possibly innate?* Primitive peoples who have had no acquaintance with internal combustion engines show no difficulty in acquiring such concepts. On Fodor's account this means that their 'language of thought' contained the concept 'carburetor' prior to their acquiring a word for that concept, even though nothing in their evolutionary history could account for how the concept 'got there' (Putnam 1988, 15; Putnam's emphasis)

How is it possible to argue against so much common sense? Can Chomsky underpin his claim to the contrary? Can he repudiate Putnam's injunction? He would not explicitly refer to it if he thought he could not: "Some, for example Hilary Putnam, have argued that it is entirely implausible to suppose that we have 'an innate stock of notions' including carburetor and bureaucrat." (Chomsky 2000 [1992], 65) But Chomsky's repudiation avoids straightforward argumentation. Instead he reverts to a parable:

> Notice that the argument is invalid from the start. To suppose that, in the course of evolution, humans come to have an innate stock of notions including *carburetor* and *bureaucrat* does not entail that evolution was able to anticipate *every* future physical and cultural contingency – only these contingencies. That aside, notice that a very similar argument had long been accepted in immunology: namely the number of antigens is so immense, including even artificially synthesized substances that had never existed in the world, that it was considered absurd to suppose that evolution had provided 'an innate stock of antibodies'; rather, formation of antibodies must be a kind of 'learning process' in which the antigens played an 'instructive role'. But this assumption might well be false. Niels Kaj Jerne won the Nobel Prize for his work challenging this idea, and upholding his own

> conception that an animal 'cannot be stimulated to make specific antibodies, unless it has already made antibodies before the antigen arrives' (Jerne 1985: 1059), so that antibody formation is selective process in which the antigen plays a selective and amplifying role. (Chomsky 2000 [1992], 65)

Thus in the very moment when we experience a particular stimulus or trigger the corresponding mental concept comes to our rescue. The trigger can be the perception of something 'real', as in the case of a carburettor, or it can be an idea, like that we are burdened by too much admin, as in the case of bureaucracy. Not all cognitive scientists will be ready to follow him that far. I myself have no idea how realistic Jerne's antibody theory is. Yet to believe that the whole infinity of future discourse objects is somehow, *in nuce*, already present in our genes seems to overstress the suggestiveness of Chomsky's charisma. Fodor himself, who shares with Chomsky the belief that concepts such as 'bureaucracy' and 'carburettor' have to be regarded holistically, and cannot be decomposed into semantic primitives, would, I think, hesitate to underwrite this claim, at least since the publication of his book *Concepts*: *Where Cognitive Science Went Wrong*. It seems now that he prefers not to become involved. Concerning the innateness of the concepts 'carburettor' and his other favourite example, 'doorknob', he only tells us: "A lot of people have Very Strong Feelings about what concepts are allowed to be innate…[T]here is, at present, a very strong consensus against, as it may be, DOORKNOB or CARBURETTOR. I have no desire to join this game of pick and choose since, as far as I can tell, it hasn't any rules." (Fodor 1998, 28) In those few instances in which he is more specific, his peculiar metaphoric way of speaking makes it hard to pin down his true position. Concerning the concept 'doorknob' he explains: "[W]hat has to be innately given to get us locked to doorknobhood is whatever mechanisms are required to come to strike us as such. Put slightly differently: if the locking story about concept possession and the mind-dependence story about the metaphysics of doorknobhood are both true, then the kind of nativism about DOORKNOB that an informational atomist has to put up with is perhaps not one of *concepts* but of *mechanisms*." (Fodor 1998, 142) Thus Fodor implies that the question whether doorknobs correspond to an innate concept is the wrong question. For him, the important issue is the innateness of the mechanism that links concept and object. Thus he leaves himself a door open. He could still agree with Ruth Millikan when she, as she recently did on a conference, pleads for first-person experiences, and not innateness, as the precondition for mental concepts to become relevant "[A]ll concepts, including logical concepts, are tested *for their very having of content* through ongoing experience... What can be gained through conceptual analysis is then only what has previously been inductively acquired through experience." (Millikan [2004], 1)

Particularly in continental Europe, the view that complex concepts cannot be reduced to basic concepts is not very popular. While on the one hand many cognitive linguists like for instance Anna Wierzbicka take universal, innate or inherited concepts for granted, they insist that there is only a rather limited number of inherited basic concepts, called *semantic primes* by her, which are the building bricks for all more complex concepts. Wierzbicka lists about 50 *semantic primes*, among them variables like *sometimes*, *someone*, *something*, verbs such as *think*, *want*, *feel*, *say*, *happen*, *move*, four adjectives: *good, bad, big, small*, nouns such as: *part, kind, kind, people*, two pronouns: *I* und *you*, and a medley of connectors like *where, above, after, if, because* etc. This, then, would be the translation of the natural language sentence "X felt guilty" into a representation by semantic primes:

> X felt something
> sometimes a person thinks something like this:
> I did something
> because of this, something bad happened
> because of this, this person feels something bad
> (*http://rhm.cdepot.net/knowledge/theory/NaturalSemanticMetalanguage/definition.html*)

Of course, Anna Wierzbicka is well aware of the problematic nature of such a theoretical construct. If there 'really' were language-independent concepts, of a more primitive or even of a complex nature, holistic or not, how would we know what they 'mean' and how they would translate into natural language? Only in Texas it may be the common understanding that English is the language of thought. Inaccessible as language-independent *semantic primes* are for us, we can only encounter them once they are translated into a natural language, and we will never be able to control the appropriateness of this translation. Furthermore they are, in translation, as ambiguous as natural language tends to be. Thus we are left in doubt whether the translation of the complex mental concept corresponding to *X felt guilty* is correct. And is it 'really' true that *bad* in 'something bad happened' is the same *bad* as in 'feels something bad'? Is *X felt guilty* really the same as *X fühlte sich schuldig*? Does it mean the same as *X had a bad conscience*, *X had pangs of conscience*, *X had a sense of guilt, X felt remorse* and *X repented*? Wierzbicka would point out that she is only sketching a model and that this model is not materially and perhaps not even functionally equivalent to the 'real' mental representation. Widely read as she is, she has repeatedly related her approach to Leibniz; for example in this quote: "Im wesentlichen geht diese Idee auf Leibniz zurück und auf seine Vorstellung 'eines Alphabets menschlichen Denkens', das heißt, "einen Katalog der Begriffe, die aus sich selber verstanden werden können, und aus deren Kombinationen unsere anderen Vorstelllungen entspringen'" [In all relevant aspects this idea is based on Leibniz and on his model of an 'alphabet of human thought', i.e. a catalogue of concepts which can be understood out of themselves, and whose combinations engender our other ideas.] (www.humboldt-foundation.de/kosmos/titel/2002_003.htm) Even if Leibniz never distanced himself from the youthful folly of his doctoral dissertation *Ars Combinatoria* (1666), it was as unsuccessful as all other endeavours in the last millennium to construct a perfect language. This is the sobering conclusion we can draw from Umberto Eco's *The Search for the Perfect Language*.

Wierzbicka's original contribution is her suggestion of a (universal) syntax that informs the relationship her semantic primes have with each other when they are composed into a mental concept. Is this how we have to imagine mentalese, the language of thought? Do we find there the same categories that we use to describe the natural languages? Do we find there finite verbs, as we have them in English, but not in Chinese? Ray Jackendoff seems to know that the system of cognitive representations, mentalese, or, in his terminology, 'conceptual structure', does not have parts of speech: "Whatever we know about this system, we know it is not built out of nouns and verbs and adjectives." (Jackendoff 1997, 31) This is not how Steven Pinker sees it.

Are there 'really' fifty semantic primes whose meaning is universal but can only be understood once it is translated into a natural language? Is there 'really' a 'conceptual structure' in which we find concepts but no parts of speech? Is there 'really' something we 'know' about life after death? Or is what we claim to know about semantic primes, conceptual structures and life after death more a conjectural hypothesis than empirical knowledge?

Apart from Wierzbicka's mental syntax, we may well compare her *semantic primes* to the semes which were at the core of the mainstream continental European semantic theories of the sixties and the seventies. Usually we identify this semantic feature theory with Louis Hjelmslev's *Prolegomena* (Hjelmslev 1963 [1943]) His phonological analysis and his concept of the phoneme became the model for semantic analysis and the concept of the séme. Bernard Pottier combined Hjelmslev's approach with the Prague school of structuralism. He was the first to call the 'distinctive semantic features of lexemes' *sémes*. This is how he describes the meaning of *chair*:

> *chair:* {s1, s2, s3, s4} ("to sit on, on legs, for one person, with a backrest"). Relative to the set containing *easy chair*, chair is defined as *without* the seme s5 ("with armrests") and so on. (Pottier 1978, 86)

Thus meaning can be analysed in term of differences, through the presence or absence of sémes. It is this focus on difference which grounds this theory in de Saussure's structuralism.

Algirdas Julien Greimas, too, uses the concept of sémes. (Greimas 1966, 22 ff.). He distinguishes between the presence of a séme, the negation of this presence ('negative séme') and a state in which a given theme is neither present nor absent ('neutral séme'). For Pottier and Greimas the séme thus is the smallest feature (*trait distinctif*) to distinguish meaning that accounts for the difference between one word such as *chair* and another, semantically related word, *easy chair* (a word belonging to the same semantic field). Sémes here are understood as heuristic constructs. In the second edition of Theodor Lewandowski's *Linguistischem Wörterbuch* (1976) we find this entry for *semantisches Merkmal* ['semantic feature']:

> Bedeutungsatom, Bedeutungskomponente, Element des Begriffs bzw. Inhalts, der als in sich (mikro)strukturiert aufgefaßt wird, Basis-Element und Konstrukt einer semantischen Theorie, das sich mit Konstrukten wie Atom, Gen usw. vergleichen lässt. Bei der Konzeption des semantischen Merkmas handelt es sich um eine Übertragung des Prinzips der distinktiven Merkmale auf den Bereich der Semantik…
> Bei Bierwisch (1967, 3) sind semantische Merkmale "certain deep seated, innate properties which determine the way in which the universe is conceived, adapted, and worked on." [Atom, component of meaning, element of the concept or of the content looked at as a (micro-) structure in a semantic theory that can be compared to constructs such as atom, gene etc. Conceiving of semantic features in this way is a transference of the principle of distinctive features onto the field of semantics… For [Manfred] Bierwisch, semantic features are "certain deep seated, innate properties which determine the way in which the universe is conceived, adapted, and worked on." (Lewandowski 1976, 3, 663)

What is interesting here is the naivety in this entry in which semantic features are fused with semantic primes. For these features are theoretical constructs within a linguistic model to which no ontological reality is ascribed. This is the understanding we find in the entry *Merkmal* ['feature'] in Lewandowski's linguistic glossary:

> Begriffliches Konstrukt, ein Begriff, der für das richtige Verstehen der Sprachstrukturierung unentbehrlich ist (Martinet); für die Konstruktion und Funktion sprachlicher Einheiten als notwendig betrachtete begrifflich-hypothetische Mikroelemente. [Theoretical construct, a concept indispensable for the proper understanding of structuring language (Martinet); conceptual-hypothetical micro-elements considered as essential for the construction and function of linguistic units. (Lewandowski 1976, 2, 446)

Would Bierwisch agree? Developing further the contention forwarded by Jerry Fodor and Jerold Katz that it were possible to "construe a meta-theory containing a list of semantic features from which we can take the theoretical vocabulary of any special semantic theory" (Fodor/Katz 1963, 208), Bierwisch explains:

> This does not mean, of course, that the dictionary of each given language must show exactly the same distinction as that of any other language. It implies only that, if a distinction is made, that property can be characterized in a nontrivial way in terms of a universal set of semantic markers. If we accept this view, then two different questions immediately arise:

What is the theoretical status of the universal semantic markers; how must they be interpreted?

What are the elements of the universal set and how can they be established?

> ... The question here is: in what way, by what type of phenomena, are they motivated outside of the structure in the narrower sense? In other words: what is the interpretation of semantic markers, how are they connected with thought? [The German version here reads: "welche Beziehungen bestehen zwischen ihnen und den kognitiven und perzeptiven Leistungen des Menschen?"] (Bierwisch 1967, 2; 1970, 270-271)

Bierwisch, it seems, is not troubled by the question whether the semantic features are theoretical constructs of the linguist which he or she derives from the analysis of a natural language. For him they are real, ontologically given, located in human cognition. Are they learnt or inherited? He is obviously very sure: "Not only is there no reasonable explication of how semantic markers are learned. It is also very difficult to explain in a natural way such well known facts as displaced speech, fictitious objects and in general all gaps between meaning and reality." (Bierwisch 1967, 3) For Bierwisch, there is no alternative to the inheritance option. This was not so unexpected at a time when the attraction of Chomsky's model had its first peak in Europe. Particularly linguists in East Germany and other Eastern European countries embraced it because it located their field securely within the sciences, outside of the *Geisteswissenschaften* with their suspected bourgeois affinities. As real scientists, linguists thus could exempt themselves from the obligation to justify their approach from a Marxist-Leninist perspective. If the language organ was real, then semantic features must be real, as well:

> There are good reasons to believe that the semantic markers in an adequate description of a natural language do not represent properties of the surrounding world in the broadest sense, but rather certain deep seated, innate properties of the human organism and the perceptual apparatus, properties which determine the way in which the universe is conceived, adapted, and worked on. (Bierwisch 1967, 3)

Bierwisch fully subscribed to this fairly pervasive Anglo-Saxon entrenchment in a realism born out of common sense which up to this day still determines, to a very large extent, analytic philosophy in America, turning relativism there almost into a swearword. The abyss between Anglo-Saxon realism and Continental constructionism, in the guise of nominalism, hermeneutics or (post)structuralism, is rarely bridged. Bierwisch's semantic features do no longer correspond to the sémes hypothesised by Poittier or Greimas. Lewandowski's dictionary passes over this crucial difference in silence, namely that sémes are viewed as the linguist's constructs derived from the analysis of a natural language by applying heuristic procedures, while the cognitivists' mental concepts are seen as ontologically real entities. Whether it makes sense to posit a given séme in order to account for the difference between two semantically closely related words can be negotiated. But whether there is an innate mental concept meaning 'with armrest' is not a matter for discussion; it has to be proven in a scientific sense.

It is this insistence on ontological reality as opposed to hypothetical models that distinguishes the programme of cognitive semantics from that of structural semantics in continental Europe. Behind innate mental concepts we find looming the postulation of universality. It is not any more a particular natural language we are analysing, but the language of thought, the common mother of all languages. Does this utopian claim really help us to reveal the mystery of meaning? What do we gain if we set out to investigate some elusive model of the workings of the mind rather than sticking to the real language data to which we share access? It does not matter how we have to imagine mental concepts. They may be holistic as in the case of Chomsky's example of the carburettor, or we may imagine them as concepts composed out of semantic primitives; there is never a kind of empirical evidence about them that could be objectified. Perhaps this is the most plausible explanation for the lack of consensus among cognitive linguists as to the nature of these mental/cognitive concepts/representations.

Natural languages leave a lot to be desired. They are, as we have learned, full of vagueness and ambiguity. They are subject to constant change, and under closer scrutiny they tend to get lost in an almost infinite diversity of regional, situational, social, genre-specific and domain-specific variation. We are always encountering language usages which seem foreign to us. For thousands of years people have been complaining about the decay of language. Often when we question language use (normally the way other people use language) we look into the past. Then we ask ourselves what the word in question 'really' means, and for *really* we could read *originally*. Apparently we long for a Golden Age when there was still a natural, uncorrupted relationship between the word and what it stood for. This explains the popularity of etymological dictionaries. But even Plato, in his dialogue *Cratylus.* , could not convincingly answer the question what makes such a relationship natural. Thus Socrates asks: "For the gods must clearly be supposed to call things by their right and natural names, do you not think so?", to which Hermogenes responds: "Why, of course they call them rightly, if they call them at all." (Translated by B. Jowett; http://bang.pmc.purdue.edu/victorian/-uploads/R00010/Cratylus.pdf) In which language do the Olympians converse when they are among themselves?

In this context it might be worth having a look at machine translation and artificial intelligence. Perhaps there we can learn why concepts seem to be so much more attractive than natural language expressions. For these are fields in which concepts have been key features from early on. But in machine translation we are confronted with a fusion of two theoretical concepts of the concept. Concepts are not only mental entities, they are also the staple fare of terminology. Terms are the expressions of concepts, and a concept is how an element, a feature or a property pertaining to a specific domain has been defined by the experts. In terminology, a concept is identical with its definition. Context and usage are irrelevant, and so is the natural language expression by which it is denoted. In terminology, just as in cognitive linguistics, the concept is universal, while the terms, the expressions will differ from language to language.

Terminological concepts can be easily processed by computers. Terms can be translated on a one-to-one basis and they can be used in information retrieval. General language words, however, pose problems. They are fuzzy and ambiguous. A word in one language hardly ever maps onto a word in another language. In each occurrence of a word, its meaning is contaminated by its context. This is what makes natural language processing do unrewarding. A solution to this problem which still has a large following in artificial intelligence and machine translation is to convert words into concepts. Thus everything disturbing, unclean, fuzzy and ambiguous can be filtered out, so that we are left with nothing but the true, authentic, uncorrupted meaning of a concept. This expectation explains the popularity of conceptual ontologies in the artificial intelligence community. They account for all the concepts of a given domain and the relationships that obtain between them. Concepts in language engineering thus are the language-independent, spiritual, angelic natures of natural language words which have become unclean through their incarnation. This is how the difference between the word and its concept is commonly described:

> Ontologies describe concepts, not the way these concepts are expressed in words in a natural language. Therefore it is usually assumed that the ontology is language-independent. (Hans Weigand (1997): A Multilingual Ontology-based Lexicon for News Filtering. www.uvt.nl/infolab/prj/trevi/trevi.ps)
>
> Concepts represent the abstract meanings of words, and lexical entries represent the surface realizations of these meanings… Concepts represent word meanings, whereas the lexical knowledge they have represents ways to express these meanings with words. (Mattias Agnesund (1997): Representing culture-specific knowledge in a multilingual ontology. svenska.gu.se/~svema/ijcai97.ps)

Concepts, it seems, are pure meanings, cleansed from the impurities which they contracted through the contingencies of change afflicting natural languages. In these 'language-independent' ontologies there is no room for doubt what is a concept, how it is defined and how it is related to other concepts. Concepts are neither fuzzy nor ambiguous. Every proposition is either correct, 'grammatical', or not. If machine translation is still unsatisfactory then only, according to this claim, because we still have problems in converting natural language sentences into their conceptual representations.

What ontology engineers such as Agnesund and Weigand tell us about the relationship between words and concepts surely accounts also for the attraction of the innateness theory. Language ceases being unruly once it has been transferred into a language-independent universal representation. As soon as a natural language sentence is translated into mentalese, we have something similar to a mathematical equation, an expression that can be decided on the basis of its formal properties. In addition, it can claim reality and universality. In formal calculi, we observe the workings of immutable laws. The linguistics of formal languages is a pure science.

Is understanding really that simple? Leaving aside, for the moment, the question whether concepts are innate, holistically acquired or composed of inherited semantic primitives, what then is the 'real' content of the concept 'carburettor'? Is it enough to know that a carburettor is an important part of an internal combustion engine which has to be repaired or replaced when it stops working, or does it include a comprehensive representation of its functionality? Is the meaning of the word *carburettor* identical with the (content of the) concept, or, if not, what is the difference? What could we gain from an analysis of the concept 'carburettor' what we could not gain from the analysis of the word *carburettor*? What is the use of mental representations of concepts for linguists? One reason, I believe, why cognitive linguists not normally ask these questions is that many of them, particularly those belonging to the camp of one-level semantics, want to avoid the issue of intentionality. They are content with a purely syntactic processing of a cognitive representation. For them, the ultimate 'meaning' of a natural language expression is the form of the corresponding mental concept, i.e. its neuronal representation in the brain tissue.

4. Corpus linguistics and cognitive linguistics: is there a common ground?

As long as we have nothing but conjectures on swampy ground for cognition, as long as there is no *bauplan* which correlates the mind to the brain, would we not do better to focus our investigation not on language as a mental phenomenon, but on language as a social phenomenon? This is the programme of corpus linguistics, The discourse, this virtual structure containing the entirety of all the verbal exchanges between the members of a discourse community, is the market into which new objects of discussion are introduced, in which the meanings of discourse objects are negotiated, by acceptance, modification or rejection what has been said about them before, by explication and paraphrase, in which we are told what is proper to say and what not. The discourse is the supermarket where we shop our attitudes, beliefs and ideologies.

Without the discourse, our minds would be blank slates. The content we have in our minds is the content we have downloaded from the discourse. The discourse has an answer to our question what it means to open a bottle of wine. There is no other way for us to know how a hearer understands what the speaker says than by exchanging texts, by entering the discourse. Mental concepts, if there are any, are but the residue of the meaning of the lexical items we encounter in the discourse. We know what *carburettor* or *bureaucracy*, means, we know what we are expected to do if we are asked to open a bottle of wine because it was explained to us in the discourse.

Can we expect that the two paradigms, corpus linguistics and cognitive linguistics, will eventually form a joint platform? Mike Stubbs has not given up this hope. He asks "whether it is possible to state causal relations between linguistic, cognitive and social patterns". (Stubbs 2003, 1) I am not worried about finding out how social patterns determine the discourse or how the discourse constructs social patterns. This is well within the realm of corpus linguistics. I am much more hesitant with a prognosis of a satisfactory account of the interaction between linguistic and cognitive patterns. Stubbs is right when he states that corpus linguists "have hardly considered the relevance of corpus evidence to questions about the mental lexicon" (Stubbs 2003, 2). But would cognitive linguists be at all interested in what corpus linguistics has to offer? They tend to use the corpus mostly as a quarry for examples that suit the points they want to make. They still cling to the notion of single words as the core units of meaning, and they do not take into consideration that corpus linguistics has rendered the alleged fuzziness and ambiguity of natural language as a mock problem that disappears once we shed the belief that meaning can be studied in single words in isolation. Can we discuss how what we get out of the discourse has an effect on what we contribute to the discourse? Corpus linguistics is providing the data for an empirical theory of 'cultural transmission', and more than that, it has its own theoretical framework, intertextual hermeneutics, to explain how new objects are entered into the discourse, how they are accepted, modified, changed or rejected, and how they are compared, across languages, cultures and times, to other discourse objects. But do we have to understand how the mind works to understand these discourse processes? Language is public. The mind is private. Only when first-person experiences are communicated as testimony they become accessible. But then they are already a part of the discourse. So what is there to gain from looking into people's heads? Would we not gain more from a serious project analysing intertextuality in a diachronic corpus?

Mike Stubbs is, I think, the one corpus linguist cognitive linguists would listen to. He is sympathetic towards their aspirations, while firmly grounded in the rationality of empirical analysis. He combines steadfastness, common sense and patience with outstanding scholarship. I hope that he will initiate an open discussion between the two camps. If anyone can turn such a dialogue into a success, it will be him.

References

Bierwisch, M. 1967. Some semantic universals of German adjectivals. In: *Foundations of Language* 3, 1-36.

Bierwisch, M. 1970. Einige semantische Universalien in deutschen Adjektiven. In: Hugo Steger (ed.): *Vorschläge für eine strukturale Grammatik des Deutschen*. Darmstadt: Wissenschaftliche Buchgesellschaft, 269-318.

Chomsky, N. 2000 [1992]. Language and Interpretation. In: N. Chomsky: *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press, 46-75.

Chomsky, N. 2000 [1994]. Language as a Natural Object. In: N. Chomsky: *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press, 106-134.

Dennett, D. 1993. *Consciousness Explained*. London: Penguin Books

Dennett, D. 2003. *Freedom Evolves*. London: Allen Lane.

Eco, U. 1994. *The Search for the Perfect Language*. Oxford: Blackwell.

Fodor, J. 1998. *Concepts*: *Where Cognitive Science Went Wrong*. Oxford: Clarendon Press.

Fodor, J. No year. Distributed Representations¸ Enough Already. www.nyu.edu/gsas/dept/philo/courses/representation/papers/-fodordistributed.pdf.

Fodor, J. and J. Katz 1963. "The structure of a semantic theory". In: *Language* 39. 170-210.

Greimas, A. J. 1966. *Sémantique structurale*. Paris: Larousse.

Grice, H. P. 1989. *Studies in the Way of Words*. Cambridge, Massachussetts: Harvard University Press.

Harras, G. 2000. Concepts in Linguistics - Concepts in Natural Language In: B. Ganter and G. M. Mineau (eds.). *Conceptual Structures: Logical, Linguistic, and Computational Issues*. Heidelberg: Springer, 13-26.

Harras, G. 2001. Sprachproduktion als kollektives Handeln: Sprachphilosophische Grundlagen. In: T. Herrmann and J. Grabowski (eds.). *Sprachproduktion*. Göttingen: Hogrefe, 899-930.

Hjelmslev, L. 1963 [1943]. *Prolegomena to a Theory of Language*. Madison: University of Wisconsin Press.

Jackendoff, R. 1987. *Consciousness and the Computational Mind*. Cambridge, Massachusetts: MIT Press.

Jackendoff, R. 1997. The Architecture of the Language Faculty. Cambridge, Massachusetts: MIT Press.

Levinson, S. C. 1996. From inner to outer space: Linguistic categories and non-linguistic thinking. In: J. Nuyts and E. Pederson: *Language and Conceptualization*. Cambridge: Cambridge University Press, 13-45.

Lewandowski, T. 1976. *Linguistisches Wörterbuch*. 3 Bände. 2. Auflage. Heidelberg: Quelle und Meyer.

McGinn, C. 1991. *The Problem of Consciousness*. Oxford: Blackwell.

Millikan, R. 1984. *Language, Thought and other Biological Categories*. Cambridge, Massachussetts: MIT Press.

Millikan, R. [2004]. On the Epistemology of Concepts and How It Implicates Metaphysical Realism. Unpublished manuscript.

Nuyts, J. and E. Pederson 1997. Overview: The relationship between language and conceptualization. In: J. Nuyts and E. Pederson (eds.): *Language and Conceptualization*. Cambridge: Cambridge University Press, 1-12.

Origgi, G. and D. Sperber [2002]. Evolution, Communication and the Proper Function of Language http://dan.sperber.com/evo-lang.htm.

Pinker, S. 1994. *The Language Instinct*. New York: HarperCollins.

Pottier, B. 1964. Vers une sémantique moderne. In: *Travaux de Linguistique et de Littérature* II,1, 107-137 (German translation: Entwurf einer modernen Semantik, in: Geckeler, H. (ed.): *Strukturelle Bedeutungslehre*. Darmstadt, Wissenschaftliche Buchgesellschaft (1978), 45-89).

Putnam, H. 1988. *Representation and Reality*. Cambridge, Massachussetts: The MIT Press.

Rose, S. 2005. *The 21st Century Brain: Explaining, Mending and Manipulating the Mind*. London: Cape.

Scaruffi, P. 2003. Thinking about thought http://www.thymos.com/tat/neural.html

Searle, J. 1980. Minds, Brains, and Programs. In: *Behavioral and Brain Sciences* 3, 417-424.

Searle, J. 1990. Is the Brain a Digital Computer? http://www.ecs.soton.ac.uk-/~harnad/Papers/Py104/searle.comp.html.

Searle, J. 1992. *The Rediscovery of Mind*. Cambridge, Massachusetts: MIT Press.

Searle, J. 1998. *Mind, Language and Society*. New York: Basic Books.

Searle, J. 2005. "Consciousness: What We Still Don't Know". In: *New York Review of Books,* Vol. LII, Number 1, 36-39.

Sinclair, J., S. Jones and R. Daley 2004. *English Collocation Studies: The OSTI Report*. Ed. by R. Krishnamurthy. London: Continuum.

Sperber, D. and D. Wilson 1998. The Mapping between the Mental and the Public Memory. In: P. Carruthers and J. Boucher (eds.): *Language and Thought: Interdisciplinary Themes*. Cambridge: Cambridge University Press, 184-200.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Stubbs, M. 2003. Corpus Analysis: The State of the Art and Three Types of Unanswered Questions. Plenary, International Systemic Functional Linguistic Congress (ISFC 29) on Systemic Linguistics and the Corpus, Liverpool 15-19 July 2002. Manuscript, version 12 February 2003.

Teubert, W. 2005. My Version of Corpus Linguistics. In: *International Journal of Corpus Linguistics* 10, 1, 2-16.

Toolan, M. 1996. *Total Speech: An Intergrational Linguistic Approach to Language*. Durham, NC: Duke University Press

Developing language education policy in Europe – and searching for theory

Michael Byram

University of Durham

Abstract

Starting from a discussion of work analysing and evaluating language education policy at European level, conducted under the aegis of the Council of Europe, this chapter shows that the Council of Europe has a policy position and influences the policy making of member states. It does so more directly than through the dissemination of ideas via networks and workshops as has been done for the Common European Framework of Reference. For in the process of producing Language Education Policy Profiles, the Council of Europe promotes a specific view of the purposes of langue education and the preferred objectives for learners and education systems.

The second, more speculative section reflects on the need for explanatory theory which relates language policy to social conditions and the education environment. There is a need to go beyond case studies, useful as these may be, to a taxonomy of polices and then to an explanation of the implementation of policies according to circumstances. The underlying question is under what circumstances Council of Europe member states might accept and implement a policy of plurilingualism. Theory which can predict this will be of significance in Europe but perhaps applicable beyond. [1]

1. Purposes

There are two related but separate purposes in this paper:

- to present a critical analysis of current work at the Council of Europe on the promotion of a policy for language education
- to emphasise the need for a theoretical perspective on this and other language policy activity which might help to explain and predict the outcomes of policy-making.

In the main part I will describe one aspect of the current work of the Language Policy Division at the Council of Europe in Strasbourg, and raise the question whether this is an example of policy making at supra-national level. In the conclusion, I will address what seems to me to be a lack of adequate theorising about language policy, and language education policy in particular. I do not propose a means of filling this gap, unfortunately, but hope identifying the gap will be a first step towards this.

2. Languages and polities

Anderson (1991) in his well-known discussion of the nation as an imagined community, points to the significance of language and argues that the close relationship between language and nation was promoted from a European, Humboldtian perspective, and was part of the 'model' of the nation-state which was borrowed - or as he says, 'pirated' - in many parts of the world. By referring not simply to language but to 'print-language' and the power of newspapers and books to create a sense of community, Anderson also emphasises the significance of literacy. A nation-state is thus *inter alia* a community of communication which needs a shared language, and usually this shared language is the one designated as the national language. Thus, linguistic identity and national identity are closely connected, wherever there is a formal, institutionalised community of communication. The connection is reinforced by schools as national institutions where one learns the national language, whatever one's home or first-acquired language.

Yet there are also other levels of community within a nation-state which are not necessarily formalised. The organisations and institutions of civil society have differing degrees of formality, and where there is freedom of speech, these communities of communication can challenge the official discourses of the state (Kennedy and Fairbrother, 2004: 296). Nonetheless, such discourses are likely to be conducted through the same national, officially recognised language, and again we see the significance of the national language and the reinforcement of the relationship between the national language and national identity.[2]

The significance of communication and interaction becomes all the more evident as the nature of polities changes. For Habermas, the model which should replace out-dated concepts of 'the classic republican idea of the self-conscious political integration of a community of free and equal persons', is a model dependent on communication flows:

> a model of deliberative democracy, that no longer hinges on the assumption of macro-subjects like the 'people' or 'the' community but on anonymously interlinked discourses or flows of information (Habermas, 1994: 32).

This applies to the evolution of the nation-state, but all the more to the evolution of democratic processes in transnational contexts. Communication flows and the 'informal networks of public communication' at a transnational level pre-suppose favourable conditions for mutual understanding.

The importance of this issue is evident from the evolution of transnational civil society in response to the trend towards global governance through such organisations as the World Trade Organisation and the International Monetary Fund. The exact nature of the organisation and nature of transnational civil society and of a democratisation of global governance is not yet clear. However, it can be argued that the present legitimisation based on the notion that experts can deliberate and come to representative consensus is inadequate and should, and will, be replaced by debate in a public sphere, where a public is understood as 'a collectivity of persons connected by processes of communication over particular aspects of social and political life' (Nanz and Steffek, 2004: 8). Nanz and Steffek argue that 'organized civil society has a high potential to act as a 'transmission belt' between deliberative processes within international organisations and emerging transnational public spheres' (ibid: 10).

Perhaps the most likely place for this to happen first is in the political and cultural space which has been created in Europe over the last half century.

The role of the Council of Europe is important in this, because of its influence in forty-five European countries. As an inter-governmental organisation, the Council of Europe does not have a policy-making function independently of its member States. On the other hand, in practice, proposals evolve from meetings and conferences and are ultimately endorsed by member States at Councils of Ministers. As part of this process, the Council of Europe has developed in the last few decades a clear language education policy position, and this was recently stated in a draft document to celebrate the 50th anniversary of the signing of the Cultural Convention. The statement is not a repetition of the many and various recommendations which have been endorsed by member States, but rather a summary of the purposes of these recommendations:

> Council of Europe language education policies aim to promote:
> *Plurilingualism*: all are entitled to develop a degree of communicative ability in a number of languages over their lifetime in accordance with their needs
> *Linguistic diversity*: Europe is multilingual and all its languages are equally valuable modes of communication and expressions of identity; the right to use and to learn one's language(s) is protected in Council of Europe Conventions
> *Mutual understanding*: the opportunity to learn other languages is an essential condition for intercultural communication and acceptance of cultural differences
> *Democratic citizenship*: participation in democratic and social processes in multilingual societies is facilitated by the plurilingual competence of individuals
> *Social cohesion*: equality of opportunity for personal development, education, employment, access to information and cultural enrichment depends on access to language learning throughout life
> (Council of Europe, 2004a).

This statement can be taken as a policy position and in the sense that it creates a consensus which is endorsed by member States, the Council of Europe can be described as a policy-making body.

Turning to the European Union, we can see a more obvious policy-making function as nation-states gradually give up some of their power and adopt a more international, or at least European, perspective. In such circumstances, the notion of a national language and linguistic identity is weakened and there is encouragement for other, 'foreign', languages to be given a status as part of the creation of identification with a community. This is made very clear in the EU's White Paper of 1995:

> Languages are also the key to knowing other people. Proficiency in languages helps to build up the feeling of being European with all its cultural wealth and diversity and of understanding between the citizens of Europe.
> (….) Multilingualism is part and parcel of both European identity/citizenship and the learning society.
> (European Commission, 1995: 67).

What we have here then is a statement where the word 'European' could be substituted by the name of almost any nationality, and the parallels with the role of language in an imagined community are clear. It is also clear that, as in the nation-state, the levels of communication are not only those which are formal and institutional, but also include those of civil society.

The subsequent recommendation for practice is that European citizens should speak their mother tongue(s) plus two other languages, and this implies that a knowledge of three or more languages – perhaps to different degrees and in different ways – will create a sense of European identity and citizenship, and a potential for participation and integration into an international/ European society and polity.

This position had changed by 2003, when a weaker statement was issued. Although it still uses the White Paper as one of its sources, the focus now is on effective participation and social cohesion; the reference to identity no longer appears:

> (1) knowledge of language is one of the basic skills which each citizen needs to acquire in order to take part effectively in the European knowledge society and therefore facilitates both integration into society and social cohesion; a thorough knowledge of one's mother tongue(s) can facilitate the learning of other languages
> (2) knowledge of languages plays an important role in facilitating mobility, both in an educational context as well as for professional purposes and for cultural and personal reasons
> (3) knowledge of languages is also beneficial for European cohesion, in the light of EU enlargement
> (4) all European languages are equal in value and dignity from the cultural point of view and form an integral part of European culture and civilisation.
> (European Commission 2002)

Here is an emphasis on mobility but the professional / economic purposes are linked to the personal, and the specific issue of enlargement from 15 to 25 countries is given prominence.

In summary, it appears that the European Union approach to language education postulates some, unclear, relationship between national language/ mother tongue learning and foreign language learning; second, a causal relationship between language learning and identity/ citizenship; and, third, a conditional relationship between language learning and participation in European society. Learning several languages is at least a pre-condition and perhaps a causal factor in the evolution of citizenship in the narrow sense of being an elector, and in the broader senses of an affective bond with an international society and a participation in the economic, political and cultural life of the society.

What makes the European situation different from nation states is that it is not expected that people should be native speakers of all the languages they might acquire as part of becoming European citizens, even though there are powerful forces encouraging people to acquire as high a level of competence as possible.[3] The success of a European imagined community of communication pre-supposes plurilingual competence so that discourses at formal level and in civil society can take place, can be extended beyond the national frontiers, to European level. Thus, association of native speaker competence with identification with a polity is put in doubt, and replaced by plurilingual competence.

The alternative, of creating a shared *lingua franca* – which at this point in history could only be English – is not politically acceptable since there would be accusations of linguistic imperialism and/or allowing unequal and unfair dominance to native speakers of English. Whether these are justified or not, a lingua franca would not be efficient. Transnational discourses cannot rely on a single, taken-for-granted, shared language and its meanings. The discourse which is necessary is not simply a matter of establishing an agreement on and/or an exchange of information such as might be achieved through a *lingua franca*. The issues which arise in social discourse are shot through with contemporary and historical nuances, and the relationship between language and thought, between language and world-view is crucial.[4] When people engage in cooperation in civil society, they do so as social beings whose social identities are embodied in the languages they speak. To use a lingua franca is reductive of their social identities and diminishes them as human beings.[5]

3. Policy Profiles - a tool for policy implementation?

We have then in Europe two supra-national bodies with their language education policy. However, policy without implementation is ineffective, and from this point I will focus on the Council of Europe and the question of whether it has this function too.

Over the recent period of two to three years, the Language Policy Division in Strasbourg has created a new 'service' offered to member States. The Language Policy Division helps member States which invite it to do so, to review and develop their policies for language education. This activity has a broad remit, to include the teaching and learning of all languages in a polity: from a learning perspective, first languages, second languages, foreign languages; or put it in sociological terms, minority and regional languages, national or official languages, immigrant languages, foreign languages. The polity may be a country but can also be a region or city with its own language education policy.

The role of the Language Policy Division is made clear in this extract from the guidelines which govern the activity:

> Language Education Policy Profiles
> The Council of Europe has launched a new activity to *assist* member States who so wish in reflecting upon their language education policy. The aim is to offer member States the opportunity to undertake a *'self-evaluation'* of their policy in a *spirit of dialogue* with Council of Europe experts, and with a view to focusing on possible future policy developments within the country. It should be stressed that developing a language education policy profile *does not mean 'external evaluation'.* It is a process of reflection by the authorities and members of civil society, and the Council of Europe experts have the function of *acting as catalysts* in this process.
> (Council of Europe, 2004b - my emphasis added)

The purpose is clear: that the Language Policy Division does not interfere in policy development but facilitates self-analysis.

The process involves several stages:

- after a preliminary organisational visit, a Country Report, by or on behalf of the authorities, is written to describe whatever issues the authorities consider important
- a group of three to five experts visits the country for a week, talking with representatives of stakeholders in education, and produces its own Experts' Report
- this Experts' Report is circulated within the country to whomever the authorities wish, including all the stakeholders whom the experts had met
- a one day round table discussion is held between all stakeholders invited by the authorities and the experts, where issues of accuracy, of comprehensiveness are raised, and where the stakeholders can exchange views
- in a final stage a Language Education Policy Profile is produced by the expert group (mainly by the group *rapporteur*) in consultation with the authorities, and published jointly by the authorities and the Council of Europe.

Throughout this process, the expert group reminds the readers of its report and of the final profile about the policy position of the Council of Europe, and about the instruments it has produced which are useful for the implementation of policy.

These include the *Common European Framework of Reference*, the European Language Portfolio, and the *Guide for Language Education Policy*. Thus the experts, as catalysts, bring to the notice of the authorities and other stakeholders the policy and instruments which all member States have endorsed, and there are a number of criteria which underpin the Experts' Report:

- that language education must be considered holistically, overcoming the separation between first, second, foreign languages
- that the promotion of plurilingualism and diversity is axiomatic in all planning
- that curriculum design and pedagogy must reflect and be determined by the holistic vision of a language education
- that language education is tied to education for citizenship in all multilingual polities, of which European countries and Europe as a whole are clear examples.

Thus far, then, the role of the Council of Europe through its experts involves in principle a catalytic function. On the other hand, in practice, it is clear that member States and stakeholders within a country are not as aware as they might be of Council of Europe policy and instruments. The catalysts in fact bring new elements to the process, which cause a re-assessment of existing assumptions, even though the *Guidelines* explicitly say that there is no external evaluation. These new elements are particularly characterised by the European perspective and not just a national, regional or local one. This includes for example an emphasis on the teaching of all languages irrespective of their social status, with a strong emphasis on diversification of language learning opportunities throughout life, resistance to the dominance of English as a lingua franca, a transversal, holistic vision of convergences in the languages curriculum among national, minority, foreign and other languages.

Further factors in this catalytic process are the role of the Experts' Report, written independently by the experts bringing both Council of Europe perspectives and their own expertise to the analysis, and also the role of the experts, in particular the *rapporteur*, in the authoring of the final profile. Both of these allow opportunity for a new and ultimately evaluative perspective on the assumptions of the authorities and other stakeholders. This perspective has in practice been welcomed and encouraged by the authorities in most cases hitherto. There is none the less a delicate balance of power to be sought in the final Profile, since it has to be acceptable to and published by the authorities and the Council of Europe.

4. The Council of Europe as a policy making body?

I pointed out earlier that the Council of Europe is an entity which has language education policies, even though in principle these are the formulation of the views of member States and not independent European policies. It also has instruments which can be used in the implementation of those policies. On the other hand, it does not have an obvious function in the implementation of policies; it cannot send out directives to member states in the way that the European Union can.

Does the Council of Europe none the less have the characteristics of a policy making body? One way of addressing this question is to use Cooper's and Ager's frameworks for analysing language policy, to see if and how the Council of Europe fits into them.

Cooper (1989) provides an ordered list of questions which can guide policy analysis, and I present these here with application to the Council of Europe and its Policy Profile activity:

Question: What actors?

> An inter-governmental organisation acting with national (or regional or local) authorities. Hitherto, actors have been individuals, groups or agencies at or below national level, but the two supranational bodies, the EU and the CoE have now become active. Ager (forthcoming) argues that the European Union is a special case where policy is formulated in general terms and not pursued in detail because of the political sensitivities. The Council of Europe is more precise in its formulation and, through the Policy Profiles is seizing the nettle of influencing member States,

Question: attempt to influence what behaviours?

> The CoE attempts to influence production of policy at national (etc) level with respect to planning language acquisition and, indirectly, attitudes towards plurilingualism,

Question: of which people?

> of national (etc) authorities making policy for language acquisition at national or regional level,

Question: for what ends?

> following Ager's (forthcoming) distinctions, at the level of unattainable but necessary 'ideals', the CoE sets ideals which are
>
> (a) language related: to increase the diversity of languages in society and the diversification of languages learnt in the curriculum
>
> (b) non-language related: to promote education for democratic citizenship, and an understanding of linguistic otherness
>
> secondly, with respect to 'attainable, but long-term objectives', the CoE provides
>
> (a) language related instruments of various kinds to support language acquisition planning (the *Common European Framework,* the European Language Portfolio, the *Guide for the Development of Language Education Policies in Europe*) and planning for regional and minority languages (*The European Charter for Regional or Minority Languages* and the *Framework Convention for the Protection of National Minorities*)
>
> (b) non-language related instruments for planning education for democratic citizenship (e.g. *Draft Common Guidelines on Education for Democratic Citizenship*) although at this level there is as yet no proper

coordination between language and non-language related activities but, at the third level, the Council of Europe does not become involved in 'short term objectives' such as curriculum planning,

Question: under what conditions?

against a background of European integration, economic mobility and human capital theory, and a move towards an international civil society; the geographical and political space in question is however no longer limited to the nation-state, but extends to Europe as the totality of all member States of the CoE,

Question: by what means?

unlike some states in both past and present, the CoE does not use force or bribery, but its authority as an inter-governmental organisation, and its reputation gained through language-related work over several decades,

Question: through what decision making process?

through the formulation of Europe-level policy by recommendations – for long-term objectives – and a process of agreement to the recommendations at meetings of ministers of education and/or Heads of State and Government; and through reference to these and the instruments for implementation in the Experts' Report and the Profile,

Question: with what effect?

in completed studies so far:

- in Norway, with impact on current education reform at the level of long-term objectives and with some impact on details of curriculum planning
- in Hungary, with input to new policies and plans countrywide for language teaching, and thus at the level of long-term objectives.

It seems therefore that of the Council of Europe, working with member States (or other polities) fulfils some of the characteristics of a policy making and policy implementing body, even though it has no direct power over implementation. If we consider the three levels of 'ends' in Ager's definitions, then the 'ideals' are present in the discourse at European level and have begun to infiltrate the discourse of national bodies. The parallel although not identical position of the European Union no doubt contributes strongly to this. At the level of long-term objectives, the use of the *Common European Framework* in planning national curricula is evident in some countries. In the case of the Country Profiles, the impact will doubtless vary from case to case and is yet to be seen over a number of forthcoming cases. At the level of short-term and immediate objectives, the Council of Europe does not expect to have impact but there may be some evidence that this happens through the Country Profiles; there are as yet too few cases to draw any conclusions.

It is thus possible to turn now to the Ager's (2001) model for analysing the motivations behind language policy-making to see if this can throw further light on the role of the Council of Europe.

Ager focuses on questions of identity and the motivation in policy making to maintain or develop (national) identity. If we take some of the elements of his model which deals with identity sequence, attitudes, and purposes, it becomes evident that the Council of Europe is acting in some respects in a way similar to the nation states with which Ager is concerned:

Identity promoted in the policy:

- what identity: the construction of European identity underpins Council of Europe policy
- what ideology: the Council of Europe promotes the equality of all languages and the correction of inequalities by supporting linguistic and regional minorities, and by promoting relations among nation states

Attitudes in the policy:

- there is emphasis on the attractiveness of plurilingualism
- there is action taken through the policy profiles and the instruments to promote plurilingualism

Purpose of the policy:

- there is an explicit pursuit of linguistic diversity, of international citizenship, and of cohesion among member states.

If we consider the sequence of events, however, there is a difference. Nation states are the realisation of a bottom-up desire of ethnic groups with an existing identity for a national identity and political power (Edwards, 1994); the Wilsonian principle of post-1918 and its presence in the Treaty of Versailles was one very important reflection of this. After 1945, the Council of Europe started from an ideal of cultural cooperation and mutual respect among nations, as stated in the summary of the Cultural Convention:

> to develop mutual understanding among the peoples of Europe and reciprocal appreciation of their cultural diversity, to safeguard European culture, to promote national contributions to Europe's common cultural heritage respecting the same fundamental values and to encourage in particular the study of the languages, history and civilisation of the Parties to the Convention.
> (http://conventions.coe.int/Treaty/EN/cadreprincipal.htm)

This has been formulated since then in terms of social inclusion, citizenship and mobility for individuals, and in the creation of a European identity: 'The Council was set up to (…) promote awareness of a European identity based on shared values and cutting across different cultures' (www.coe.int/T/EN/Com/About_COE/). In this it is joined by the European Union, even though the latter had a different starting point.

There are also other differences between the Council of Europe and the nation-state as represented in Ager's model, in particular the relationship with other comparable entities, and the possible integration with other entities. There are no comparable entities which, in the case of nation states, are the external 'threats' which help maintain internal unity. Moreover, whereas a nation-state will usually seek to identify and promote one language – so that it becomes or remains a national language – this is clearly not the case with the Council of Europe, which substitutes plurilingualism for monolingualism or, in multilingual territories, the dominance of one language.

There are then similar but not identical processes and purposes at work here to those which Ager identifies, and the points in which his model does not apply are the indications of difference.

5. Conclusion

The Council of Europe has, in short, the characteristics of a policy-making and a policy-implementing body. It is comparable to the nation state and other policy bodies in this respect even though it operates at different levels and has different means at its disposal. The question which then arises is whether it is possible to predict the outcomes of this activity and it is here that I fear there is a gap to be filled.

Our need in the case in question is to predict whether a policy of plurilingualism and diversification as proposed by the Council of Europe will be accepted and implemented by member States. They have endorsed it in principle but principle does not necessarily end in practice. Theory which helps to predict whether a policy will be successfully implemented in a given set of circumstances might also allow us to identify inhibiting factors, and to change these in order to facilitate implementation.

One example exists with respect to the teaching of foreign languages, a necessary but not sufficient aspect of Council of Europe policy. This is a paper by Trim (1994) in which he identifies a range of different conditions which are more or less likely to lead to successful policies for foreign language teaching. This could perhaps be extended to encompass plurilingualism, diversity and diversification in the curriculum. Implicit in Trim's paper is an attempt to produce a taxonomy of language situations. A taxonomy is crucial to prediction, but needs to embrace the multilingualism within a polity in a holistic way if it is to help in prediction of the success of policies of plurilingualism. This might lead to predictions of there following kind:

- in language situations of type A, plurilingualism can be attained by implementation of a policy of type X.

However, language situations need to be theorised in a sociological perspective too, since policies exist in the interplay of entities holding power. One approach would be through Bourdieu's theory of social reproduction through education, which would suggest that language education policies are subject to the efforts of certain groups in society to maintain their cultural and social capital through education and education policies.

The potential for successful implementation of language education policies can also be analysed using economic theory. Grin (2004) has led the way in applying economic theory of costs and benefits to policies, as a means of helping authorities and other stakeholders to make decisions. It is also possible to envisage an analysis of language and education policy from the perspective of the debate about the marketisation of education, and whether education should be treated as a public good. This debate has been particularly vehement in anglophone countries.

These and other approaches need to be explored not only in the context of Council of Europe policy work, but perhaps the significance of this work makes the need all the more urgent.

Notes

1 I am very grateful to Dennis Ager for comments on and suggested additions to a draft of this paper. I remain of course responsible for its contents.

2 This analysis is deliberately simplified, and has to be modified *mutatis mutandis* for nation states where there are more than one national language or where the speakers of a minority language are accorded legal rights to use the language in public discourse.

3 All the work currently being carried out in the EU on the development of a 'Europass' for languages or in Strasbourg on a 'European Language Portfolio' is a sign of the recognition by European authorities, and the national authorities which support them, that plurilingual competence of some kind is crucial.

4 I take a 'weak' Whorfian/Humboltian position which cannot be developed and defended here but for which there is supportive empirical evidence in Levinson (1997).

5 See Breidbach (2003) for a further discussion of levels of public fori and the language combinations which might be required.

References

Ager, D. 2001. *Motivation in language planning and language policy*. Clevedon: Multilingual Matters.

Ager, D. (forthcoming). *Language policy and language planning*. Strasbourg: Council of Europe.

Anderson, B. 1991. *Imagined communities.* 2nd ed. London: Verso.

Breidbach, S. 2003. *Plurilingualism, democratic citizenship in Europe and role of English*. Strasbourg: Council of Europe.

Cooper, R.L. 1989. *Language planning and social change.* Cambridge: Cambridge University Press.

Council of Europe, 2004a. *Plurilingual education in Europe*. Draft document presented at the Language Policy Forum, June 2004.

Council of Europe 2004b. *Language education policy profiles. Guidelines and procedures*. www.coe.int/lang .

Edwards, J. 1994. *Multilingualism.* London: Routledge.

European Commission 1995. *Teaching and learning: towards the learning society*, Brussels: European Commission.

European Commission 2002. Council Resolution of 14 February 2002, Official Journal C 050, 23. 02.2002.

Grin, F. 2004. Towards a public policy approach to language scenarios for the EU. Paper at the conference 'Language and the Future of Europe', Southampton, July 2004.

Habermas, J. 1994. Citizenship and national identity. In B. van Steebergen (ed.) *The Condition of citizenship* London: Sage. 20-35.

Kennedy, K.J. and Fairbrother, G.F. 2004. Asian perspectives on citizenship education: postcolonial constructions or pre-colonial values? In W.O. Lee et al.(eds.) *Citizenship education in Asia and the Pacific. Concepts and issues*. Hong Kong: Comparative Education Research Centre and Kluwer Academic. 289-301.

Levinson, S.C. 1997. From outer to inner space: linguistic categories and non-linguistic thinking. In S. Nuyts and E. Pederson (eds.) *Language and conceptualization.* Cambridge: Cambridge University Press. 13-45.

Nanz, P. and J. Steffek 2004. Global governance, participation and the public sphere. *Government and Opposition* 39, 2, 314-35

Trim, J.L.M. 1994. Some factors affecting national foreign language policymaking in Europe'. In R. Lambert (ed) *Language planning around the world: contexts and systemic change.* National Foreign Language Center Monograph Series. Washington D.C. NFLC. 1-15.

The semiotic patterning of Cædmon's Hymn as a 'hypersign'

Wolfgang Kühlwein

University of Trier & Université du Luxembourg

Abstract

Cædmon's Hymn has been researched extensively with a view towards its assumed significance as an early key document of English literary and sociocultural history, and as an (often questioned) 'masterpiece' of the evolving skills in the use of Anglo-Saxon poetic devices, such as metre, rhythm, alliteration, variation etc. However, accounts of its overall structure have remained scarce.

Taking for granted the high intensity of both its emotive and its appellative load as explicitly commented on by Bede himself, an approach that is methodically based on (Peircean) semiotics lends itself for a descriptive analysis of its overall patterning.

Basically, the Hymn presents itself as an interlace of the intricate network of the dyadic interrelationships between 'God and Creation', 'Creation and Mankind' and 'God and Mankind'. However, it is God's act of GIVING that elevates it above a mere accumulation of dyadic patterns. It is the explanatory claim of this semiotic approach to the overall patterning of this 'hypersign' to account for the alleged semiotic thrust of the Hymn.

1. Dedication

Professor Michael Stubbs, whose 60th birthday this Festschrift celebrates, is enjoying world-wide reputation in many areas of the multifaceted linguistic landscape; in particular, he may well claim to have been a pioneer for present-day corpus linguistics and to be one of its prime representatives today.

As to his relationship to the author of this contribution to the Festschrift: both served as Presidents of national affiliates of the International Association of Applied Linguistics (AILA), Mike Stubbs of BAAL, the author of GAL; and as Professors in the Department of English at the University of Trier/Germany both have in the literal sense of the word been 'next-door' neighbours for these last 17 years – a neighbourhood that enhanced opportunities for informal talks about all the diverse decision-making procedures in university management and for many ad hoc discussions of scholarly matters.

What, therefore, seemed to the author to be an appropriate token to show his gratitude for all that Michael Stubbs has been giving to him, had, of course, to be a study that is corpus-based and that chooses a "gift" for its theme – the only difference being that its giver is not Mike Stubbs but God Almighty, and the receiver is not Wolfgang Kühlwein alone, but all Mankind.

2. The research object

As opposed to the large size of corpora drawn upon by Professor Stubbs, the corpus we are going to analyse happens not to exceed 9 lines! It will be set into relief, however, against the total corpus of Old English poetry wherever hypotheses based on a work-immanent view call for further evidence, be it for purposes of their verification or of their falsification. Among scholars of English and Germanic Philology, Historical English Language Study, Anglo-Saxon Literature, and Theology of Early Christendom likewise, these 9 lines have become well known as *Cædmon's Hymn* (henceforth *CH*), '*Kædmonischer Schöpfungshymnus*'.[1]

That *Hymn* will have been composed at the Northumbrian abbey of *Strenæshalc* / northumbr. *Streunaes Halh* (= *Whitby*) sometime between 657 A.D. and 680 A.D., i.e. during the term of office of the reliably recorded abbess Hild, the ruling Northumbrian King's sister.

The *Hymn* has been handed down to posterity about half a century later by the Anglo-Saxon chronicler Venerable Bede (= *Beda Venerabilis*: b. ca. 672 A.D.; visitor to Yorkshire and in all likelihood to the monastic and scholarly centre *Streunaes Halh* at least once, in 733 A.D.; d. 735 A.D.). His *Historia Ecclesiastica Gentis Anglorum* (Book IV, Chapter XXIV) offers the by now famous account of how that *Hymn* was conceived by the sleeping cowherd Cædmon at the instigation of a miraculous nocturnal apparition (*sum mon* i.e. some 'man' some 'living being': mostly interpreted as 'an angel', or even as 'the Almighty' [contested]) who induced him to –very reluctantly – sing a song, a song of the beginning of existence.

Bede offers the Hymn in its Anglo-Saxon version, furnishing a translation into Latin.[2]

In its Northumbrian version of MS Cambridge University Library Kk, 5.16 the text of *CH* runs as follows:

1	Nu scylun herʒan		hefaenricaes uard,
2	metudæs maecti		end his modʒidanc,
3	uerc uuldurfadur		sue he uundra ʒihuaes,
4	eci dryctin,		or astelidæ;
5	he aerist scop		aelda barnum
6	eben til hrofe,		haleʒ scepen,
7	tha middunʒeard		moncynnæs uard;
8	eci dryctin		æfter tiadæ
9	firum foldu,		frea allmectiʒ.

Primo cantauit Caedmon istud Carmen**.**

> (*Now we shall praise the Guardian of the heavenly kingdom, the power of the Almighty, and His spirit and thought, the achievement of the Father of Glory, because He, the Eternal Lord, set the beginning for any kind of wonders; in the beginning He, the Holy Creator,*

> *Guardian of Mankind, shaped heaven as the roof for the children of men, and the earth to live on; afterwards the Eternal Lord, the Almighty Ruler, adorned the world for Mankind.*)[3]

The prime legitimation for our analysis springs from the multitude of scholarly evaluations of *CH*. Be it from a linguistic, from a literary, from a theological stance, they do not merely differ, but they are as incompatible as can be, oftentimes irreconcilably contradictory. To provide a few examples out of many:

(a) "…Cædmon's Hymn appears to display no great originality, for, though it is technically accurate, nine or more of its eighteen half-lines can be paralleled in other poems" (Smith 1933: 14f.)[4]

(b) "…it has qualities of balanced and rhythmic grandeur…" (Wrenn 1947: 9)

(c) "The Hymn has hardly enough literary merit to allow of discussing it at any length as a piece of poetry…" (Kane: 1948 250f.)

(d) "…the *Hymn* is made up entirely of formulas or systems of formulas, in a word, …its language is quite traditional" (Magoun 1955: 53)

(e) "a technically miserable performance" (ibid.: 57)

(f) "…his [Bede's] attitude toward poetry is Augustinian…. Since God is the source of all beauty, it follows that for Bede the *Hymn* had some relation to the divinely inspired poetry of Scripture. Angelic inspiration implies revelation: The angel brings to a chosen vessel, characteristically humble, the obligation to receive and to be the first to communicate God's word in English poetry. In consequence…. Caedmon's *Hymn* must for the believer have seemed as nearly perfect as man's work may be; either the poem was beautiful to the eyes of faith, or there was no miracle. It is impossible that God should have inspired what is inferior or merely workmanlike. Since the demands on the little poem were very large, Bede must have seen in it much more than the best disposed modern is likely to allow" (Huppé 1959: 102f.)[5]

(g) "one of the greatest landmarks in the history of our English poetry" (Wrenn 1968: 57)

(h) "Cædmon's Hymn appears to display no great originality" (Smith 1968:14f.)

(i) "freshness and originality" (ibid: 15 !)

(j) "a pleonastic tour de force" (Bessinger 1974: 93)

(k) "Cædmon's Hymn…shows an adaptation of traditional style and Christian content (if the usual view of the poem's heroic diction is right) so perfect and so comfortable that one cannot help finding the poem –as Bede did– 'miraculous'" (Gardner 1975: 7).

In view of these discrepancies among both the linguistic and the literary evaluations, what is indicated is the search for an approach, which is superordinate to both the particularizing studies of literature and of language. This cannot but be an approach based on the theory of signs – signs as encompassing literature, language, and to a large extent, theological exegesis, too: i.e. the theory of General Semiotics.

3. The theoretical and methodical toolkit for the analysis

As to the reasons that were responsible for our decision in favour of *the* specific semiotic theory (*vs*. competing ones), to base our analysis on, we refer the reader to Kühlwein (2006a: 105-108). With a view to our specific object of research, *CH*, we opted for Charles Sanders Peirce's semiotic theory as an adequate research tool. Admittedly, Peirce himself never made use of it to describe and explain the semioticity of literary texts, let alone of texts from periods as remote as the Anglo-Saxon one. His own applications are mainly devoted to explain phenomena from the realms of philosophy, theology, mathematics, natural sciences, above all from logics. However, Peirce postulates: "Logic, in its general sense, is… only another name for *semiotic,* … the quasi-necessary, or formal doctrine of signs" (Peirce 2.227). Both his *Phenomenology* and his *Elements of Logic* are concerned with relations – and it is a set of interrelational patterns that our semiotic analysis intends to cast some light upon, the relations between God and Creation, between Creation and Mankind, and between God and Mankind, as well as the intricate relational interlace of these three relations among each other.

As a full presentation of Peircean semiotics would be out of place here, we shall confine the presentation of our Peircean theoretical-methodical 'toolkit' to a sketch of those elements that come to bear in the analytical section below:[6]

In his *Phenomenology* Peirce distinguishes as the three elements of phenomena 'quality', 'facts', and 'thought'. His *Elements of Logic* likewise centre on these three properties, each one being triadic in turn. In 2.233ff. he distinguishes three general kinds of triadic relations:

Relations		**Their nature**
Triadic relations of comparison		Logical possibilities
Triadic relations of performance		Actual facts
Triadic relations of thought		Laws

Any triadic relation has **three correlates**:

1st correlate (= **'Firstness'**):		Simplest nature	Representamen
2nd correlate (= **'Secondness'**):		Middling complexity	Object
3rd correlate (= **'Thirdness'**):		Most complex	Interpretant

Phenomenological definitions and semiotic examples:

Firstness: a phenomenon, whose essence is determined by mere strength of itself, e.g. the relationship of a phenomenon to itself as representamen

Secondness: a phenomenon, whose essence is determined by strength of its relation to something other than it is itself, a 'second', e.g. the relationship of a phenomenon to the object which causes it

Thirdness: a phenomenon, whose essence is determined by strength of relating a 'first' to a 'third' via a middling 'second', e.g. a phenomenon as representing a certain object in such a way as to cause a certain interpretation.

As noticed above, for Peirce logic in a wider sense is semiotic. The "phenomena" of semiotic are signs. It follows, that signs, too, are subject to these triadic relations and their correlates. According to 2.243ff. ('*Trichotomy of Signs*'), signs can be subclassified by three categories:

I. *According as the Sign in itself is*

a. A mere quality		*Quali*sign (= a quality which is a Sign) – 'First'
b. An actual existent		*Sin*sign (= an actual existent thing/event which is a Sign) –'Second'
c. A general law		*Legi*sign (= a law that is a Sign) – 'Third'

II. *According as the relation of the Sign to its Object consists in*

a. the Sign's having some character in itself		*Icon* – 'First'
b. the Sign's having some existential relation to that Object		*Index* – 'Second'
c. the Sign's having a relation to its Interpretant		*Symbol* – 'Third'

III. *According as its Interpretant represents it as a Sign of*

a. Possibility		*Rheme* (= a Sign of qualitative possibility) – 'First'
b. Fact		*Dicent* (= a Sign of actual existence) – 'Second'
c. Reason		*Argument* (= a Sign of law) – 'Third'

A few explanatory comments will be required here:

As to (II a), an *Iconic Sign* refers to the Object that it denotes, merely by virtue of inherent characteristic features of its own

As to (II b), an *Indexalic Sign* refers to the Object that it denotes, by being really affected by that Object

As to (II c), a *Symbolic Sign* refers to the Object that it denotes, by virtue of a law, usually an association of general ideas

As to (III a), a *Rhematic Sign*, for its Interpretant, is understood as representing such and such a kind of possible Object; it represents its Object in its characters merely

As to (III b), a *Dicentic Sign* is a sign, which, for its Interpretant, represents its Object in respect to actual existence; therefore it cannot be an Icon, which affords no ground for an interpretation of it as referring to actual existence; the Dicentic Sign affords ground to judge whether what it expresses is true or false

As to (III c), an *Argument* is a sign which, for its Interpretant, is a sign of law; it is understood to represent its Object in its character as a Sign; the Interpretant of an Argument represents it as an instance of a general class of Arguments.

Thus any sign is constituted by the three relationships (I), (II), and (III), i.e. by its relationship to itself, to its Object, and to its Interpretant.

Each one of these three relationships allows for its further subtle classification according to the three triads (a), (b), and (c). Theoretically this would yield 66 classes. However, only 10 actually occur, because both the basic trichotomy on the one hand (Representamen, Object, Interpretant) and the respective subclassifications (quali, sin, legi; icon, index, symbol; rheme, dicent, argument) on the other hand are hierarchically structured: Thirdness involving Secondness and Secondness, in turn, involving Firstness; e.g. a legisign is materialized by its occurrence in actual existence, in other words as a sinsign, which in its turn can only be perceived as a qualisign: In Peircean terms: a qualisign that embodies a sinsign is a '*replica*' of the latter one; and a sinsign that manifests a legisign is the latter one's '*replica*' likewise.

Thus,

- an argument sign cannot but be both symbolic and legi
- a dicentic sign can neither be iconic nor quali.

Technically speaking, the 'path' in the direction from (I) via (II) to (III) can never lead 'upwards':

	(I)	(II)	(III)
a.	legi	symbol	argument
b.	sin	index	dicent
c.	quali	icon	rheme

Thus e.g.

a sign with the properties I c – II c – III c can exist: a quali - iconic – rheme sign;

a sign with the properties I b – II c – III c can exist: a sin - icon – rheme sign;

a sign with the properties I b – II b – III b can exist: a sin - index - dicent sign;

but a sign with the properties I c – II b – III c cannot exist;

nor can a sign with the properties I b – II b – III a, etc.

4. Base and aim of the analysis

4.1 The Base

We have tried to pave the way towards this analysis by means of two preceding ones.

The first study, Kühlwein 2006a, gave a semiotic in-depth analysis of Bede's narrative on how Cædmon conceived of the Hymn and of its effects on the hearers. The concept of 'gift' is the hinge for both Bede's narrative and for the Hymn likewise: God's gift to Cædmon to praise (according to Bede's narrative) God's gift of Creation as bestowed upon all Mankind (according to Cædmon's Hymn); two acts of creative shaping are thus intertwined. Semiotically God's bestowal of the Hymn to Cædmon reaches beyond random factuality (=Secondness), the relationship between the giver and the recipient rather turned out to be a rule-governed one, involving a law, i.e. semiotically the mind. This elevates that act of giving beyond a sum of two factual dyads , (1) "God gives to Cædmon" and (2) "Cædmon receives from God" to the level of a semiotic triad, involving (3) the recipient's responsibility as to his future appropriate use of the gift, and this is exactly what Bede subsequently expounds: arisen from Cædmon's pure feeling (Firstness), raised to an appeal (Secondness) to kindle the listeners' minds to turn to the continuous quest for heavenly bliss (Thirdness).

Finally, the semiotic relationship between the narrative and the Hymn itself led to the tentative hypothesis of the Hymn being a Dicentic Symbol.

The second study, Kühlwein 2006b, traced that hypothesis from the point-of-view of the semiotics of the three key concepts of the Hymn. It investigated the lexemes used in the Hymn to designate God, Creation, and Mankind. The semiotic relations holding for all lexemes used for each one of these three concepts were traced separately (i.e. '*intra*conceptually'), not yet, however, '*inter*conceptually'. That intraconceptual analysis indicated that the overall semiotic structure of the Hymn in its entirety might well be much more complex and considerate than many critics have assumed hitherto – an observation that would tie in with the previously assumed character of the Hymn as a Dicentic Symbol.

4.2 The Aim

It is the aim of our following analysis (part 5) to ultimately verify or falsify that hypothesis, which claims that the Hymn in its entirety is one Dicentic Symbol.

What is required to achieve that goal, is

- the textstructural semiotic analysis (part 5.1) of the Hymn in its entirety from the *inter*conceptual stance, i.e. the analyses of
  (1) the interrelationship between God and Creation (part 5.1.1)
  (2) the interrelationship between Creation and Mankind and (part 5.1.2)
  (3) the interrelationship between God and Mankind (part 5.1.3) and

- its collation with the results of the above-mentioned lexical analysis under the auspices of the semioticity of the Hymn when viewed as one whole entity (part 5.2).

5. The semiotic analysis

5.1 Textstructural semiotic evidence

5.1.1 The Interrelationship between God and Creation

The perspective under which to experience Creation is set by ll. 1 –3a: Cædmon's exuberant praise of God. These five half-lines comprise as many as four representamens of God, followed even by a fifth, pronominal one in the sixth half-line. It is as late as in l. 3 that Creation appears as a theme: strictly speaking, Cædmon's presentation of process and product of Creation starts with l. 3b. This thematic divide, that we make between l. 3a and l. 3b seems to be evident on the sentence-semantic plane, likewise.

The hinge of the matter is the semantics of *sue*- hitherto unresolved in this passage. In Modern English translations of *CH* it is rendered in three different ways:

(a) Some editors choose 'in this way, thus' for its equivalent. This yields a reading for ll. 1 – 4 "Let us now praise God,…… *In this way* He set the beginning of all wonders…" i.e. God's doing is described.

(b) Sometimes it is translated as 'how'. This yields a reading for ll. 1 – 4 "Let us now praise God,…… , *how* He set the beginning of all wonders…" i.e. again God's doing is described.

(c) Despite Bede's translation-equivalent for *sue* as lat. *quomodo*, e.g. Mitchell (1967:204) proposes a causal connection as a third possibility, which would render *sue* as 'because, inasmuch as', thus yielding a reading for ll. 1 – 4 "Let us now praise God,…… *because* he set the beginning of all wonders…"; i.e. the reason for God's praise is emphasized.

Semiotically, it seems, the reading that implies a certain amount of causality beyond mere description, can be supported on the basis of the textual structure Hymn in its entirety:

The semiotic entailment of readings (a) and (b) is that the praise of God in his unquestionable 'Firstness' would reach far beyond ll. 1 – 3a , and would, actually, be extended right to the end of the entire text.[7] The semiotic entailment of reading (c), on the other hand, equals reading (a) and (b) insofar as it allows the praise of God in his 'Firstness' in ll. 1- 3a likewise; however, it differs insofar as it emphatically ('because'!) proceeds from the praise of God's 'Firstness' to His 'Secondness', i.e. to God in his 'factual' role as the agent, the doer, the creator.

Both roles are different but inseparable.[8]

As will be become apparent below, the emphasis on that passage from God's 'Firstness' to His 'Secondness' is a decisive feature of *the* constitutive theme of *CH* in its entirety, contributing towards making it what we call a semiotic 'Hypersign'.

Conceiving of *sue* as a causal connector – be it with causality as the only and primary sense, be it as a secondary overtone merely - *sue* would mark a cesure between l. 3a and l. 3b considerably more distinctly than the senses 'thus' or 'how' would do; and, actually, it is from l. 3b onwards that God is presented in His 'Secondness'.[9] Very much like *Nu* in l. 1a in its function as a textsemantopragmatic opener to ll. 1 – 3a, *sue* as initiating l. 3b would function as the textsemantopragmatic opener to all that follows: both ll. 3b – 4 in general and ll. 5 – 9 in particular. The semiotic reason, why we would like to emphasize a clear recognition of that passage between God's 'Firstness' to His 'Secondness' – along with the markedness of the corresponding formal cesure between l. 3a and l. 3b - will become apparent below.

In any case, our look at the interrelationship between God and Creation should, definitely, set in with l. 3b. We shall move from form to function.

Formally the part ll. 3b – 9 of the Hymn has to be divided into three sections:

(a) ll. 3b – 4 are a sign for the relationship between God and Creation in general;
(b) ll. 5 – 7 are a sign for the interrelationship between God and Creation in specific;
(c) ll. 8 – 9 (if not interpreted as a mere follow-up of scop in l. 5) are a sign for God's making Creation inhabitable for Mankind.[10]

All three sections, thus divided, evince an identical basic semiotic pattern. Each occurrence of Creation in general (see (a)) or of each specific element of Creation (see (b)) is conjoined with its Creator by a twofold link:

(a) or (*uundra ʒihuaes*) is caused (*astelidæ*) by *he* + *eci dryctin*
(b1) *eben* is caused *(scop)* by *he* + *haleʒ scepen*
(b2) *tha middangeard* is caused (*scop*) by *haleʒ scepen* [as resuming *he*] + *moncynnes uard*
(c) *foldu* is caused (*tiadæ*)[11] by *eci dryctin* + *frea allmectiʒ.*

This overall pattern is extremely pervading and, as a consequence, cannot but be semiotically significant.[12] The apparent iconicity of this recurrent pattern calls for an equally iconic visualization, see Figure 1.

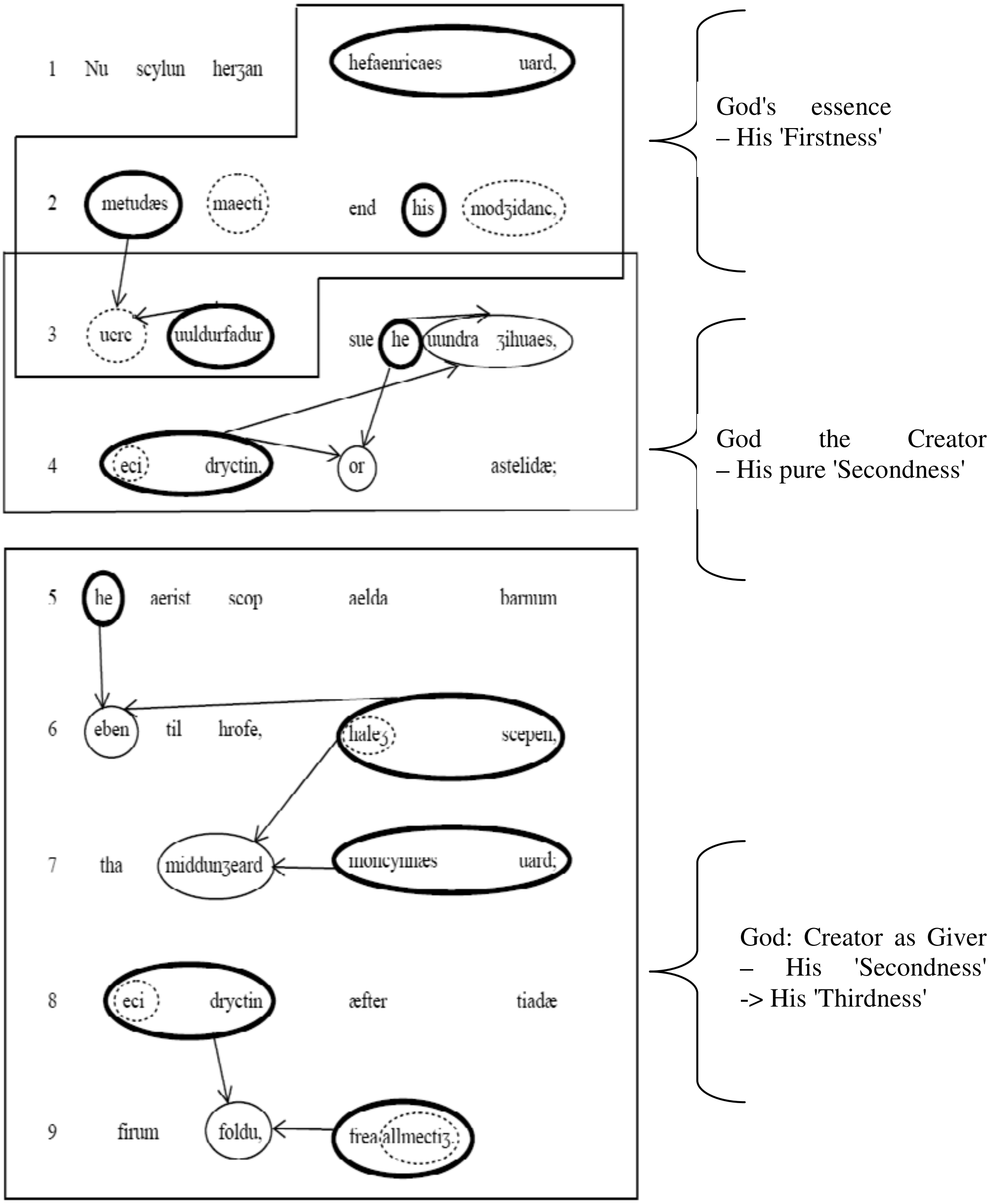

1 Nu scylun herʒan hefaenricaes uard,
2 metudæs maecti end his modʒidanc,
3 uerc uuldurfadur sue he uundra ʒihuaes,
4 eci dryctin, or astelidæ;
5 he aerist scop aelda barnum
6 eben til hrofe, haleʒ scepen,
7 tha middunʒeard moncynnæs uard;
8 eci dryctin æfter tiadæ
9 firum foldu, frea allmectiʒ.
God's essence – His 'Firstness'
God the Creator – His pure 'Secondness'
God: Creator as Giver – His 'Secondness' -> His 'Thirdness'

Figure 1: God – Creation
(References to God: bold-face type; to attributes of His: dotted; to Creation: regular face type)

These charts bring to light another special feature shared across (a) – (c). Each one of the references to Creation is framed by one reference to God that precedes it, and one reference to God, that follows it:

- *eben* is enclosed between *he* and *haleʒ scepen*;
- *middunʒeard* is enclosed between *haleʒ scepen* (as intensifying *he)* and *moncynnæs uard*;
- *foldu* is enclosed between *eci dryctin* and *frea allmectiʒ.*

Semiotically we can, legitimately, conceive of this textstructural feature as a sign of its own right. It is a sign of high iconic load, in other words a very elementary one: it is in a kind of photographic way that it shows God as the beginning and as the end of whatever is created. Creation in its entirety (*uundra ʒihuaes*) is embedded in the Creator's essence and it is He who literally embraces each specific constituent of Creation (*eben, middunʒeard, folde*).

But in addition, this 'photographic picture' indicates a relationship of a higher semiotic order, too. Let us, therefore, still conceive of ll. 5 –9 as being *one sign* and ask how it refers to the object it stands for. The continuity of Cædmon's interlocking pattern between signs denoting God and signs (syntactically: phrases) denoting the act of Creating indicates, that there is nothing that might intervene between God's act of creating and the coming into being of Creation itself. The act of creating pertains to God as such. Semiotically we might say, it is an emanation of God's Absolute Firstness, i.e. His 'essence'. However, it makes Him the (active) subject and it makes Creation the immediately affected (= passive) subject, within a mutual relationship, which holds between Him and Creation[13] – and this reveals God in His Secondness (= His creatorship). Linguistically God is topic throughout, Creation is comment. As far as that we are faced with a prime example of a semiotic *dyad.*[14]

Cædmon cannot but must have felt that dyadic character: quite obviously, it was impossible for him to think of Creation without referring to God at the same time. Ll. 1–3a provide further evidence: praising Creation (*uerc*) is equivalent to praising the Creator. Cædmon leaves no doubt either, that the two subjects, which constitute this dyad 'God/Creation', differ as to their semiotic status. God is First, Creation Second: even a simple numerical account can be regarded as a mirror of this gradation: ll.1–3a invoke God as First four times, and Creation once (*uerc*); ll.3b – 9 refer to God as First seven times, to Creation four times.[15]

The Hymn in its entirety reveals a further highly iconic feature of that dyad. The signs denoting Creation and its individually mentioned elements are not only inseparably embedded by those, which denote God, by strength of the dyadic relationship, but the Hymn sets out with those many invocations of God, and its last line 9b ends up with a reference to God, too. The appropriate iconic representation of this will have the shape of a ring. There can only be one object which this ring will designate: God , who has no beginning and no end, the *eci dryctin*.

This simile of an encompassing iconicity is indicative of great compositorial artfulness: its accomplishment is the weaving together of all single strands (here: all signs) that constitute the Hymn, into a texture of oneness. It is this very integrity that makes the Hymn what we choose to call a 'hypersign'.

5.1.2 The interrelationship between Creation and Mankind

As to the relation between Creation and Mankind two semiotically relevant patterns emerge: (5.1.2.1) as mirrored in (5.1.2.2) below. They are similar to the ones which reflect Cædmon's conception and presentation of the interrelationship between God and Creation.

The references to Mankind are syntactically 'enclosed' and semiotically 'embraced' by signs which denote God.[16] What might, therefore, seem to be incongruent at first sight: the introductory part of the Hymn, i.e. ll. 1 – 4, does without any mention of Mankind at all. However, this is what is semiotically to be expected qua the character of that introductory part ll. 1- 4: it is a praise God's Absolute Firstness. It cannot but be along with the references made to non-human Creation (*eben*, *middunȝeard*, *folde*), that Mankind, too, enters the Hymn, i.e. as late as in its second part - because it is this part that reveals God His (creative) Secondness.

What, nevertheless, seems to be semiotically significant, is the fact that no specific reference is made as to the sixth day of Creation, i.e. the very Creation as such of Mankind. Whereas *Genesis* is satisfied with presenting the mere sequence of creative acts, ending up with the Creation of Adam and Eve, Cædmon abstains from merely recapitulating that time- sequence. Instead, he provides a purposeful underpinning for God's creative doing from the very outset of Creation: that Creation is intended for the benefit of Mankind is expressed by Cædmon as early as in l. 5b (*aelda barnum*). Thus this destination initiates his presentation of the entire act of Creation. In *Genesis* God gives this purpose to His Creation as late as on its sixth day:

> "dixitque Deus ecce dedi vobis omnem herbam adferentem semen super terram et universa ligna quae habent in semet ipsis sementem generis sui ut sint vobis in escam/ et cunctis animantibus terrae omnique volucri caeli et universis quae moventur in terra et in quibus est anima vivens ut habeant ad vescendum et factum est ita" (*Gen* I. 29-30; Biblia Sacra Vulgata).
> ["And God said, Behold, I have given you every herb bearing seed, which is upon the face of all the earth, and every tree, in the which is the fruit of a tree yielding seed; to you it shall be for meat. / And to every beast of the earth, and to every fowl of the air, and to every thing that creepeth upon the earth, wherein there is life, I have given every green herb for meat: and it was so" (King James Bible)].

Considering that *CH* is a 7th century poem, this difference is amazing. Semiotically it bears evidence of the great significance, which Cædmon must have attributed to the idea of Creation as constituting a gift bestowed upon Mankind (– and thus paralleling the gift which Cædmon himself had received from the angel in his dream, i.e. the Hymn).

This concept of 'gift' runs through the entire part ll. 5 –9 of the Hymn in a consequent pattern:

- *eben* aims at the dative *(aelda) barnum* as its recipient
- *middunȝeard* likewise aims at *aelda barnum*
- *folde* aims at the dative *firum* as its recipient.

Figure 2 indicates this iconicity.

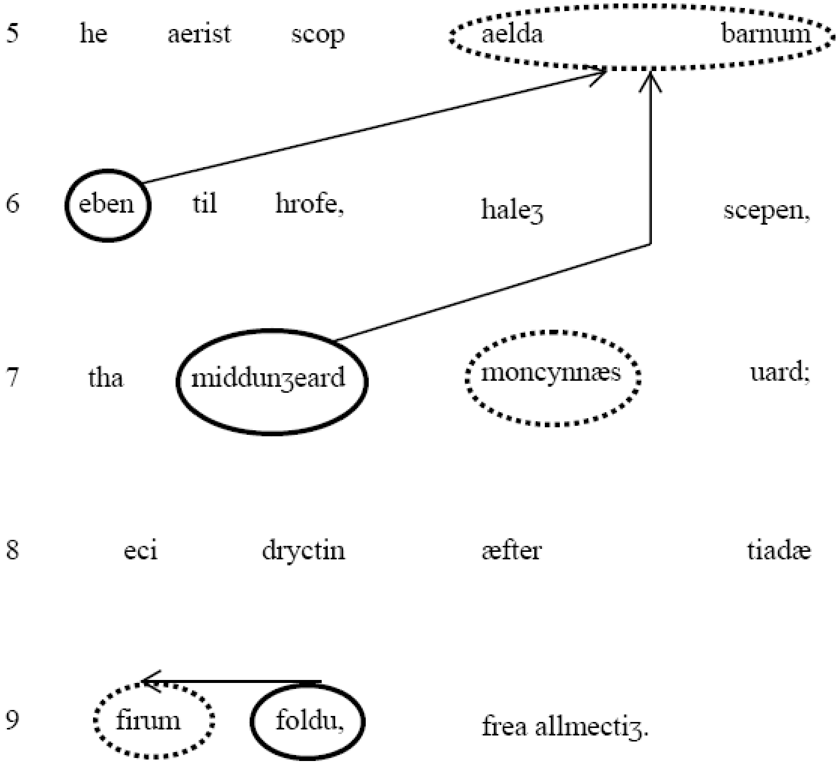

5 he aerist scop aelda barnum
6 eben til hrofe, haleȝ scepen,
7 tha middunȝeard moncynnæs uard;
8 eci dryctin æfter tiadæ
9 firum foldu, frea allmectiȝ.

Figure 2: Creation – Mankind
(References to Creation: bold-face type; references to Mankind: dotted lines)

In each case the recipients are mentioned first, the gift itself second. In the last instance, *firum foldu* in l. 9, both are 'consociated' with each other to the utmost density. This consociation between the non-human and the human part of

Creation does not make them a dyad. The contiguity between both rather is suggestive of a syllogism.

(a) Earth has a certain relation to God (qua having been created and shaped by God; *scop*, *tiadœ*)

(b) Mankind has a similar relation to earth (qua having been created and shaped from earth): a Biblical reference that will have been well-known at a religious centre like Whitby, where Cædmon had served as a minor brother in the monastery)[17]

The result:

(c) Mankind has a similar relation to God (qua having been created and shaped by God).

Thus, this contiguity that exists between non-human existence and human existence in the second part of the Hymn, culminating in its last line, provides the answer to the question, why the Creation of Mankind, i.e. the sixth day of Creation, is not mentioned as a creative act of its own right. Pragmatically however, it *is* mentioned: by means of implicature.

As seen from a semiotic perspective, this fact might well mirror Cædmon's 'hierarchy' in as far as the mere act of creating is concerned, i.e. in as far as God's Secondness is manifested:

God ranks first, Creation second, Mankind third.[18]

And once again, just like in the case of the interrelationship between God and Creation, a sheer numerical count of the references to these three concepts (God, Creation, Mankind) provides further evidence for Cædmon's 'hierarchy': among these three concepts Mankind is the one, to which the fewest references are devoted; actually, two altogether. There are *aelda barnum* and *firum* only; at best, *moncynn* in *moncynnæs uard* can be added. A semiotic look at *moncynnæs uard* at that very place (l. 7 b) in the Hymn might well give further support to our stance: this syntactic construction consociates God and Mankind to the closest extent possible – and two lines after that it is in l. 9a, that Creation and Mankind are consociated in an equally close way: *firum foldu*, i.e. not via the stylistic device of variation but directly within one half-line; semiotically: God takes care of Mankind (*uard*) → God created and adorned (*tiadœ*) the earth (*foldu*) as now existing for the benefit of Mankind.

5.1.3 The interrelationship between God and Mankind

As Mankind does not enter the Hymn until the specific part of Creation as beginning with l. 5, the text-based analysis of the interrelationship between God and Mankind should set out from there.

As indicated in 5.1.2 above, semiotic parallels with Cædmon's conception of the interrelationship between God and Creation are obvious. Like there, both references to Mankind are joined with its Creator by a twofold link:

(a) *he* and *haleȝ scepen* (reemphasized by *moncynnæs uard*) are the benefactors for *aelda barnum*

(b) *eci dryctin* and *frea allmectiȝ* are the benefactors for *firum*.

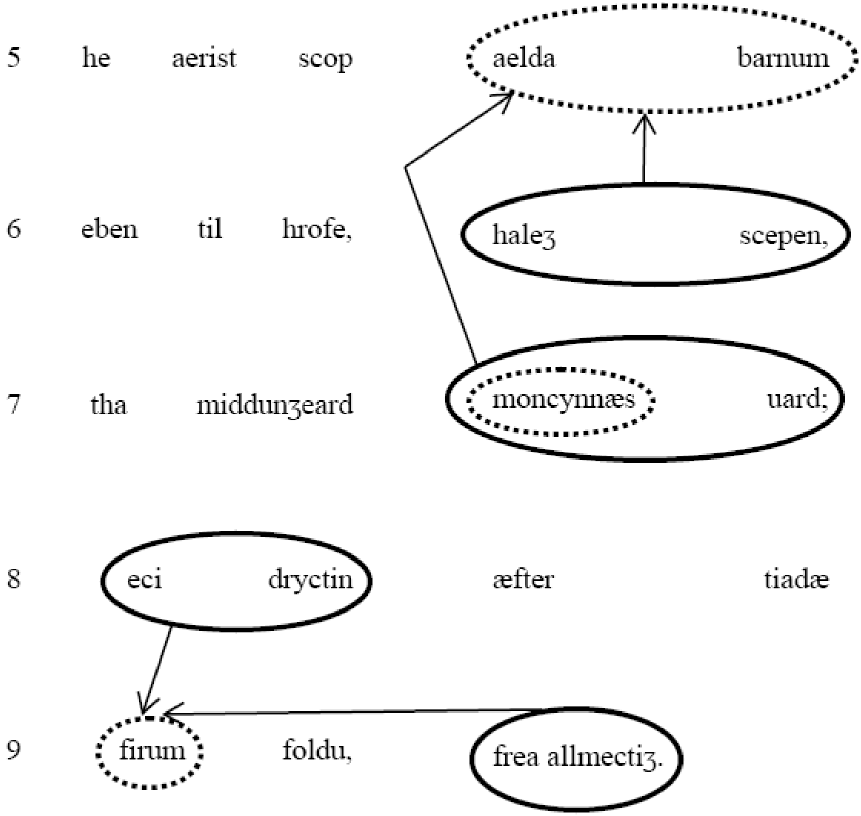

5 he aerist scop aelda barnum
6 eben til hrofe, haleȝ scepen,
7 tha middunȝeard moncynnæs uard;
8 eci dryctin æfter tiadæ
9 firum foldu, frea allmectiȝ.

Figure 3: God – Mankind
(References to God: bold-face type; references to Mankind: dotted lines)

To be more specific: each one of the two explicit references to Mankind is framed by one reference to God, which precedes it, and another one, that follows it:

- *aelda barnum* is 'embedded' between *he* and *haleȝ scepen*
- *firum* is 'embedded' between *eci dryctin* and *frea allmectiȝ.*

In addition, as we have seen above, in *moncynnæs uard* both, God and Mankind, enter into an even closer relationship.

The semiotic conclusions which we drew from the patterning of the interrelationship between God and Creation of the non-human world, obviously, hold for God's relation to the human part of Creation, i.e. Mankind, too. Semiotically Cædmon drew on a high degree of iconicity throughout. We tried to indicate that pervading feature in our figures above, consistently emphasizing that pervading pattern of 'embedding' / 'embracing'.

As to the interrelationship between God and Mankind, however, there is one essential difference:

In His relation to nonhuman Creation God is designated as the 'maker' the Creator, i.e. in His Secondness.

However, one should mind that in His relation to Mankind His property of being its Creator is merely *implied* (s.a.). Why?

It rather is God's property as a 'giver', that constitutes His relation to Mankind in the Hymn.

This difference has quite a strong semiotic impact: 'giving' as a sign is, as has already been indicated above, not Firstness, nor is it pure Secondness. It is genuine Thirdness.

5.1.4 Intermediate Conclusions

The semiotically relevant observations as gained from textual structure are:

(a) the conception of God as setting out from His Absolute Firstness via His Secondness to His Thirdness

(b) the theme of temporality as materializing in the above-mentioned icon of a ring

(c) the hierarchical sequence 'God first, then Creation in general, then Mankind specifically'

(d) the Hymn in its textual entirety as constituting *one* unitary sign, signifying: The exhortation to God's praise for His gift of Creation for the benefit of Mankind.

5.2 Lexical semiotic evidence

Are these observations supported, questioned or contradicted by the lexical choices concerning the key concepts of the Hymn? [19]

5.2.1 Ll. 1– 3a: hefaenricaes uard – metud – maecti - modʒidanc – uerc – uuldurfadur

The initiating 'opener' *Nu scylun herʒan* ('Now ought we to praise') is the verbal resonance to the strong urge towards making an appeal as felt by Cædmon when he was stirred by that nightly angelic apparition. It is an imperative, signifying a command, that is directed to himself and the angel first, to his brethren next, to all mankind at last; it is an index.[20]

Most other 'openers' met in the corpus of Germanic poetry are quite different. They would denote circumstances that can – be it historically, be it fictionally – be either verified (e.g. according to shared experiences, beliefs, assumptions) or falsified.[21] Semiotically, they even come close to being arguments, employing reasoning, i.e. Thirdness.

The opening to *CH*, definitely, is on a considerably lower semiotic level. As reported by Bede, *CH* arose from mere personal feeling and perception – i.e. in Peircean semiotic concepts: elementary Firstness – a feeling, that explodes into an imperative burst of appeal/command. This spontaneous outcry is a Rhematic Indexical Sinsign. It is indexalic, because this singular utterance directs the addressees' attention to the very object by which its occurrence is caused, i.e. God's Creation for Mankind; it is rhematic, because at that initial stage nothing definite is said yet about that object Creation. As a Rhematic Indexical Sinsign this opener will have secured the listeners' attention and will have aroused their desire to hear more about what will have caused the occurrence of a sign of that kind.

To sum up: the opener has its roots in Cædmon's state of pure feeling (quali – icon – rheme : Firstness), and slightly raises that semiotic level towards Secondness (sin – index – rheme).

This is the very semiotic level that is maintained in the following ll. 1b – 3a; now, however, projected onto God: (His) Firstness, now likewise slightly shading over into (His) Secondness.

The phrase ***hefaenricaes uard*** as it stands is a sign of essence (rheme). If there were a copula ('he *is* the warden of the heavenly kingdom'), the statement would be open to being questioned (dicent). The elementary quality of that sign (quali) *hefaenricaes uard* is a safeguard against such –inappropriate – questioning of God. As such it is a sign of the simplest nature, of mere (unbounded) possibility - introducing God in his quality as Absolute and Irreducible Firstness.

It is in ll. 2a – 3a that this elementary possibility is to gain a profile: the essence of God's *maecti*, *modʒidanc*, and *uerc* is presented gradually and, as there is no verb that would allow for debate, cautiously. In other words, there cannot be any questioning of God's might, his mind and thought, his work.

All three essences are associated with each other in a sequence, pointing out from iconicity towards more indexality in two directions: towards mankind and by means of their constant association with God (*metudes; his; uuldurfadur*) towards God Himself, too: in Anglo-Saxon poetic diction *uard* evokes both 'protection' and 'lordship'!

In many instances *maeht/**maecti*** denote 'power' as manifesting itself in causing, achieving, even creating and providing protection, and the related verb *magan* shades over into 'to help, provide benefit': a strongly intensifying index as to the *uard.*

Metud, related to the verb *metan* 'to measure', might in the 7th century still have had that connotative ring of 'measuring out [i.e. the span of life (as the North-Germanic norns had done)]'. In this capacity it is the proper link between the component 'Lordship' in the preceding *uard* and its subsequent further profiling in *maecti.*

Correspondingly, *(ge)danc* in ***modȝidanc*** as 'planning activity of the mind', frequently yielding 'pleasure, something to be thanked for', associates the protective side of God's *uard*ship.

The first constituent of the compound *modȝidanc*, *mod-*, modifies *-ȝidanc* ambivalently. Its semantic centre 'mind, will-power' via 'soul, spirit' extends as far as into the domain of feelings: both good ones (e.g. 'courage') and bad ones (e.g. 'pride', 'violence') – just as the related adjective *modiȝ*. Thus *mod* semiotically signifies a possibility. Here the semiotic correlate of this ambivalence of *mod* is the re-emphasis of God's Absolute Firstness: being *mod ȝidanc*, His *ȝidanc* cannot be subjected to any kind of influence. It rather is plain existence and as such beyond any human evaluation as to being good or bad. Therefore, the compound *modȝidanc*, once again, semiotically links up with *uard*, here, however with the *uard*'s second property, His 'Lordship'.

Textually a mediating position can be attributed to ***uerc***. *Uerc*, denoting a product, is directly conjoined with the process *modȝidanc* as its corporeal manifestation, and indexically it points back to the *uard* in His capacity as its 'creator/protector', too. At the same time, however, it points forward towards all the praiseworthy 'wonders' (*uundra ȝihuaes*), which have come into being. Thus, on the one hand, *uerc* rounds up the first part of the *Hymn*, i.e. the part that invokes God in His Absolute Firstness; the opener is not followed by any verb, i.e. up to here there is no 'action' on God's part; semiotically, this is not at all surprising; Firstness is *per se* not a matter of cause and effect; Firstness is by its very nature static. On the other hand, *uerc* provides the basis for ll. 3b – 9: the Hymn can now proceed to the dynamic point-of-view in naming the processes of Creation.

Uuldurfadur is complementary to *metud*. Whereas the latter invoked the *uard*'s essence as 'Lord', *fadur* denotes the 'Protector'.

As to *uuldur*, it comes close to the status of a near-synonym of 'God', implying the association of 'heaven'. Thus it refers back to to the whole introductory phrase *hefaenricaes uard*. On the other hand, via its use for 'splendour on earth', 'splendour of the universe', 'glorious things to be enjoyed by man', it points forward towards what is going to be praised in the remaining lines of the Hymn. The same textual-semiotic status of both pointing backwards and forwards was diagnosed for the sign *uerc*, too. This indexalic equivalence makes *uuldur* a perfect match for *uerc*, with which, in addition, it makes up one syntactic collocation in the Hymn.

It is beyond the scope of this paper, to set the preceding semiotic findings into relief with prevailing theological interpretations of *CH*. Therefore, one reference will have to suffice here: Huppé (1959: 111) argues in favour of a patristic interpretation of the triad *maecti*, *modȝidanc*, and *uerc*:

> In the representation of the Trinity through the creation, God the Father is the Power or Might, the Son is the Shaping Wisdom, the Holy Ghost, the Perfection of the Work…The three phrases [i.e. triad *maecti*, *modȝidanc*, *uerc*] reflect the traditional division of the Persons

> of the Trinity as they are revealed in Genesis: the Might of God, the Creator, would represent the Father; the Thought of the Father, His plan and disposition of creation, the Son; the Work, the Holy Ghost.

In many respects our semiotic analysis above ties in with that theological pattern of the Fathers' reasoning (with which Bede had, definitely, been familiar). Conclusion: if this is what had been on the back of the composer of the Hymn's mind, both his lexical choices and their textstructural/textsemiotic use in the native Anglo-Saxon tongue deserve unlimited appreciation and cannot but have been deliberate; if, however, the alleged cowherd–composer had been unaware of these –in his days- patristic "commonplaces",[22] Bede's admiration for him is justified the more.

5.2.2 Ll. 3b – 9: uundor – eci – dryctin – scop – aelda barn – eben – haleȝ – scepen - middunȝeard - moncynnæs uard – folde – frea allmectiȝ

In Peircean terms, Creation is a sign of Secondness *par excellence*: it is constituted by a relation between cause and effect. Here God is cause, is 'agens', Creation is effect, is 'patiens'.[23]

Ll. 3b – 4 fulfil what has been indicated by l. 3a: they present God as having 'set the beginning' (*or astelidæ*) of *uundra ȝihuaes*. Like *uard,* that turned out to be the semiotic hinge for ll. 1 – 3a, *wundru*, initiating the second part, serves as the hinge for all the processes and products of Creation in the remainder of the Hymn. In its wider sense ***uundor*** can designate both very positive matters and events and absolutely negative ones, too. In this respect it links up with *mod*, i.e. with the first part of the Hymn, where God's essence as being beyond human categorizing and evaluating is presented. Textsemiotically *wundru* is a continuation of *uerc*, pointing both backwards and forward in the text: backwards via that relation to *mod*, forward by means of its introducing God as creator, i.e. God in his Secondness. Thus, like *uerc* in the same line, *wundru* is semiotically middling between both parts of the Hymn, ll. 1 – 3a on the one hand, and ll. 3b – 9 on the other hand. It follows: God in His Secondness, i.e. as the creator, is presented under the auspices of God in his Firstness.

It is only as late as in the second half-line of l. 4, that God's Secondness starts being revealed. This is done under the auspices of temporality: *or astelidæ*. Beginning associates ending, too: Creation is transitory by nature.

In l. 5a, *aerist* follows suit, and right in l. 5b that notion of temporality and transitoriness is intensified by the choice of ***aelda barnum*** to denote Mankind here; via the word-family of *aelda* (*ielde* 'people'; *ield* 'period of time, duration of life, age', *ealdor* 'life'; *yldran, ældrean* 'parents') and via the relation of *barn* 'to bear', the 'born' one, the 'child', too, the cycle of coming into being and passing away from being, the transitory nature of human being, is associated. It is within the very same phrase that God is set apart from any transitoriness: the ***dryctin***, the 'Lord of the followers' is ***eci*** 'eternal'.- The re-occurrence of *eci dryctin* in l. 8 is a reminder. In l. 4a *eci dryctin* causes the first phase of Creation, in l. 8a *eci dryctin* is rounding up its last phase: ***æfter tiadæ*** 'he finally he adorned it'. The semiotic impact of this resumption is to reemphasize God's eternality – as opposed to the old pagan Germanic deities, who had been mortal.

It is only after the general presentation of God as the originator of everything, as well as after the first appearance of Mankind in the Hymn (l. 5b), that He can be referred to as ***haleʒscepen*** ('holy creator'): the quality of *eci*, as in *eci dryctin*, is independent of man, even contrasts with the nature of man.[24] The attribution of a quality like *haleʒ*, however, admits of human judgment.

And it is human existence, too, that might well account for the seemingly tautological construction '***scop*** + scepen' in ll. 5f. Textstructually one should mind the iconicity: '*aelda barnum*' is embraced by the process of creating ('*scop*') and its originator ('*scepen*'); once again: the *uard* in His essence of creator as the protector of Mankind. Furthermore, a corpus-analysis of *scop/scepen* reveals many instances where their general meaning 'to create, to produce; Creator, producer' is specifically qualified as to the notion of 'forming, gaining shape; shaper'. In early Anglo-Saxon poetic diction *scop* even is occasionally profiled as far as to, 'to assign', 'to determine', and 'to destine'. If we concede this shade of meaning here, it greatly impacts on the semioticity of *CH*: with regard to the close syntactic contiguity of *scop/scepen* to Mankind in ll. 5f. and with a view to the semioticity of 'giving' as the overall pattern of the Hymn, ll. 5f. are the first explicit indication of what is going to be made clearer and clearer up to the ending of the Hymn: what the *scepen* had created was not mere Creation as such (='Secondness'), but 'functional Creation' with a view to the weal of Mankind (and thus for the first time in the Hymn pointing out to 'Thirdness').

However, there are two caveats: the giving *scepen* is ***haleʒ***. Wietelmann (1952: 6-17) convincingly argues that in Cædmon's time its sense oscillated between *mysterium fascinosum* and *mysterium tremendum* – a finding that semiotically links up with our analysis of *modʒidanc* above, i.e. with Cædmon's basic admonition to praise God in his Firstness. Secondly, in the 7th century *haleʒ* might still have conveyed the old connotation of 'perfection', here as setting God apart from the *aelda barnum*, who must needs be the contrary.

The designation of God as ***moncynnæs uard*** in l. 7b could well be interpreted as a third reminder (caveat) to be aware of God in His Firstness when praising Him in His Secondness, i.e. in His capacity as the creator. Evidently, this phrase refers back to His designation at the very outset of the Hymn as *hefaenricaes uard*. Textsemiotically, however, the two phrases are not at all simply variations of each other (as purely philological assessments of *CH* hardly ever recognized). L. 1b is a sign of God's Firstness in His incogitable Heavenly Kingdom. As such it cannot but show up in no interdependence with Mankind whatsoever. By having come to l. 7b, however, God has by now been revealed in His Secondness as the creator, and even the perspective to His Thirdness as the giver, has meanwhile been indicated by *aelda barnum*. And in l. 7b it is this very specification of His *uard*ship as to that part of Creation that is Mankind, that makes the syntagma *moncynnes uard* the semantically and semiotically appropriate expression at that very place within the entirety of the Hymn. Its reference to Mankind there (once again) is an index that points both backwards and forwards. Backwards it resumes the recipients of God's gift, *aelda barnum*, serving as an intensifier.[25] Forwards it paves the passage towards the immediately following ll. 8-9. And it is these last lines that finally consummate the passage from the Creator-God in His Secondness as presented in the middle, semiotically middling part of the Hymn towards His Thirdness as the Giver-God

CH employs three lexemes to present those biblically attested products of God in his pure Secondness: *eben, middunȝeard, folde.* What all three of them stand for is actual existence, is the facticity as attributed to Gen I.1: *In principio creavit Deus cælum et terram.* It is verifiable: the Hymn here ascends from being rhematic to being dicentic.

Due to *til* the metrical structure of l. 6a ***eben til hrofe*** strikes the listener's ear as being a specific one within the Hymn. This insertion of *til* was required because of *hrofe* ('roof'). 'Roof' is not part of Gen I.1. Why then did it get into *CH*? Patristic literature - on which Bede could draw- distinguished between eternal Heaven and Temporal Heaven, the former one being created first and unseen by man, the latter one being created afterwards for Mankind. It may be the case, as argued by Huppé (1959:113ff.), that with *uundra ȝihuaes...or astelidæ* the Hymn evokes the association of that primeval eternal Heaven and with *eben* it refers to the temporal one. There is semiotic evidence for that view: *eben* is conjoined with the first appearance of a sign that denotes Mankind, *aelda barnum*, and *eben* is semanto-syntactically squeezed in between *scop* and *scepen*, both of which do not merely stand for 'coming into being of something new' but for functional Creation, i.e. Creation with a view to Mankind. Bede will have been familiar with metaphors like 'House' or 'Room', applied by the Early Fathers when expounding 'Heaven'. Thus the Hymn might well have drawn upon that metaphor of the *hrof*: it is an icon that represents Heaven's shielding, protective quality for Mankind. What God created on the first day of Creation came into being with the very prospect of events on the 6th day, the Creation of Mankind itself.[26]

The semantic feature of temporality of the non-eternal heaven ties in with semantic features that qualify the lexeme chosen to designate 'earth' at this very place in the Hymn: ***middunȝeard.*** It implies the feature 'inhabited by Mankind', which it had denoted in Germanic pagan mythology, and which it may have connoted still in the 7th century. On that ground *middunȝeard* is the perfect complement for *eben* in its function as *hrof.*

The third reference to elements of Creation that can substantially be experienced is ***foldu***, the accusative form of *folde*. It primarily denotes earth in its material substance, but does not have 'inhabited' as one of its necessary semantic features (*vs. middunȝeard*). But a corpus analysis of Old English poetry indicates that it shares with *middunȝeard* rather more positive than negative connotations. It lends itself for referring to a dwelling-place, often occurs with a vegetative component (thus associating Creation), and even lends itself to designate 'paradise' and it is in that final line of the Hymn that Cædmon relates Earth to Mankind even *expressis verbis*: via the sign *firum.* This latter sign designates Mankind in a wide sense, and frequently implies the specific relationship between

Mankind and Creation, which grants to man the prime role among all living beings. In that extolled position as the concluding line of the Hymn undoubtedly is, Cædmon literally zooms in on the benefactory function of Creation for Mankind: both lexemes are presented in direct contiguity (*firum foldu*). Each one of these two signs *per se* is a sign of Secondness, each one refers to objects that are effects of a cause. Their immediate contiguity, however, makes these two signs one unique sign of a higher order: *foldu* as the gift to *firum-* and as pointed out above, semiotically 'giving' is a genuine sign of Thirdness- and as such the relationship between the giver, the recipients and the gift itself- is not a mere existential fact but has the status of a generally holding law.

L. 9, being the architectural keystone of the Hymn, extends this contiguity by naming the giver without any linguistic filler between him and the preceding signs for the receivers and for the gift itself*: firum foldu,* ***frea allmectiʒ.*** The designation *frea*, used here for God, and as being modified in particular by *allmectiʒ*, evokes the same fundamental values as *dryctin*, such as leadership and of loyalty, allegiance, and togetherness. But it reaches beyond them in two respects. Corpus analysis shows that it puts a stronger emphasis on the supremacy of God's position – in this line, in relation to both Mankind and Creation. In addition it may well have connoted semantic features that had still been denotative ones with contemporaneous cognate lexemes of its word-family, such as *fruma*, in which two features amalgamated: 'ruler, leader, king' and 'beginning, origin, cause, creation, creator, founder'.

If *frea* is a sign, which essentially features supremacy and creativeness, its modifier *allmectiʒ* is its perfect semiotic match. It is obvious that it links up with *maecti* in l. 2a. From what had been observed there about this word family, *allmectiʒ* highlights the supremacy of *frea* qua giving emphasis to his *potestas*, and furthermore, it indicates that the *frea* exerts his supreme power to the benefit and protection of Mankind. In addition, it lends itself for an emphatic conclusion of the Hymn: in Anglo-Saxon poetry *allmectiʒ* is reserved to God exclusively![27]

This final burst of praise in l. 9 thus links up with the exuberant exhortation in the introductory line 1a. Looking for a metaphor to cover the semiotic structure of the Hymn in its entirety, once again, a 'ring' would lend itself as an icon to depict its rhetoric. *CH* is not a 'Song of Creation', not a 'Schöpfungshymnus', but a 'Song of the Creator', a 'Hymnus auf den Schöpfer', who, like the ring, has no beginning and no ending.

6. Results: The Hymn as a Dicentic Symbol

It was the aim of the preceding analyses to come to an assessment of *CH* as a whole when reviewed from a semiotic stance. For that purpose the semioticity of the interconceptual relationships, that hold between God, Creation, and Mankind in the Hymn was studied. This was done from a textstructural point-of-view, that was complemented by a likewise semiotically oriented lexicological analysis.

It turned out that the results of the latter investigations supported the ones gained from the former ones to an unexpected degree. In particular as to the inner cohesion of the Hymn in its entirety the lexis chosen by the composer reinforced the textual structure.

Semiotically the structure of the Hymn in its entirety presents God in each one of the three phenomenologically possible modes:

(1) His simple (though impenetrable) essence in itself, i.e. semiotically God in His 'Firstness'
(2) His being creator of the actually existing and verifiable objects of creation, i.e. semiotically His 'Secondness'
(3) His being represented in Creation in such a way as to cause the interpretation, according to which it lawfully holds that it was God's will to make Creation serve Mankind, that He intended Creation to be His gift to Mankind., i.e. God in His 'Thirdness'.

These three modes are hierarchically structured: (1) being the simplest one, (3) being the most complex one, (2) middling between both. And the textual structure of *CH* mirrors this very sequence to a nicety. It ascends from (1) via (2) to (3).

The three modes of God's revelation in the Hymn are not, however, piled upon each other as isolated bricks. Instead, the transitions from (1) to (2) and from (2) to (3) are of a processual kind with semiotic markers of overlapping to mark the gradation.

And this is where the lexical elements as chosen by the composer reveal a textlinguistic potential that was unexpected by us. All interpretations and translations of *CH* which we have come across, use to concentrate on its key lexemes on word-semantic, on literary, on cultural-historical, etc. grounds exclusively. Hardly any attention has been paid to potential textstructural significance, let alone from a semiotic point of view. Our projection of the lexicological analysis onto the preceding textstructural one, however, reveals right through the entire text that lexemes are chosen in such a way as to ease the passages from the presentation of God in His Firstness toward God in His Secondness and from there toward His Thirdness. They are the joints, whose cohesive power achieves coherence across (1) to (3). This is not always achieved by means of their denotative meaning, but rather by connotative-associative components of meaning that might well have still been alive in the 7th century.

In its entirety CH can be seen as a ring. Its conclusion in ll. 8 - 9 shows God as a 'giver' in His Thirdness, but simultaneously it resumes God in His Firstness, which is the mode in which ll. 1 –3a introduced Him in the beginning.

Like the ring, God's eternal essence sharply contrasts with the transitoriness of human being. This contrast is mirrored by the opposition eternality vs. temporality, that pervades the entire Hymn.

The metaphor of the ring indicates, too, that the overall patterning of the Hymn is highly iconic. The textual as well as the lexicological-semantic distribution bear witness of this iconicity.[28]

Semiotically viewing the Hymn as an icon, however, can imply that it is an instantiation of an index, a *replica*. And textual evidence shows, that, actually, it *is* an appeal, i.e. a sign that directs the listeners' minds (index!) to praise God (1st section of the Hymn) on the basis of their experience of Creation (2nd section of the Hymn), which God from the very outset of His act of creating designated to be His gift to Mankind (3rd section of the Hymn). In calling up in the listeners' minds he image of the Creator which it suggests to the minds, the Hymn as an entirety acts upon a *symbol* already stored in the minds. This very stored symbol is *God's Creation for Mankind*. Thus the Hymn as an index is a replica of the Hymn as a symbol, just like the Hymn as an icon is a replica of the Hymn as an index.[29] In its relation to the interpretant this symbol can be verified on the basis of the experience of Creation, i.e. it is dicentic.

Summing up, the semiotic analyses above support the semiotic hypothesis according to which *CH* is a Dicentic Symbol.

With regard to the overall textstructural semiotic patterning, held together by its lexis in a supporting function, furthermore with regard to the encompassing iconicity - as a *replica* of its indexality-, in turn being a *replica* of its symbolicity, with regard to the result of the preceding analysis, that revealed that all of these properties cause all individual signs of the Hymn to relate to each other is such a way as to make it one unitary sign –we, therefore, name it a 'Hypersign'- we leave it to the reader to side with whatever assessment she/he considers as being appropriate among the assessments quoted to in part 2 above.[30]

Notes

[1] Actually, a misnomer as we shall see below.

[2] As a detailed discussion concerning (a) place of origin of *CH*, (b) its exact date of origin, (c) the relationship among the extant 17 manuscripts, in which *CH* has come down to us, and (d) the relationship between the Anglo-Saxon and the Latin versions is out of place here, we refer the reader to the state of the art of research as summarized in Kühlwein (2006a: 101f.): specifically there in detail as to (a): p. 101 FN 1; as to (b) p. 101 FN 2; as to (c) p. 101 FN 3; as to (d) p.102 FN 4. - As to (d) we strongly side with the sequence 'Anglo-Saxon version first, translation into Latin next' (*vs.* vice versa!) for manuscript-based reasons given there.

[3] Due to the limits imposed by translating a text, in particular a poetic one that is nearly one and a half millennia remote, this translation into Modern English prose merely serves as a preliminary working- translation; it should be one supererogatory consequence of the following analysis to lay bare its shortcomings.

4 The problem with this evaluation and some of the following ones is, that these parallels are rightly observed; however, as no pre-Cædmonian Old English Christian literature has come down to us, the question whether Cædmon' use of poetic language *traditionally* drew on predecessors or extant authors of later date drew on him, remains unresolved. Furthermore, as to the explanatory strength of comparative frequency counts of Anglosaxon oral-formulaic poetry, the constraints imposed by metre and alliteration will have exerted much less influence upon thematic and lexical choice than had previously been asserted (for evidence c.f. Creed 1959, Greenfield 1972, Fry 1974 and 1979, Miletich 1983, O'Keeffe 1987, Kühlwein 2006b).

5 Strangely enough, hardly ever quoted in subsequent linguistic and/or literary critics.

6 For some less concise outlines cf. Hervey 1982: 17-37, Nöth 2000: 33-46.

7 As a result, this conception would have to cause editors' punctuation to have a comma or at best a colon following *astelidae*, because all that follows would cognitively merely serve the purpose of a continuation, which specifies the preceding 'thus' or 'how'.

8 *"I think we must regard Creative Activity as an inseparable attribute of God" (Peirce 6.506).*

9 Punctuation in this case would better have a semicolon or even a full stop following *astelidae* and thus give special causal emphasis to the acts of His creative doing as following in ll. 5 – 9.

10 The adequate punctuation indicating that division is either a semicolon or even a full stop after *uard* in l. 7.

11 For the semioticity of *tiadae* as not simply to be interpreted as a pure synonym of *scop* 'created' but beyond that as 'creatively adorned', as had been proposed by J. B. Bessinger Jr as early as in 1974 cf. Kühlwein (2006 b: 76).

12 This very consistency of that pattern may well be taken as an argument in favour of the punctuation followed by us. It is on these grounds that we reject a punctuation that closes l. 6 with a Semicolon, and separates half-line 7b from 8a by a mere comma, as some editors do, e.g. the Aelfredian version as taken from the Tanner text (MS Bodleian Tanner 10) as edited by Krapp in ASPR VI, 105f. or the punctuation offered e.g. by www.heorot.dk/bede-caedmon-i.html, by www.georgetown.edu/-labyrinth/library/oe/texts/a32.1.html (for the Northumbrian version) and www.georgetwon.edu/labyrith/library/oe/texts/a32.2.html (for the West Saxon version likewise). That latter punctuation would move *tha*

middunȝeard to a front position with relation to the two following references to its Creator (*moncynnes uard* and *eci dryctin*), i.e. Creation – and not the Creator - would syntactically become the topic, the two following references to the Creator (*moncynnæs uard* and *eci dryctin*) would become comment. Such a view would run counter to (1) the overall semioticity of the Hymn as well as (2) to the semiotic gist of Bede's narration that encompasses it in the way as semiotically shown in Kühlwein (2006a: 112ff.), furthermore (3) it would disregard the fact that in all other cases at least one reference to the Creator, actually, precedes the respective reference to Creation (or some part of Creation), and (4) finally, the immediacy of the ensuing conjoining of the two epithets for God (*moncynnæs uard*, directly followed by *eci dryctin*) within what then would have to be seen as the same syntactic sequence, would not at all be in line with the poetic way in which the balance between topic and comment is achieved in all other parts of the Hymn. [As a result the syntactic sequence would have been fairly odd: *"… holy creator**;** the inhabited earth, *the protector of Mankind, the eternal Lord* afterwards adorned, for Mankind the ground, the almighty Leader"].

13 By strength of this kind of relationship God and Creation, actually, determine each other: God enforces the existence of Creation, which in turn, makes Him the Creator.

14 Peirce himself provides two examples from thematically related fields for what he calls a *dyad*. One example is *Genesis* I. 3 "*Dixitque Deus: Fiat lux. Et facta est l*ux" ('God said, Let there be light, and there was light.'): "We must simply think of god creating light by fiat. Not that the fiat and the coming into being of the light were two facts; but that it is in one indivisible fact. God and light are the subjects. The act of Creation is to be regarded merely as the suchness of connection of God and light. >The dyad is the fact…..pure dyadism is an act of arbitrary will" (Peirce: 1.327f.).- For a second example Peirce (2.316ff.) draws on the proposition 'Cain kills Abel', which likewise has two subjects, "Cain" and "Abel"; though it relates to the real existence of either one of them, it nevertheless "may be regarded as primarily relating to the Dyad composed of Cain, as first, and of Abel, as second member. This Pair is a single individual object having this relation to Cain and to Abel, that its existence *consists* in the existence of Cain and in the existence of Abel and in nothing more. The Pair….. [is] just as truly existent as they severally are."

15 Even if one includes the two references to 'Mankind' within the ones for 'Creation' the ones to 'God' still outweigh that sum.

16 For a detailed analysis s.b. 5.1.3.

17 Gen II. 7: "Formavit igitur Dominus Deus hominem de limo terrae…" (Biblia Sacra Vulgata). [And the LORD God formed man of the dust of the ground…] (King James Bible).

18 Cf., however, the next semiotic level as approached in 5.1.3, where Creation will be looked at under the perspective of God's intention, revealing God in His Thirdness.

19 Space forbids to enlarge upon corpus-based evidence outside *CH* in detail. For the base of statements concerning the semanticity (including possible etymological and connotative rings) of the key-lexemes, whose semioticity will be discussed below in 5.2, we refer to Kühlwein (2006b: 70-85), that is based on the entire corpus of Old English poetry.

20 "Because compulsion is essentially *hic et nunc*, the occasion of the compulsion can only be represented to the listener by compelling him to have experience of that same occasion. Hence it is requisite that there should be a kind of sign which shall act dynamically upon the hearer's attention and direct it to a special object or occasion. Such a sign I call an Index. It is true that there may, instead of a simple sign of this kind, be a precept describing how the listener is to act in order to gain the occasion of experience to which the assertion relates." (2. 334f.)

21 Cf. e.g. Beowulf, Nibelungenlied, Scaldic Northern sagas.- On the other hand, an opener in the 1st pers. sing., as met in Old English elegies would have been inappropriate: Cædmon's personal feeling is the ground, but not the theme; he merely is the angel's transmitting instrument.

22 Huppé 1959:111.

23 This relationship is syntactically mirrored: whereas in ll. 1 – 3a the references to God are in the genitive or accusative, in ll. 3b – 9 all of them are in the nominative.

24 Cf. Wietelmann (1952: 23): "Das Verhältnis zwischen der Ewigkeit und der Zeit gleicht dem Abstande zwischen dem Schöpfer und seinen Geschöpfen, d.h. zwischen Gott und den Menschen."

25 As to intensification (vs. logical precision) as a major characteristic of Anglo-Saxon poetic diction cf. generally Kühlwein (1967: 42ff.).

26 Cf. 5.1.2 above.

27 A certain amount of semiotic significance has to be attributed, too, to the fact, that amongst all signs which the Hymn uses to designate God, this final occurrence of such a sign is the only one, where it has a postmodifier (*allmectiȝ*). A metrical analysis that employs the subtlety of John C. Pope's studies of Anglo-Saxon poetic rhythm reveals, that the collocation in its

form '*frea + allmectiȝ*' is enjoying a degree of metrical emphasis unparalleled in the entire hymn.

28 For evidence cf. the Figures 1 – 3 above.

29 "…the most perfect of signs are those in which the iconic, indicative, and symbolic characters are blended as equally as possible." (Peirce 4.448).

30 Our own vote would be in favour of (i), (f), (g) in that sequence.

References

Texts and Works of Reference

[Bede]. Miller, T. (ed.) 1891, repr. 1959. *The Old English Version of Bede's Ecclesiastical History of the English People.* EETS, Original Series No.96. London, New York, Toronto: Oxford University Press. 2 vols.

[Bede]. Colgrave, B. and R.A.B. Mynors (eds.) 1969. *Bede's Ecclesiastical History of the English People.* Oxford: Clarendon Press.

Krapp, G. Ph and K. van Dobbie (eds.) 1931 – 1953. *The Anglosaxon Poetic Records.* 6 vols. New York: Columbia University Press.

www.heorot.dk/bede-caedmon-i.html

www.georgetown.edu/labyrith/library/oe/texts/a32.1

Biblia Sacra Vulgata(http://www.biblegateway.com/versions/)

The Bible, King James Version. Old and New Testaments, with the Apocrypha.

The Electronic Text Center, University of Virginia Library (http://etext.virginia.edu/kjv.browse.html).

Toller, T. N. (ed.) 1898, repr. 1964. *Bosworth, Joseph. An Anglo-Saxon Dictionary.* Oxford: Oxford University Press/London: Humphrey Milford.

Toller, T. N. (ed.) 1921. *An Anglo-Saxon Dictionary Supplement. Oxford: Oxford University Press.*

Grein, C.W.M. et al. 1912; 2nd ed. repr. 1974. *Sprachschatz der angelsächsischen Dichter.* Heidelberg: Universitätsbuchhandlung Carl Winter.

Further Works of Reference

Bessinger, J. B., Jr. 1974. 'Hommage to Caedmon and Others: A Beowulfian Praise Song', in Burlin, Robert B. and Edward B. Irving, Jr. (eds.), *Old English Studies in Honour of John C. Pope,* Toronto and Buffalo: University of Toronto Press. 91 – 106.

Creed, R. P. 1959. 'The Making of an Anglo-Saxon Poet', *ELH,* 26: 445-454.

Fry, D. K. 1974. 'Cædmon as a Formulaic Poet', *Forum for Modern Language Studies,* X : 227-247.

Fry, D. K. 1979. 'Old English Formulaic Statistics', *In Geardagum*, III: 1-6.

Gardner, J. 1975. *The Construction of Christian Poetry in Old English*. Carbondale and Edwardville: Southern Illinois University Press.

Greenfield, S.B. 1972. *The Interpretation of Old English Poems.* London and Boston: Routledge and Kegan Paul.

Hervey, S. 1982. *Semiotic Perspectives*. London: George Allen & Unwin.

Huppé, B. F. 1959. *Doctrine and Poetry*: Augustine's Influence on Old English Poetry. New York: State University of New York.

Kane, G. 1948. 'Review of C.L. Wrenn, *The Poetry of Cædmon*. London.1947', *MLR* XLIII : 250-252.

Kühlwein, W. 1967. *Die Verwendung der Feindseligkeitsbezeichnungen in der altenglischen Dichtersprache*. (Kieler Beiträge zur Anglistik und Amerikanistik vol. 5). Neumünster: Karl Wachholtz.

Kühlwein, W. 2006a.., 'Bede's Narrative on Cædmon: A Semiotic Analysis', in: Cho, See-Young – Erich Steiner (eds.) *Information Distribution in English Grammar and Discourse and Other Topics in Linguistics.* Festschrift for Peter Erdmann on the Occasion of his 65th Birthday. Frankfurt / M.: Peter Lang, 99-124.

Kühlwein, W. 2006b. 'The Semioticity of God, Creation, and Mankind in Cædmon's Hymn', in: Rösel, Peter (ed.). *English in Space and Time. Englisch in Raum und Zeit.* Forschungsbericht zu Ehren von Klaus Faiß. Trier: WVT Wissenschaftlicher Verlag Trier, 60 – 89.

Magoun, F. P. Jr. 1955. 'Bede's Story of Cædmon: The Case History of an Anglo-Saxon Oral Singer', *Speculum*, XXX: 49-63.

Miletich, J. S. 1983. 'Old English "Formulaic" Studies and Caedmon's Hymn in a Comparative Context', in: in Matešić Josip and Erwin Wedel (eds.) *Festschrift für Nikola R. Pribić*. Neuried: Hieronymus: 183-194.

Mitchell, B. 1967. "'Swa" in Cædmon's "Hymn" line 3', *Notes and* Queries n.s., XIV: 203-204.

Nöth, W. 1985, 2nd ed. 2000. *Handbuch der Semiotik.* Stuttgart: Metzlersche Verlagsbuchhandlung.

O'Keeffe, K. O'Brien 1987. 'Orality and the Developing Text of Cædmon's Hymn', *Speculum*, LXII : 1-20.

[Peirce, Charles Sanders] 1931- 1958, 3rd printing 1974. Hartshorne, C. and P. Weiss (eds.) *Collected Papers of Charles Sanders Peirce.* Vols. 1-6; Hartshorne, Charles – Paul Weiss (eds.) ; A. W. Burks (ed.), vols 7-8. Cambridge, MA: Belknap Press of Harvard University Press.

Pope, J. C. 1942. *The Rhythm of Beowulf.* An Interpretation of the Normal and Hypermetric Verse-forms in Old English Poetry. New Haven: Yale University Press.

Smith, A.H. (ed.), 1933, 2nd ed. 1968. *Three Northumbrian Poems: Caedmon's Hymn, Bede's Death Song, and The Leiden Riddle.* London: Methuen.

Wietelmann, I. 1952. *Die Epitheta in den "Caedmonischen" Dichtungen*. PhD Diss. Göttingen [typescript].

Wrenn, C. L. 1947. *The Poetry of Cædmon*. London.

Wrenn, C. L. 1968. 'The Poetry of Cædmon' in: Bessinger, J. and S. Kahrl (eds.), *Essential Articles for the Study of Old English Poetry.* Hamden: Connecticut, 407-427.

Traditional grammar and corpus linguistics
'*with critical notes*'

David A. Reibel

Tübingen and York[1]

Abstract

After Robert Lowth published his A Short Introduction to English Grammar: With Critical Notes *in 1762, no one who took a serious interest in the subject could not have seen that he had changed the definition and practice of this subject forever. The purpose of this study is to show how he did it.*

I defend Lowth against the oft-levelled charges of lack of grammatical competence and acumen, arbitrariness, and disregard for usage; above all, for his desire to 'regulate' the English language by prescribing arbitrary rules, which would at the same time proscribe errors.

He is shown as highly competent in the field of grammar and literary criticism, and displays considerable originality, ingenuity and skill in the fashioning and application of his rules, based on the meta-principle of Strict Construction. Far from imposing a Latinate grammar on English, he sought to eliminate, among other constructions, the non-native Latinisms, imported into English during the English Renaissance (1550-1660) through the medium of the Periodic Sentence. He also judged improper those native English syntactic forms which also violated the principle of Strict Construction. In this regard he represented the 18thC purist view of English that replaced the looser construction of earlier generations with a more refined, more construable prose, epitomized by Samuel Johnson.

Lowth is far from perfect, and neither is his English Grammar, but most present-day critics write about myths and inventions of their own, instead of studying Lowth's life and works for what they represented to the scholars and educated classes of his day, who regarded him highly as a respected officer of the Church and a distinguished man of letters.

> Ask, and it shall be given you;
> seek, and ye shall find;
> knock, and it shall be opened unto you.
>
> For everyone that asketh receiveth;
> and he that seeketh findeth;
> and to him that knocketh it shall be opened.
>
> Jesus, *Sermon on the Mount*
>
> Matthew 7:7-8; Luke 11:9-10

I shall only remark here, how easily men take upon trust, how willingly they are satisfied with, and how confidently they repeat after others, false explanations of what they do not understand.[2]

> *Dans les champs de l'observation,*
> *le hasard ne favorise que les esprits préparés.*
> 'In the field of observation, chance favours only the prepared minds. '
> *Freely:* "In the empirical sciences, only prepared minds are favoured by chance discoveries."
> — Louis Pasteur (1822-1895), French chemist and biologist. Address given on the inauguration of the Faculty of Science, University of Lille, 7 December 1854.

Sat. 16 [June 1770]— … In the afternoon I looked over Dr. Priestley's English Grammar. I wonder he would publish it after Bishop Lowth's.[3]

Preface

In this little *jeu d'esprit*, I defend Robert Lowth against the oft-levelled charges of lack of grammatical competence and acumen, arbitrariness, and disregard for usage; above all, for his desire to 'regulate' the language, i.e., set up rules for it (cf. Latin *regula* 'rule'), to prescribe English usage by arbitrary rules, which would at the same time proscribe errors.[4]

He is shown as highly competent in the field of grammar and literature, and displays considerable originality, ingenuity and skill in the fashioning and application of his grammatical rules. Far from imposing a Latinate grammar on English, he sought to eliminate, among other constructions, the non-native Latinisms, imported into English during the English Renaissance (1550-1660), that, as he thought, rightly or wrongly, disfigured the language, especially of the earlier generation of post-Restoration writers, even the most eminent. He also judged improper those native English syntactic forms that violated the principles of Strict Construction. In this regard he represented the 18thC purist view of English that replaced the looser construction of this and earlier generations with a more refined, more construable prose. Samuel Johnson epitomizes this carefully crafted new prose style, based on the periodic sentence.

Lowth is far from perfect, and neither is his *A Short Introduction to English Grammar: With Critical Notes* (1762), but most present-day critics, from the depths of their abysmal ignorance of what Lowth actually says and does, and their *a priori* prejudices and lack of analytical understanding, write about myths and inventions of their own,[5] instead of studying Lowths life and works for what they represented to the scholars and educated classes of his day, who regarded him highly as a respected officer of the Church and a distinguished man of letters.

Among the many practitioners of Corpus Linguistics, the name of Robert Lowth (1710-1787) is not likely to be mentioned. But he followed the old-fashioned time-honoured method of collecting examples from a body of literature, probably on file slips made from marginal pencil-markings on the pages of his daily reading-matter, as did Samuel Johnson for his *Dictionary* (1754). 'His temper was generally cheerful, though sometimes irritated by the vexations of office, and the disappointments and provocations of a life of literary popularity. It is said that, like George Stevens and Professor Porson, he never read a book, without a pen or pencil in his hand.' (Hall 1834: 40-41)[6]

Thus the compilation of 'improprieties' or 'inaccuracies' (Preface, 1762:viii) in his *English Grammar* was based just as surely on an open-ended random corpus of texts as any similar present-day compilation,[7] with this important distinctive difference: Lowth had already formulated the general conclusions to be drawn from the examples in his corpus before he ever started on this enterprise. As he says in the 'Preface' to the *Short Introduction*:

> The principle design of a Grammar of any Language is to teach us to express ourselves with propriety ['1. Accuracy; justness.' (Johnson)] in that Language, and to enable us to judge of every phrase and form of construction, whether it be right or not. The plain way of doing this, is to lay down rules, and to illustrate them by examples. But besides shewing what is right, the matter may be further explained by pointing out what is wrong.[8] (1762: x)

In his typical way, when he was commissioned to write this grammar,[9] he saw at once that there was a gap in the coverage of all previous works, and came up with a plan to base the new section on syntax, which he entitles 'Sentences', on a treatment of the faults of English along with the facts.

Thus the corpus consists virtually exclusively of 'improprieties'. How are 'improprieties' to be identified? They cannot come from the lower orders, who do not speak or write standard English, nor from writers known to write in an out-dated style, full of archaisms and similar constructions that have been superseded, nor from those whose writing is said to contain 'inaccuracies', i.e., grammatical solecisms. So the corpus is composed of instances from reputable writers which nevertheless, he says, quoting Swift, ' "offended against every part of Grammar." ' (1762: ii).

The second part of the procedure was to find a grammatical meta-rule according to which the appropriate 'Rules' might be 'laid down'. For this we need go no further than Lowth's definition of 'Sentence', at the beginning of the section on 'Sentences', or syntax:

> A SENTENCE is an assemblage of words, expressed in proper form, and ranged in proper order, and concurring to make a complete sense. (1762: 94)

To understand all this, we need to have recourse to Johnson's *Dictionary*, which gives the 18thC senses of the key words.[10] It should be remembered that the largest grammatical unit recognized from antiquity down to Lowth's day was the *period*, or 'periodic sentence', the universally practised classical sentence-form, from Greek *períodos*, 'meandering road' — not a bad description of the feeling one has when making one's way through one of the longer instances. Here are other senses of 'Sentence' from Johnson:

1. Determination or decision, as of a judge, civil or criminal.
2. It is usually spoken of condemnation pronounced by the judge; doom.
3. A maxim; an axiom, generally moral.
4. A short paragraph; a period in writing.

As in most dictionaries, looking up the meaning of the key terms in a definition can only lead to circularity, as in this from Johnson's list of senses of 'Period':

7. A complete sentence, from one full stop to another.

This is true enough, as long as one knows where and how to place the 'full stops'. Thus Lowth's definition of 'sentence', taken as a whole, must be considered wholly new and original, and, as far as can be determined, not paralleled or repeated by subsequent traditional definitions.[11]

The next two members of the definition, 'expressed in proper form, and ranged in proper order', probably come from Quintilian's *Institutio Oratoria* ('Principles of Oratory'), Book VIII, Chapter ii, § 23, in his definition of *perspicuitas* 'perspicuity', 'clarity': *propria verba*, *rectus ordo*. It is clear from the discussion that follows in the *Institutio* that Quintilian is thinking of *propria verba* as 'appropriate diction', and *rectus ordo* as 'straightforward arrangement'. Lowth has split the sense of propria verba, first, into 'assemblage': '1. A collection: a number of individuals brought together.' (Johnson); that is, not a mere fortuitous, random selection or collection; and second, into 'expressed in proper form'. 'Form' must mean '*grammatical* form', and 'proper', '6. Exact; accurate; just.' (Johnson). So the words must have the correct grammatical or morphological form required by the construction. *Rectus ordo* now means 'ranged (lined up) in grammatically correct order'. Cf: 'To Range. 2. To be placed in order; to be ranked properly…' 'To Rank. 3. To arrange methodically.' (Johnson). So Lowth has taken Quintilian's terms and given them new senses.

Finally, the words must 'concur to make a complete sense.'[12] This is usually misunderstood both by later critics of traditional grammar as well as by its practitioners as meaning that a 'sentence' is *any* assemblage of words that makes (a) complete sense. Or else that in order to make complete sense it must be a grammatically complete sentence. Or that a grammatically complete sentence makes (a) complete sense. This would be Johnson's tenth and last sense of 'sense': 'Meaning; import'. But Lowth means *grammatical* sense: cognate parts of cognate constructions within a sentence must have constituent parts that concur. Forms of words that fulfil identical functions within cognate constituents of sentences cannot have their grammatical form determined locally, but must agree with each other in their grammatical — morphological and syntactic — features across unbounded dependencies. This leads to the principle which I have called 'Strict Construction', which has very wide-spread applicability.

For example, suppose we have a general rule that if a pronoun is the grammatical subject or part of the grammatical subject of a sentence, i.e., of the verb, it must be in the nominative case. Expressions such as 'Us adults are going to have a party' is ungrammatical because 'us', which is part of the subject of the verb, is in the objective and not the subjective case of the first person plural personal pronoun *we* in English. Selecting the form locally, say by some rule that says that only when the pronoun is in absolute subject position directly before the verb *must* it be in the nominative and not the subjective case. Both versions of English grammar agree that it must be: 'We are going to have a party'. No one says 'us are'.

By the same disallowed rule, such expressions as 'Him and me / Me and him went'; 'Me and my brother / My brother and me are twins' — found in all forms of non-standard English, not treated by Lowth or other traditional grammarians until later in the 19thC; cf. the later use of the term 'low expression' — are by the meta-rule of Strict Construction disallowed in Standard English. The rule of local determination says that neither *him* nor *me* is in *absolute* subject position; the grammatical subject in direct construction with the verb is the superordinate NP dominating the conjoined *him-and-me*, etc.[13]

Having set up his criteria and found his texts,[14] Lowth now has to set about writing his grammar. Of the many criticisms levelled at earlier traditional grammarians, none is more critical or crucial than the assertion that they had no qualifications for the job. But Lowth was a 'classic': a man learned in languages: Latin, Greek, Hebrew. From 1741 to 1750 as Professor of Poetry at Oxford, a post that was awarded on the candidate's 'Latinity' — being well-versed in the Latin Language — as much as for any other form of learning. Lowth gave his *Lectures on the Sacred Poetry of the Hebrews* (Latin 1753; English 1787), with the requisite Ciceronian style that has been independently judged by three Latinists at the University of York to be very good and typical.

In Lecture XIX,[15] 'The Prophetic Poetry is Sententious' ('Sententious. 2. Comprising sentences.' Johnson), he finds the solution that had evaded all previous attempts to find the structural basis of the Hebrew poetry of the Hebrew Bible. First he asserts that the basic unit is a sentence, and that it is parallelism of sentences and the (often contrasting) parallelism of their import that is the basic principle.[16] Without so much as a warning, he now uses the technical term 'sentence' in its present-day sense.

> The poetical conformation of the sentences, which has been so often alluded to as characteristic of the Hebrew poetry,[17] consists chiefly in a certain quality, resemblance, or parallelism between the members [clauses] of each period [complete sentence]; so that in two lines (or members of the same period) things for the most part shall answer to

> things, and words to words, as if fitted to each other by a kind or rule or measure. This parallelism has much variety and many gradations; it is sometimes more accurate and manifest, sometimes more vague and obscure: [Lowth-Gregory 1787.II: 34. Analysis of the three kinds of parallelism omitted.]

In discussing the first species of the three forms of parallelism that he identifies, the synonymous parallelism (Lowth 1753: 180; Lowth-Gregory 1787.II: 35) (the other two are the antithetical parallelism, and the synthetic or constructive parallelism), Lowth observes:

> Saepe deest aliquid in posteriore membro, e priore repetendum ad explendam sententiam, [...] (Lowth 1753: 185)
>
> 'There is frequently something wanting in the latter [second] member [clause], which must be supplied from the former to complete the sentence [sense and/or clause]:'
> "Kings shall see him and shall rise up:"
> "Princes [GAP], and they shall worship him;"
> [Isaiah XLIX.7] (Lowth-Gregory 1787.II: 41)

In other words, to complete the 'sentence' (Latin *sententia*) or 'sense' (NB equivocation), the VP of the first line, 'shall see him and rise up', — just two words in the Hebrew — must be interpolated into the second line after (or perhaps before) the subject NP 'Princes', filling the 'gap'. It may fairly be said that Lowth discovered gapping, a distinction in the various mechanisms for shortening consecutive conjoined constituents by deleting repeated terms or constituents, generally credited to Hudson 1976; see also van Oirsouw 1987. In fact, most traditional grammars say something about this process, albeit usually in very general terms.

Lowth is less interested in the grammatical generalization than he is in accounting for the role that it plays in the structure of successive lines of Hebrew poetry.

1. Asymmetrical Conjunction

The best way to illustrate Lowth's method is to present one of his collections of instances of an improper construction, and to set the reader the task of setting up a rule of grammar which, on the face of it, seems an unexceptionably linguistic commonplace, but which can at the same time be used to rule out the assembled instances as violations of it, and, therefore, as 'improper', or ungrammatical.

The 'data' are an assemblage of Lowth's own compilations,[18] taken from various editions of his *English Grammar*. His square brackets, or 'Crotchets', as he calls them, enclose the elided word, which he has supplied. Biographical and bibliographical information has been added in parentheses or square brackets by DAR as well as occasional editorial clarification.

As you read through examples 1 to 11, try your hand at formulating the rule that Lowth formulated and which excludes these expressions or constructions from the canon of grammatical sentences or constructions of English. Formulate also an alternative rule that allows them.

1. Forasmuch as it hath pleased Almighty God of his goodness to give you safe deliverance, and [who] hath preserved you in the great danger of Childbirth:—Liturgy. [*The Book of Common Prayer* (1662); revised edition of the Prayer Book of Edward VI (1549; 1552), where this originates. 'The Thanksgiving of women after Childbirth, commonly called, the Churching of Woman.']
2. If the calm, in which he was born, and [which] lasted so long, had continued. Henry Hyde, second earl of Clarendon (1638-1709), Life (1668-1670; 1672 ff.; published 1759), p. 43.
3. The Remonstrance which he had lately received from the House of Commons, and [which] was dispersed throughout the Kingdom. Clarendon, Hist. (1702-1704) Vol. I. p. 366. 8^{vo}.
4. These we have extracted from an Historian of undoubted credit, a reverend bishop, the learned Paulus Jovius; and [they] are the same that were practised under the pontificate of Leo X. Pope (1688-1744), Works, Vol. VI, p. 201.
5. A cloud gathering in the North; which we have helped to raise, and [which] may quickly break in a storm upon our heads. Jonathan Swift (1667-1745), Conduct of the Allies (1711).
6. A man, whose inclinations led him to be corrupt, and [who] had great abilities to manage and multiply and defend his corruptions. [Swift,] Gulliver (1726), Part I. Chapt. vi.
7. My Master likewise mentioned another quality, which his servants had discovered in many Yahoos, and [which] to him was wholly unaccountable. Gulliver, Part IV, Chap. vii.
8. This I filled with the feathers of birds I had taken with springes [snares] made of horse hairs, and [which] were excellent food. Ibid. Chap x.
9. Osyrus, whom the Grecians call Dionysius, and [who] is the same with Bacchus. Swift, Mechan, Oper. of the Spirit, Sect ii (1704).

Two further examples were added in some edition later than The Second Edition, Corrected (1763):

10. Which Homer might without a blush rehearse,
And [which] leaves a doubtful palm to Virgil's verse.
Dryden (1631-1700), Fables (1700), Dedication.

["The 'Fables' again show Dryden's energy of thought and language undiminished by age." Article on Dryden by Sir Erasmus Henry in *DNB*.]

11. Will martial flames for ever fire thy mind,
    And [will it, thy mind,] never, never be to heav'n resign'd?
    [Pope,] Odyssey, xii. 145.

What would be the first step? Most likely to sort the examples into different *classes* or *types* of construction, with a brief piece of observational analysis. They all seem to involve pairs of conjoined sentences or clauses, with an elided subject in the second clause whose antecedent is some kind of object, often preposed, in the first clause.

> Type 1. A subject RelPn in the second clause is coreferential with an object NP in object position in the first clause: Exs. 1, 11. The two examples are otherwise distinct in construction.[19]
>
> Type 2. A subject NP in the second clause is coreferential with a fronted object NP in the first clause: Ex. 4.

Pairs of Conjoined Relative Clauses:

> Type 3. The subject RelPn in the second RelCl is coreferential with the object RelPn in the first RelCl: Exs. 3, 5, 7, 8 (with elided object RelPn in the first RelCl): Exs. 9, 10.
>
> Type 4. A subject RelPn in the second RelCl is coreferential with a RelPn in a PrepPh in the first RelCl: Ex. 2.
>
> Type 5. A subject RelPn in the second RelCl is coreferential with the possessive RelPn whose in the subject NP of the first RelCl: Ex. 6.

Here is what Lowth says about Ex. 1 in his 'Critical Note' (footnote) (1762: 122-123):

> The Verb *hath preserved* hath here no Nominative Case; for it cannot be properly supplied by the preceding word *God*, which is in the Objective Case. It ought to be, "*And He hath preserved* you;" or rather, "*and to preserve* you."[20] Some of our best Writers have frequently fallen into this [Swift is represented many times], which I take to be no small inaccuracy: … [Here follow the examples above.]

By the term 'supplied', Lowth means no more than that the gap or missing or elided portion of the expression as it stands is to be filled with morpho-syntactically *identical* cognate terms (copies) from the preceding cognate constituents of the overall construction. But this, as he points out, is impossible, because the gap in the second member of the construction requires a term with different morpho-syntactic features from those of its cognate term in the first member.

'Cognate' is to be understood in the appropriate sense: 'coreferential' and/or 'structurally parallel'. Examples 1 and 11 require only that they be coreferential; the others that they be both coreferential and structurally parallel, that is, initial in their syntactic category. But in all of these cases, the principle of Strict Construction has been violated: Morpho-syntactic features of gaps cannot be only locally specified, but must agree with those of their antecedents. The only solution is to restore the elided elements supplied by Lowth in his 'Crotchets', that is, supplying them with their overt local morpho-syntactic features, obviating illegally supplying (copying) them from their antecedents with the wrong morpho-syntactic features.[21]

A diverse range of grammarians have claimed that these constructions are nevertheless indeed ungrammatical, despite the fact that they have been in English ever since OE times. Their tendency to appear almost at random in a wide variety of historical texts is however well documented. (See Visser 1963-1973.) Example 12 is from the story of Cædmon in the OE Bede (*Ecclesiastical History of the English People*). While it is not of the asymmetrical type, it does show the elision of the relative pronoun in the second of two conjoined relative clauses. Further, the number on the gapped relative pronoun is determined locally, singular instead of plural, like its antecedent. Relating how Cædmon employed his gift of poetry, the following statement appears:

12. Ond he forþon næfre noht leasunge ne idles leoþes wyrcan ne meahte, ac efne þa an þa [*neuter plural*] ðe to æfæstnisse belumpon [*plural*], ond [GAP; supply *ðœt ðe* that which: *singular*] his þa æfæstan tungan gedeofanade [*singular*] singan.[22]
13. 'And he for this reason [he had not been taught poetry but had received it as a divine gift] never could compose any falsehoods or idle songs, but those alone which pertained to piety, and [GAP] suited his pious tongue to sing.'

That this is an original OE creation is shown by the Latin original, which is different in construction:

14. Unde nihil umquam frivoli et supervacui poematis facere potuit, sed ea [*plural*] tantummodo quae [*plural*] ad religionem pertinent [*plural*] religiosam eius linguam decebant [*plural*].
15. Whence he never could compose anything (of) frivolous or vain poetry, but only those [things] which pertained to religion were suitable for his pious tongue.

Where the Latin has two conjoined clauses, the second incorporating a relative clause, the OE splits the second clause into two relative clauses.[23]

What is the explanation for this strange state of affairs? The disharmonious case relationships and the asymmetry of the types demands some analysis. In what follows, a very simple form of constituent structure is used heuristically and a configurational pattern is posited as the explanation, without case relationships being relevant. The level that is attempted to be attained is Chomsky's Observational Adequacy.

Let us assume that every time a constituent is preposed to the left of a sentence, a new superordinate sentence node is created, with a gap left behind where the moved constituent comes from, thus:

16. The man S[(whom) S[we invited [GAP] to dinner]]
    < the man S[we invited the man to dinner]

No. 16. shows that when the object NP *the man* is moved to the left, the S-node dominating S[we invited the man to dinner] is expanded into another superordinate S-node with S[whom > the man dominating the lower S-node that now contains a gap: [we invited [GAP] to dinner].

Let us call each type of S-node a 'projection (of S)'.[24]

An independent or else lowest S-node that does not dominate any other S-node whether or not it is dominated by another S-node is a *minimal projection*. A superordinate node that dominates an S-node and is not dominated by another S-node is a *maximal projection*. S-nodes that dominate S-nodes and in turn are dominated by S-nodes are *intermediate projections*. In this way chains of minimal, intermediate, and maximal projections of S can be built up. (Intermediate projections do not play a role in this analysis.)

Now, in the expression, *the man didn't come*, the S-node dominating it is a minimal projection of S, because it does not dominate any other S-node.

Gapping of the second relative pronoun in a pair of conjoined relative clauses occurs when an antecedent relative pronoun invades the second of two conjoined S-nodes looking for a coreferential node to delete. It is a kind of search-and-destroy mission. But it can only destroy coreferential nodes that are in parallel or cognate positions in configurationally similar S-nodes.

These conditions are met in the first, acceptable construction, *The man we invited to dinner but didn't come*.

The head NP of the whole NP, *the man*, has a pair of conjoined relative clauses dominated by an S-node, as a post-modifier. Restoring the elided preposed object relative pronoun in the first relative clause, we have *the man whom we invited* [GAP] *to dinner. Whom we invited* [GAP] *to dinner* is a maximal projection. It has the structure:

17. S[whom S[we invited to dinner]].

Now the *whom* sets off on its search-and-destroy mission in the second, conjoined relative clause *who didn't come*, which is a minimal projection. It is a maximal projection only by default, because it does not dominate any other S-nodes. The object relative pronoun *whom* can destroy the subject relative pronoun *who* in the second relative clause because they are both initial in their syntactic category and are coreferential. The fact that *who didn't come* is not a maximal projection (except by default) is irrelevant: the pronouns are in the same initial position with no superordinate S-node. If this laborious deduction is correct, it confirms that a configurational account is acceptable.

Now compare this with the situation in the ungrammatical *the man who came to dinner but* [GAP] *we didn't invite* [GAP]. The first relative clause is, as we have stated, a minimal projection. It is a maximal projection only by default, because it does not dominate another S-node. The subject relative pronoun *who* in the first relative clause now sets off on its search-and-destroy mission, looking for a coreferential subject relative pronoun in a minimal projection of S in the second relative clause S[whom S[we didn't invite [GAP] to dinner] in a minimal projection of S. The only candidate for a minimal projection of S is [we didn't invite [GAP] to dinner] which has as its subject *we*, not *who*. The *who* and the *we* are not coreferential, and the *who* cannot destroy the *we*. The mission is aborted. There seems to be a meta-rule that only one search-and-destroy mission is allowed. If not, then the *who* could continue its search in the superordinate S-node, S[whom S[…]] and successfully destroy the accusative *whom* without any further conditions, because this syntactic process does not seem to be sensitive to case-relationships.

The crucial difference between the permitted and the proscribed constructions is their configurational differences. Now all this may seem arbitrary and *ad hoc*, but it has at least Formal Adequacy, the level below Observational Adequacy: it works. It makes use of very simple geometrical configurations that are not sensitive to case, agreement, or government relations, but only to positional, that is, configurational, relationships.

For Lowth, it is the syntactic relations and the case relations that matter. The orphaned [GAP] in the second relative clause could not find a cognate relative pronoun in the first relative clause, so the construction was improper.[25]

Lowth did not give a complete account of the phenomenon, but must be credited with its initial discovery. He was interested only in showing that perfectly unobjectionable self-evident rules of English grammar could be set up that, using his definition of Sentence and the Principles of Strict Construction, could eliminate faults in the construction of English sentences.

Given his complete body of data, collected initially quite randomly, and asked to classify them into fault-types and to provide English grammatical rules that would judge the acceptable cases to be acceptable, and to show the fault in the faulty ones, there are very few today who could accomplish this task.

2. Casus pendens & nominativus pendens

Whereas the first type of construction proscribed by Lowth is a naturally occurring English construction-type found throughout the known history of the English language, the second construction, known as *casus pendens* or *nominativus pendens* ('dangling case' or 'dangling nominative'), is one of those Latinisms that Lowth considered improper in English because it did not construe according to the interpretation of grammaticality or 'propriety' dictated by the principles of Strict Construction. Lowth does not offer any definition of this phenomenon, or name it as such, because he was focusing on the facts of English and their interpretation according to the precepts and principles that he was using.[26]

The dangling nominatives in the two first examples offered by Lowth under the first rule to disallow them are indicated by him by italics. Very briefly, whereas all the other cases (genitive, dative, accusative, ablative) are *governed* cases, the nominative is the ungoverned case. It is not governed, but governs. Here are Lowth's instances (1762:123-124):

1. *Which rule*, if it had been observed, a Neighbouring Prince would have wanted a great deal of that incense, which hath been offered up to him by his adorers. Francis Atterbury (1662-1732), Vol. I. Serm I. [1762:124]

In some later edition, this additional example was added:

2. We have no better materials to compound the Priesthood of, than the mass of Mankind: *which*, corrupted as it is, those who receive holy Orders must have some vices to leave behind them, when they enter into the Church. Swift, Sentiments of a Church of Englandman [with respect to Religion and Government] (1708)

The following two examples are cited as improper in the Critical Note (footnote) under the rule for the case of the relative pronoun which has the same form as the Latin rule, but applies equally to English (1762:134-136).

1. "*Who*, instead of going about doing good, *they* are perpetually intent upon doing mischief." John Tillotson (1630-1694), Archbishop of Canterbury (1691-1694), [Works.] Vol. I. Serm. 18.[27] [1762:135]

Lowth's analysis reads: 'The Nominative Case *they* in this sentence is superfluous; it was expressed before in the Relative *who*.'

Also added in some later edition:

2. Commend me to an argument, *that*, like a Flail, there's no Fence [sc. defence] against it." Richard Bentley (1662-1742), Dissert. on Euripides's Epistles, Sect. i.

Lowth's analysis reads: 'If that be designed for [intended as] a Relative, it ought to be which, governed by the preposition against, and it is superfluous: thus, "*against which* there is no fence:" but if *that* be a Conjunction, it ought to be in the preceding member, "*such* an Argument[,] [that]." ' (1791:122)

The following is from Lowth's own prose (italics added):

> *The longer*$_i$ [Hebrew verses], though *they*$_j$ admit of every sort of Parallelism, yet belonging for the most part to the last class, that of Constructive Parallels, I shall treat of *them*$_k$ in this place, and endeavour to explain the nature, and to point out the marks of them, as fully and exactly as I can. (*Isaiah. A New Translation* (1778), 'Preliminary Dissertation')

The subscript indices *i,j*, *k* identify the relevant noun phrase *The longer* and the anaphoric pronouns *they* and *them* referring back to it. The preposed object *The longer* is pleonastically repeated in the resumptive pronoun *them*. It is evident that the noun phrase *The longer* has been moved from object position after the prepositional verb *treat of* and placed in initial position at the front of the sentence, focusing attention on it as it picks up the previous argument. This is a common feature of the syntax and pragmatics of the functional sentence perspective of English style. However, this noun phrase should have left a gap after its governing verb, but this position has been filled with the resumptive, or pleonastic, pronoun *them*, leaving the noun phrase *The longer* dangling at the front of the sentence, a *casus pendens*, i.e., an accusative without a governing verb.

In addition, the pair of conjoined infinitive phrases, 'to explain the nature, and to point out the marks of them', with their shared constituent, 'of them', is felt by some grammarians or rhetoricians to lack 'grace and beauty' at best, and to be 'improper', or ungrammatical, at worst.

After perusing these examples and deciding on their fault and what rule might be proposed to solve the problem of proscribing them which is at the same time an unexceptionable rule of English grammar, you may read footnote[28].

The pleonastic resumptive pronoun is superfluous; the accusative has already been expressed at the beginning of the construction, to which the object NP has been moved. If the pleonastic resumptive pronoun is retained then the initial accusative is a *dangling case* without a governing verb, and the pleonastic object *them*$_k$ must be removed.

That these constructions originate as a Latinism is clearly expressed in the trenchant critique by Anselm Bayly 1772. There Bayly provides a running commentary, mostly in the form of quibbles, on Lowth's *English Grammar*. His critique is interesting as an example of an older idea of the standard of English, and for his ingenious and well-meaning, if often incoherent or even inept or wrong-headed analytical proposals, which give some insight into how not only English but also classical texts must have been construed in order to make sense out of what were for the scholars of that time inexplicable vagaries of the syntax of the classical languages compared to English. Here is Bayly's passage on the *nominativus / casus pendens*, where he jumps in at the deep end with quotations from Cicero:

> "Labour to put an end to this horrid war; *which* if it can be accomplished, you will do eminent service to your country, and gain immortal honour yourself; I have been waiting with daily expectation of receiving messengers from you with letters, *who* if *they* come, I shall then be able to judge how to act: *which if they* should be written every one—" [See Bayly's Latin originals below.] In these sentences the relatives *which* and *who* are certainly the nominatives before the verbs *can be accomplished*, *come*, *be written*, not *it*, *they*, which are redundant. This manner of expression, though very common, the

> author of the short introduction [Lowth] judges to be improper, from a supposition, that *it* and *they* being the nominatives, *which* and *who* are left by themselves without a verb; but I should apprehend he will be of another opinion upon reflection, that this form of expression is purely Grecian and Roman, frequently used by Cicero:* And if the phrase is neat and correct in Greek and Latin without a pleonasm, certainly that figure cannot make it improper and mean in English. The elegance of the expression at least will appear from the flatness of the correction. [With the dangling nominatives removed:] "If it or this can be accomplished—If they come"—The Latin form, if it must be excluded by the decisive authority of this literal grammarian [!], may be expressed by other turns rather than that proposed; "which, if it can be accomplished, will bring eminent service to your country, and immortal honour to yourself—So soon as they come, I shall be able"—"Which rule, had it been observed, would have taken from a neighbouring prince a great deal of that incense, which hath been offered up to him by his adorers:" Short Introd. [1762:] 124. (Bayly 1772: 82-83)
> [Footnote to p 82:] **Quod si erit factum*, et rempublicam divino beneficio affeceris, et ipse æternam gloriam consequere. Cicero Planc. Fam. 10 4. Nos quotidie tabellarios vestros expectamus; *qui si venerint*, fortasse certiores quid nobis faciendum sit. Fam. 14. 22.

Bayly does not mention that the relatives *quod* and *qui* have been moved (extraposed) to the left out of the clauses within which they originate. This is impossible in English, and explains the resumptive pleonastic pronouns: the clauses would not construe without them.[29]

The internal evidence is that Bayly's linguistic intuitions are at least a generation behind Lowth's. He does not see that in Latin, unlike in English, one can move an item like the subjects *quod* 'which' and *qui* 'who' out of their clauses to the left of the complementizer or connective *si* 'if'. The inflection on the verb in the clauses out of which the *quod* and the *qui* have been moved serves the function of the overt subject. Why does Bayly not see this?

Notes

1 This account of the method of Traditional Grammar is offered to Mike Stubbs in recognition of his contribution to the methodology of present-day linguistics, and to the study of the English Language.

To set the mood for this piece one could do no better than to read, or to listen to Robert Schuman's setting of, Heinrich Heine's poem, *Die alten, bösen Lieder*, from *Buch der Lieder* (1817-1826).

2 John Horne Tooke (1736; 1792), *ΕΠΕΑ ΠΤΕΡΟΕΝΤΑ* [Épea Pteróenta ('winged words')]. *Or, The Diversions of Purley*. London, 1786-1805. Two Parts [Volumes]. Cited from: Part I, Chapter V 'and' III, 'Etymology of the English Conjunctions: *AND*.' [Tooke derives *and* from the verb 'to add'.] Tooke here criticises Lowth for stating that: 'THE Conjunction connects or *joins together* Sentences; so as out of two to make one Sentence.' (1762:92) Tooke points out that in the sentence, *John and Jane are a handsome couple*, the individual noun phrases *John* or *Jane* cannot each appear alone with the predicate, *is a couple*: 'Is John a couple? Is Jane a couple?' He gives other examples as well. He cites in support the Latin examples in the *nota* added (1714) by Jacobus Perizonius *né* Voerbroek (1651-1715) to the edition by Gaspar Scioppius (1576-1649) of the *Minerva sive de causis linguae latinae* (1562) of Franciscus Sanctius (1523-1601). Tooke cites the examples adduced by Perizonius to refute Sanctius' assertion also that conjunction results from syllepsis of two sentences: *Emi librum .x .drachmis et .iv. obolis.* Saulus et Paulus sunt iidem. This particular construction was known also to such grammarians as George Oliver Curme (1860-1948) (*Grammar of the English Language*, Part III, *Syntax*, 1931), and is today termed 'phrasal conjunction', rediscovered as if for the first time at the beginning of the heyday of the first era of generative-transformational grammar in the mid 1960s.

3 *The Journal of The Rev. John Wesley* (1703-1791). Edited by The Rev Nehemiah Curnock (1840-1915). Standard Edition. Eight Volumes. London: Charles H. Kelly, 1909-1916. Volume V, 1914:370. The reference is to: Joseph Priestley.(1733-1804) 1761. *The Rudiments of English Grammar; adapted to the Use of Schools. With Observations on Style*. London: Printed for R. Griffiths.

Wesley is probably reading a copy of the 'much expanded' second edition of 1768. By the English Presbyterian minister, schoolmaster, controversial religious writer, chemist and physicist, and polymath.

Wesley does not seem to have noticed, nor does it matter, that Priestley's work was published a year before Lowth's. The significance of Wesley's remark is that Priestley's grammar, while much praised by present-day students of the history of English traditional school grammar, from Lowth on, for his support of the primacy of usage over putatively arbitrary rules, is otherwise very conventional in content and lacks the comprehensiveness and originality of Lowth's, as Wesley seems to have observed.

4 The best compilation that I know is by Pullum 1974, in what was originally one of three essays completed in the academic year 1970-1971 or 1971-1972 as part of the requirements for the three-term course, 'History of the English Language' *aka* 'HEL', in the Department of

Language, now Department of Language and Linguistic Science, University of York (UK).

The task was to take a good, representative traditional grammar from R. C. Alston's reprint series, *English Linguistics* 1600-1800, and to compare it with the compilation made earlier of typical strictures about such grammars and their authors in typical textbooks of the History-and-Structure of English type.

I forbear to quote from Pullum's article lest readers inadvertently conclude that I concur in the strictures enumerated there.

In a bizarre example of attributing to Lowth not only prescriptive and proscriptive practices but also the ability to dictate the course of development of the grammatical usage of a whole generation of Standard English speakers and writers and their descendants, he is credited with having introduced into English the rule that 'two negatives make a positive'. (For a good example of multiple negation in OE see example 12 above.) The *locuis* is usually given as the first edition of the *English Grammar* (1762), and a reference to the section on the Adverb in the Section on Words (Morphology, or Etymology), where it does not ever appear, with inaccurate page references (1762: 90-91).

In the first edition, and then repeated in later editions as an introductory statement to the now extended text, all Lowth has is the laconic: 'ADVERBS have no Government.' (1762: 126). There are no illustrative 'critical notes'.

In fact, the rule was added in *The Second Edition, Corrected*, in the section on 'Sentences' (or syntax), in the passage dealing with Adverbs (1763: 138-140).

Two Negatives in English destroy one another, or are equivalent to an Affirmative: as,

"*Nor* did they *not* perceive the evil plight
In which they were, or [*sic*] the fierce pains *not* feel."
Milton, P. L. i. 335[-136]. (1762-139-140)

There are two further examples (1763: 139-140 from Shakespeare, and two from Richard Bentley (1662-1742).

Lindley Murray, in his version of this rule (Rule XVI) of Syntax in his *English Grammar*, in order to make the import of the rule and the example from Milton crystal clear, adds the gloss: 'that is, "they did perceive him."' This suggests very strongly that those critics who give this rule and this reference have not looked into the 1762 or any other edition of Lowth's *Grammar*.

I have gone into this at some length in order to point out that Pullum is the only person known to me among several generations of linguists who has actually studied in depth Lowth's English Grammar.

In fact, it is a commonplace of Logic, one of the Seven Liberal Arts, the Scholastic curriculum, that *duplex negatio affirmat*, 'double negation affirms'. It is quite ancient and is found in logical systems throughout the ages, including in texts in Sanskrit, which has double negation. See Mates 1961: 31-32; 95.

Multiple negation had in fact already virtually disappeared from educated (literate) English by 1600 (Queen Elizabeth's letters show only a few traces), beginning with the English Renaissance (1550-1660), possibly in translating legal texts from Latin into English, in order to avoid potential ambiguity. But this seems to have begun as a natural process, not motivated by the force of observing arbitrary grammatical strictures.

Wittgenstein has commented (*Philosophische Grammatik* (1969); *Philosophical Grammar* (1974), both Oxford, Blackwell, *passim*) that the formula, $\sim\sim P \supset P$; or: $\sim(\sim P) \supset P$, is not in fact a rule of logic or grammar at all, but merely a consequence of the behaviour (interaction) of symbols such as ~, P, and ⊃.

5 A pair of complementary assertions often forms part of the uninformed critiques of so-called traditional grammarians. The first is that they studied writing instead of speech. So, as it turns out, has nearly everybody else. It is sufficient to look at the vast majority of descriptive English grammars, whether by linguists or textbook writers, to see that there are virtually no English grammars written on the basis of speech alone or in part, except perhaps Fries 1952, where it is hardly noticeable, or the grammars of English by Quirk et al., which use the corpus of tagged spoken texts from the Survey of English Usage in the English Department of University College, London.

The second is that they did not even examine the language, but rather some incorporeal idealized abstraction of their own invention, failing to describe even the actual usage of the written form. This may be true of the vast majority of modern scientific studies of English grammar, where the data so often consist of non-attested arbitrarily constructed examples made up *ad hoc* for illustrative purposes, often called 'intuitive data', but which might better be called *sentoids*. They are not 'data' in any natural language or natural science sense of the term, obtained by observation or experiment, and their structural or formal properties are therefore not 'facts'. This circumstance is the rationale for present-day Corpus Linguistics.

However, it is sufficient to look at the long line of compendious English grammars, often referred to, rather admiringly or affectionately, bordering on the patronizing, as 'scholarly traditional grammars', from Fiedler and Sachs (1861-1877), Mätzner (1880-1885), Koch (1878-1891), Poutsma (1914-1929), Kruisinga (1925), Kruisinga and Erades (1935; 1953-1967), Jespersen (1909-1949), Zandvoort (1957 ff.), to Curme (1931;1935), *et multi al.* (see McKay 1984, which is not complete) to see that the natural practice of these grammarians was to use a vast corpus of classified citations from literature, sometimes newspapers and other writing. Certainly H. W. Fowler's *A Dictionary of Modern English Usage* (1926) is devoted entirely to real examples, classified and analyzed in detail, from newspapers and other printed sources. And of course Lindley Murray's *English Grammar* (1795 ff.), based on Lowth's *Short Introduction*, is well-illustrated with edited quotations of good and bad usage from numerous good and bad writers that he took over from Lowth and supplemented with others. Nor does Murray consider only the Standard English of the educated writer. His *Exercises* are mainly instances of improper (ungrammatical) usage from the 'lower orders', what were commonly called 'low expressions'.

The only English grammar to examine non-standard English in detail is Fries American English Grammar. The Grammatical Structure of Present-Day American English with especial Reference to Social Differences or Class Dialects (1940), based on the corpus of correspondence from the First World War in the US War Office in Washington, DC.

6 'Introductory Memoir' (pp. 1-42), pp. 40-41. George Alexander Stevens (1710-1784). English novelist and humorist; Richard Porson (1759-1808), Greek classical scholar and regius professor of Greek at Cambridge (1792), one of the founders of modern classical scholarship; renowned for his remarkable memory and facility of recall. His ms. Greek hand is the basis of all present-day Greek typography.

7 It would be pointless to assemble a finite corpus and study that, as one cannot be sure that the relevant instance will be represented. It would be equally pointless to use a promiscuous or random, putatively

representative selection or assembly from all the writers of the day. All that inferior writers could contribute is that they are ignorant of grammar, by definition. Lowth believed that it is sufficient to show the state of the language if one uses the language of 'some of our best writers'. These are men such as Bentley, Clarendon, Tillotson, Swift, and others, all greatly admired writers of their day. The thought behind this is that the educated gentleman and scholar, the 'man of taste', embodies the best and most cultivated form of polite society: in manners, morals, taste, the arts and sciences, religion, politics, and, of course, in language. If the English language, as it is written by 'some of our best writers', is not ruled by grammar, then the language is indeed in need of those rules that will ensure that the language is so ruled, in other words, so that it does not, as Swift says, 'offend against every part of Grammar.' Lowth's sources are therefore selected both to illustrate the present state of the language, and to illustrate the application of the rules designed to bring that language into conformity with the precepts of grammar. Lowth's discussion of this point, like the other matters that he considers in his *Preface*, is admirably clear.

8 Lowth may have initially come across a different version of the idea of showing the application of a rule by showing not only its application ('what is right') but also its misapplication ('what is wrong') when he was a scholar at Winchester College from 1722 until he went up to New College, Oxford in 1729. He must have used the exercises in Latin composition by translating sentences from English into Latin by John Clark(e) (1687-1734). An early edition is entitled *An Introduction to the Making* [composition] *of Latin*, etc., 3rd edition, 1721, by John Clarke [*sic*].
In three A3 pages of hand-written notes about the curriculum ('Business at Winton. College 1756-1757') compiled in *c* 1800. amid the plethora of Greek and Latin authors and the repeated 'Grammar' of a skeleton timetable, the name 'Clark' appears once. (This information is due to Suzanne Foster, Winchester College Archivist.)

In the *Exercises*, the English sentences and a Latin vocabulary are arranged in parallel columns, English and Latin, under various rules of grammar and longer texts. The English sentence is provided with a parallel string of Latin words in the adjacent column in their dictionary entry form in approximately correct order with which to make a Latin sentence The Latin words must be converted into the correct inflectional form required by the Latin construction. An earlier work (details omitted) with this design, from which Clark must have got the pattern, was published in 1706 by Nathan (*aka* Nathaniel) Bailey (*d* 1742), better known as the author of *An Universal Etymological English Dictionary* (1721).

Both Clark and Bayley are mentioned on the synoptic title-page to the 1750 edition of *A New Grammar: Being the most Easy Guide to Speaking* *and Writing The English Language Properly and Correctly* … (1745), which went through at least thirty-four editions to 1800, by Ann Fisher (1719-1778), maiden name of Mrs Thomas Slack, wife of the Newcastle printer Thomas Slack: '[Part] IV. Syntax, or the Order of Construction; which shews how to join Words aright, in a Sentence or Sentences together. To which are added, [Chap. IV. & V, 5½ pp] Exercises of Bad English [under all the Rules of Syntax, as recommended by the author of the before mentioned Letter (the introduction, signed 'A. B.': 'Anselm Bayly?)], In the Manner of Clark's and Bailey's Examples for the Latin, to prove [test] our Concord by' (1750: 127).

Fisher states in a footnote on the first page of Chapter IV: 'Some of these Examples we set right, lest the learner, expecting them always wrong, should alter them by Guess.' This observation must have been made by an experienced teacher.

Cf. this entry from Chap. V, 'Promiscuous Exercises: or, examples under all the Rules': 'Thou and me is both accused of the same Fault. (1750: 129).

9 Whether Lowth was 'commissioned' to write this grammar, or merely presented or was presented with the proposition, is immaterial. The facts are that the publisher Robert Dodsley (1703-1764), of humble origins, but who was nevertheless accepted and respected by his betters in breeding and education, had a major hand in its genesis and publication. It could well have been his initiative that led to Lowth's authorship. The correspondence on this between Lowth et al. is to be found in Tierney 1988. This work unfortunately ends with Robert's death. There must be more from Lowth in the subsequent correspondence with Robert's brother James (1724-1797), his successor, but this has yet to be published. See also Straus 1910 for details of publishing history; also Solomon 1996.

10 Trying to retrieve this information from the OED is futile, because all the data have been pooled, leading to a kind of muddy-brown mass of information (not unlike what you get if you mix together all the colours of the paint-box) from which all the relevant chronological information has been removed except the dates of the citations. It might make more sense to list them chronologically by birth date of the author. What would be required is a *variorum* dictionary, giving the senses as found in an historical succession of dictionaries. Illustrative quotations from texts contemporary with the dictionaries would then have far more illustrative power.

11 On this point see Fries 1952, Chapter II, What is a Sentence?, which discusses a multitude of attempts by 'traditional' grammarians to define 'Sentence'.

12 Cf. the following, Rule XXII, the last rule of 'Syntax', from Murray's *English Grammar* (1795):

> ALL the parts of a sentence should correspond to each other: a regular and dependent construction, throughout, should be carefully preserved.
>
> The following sentence is therefore inaccurate: [Example of improper construction omitted.]

This is as far as I know the first clear statement of the principle of Strict Construction. The difficulty in applying the rule as seen by Lowth's and Murray's contemporaries is well expressed in the following note from West 1953/1996:

> This rule, as Murray admits, 'may be considered as comprehending all the preceding ones', but he justifies its inclusion by giving a large number of examples which he hopes will 'afford some useful direction, and serve as a principle to prove [test] the propriety or impropriety of many modes of expression, which the less general rules cannot determine.' These examples make up the rest of the observations on this rule. It was quoted by John Kigan (*Remarks on the Practice of Grammarians …* 1823: 88) as showing Murray's consciousness of the inadequacy of his own rules; and Kigan also criticises its vagueness. 'How to resolve or divide a sentence into those parts that should thus correspond', he says, 'or, in what this regular and dependent construction consists, he [Murray] has not shown. So that after the drudgery of committing these rules to memory, and our endeavours to digest them, we are obliged to learn the true construction of a sentence from a long continued attention to the practical use of words.'

13 What Lowth is offering is only the definition of and the procedure for establishing grammatical propriety. It is not a recipe for defining Standard English, as he has already taken the decision to collect his data from reputable writers with a reputation for 'accuracy': grammatical propriety. It had to wait for George Campbell's *Philosophy of Rhetoric* (1776) for the additional criteria of Standard English to be established. See Book II, 'The Foundations and Essential Properties of Elocution', of the doctrine of 'reputable, national, and present use … which gives law to language'

Nothing, however, is always as it seems. In Chapter III, 'Of Grammatical Purity', Section I, 'The Barbarism'; Section II, 'The Solecism'; Section III, 'The Impropriety', Campbell shows how any use that violates the purity of the language by containing any one of these three faults, is improper:

The barbarism is an offence against etymology [morphology], the solecism against syntax, the impropriety against lexicography [diction; choice of/proper words]. (1776: 190)

This summary statement is sufficient to show that the putative primacy of use (usage) is, in the view of the normative grammarian, in fact subject to the laws of grammar.

14 It is notable that virtually all Lowth's texts come from the previous generation of post-Restoration authors. Many historians of English literature say that there was a distinct change in English style around 1700. A compilation of the authors represented and the number of instances of improper usage from each cited by Lowth in his 'Critical Notes' shows that Swift is quoted far more than any other writer. See footnote 23 below.

15 A more fully developed version of Lowth's proposal will be found in the 'Preliminary Dissertation' to his *Isaiah. A New Translation* (1778: x-xxxiv). Finding the metrical basis of Hebrew poetry was considered essential especially to the translating of the Psalms, There was some considerable correspondence on this matter in the *Gentleman's Magazine* in the 1740s, complete with pointed Hebrew examples, which Lowth would as a matter of course have read. Lowth saw at once that the metrical basis of the Psalms and the other poetical books and passages of Hebrew Scripture could not be reconstructed because the original pronunciation of Biblical Hebrew had been irretrievably lost. The Masoretic text of the Hebrew Bible and its system of pointing he dismisses as 'the Jews' interpretation of the Old Testament'. (Lowth 1778) So his Oxford 'Lectures' could be considered, like his *Short Introduction*, his proposed solution to a generally recognized problem.

16 This has misled some enthusiastic but not very observant students of Biblical poetry to say that it is *semantic* parallelism, which had in fact been noted before. It is the 'sententious' nature of the poetry that is Lowth's real discovery, whatever later embellishments have flowed from it.

17 Lowth is being disingenuously generous to his predecessors. It is the 'conformity' i.e., the parallelism that has many times previously been noted, but not the sentential basis of this 'conformity'. No one until Lowth had proposed a sentential solution based on this 'conformity' or parallelism. In particular, his discovery of the need to repeat matter from one sentence to complete the sense of the next sentence by filling the gap there was wholly original with him.

18 It would be an interesting exercise to try to construct an algorithm for finding these constructions in any finite corpus. There are many reasons for thinking that this is in fact impossible, because of the infinite variety of

the long-range dependencies involved. Even looking for *and-which* constructions conjoined to preceding adjectival phrases etc. requires hand-sorting of the finds into hits and misses. Even then, potential candidates would fall through the net because the relative pronoun will have been elided, leaving only the *and* behind.

19 The elision of a subject NP or Pn in the second of two conjoined clauses where the antecedent is not the subject of the first clause is also allowed in earlier forms of English; see Ohlander 1938 and Burnley 1983.

20 The second emendation preserves the parallelism. A colleague in the Department of Mathematics at the University of York, with a keen interest in language, when shown Lowth's example, made the same suggestions, and with the same reasoning.

The method of correcting or reinterpreting unconstruable or 'faulty' construction by rearranging the words into a syntactically new or different, acceptable form, as if that were what was originally or ought to have been intended, is a common procedure among amateur linguists, who sometimes tend to treat the original almost as if it were a misprint. This is what might be called the 'patch-up' procedure of construing.

21 When DAR was on his way to the University of California, San Diego, to give a talk on just this topic, he was asked what he was going to talk about by a person with no special expertise in English Grammar. When given the expressions, 'The man we invited to dinner, but didn't come' *vs* 'The man who came to dinner but we didn't invite', they immediately exclaimed, 'Oh, I see — the second is ungrammatical.'

On an earlier occasion, while waiting for a taxi at the railway station on our way to a meeting, DAR was asked by another waiting colleague what he was working on at the moment. When he produced the same pair of contrasting expressions, his interlocutor retorted, 'They're both ungrammatical.' DAR rejoined: 'Have you ever read any Swift?' The retort was swift and sharp: 'Oh. — Swift!'

The very wide-spread idea is that in earlier forms of English, anything is possible, and we are not obliged to take notice of it.

22 This example is due to Bruce Mitchell, who also supplied references to a number of other instances of symmetrical and asymmetrical conjunction of this type in OE.

23 Cf. this PDE example:
In this context, granting concessions over Cyprus, which the EU is set to demand, but [*which*] would be incendiary to the nationalists, may be

practically impossible. [Deleted object relative pronoun in second relative clause restored in '[ ]'.]
(Ankara's EU project is in danger of collapse. *The Independent*, Europe, Analysis, by Daniel Howden, Wednesday 24 May 2006 p 18 *f*)

24 The distinctions drawn here between the types of projection are probably what Chomsky has termed an 'epiphenomenon'. It is the automatic consequence of the operation of the rules of iterative left-dislocation. The parser automatically recognizes the type of projection from the syntactic configuration.

25 When I gave a talk on this subject at the Neuphilologische Fakultät at Tübingen, Uwe Mönnich commented that his grandfather used to use this permitted English-type of conjunction in German, and he had often wondered about it. It now seems to have died out in favour of a more construable alternative: *der Mann, den wir zum Abendessen eingeladen hatten, der aber nicht erschien*. And *vice versa…*

The very strong sense of case in German does not like local determination of case, although it is sometimes found, as in the following newspaper example: *Viele Firmen wurden in die* [accusative singular GAP] *oder an den Rand der Pleite* [genitive singular] *getrieben*. 'Many firms were driven into or to the brink of bankruptcy.'

Local case determination seems to be permissible if the shared item has the same form, as in: *Wenn sich der Mann überlegte* [takes the dative of *sich*] *und endlich entschieden* [takes the accusative of *sich*] *hatte*, … 'When the man had reflected and finally decided, …' This example is quoted from a late 19thC book by an author who styles himself *Der Sprachwart*, 'The Guardian of Language' (cf. *Torwart* 'goalkeeper'), who condemns it on the grounds that the single *sich*, which he says quite rightly is dative by its initial position with *überlegt*, cannot supply the missing accusative gapped *sich* required by *entscheiden*. Independently of Lowth, and using only the principle of Strict Construction, he comes to the same conclusion, and with the same reasoning.

26 This is typical of his approach in all his work: not to engage in sterile explication of the obvious or to refer to the work of others as if treating their views instead of expounding his own. His straightforward expository style suits this mode of presentation very well, and lends it an authority and force that Lowth's argument would otherwise not possess.

27 '[Tillotson] was perhaps the only primate who took first rank in his day as a preacher, …' (Article on Tillotson by Alexander Gordon (1841-1931) in *DNB*.)

28 Every Nominative Case, except the Case Absolute [one use of the ablative case in Latin, but Lowth says it should be the nominative case in English (presumably because it is ungoverned)], and when an address is made to a Person [vocative], belongs to [governs] some verb, either expressed or implied; … (Lowth 1762: 123-124).

That the nominative governs the verb and not *vice versa* is shown by the agreement between the person and number on the verb with that of its nominative case, or subject.

29 This Latinism - extraposed constituents out of relative clauses, to the left of the RelPn- occasionally appears in English Renaissance verse and prose. I have not found any discussion of it in Lowth or any contemporary grammarian. Cf. this example from Shakespeare's Cymbeline, Act 2, Scene 3, 19-22:

[Musician] (sings)
Hark, hark, the lark at heaven gate sings,
And Phoebus gins arise,
His steeds to water at those springs
On chaliced flowers that lies, ...

The construction of 'at those springs' etc. is:

PrepPh[at NP[those springs RelCl[S[PrepPh[On chaliced flowers]PrepPh]S
S[that lies PrepPh[GAP]PrepPh]S]RelCl]NP]PrePh

It should be evident that the PrepPh 'on chaliced flowers' has been extraposed out of the RelCl to the left of the RelPn 'that', creating an adjacent S-node, and leaving a GAP behind.

References

Bayly, A. 1772. *A Plain and Complete Grammar of the English Language; to which is prefixed The English Accedence: with Remarks and Observations on a Short Introduction to English Grammar* [Lowth 1762]. London: Printed by G. Bigg.

Burnley, J, D. 1983. *A Guide to Chaucer's Language*. Basingstoke: Macmillan Educational Ltd. The Language of Literature; Norman, Oklahoma:

University of Oklahoma Press. OUP paperback edition 1994. Since 1989 titled: *The Language of Chaucer*.

Campbell, G. 1776. *The Philosophy of Rhetoric*. Two Volumes. London & Edinburgh: W. Strahan; T. Cadell; W. Creech.

Fries, C. C. 1952. *The Structure of English. An Introduction to the Construction of English Sentences*. New York: Harcourt Brace & World, Inc.

Gazdar, G. J. M., G. K. Pullum, I. A. Sag. 1985. *Generalized Phrase Structure Grammar*. Oxford: Blackwell Publishing; Cambridge, MA: Harvard University Press.

Hall, P. 1834. *Sermons, and Other Remains, of Robert Lowth, D.D. …* London: J. G. & F. Rivington.

Hudson, R. A. 1976. 'Conjunction Reduction, Gapping, and Right-Node Raising.' *Language* 52: 535-562. ['Right-Node Raising' is also better called 'Shared Consituent Construction'.]

Lowth, R. 1753. *De Sacra Poesi Hebræorum Prælectiones*. Oxonii: e Typographeo Clarendoniano.

Lowth, R. 1762. *A Short Introduction to English Grammar: With Critical Notes.* London: A. Millar and R. and J. Dodsley.

Lowth, R.. 1763. *A Short Introduction to English Grammar: With Critical Notes.* The Second Edition, Corrected. London: A. Millar and R. and J. Dodsley.

Lowth, R. 1778. *Isaiah. A New Translation*. London: J. Dodsley and T. Cadell.

Lowth, R. 1787 *Lectures on the Sacred Poetry of the Hebrews*. English Translation from the Latin by G. Gregory, 1787: Two Vols. London: J. Johnson.

Lowth, R. 1995 *Robert Lowth. The Major Works*. D. A. Reibel (ed.). London: Routledge/Thoemmes Press. [Enumeration of volume titles omitted.]

McKay, J. C. 1984. *A Guide to Germanic Reference Grammars. The Modern Standard Languages*. Amsterdam/Philadelphia: John Benjamins Publishing Company. Amsterdam Studies in the Theory and History of Linguistic Science. Series V. Library and Information Sources in Linguistics. Volume 15.

Mates, B. 1961. *Stoic Philosophy*. Berkeley and Los Angeles, California: University of California Press; London, England: Cambridge University Press. University of California Publications in Philosophy. 1953. Second Printing (with new Preface).

Oirsouw, R. R. van. 1987. *The Syntax of Coordination*. London; New York; Sidney: Croom Helm. Croom Helm Linguistics Series.

Ohlander, U. 1938. *Studies in Coordinate Expressions in Middle English*. Lund: C. W. K. Gleerup; London: Williams & Northgate, Ltd.; Copenhagen: Levin & Munksgaard—Ejnar Munksgaard. Lund Studies in English V.

Pullum, G. K. 1974. 'Lowth's Grammar: A Re-Evaluation.' *Linguistics: An International Review* 137:63-78. The Hague: Paris: Mouton.

Straus, R. 1910. *Robert Dodsley: Poet, Publisher and Playwright*. London: John Lane; New York: John Lane Co.

Tierney, J. E. (ed.). 1988. *The Correspondence of Robert Dodsley*, 1733-1764. Cambridge: Cambridge University Press. Cambridge Studies in Publishing and Printing History.

Visser, F. T. 1963-1973. *An Historical Syntax of the English Language*. Three Parts in Four Volumes. Leiden: E. J. Brill.

West, C. E. 1953/1996. *Lindley Murray — Grammarian*. Leeds University MA Thesis, 1953. Reprinted in: D. A. Reibel (ed.). 1996. *Lindley Murray. The Educational Works*. London: Routledge/Thoemmes Press. [Enumeration of volume titles omitted.]

Travelogues in time and space:
a diachronic and intercultural genre study

Andrea Gerbig

University of Bochum

Abstract

On the basis of a recently compiled corpus of travel literature from the 16th to the 21st century,[1] *this paper investigates both synchronic and diachronic variation. The synchronic investigation uses a subcorpus of texts from the 21st century, which shows fascinating language choices reflecting shared values and attitudes among travellers as an intercultural group, negotiating facets of their identity as independent and adventurous people. The study will show how a local evaluative schema of use develops for words which were found to be key words in a statistical sense in the subcorpus. The second case study looks at the pragmatic extension of the prepositional phrase* `'in the middle of'` *in the diachronic development of the language of travel writing covered in the corpus, in a span of 600 years. Some implications of working with such a multifaceted corpus for researchers and students of linguistics, literary studies and cultural studies are discussed. Potential applications to language learning and teaching are briefly covered, suggesting that/showing how a quantitative and qualitative approach to authentic data in computer-readable format can help language learners to cope with the phraseological nature of language.*[2]

1. The Corpus – data and potential routes of exploitation

The corpus I use for the following analyses contains travel literature from the 16th to the 21st century.[3] It comprises about half a million words per century, as evenly distributed across the centuries as possible. The corpus will be further extended in order to fill existing gaps. The subcorpus of the 21st century consists of texts which are all published on the internet, on a well-structured and well-edited web site, rather than in interactive weblogs, which are mostly of a highly colloquial style.

The travel corpus offers broad analytical and interpretative potential. At the macro-level, it is a repository of cultural knowledge and stories about intercultural encounters. The corpus can tell us about the role of travel in societies through the centuries; what it has meant for people to travel and to hear about travels. It provides us with information about the people who were able to travel and about their status in society, in political as well as economic respects. Naturally, the regions travelled to as well as the means of travelling have changed in significant ways throughout time. These factors have of course also influenced the kind of contact that developed between the traveller and local people.

At the micro-level, the corpus enables researchers to carry out diachronic and synchronic studies of the language of travel writing. We can aim at a broader description of the genre, investigate diachronic changes in style and content, and, more concretely, trace developments on the level of lexis, phraseology and structure. Two case studies will illustrate some of these possible approaches.

2. Theory and Methods

The methodological and theoretical approach underlying the present analyses starts from the assumption that we are interested in how language creates meaning. This needs to be investigated not in abstract terms of 'language', but in more concrete terms of 'actual language use', because this is the observable aspect of language. The regular and frequent way of representing something shapes our understanding and our way of dealing with it. This mildly constructivist concept of utterances in relation to social agency assumes that habitual forms of language use construe topics. The argumentative framework here goes back to Foucault's concept of discourse formations (e.g. 1980) and work by the language philosopher Searle (e.g. 1995), who proposes a major role for language in creating social and institutional facts. (For a more comprehensive discussion of the relationship between linguistic evidence and socio-cultural conceptualization see Gerbig 2003. Implications of work in this field for linguistic theory are discussed by Stubbs 2007).

Take as an example the alleged destruction of the ozone layer. We cannot perceive the ozone layer – or its absence – with our senses. We are presented with data which are interpreted for us by scientists. The 'raw' data would not help us much. Whatever we know about the ozone layer, we know from discourse. Whether countries reduce their CFC emissions or not is regulated discursively. Those arguments which are more powerful create realities. Such realities, by implication, are construed and therefore fragile.

'Representation' is a concept that helps us link people's experience and cognition to linguistic encoding. Representations construe versions of the world (Hall 1997), they construe views on how a culture, or in the present context, a sub-culture, 'functions'. Although this might be a contested view of culture, it can be revealing to investigate pervasive discourse that shapes – and is shaped by – the ordinary way a sub-culture and its individual participants function. With respect to travelling, a rather abstract, though linguistically transmitted, 'way of life' or category of experience is visible. It involves, for example, backpacks, buses, uncomfortable sleep, but also an experience of belonging to a group of independent, adventurous and outgoing people. Travellers semiotically not only create a system of rules and rituals that other travellers take as a starting point for their own behaviour, but they also create expectations about, and in a way even give rise to, those institutions which cater for travellers.

This is a reciprocal relationship between the cultural-institutional setting and the individual-cognitive development. Against the background of the community and its shared language use, the individual develops his or her social and linguistic competence, including ideas or beliefs and means of communicating about travel. The individual's language use, within the boundaries of the language norm, then not only systematically reinforces, but also gradually modifies existing institutions and the language system; continual use, with its gradual variation, shows the diachronic perspective (cf. Halliday 1992 and 1993). This view also explains variation in language use, an issue which often troubles teachers and learners of a language (cf. Sinclair 2004). The data seem to suggest that a particular meaning is generated by repeatedly using particular choices around a node. These choices are then reinforced and thereby set off from other systematic uses.

Such intricate relations between language, cognition (knowledge) and culture (in a broad sense) can be accessed via instances of language use in texts. These are most conveniently handled in the computer-readable format of corpora. The travel corpus covers a portion of the genre, diachronically and synchronically. It provides a particular view of variability and regularity of the language system. In terms of a frequency distribution, we can see cultural routines and conventions emerging from the collected utterances. The corpus therefore provides concrete material to investigate conventional linguistic behaviour that, given its frequency, is presumably significant within the group of speakers and inseparable from their shared ways of conceptualisation.

3. Synchronic example – building group identity

This case study is based on a synchronic subcorpus of the travel corpus, made up of texts published by native English speakers as travelogues on the internet, all from the 21st century. The areas travelled to are Africa, Australasia, Asia, the Middle East and the Caribbean. This subcorpus comprises 485,000 words in total. The texts are the travellers' reports and stories about their daily experiences, pleasant and unpleasant, and their adventures during their trips. Because the texts all appear on a web site maintained by a chain of travel equipment, they are written for like-minded people, to prepare them for what to expect on their trips. (Although there does not appear any obvious advertising on this web site, the chain of stores probably also offers this platform to encourage users to return to displays of their merchandise).

Surprisingly, there are hardly ever descriptions of interaction with the local people. In some regions, this is clearly due to a language problem. But even in Australia or New Zealand, the travellers in general spend too little time in one place to come into meaningful contact with locals. There is no distinctive British, Australian or other cultural focus; rather, the writers address English-speaking fellow-travellers, which is the group they want to belong to. This produces in some important respects a 'local' phraseological use that is different from a 'norm' of use, as described on the basis of a mix of genres in a large background corpus, such as the BNC. I want to show how, in the travel corpus, travellers negotiate their identity as members of groups of like-minded 'globe-trotters' in quite subtle ways. For this purpose, I will concentrate on the lexical level, by investigating keywords in their context.

I use keywords here in the sense of Mike Scott's (1997, 2006) approach**:** keywords are those words that are relatively frequent in a text, compared to their frequency in a background corpus that contains a representative mix of everyday language. If words appear significantly more frequently in the text under investigation than in other texts, they seem to be of some importance. In general, lexical items are the main hubs in the creation of meaning around which structures are built; therefore they obviously receive particular emphasis in analysis. Keywords give information about frequent topics and about their evaluation. Keywords for the subcorpus were extracted using Wordsmith Tools, with the BNC as the background corpus.

The 20 most frequent keywords in the texts written by travellers visiting Africa and Asia respectively are listed below. Function words are ignored. The occurrences are given in order of frequency:

Africa

`We, Bus, Our, Us, Tour, Tent, Sleep, Malawi, Trip, Africa, Zimbabwe, Park, Truck, Camp, Group, Uganda, Tanzania, Food, Zanzibar, Namibia`

Asia

`We, Driver, Trip, Bangkok, Rice, Bus, Thai, Korea, Tourists, Sleep, Road, Kathmandu, Restaurant, Taxi, Truck, Us, Cambodia, Ride, Lodge, Guesthouse`

The observer can quickly recognize from these lists a particular emphasis on the group rather than the individual; there are no occurrences of `I, Me, My` but many uses of `We, Our, Us`. Obviously, the regions travelled to appear prominently, but most significantly, people write in detail about how they travel**,** which is mainly by bus or by truck, where they sleep**,** and where and what they eat.

3.1 The keyword `bus`

I chose two keywords, `bus` and `sleep`, which capture the main concerns of these travellers. The following quote from the corpus epitomizes the paramount role of buses as a means of transport for 'real' travellers, as they are defined within the group.

> "We've rented a jeep." "A jeep? You don't need a jeep. Go by bus!" she jeered. I guess we weren't REAL travelers going by jeep

As I will show in the following analysis, the keyword `bus` in the texts about Africa and Asia shows an almost conventional association with 'difficulty', 'unreliability', 'danger' and 'lack of comfort'. Interestingly, in this sub-corpus (21st century, writings about Africa and Asia), these are not necessarily outright negative aspects. Rather, the interpretation of most of the occurrences suggests a voicing of the travellers' in-group status, of having successfully mastered the local pitfalls of transportation. This points to established in-group conventions of evaluating situations and behaviours with respect to people's projected image of being a 'traveller'.

Such conventions can be seen in habitual collocations around the keyword investigated. The concept of 'collocation' I use here is a statistical one, of two or more words co-occurring frequently with each other in running text, within a span of several words. Frequency of co-occurrence is of particular interest, as "corpus linguistics is based on the assumption that events which are frequent are significant" (Stubbs 2001, 29). The collocates can often be grouped semantically into sets. This marks the 'semantic preference' of the word investigated which "is the relation, not between individual words, but between a lemma or word-form and a set of semantically related words, and often it is not difficult to find a semantic label for the set" (Stubbs 2001, 65). Semantic sets very often share a particular pragmatic attitude and evaluation, visible in "discourse prosodies" (Stubbs 2001, 65/6), i.e. habitual evaluations marked in sets of semantically related collocates of a node. A discourse prosody can be understood as the pragmatic motivation for choosing the particular word/multi-word unit in the first place. In its pragmatic function, this evaluative element can be compared to the illocutionary force of speech acts. Therefore, it plays an important part in structuring the communicative competence of the members of a community.

With the help of WordSmith Tools (mentioned above), I searched the co-texts in which the keyword `bus` occurs in the texts about Africa and Asia, grouped the collocates and collocating phrases (i.e. the words and multi-word-units co-occurring with bus in a span of up to ten words to the left and right) into semantic sets and sorted them according to frequency. Once these sets of semantically related words were sorted, their particular evaluative directions became visible, as can be seen in the following examples. For reading convenience, the results detailed below are here summarized in a table.

Table 1: Keyword Bus: Collocates and collocating multi-word units

Semantic categorization of collocates	% of all concordance lines of BUS	Examples
Uncomfortable situations	22,5%	crammed, full to capacity, squashed, trying to secure seats, grab seats, no idea where we were going, oven-like conditions, roasting, chilly and bumpy
Danger	20,4%	soldiers search, weapon, concerns,knee-smashing, worried for my life, ravines looming, dust and fumes, battered tin can bus, hazardous, life-threatening
Difficulty managing the itinerary and other tour details	16,6%	intricacies, difficult, enigma, figure out, late
Noteworthy local habits	12,8%	children selling, pig bound, passed into, through the windows, jogged alongside, slapping the sides, vultures

The most frequent set of collocates expresses **uncomfortable situations** (in 22.5% of all concordance lines around bus). This can be illustrated again with a quote capturing a prototypical situation of discomfort on travels:

> Malawians have bladders of steel, the only chance to spend a penny is if the bus stops at an actual bus station, a rare occurrence.

The most frequent single-and multi-word collocates of the node bus are the following:

<crammed, full to capacity, squashed, trying to secure seats, grab seats, no idea where we were going, oven-like conditions, roasting, chilly and bumpy>

The following selection of concordance lines shows how these and other collocates combine with the node in wider co-text to form an evaluative semantic mood expressing the travellers' experience of situations on and around buses. The node word is in bold, the relevant collocates have been underlined:

- just <u>kept on selling tickets</u> until the **bus** <u>burst out of the seams</u>
- already knowing that <u>too many people</u> would be on the **bus.**
- and the **bus** was <u>packed to the ceiling with bodies</u>.
- and then catch a **bus** to P that was already <u>heaving with people</u>.
- watched as the **bus** <u>filled up until it was overflowing</u>.
- <u>terrifically cold</u> **bus** ride, <u>couldn't</u> speak or look at the scenery; just <u>burrowed</u> as
- <u>squashed</u> for four. People in the **bus** <u>complained</u>, but were just ignored.

It is clear from these concordances that `bus` in the present texts prefers the company of expressions from a particular semantic range, i.e. `bus` shows a semantic preference. The regular negative connotation of the node-collocate choices forms the discourse prosody. Preference and prosody are thus related concepts. In comparison, in the BNC, contexts around `bus` provide a different picture: there are a few delays and minor discomforts, but no substantial shortcomings. Similarly to the BNC, the texts about Australia and New Zealand from the travel corpus do not show any unusual collocations. The authors write about group dynamics within the bus, about the scenery, stops and sightseeing but there is nothing substantially dramatic.

Overall, in comparative data, the occurrence of `bus` is much less frequent. Otherwise, `bus` would not have been noted as a keyword in the present travel data. The sheer frequency then suggests that travellers recognize this means of transportation as highly noteworthy. The space they give to its negative contextualization, without ever voicing considerations of alternative transport - because the situation as given would seem unacceptable to the travellers - seems remarkable for interpretation of the travellers' style. A natural reaction one would expect to the described dangers and inconveniences would be a change of behaviour: for example to share cars or to switch to organized tourist transport. At least, one would expect a word of warning to fellow travellers. But nothing of this is visible in the data. Instead, judging from other, also frequent co-occurrences of bus with contexts of adventure and a certain appreciation of local colour, what is pragmatically implied, is that 'real' travellers have to endure such hardships. Telling about them is proof of one's in-group status.

The second most frequent set of collocates around `bus` concerns situations of **danger** (20.4% of all concordances), as epitomized in the following quote:

```
never sit near the front of the bus, it is better not to
witness how close the bus comes to being scrap metal
unless you fancy raising your blood pressure
```

The most frequent collocates of bus capturing such worries about 'danger' are:

```
<soldiers search, weapon, concerns, knee-smashing, worried for my
life, ravines looming, dust and fumes, battered tin can bus,
hazardous, life-threatening>
```

Longer co-texts as in the following concordance lines show the discourse prosody of fear and worry very clearly:

- truck had nearly been <u>run off the road</u> by a passing **bus**;
- iron bars across the <u>glassless windows</u> and metal shutters on the **bus**'s
- the **bus** had two <u>wheels</u> still on the road and the other two <u>over the cliff</u>.
- X's **bus** drivers had been <u>reckless</u>, but Y's drivers are truly on a <u>suicide business</u>.

- the **bus** did sharp turns around bends where there were no safety barriers and
- Our kamikaze **bus** driver was on a mission to get to B. as quickly as possible;
- this **bus**, an unroadworthy steel wreck, had been painted on the front a

The third most frequent set of semantically related collocates around bus concerns **difficulties managing the itinerary and other tour details** (16.6% of all concordances), as illustrated nicely by this quote from the corpus:

> they assured us that it was no problem to get a bus there (i.e. there are buses every 20 minutes, and it only takes half an hour - translation: there are buses every two hours, which take about one hour and which stop 5 km from the village, leaving you to walk the rest of the way)

More concordance lines clearly show contexts of difficulties and people's apparent unhappiness with such unreliable situations:

- the intricacies of the transport system. **Buses** are often difficult to
- Local **buses** are often an enigma of travel
- I thought I'd figure out the **bus** system.
- The expanse of land that served as a '**bus** station' was teeming with vultures tou
- The **bus** left at 8:00am (an hour late), to trundle a mere three miles down the road
- the **bus** was only two hours late leaving the station (and by 'station' I mean

If travel arrangements then turn out to be according to schedule or a smooth experience, this is stressed as noteworthy, as shown in the following set of semantically related collocates and concordance lines:

<amazingly; surprisingly; unexpectedly; left and arrived on time; smooth; fast>

- pleasantly surprised when a number 3 **bus** pulled up only a few minutes later.
- For some reason I got the most amazing **bus.** It was like flying first class, huge
- one of the smoothest **bus** trips we took - left and arrived on time

The last recognizably coherent and still reasonably frequent set of collocates revolves around **local colour** (12.8% of all concordances), illustrated with a prominent quote and several concordance lines:

> Ordinary life is conducted through the bus windows at the stops - live chickens, fish, cabbages, onions,…

- children come over to the **bus** selling bags of cool water, peanuts, baked as
- saw a pig on the **bus**, bound to one of the rear seats
- samosas from the street vendors, passed into the **bus** through the windows.
- crowd of locals jogged alongside the **bus**, slapping the sides and trying

In total, collocates from the above four sets make up 72% of the co-texts around the keyword. The remaining 28% deal with less spectacular situations. So, the majority of the occurrences of bus are with a restricted set of semantically related context words or multi-word-units. There are dangerous, overcrowded, unreliable buses, and curiously 'local' things. The discourse prosody (i.e. the evaluation) is mixed; more than a third of these collocates are negative, most of the others are more or less ironically conceding that taking a bus is part of the adventure. Interestingly, although descriptions abound about how uncomfortable and dangerous bus trips in various regions obviously are, there is never a serious suggestion of not taking the bus.

From these representations, we can observe a local schema emerging. Obviously, the evaluation of bus seen in the presented data is not, or only to a very small degree, shared in other contexts. The uses shown above do not occur in any comparable way in the BNC, and not even in the other travel sub-corpora. In the subcorpus of texts on Africa and Asia, the problems that are reported on are regularly downtoned by expressions and evaluations foregrounding the excitement about 'local colour, 'adventure' and – in terms of the travellers' ethos – 'doing the appropriate thing'. This makes up the characteristic semantics of the word in this subcorpus of the travelcorpus. It is interesting to see how widespread this evaluation is, given that over 70% of all contexts of bus could be classified accordingly, i.e. 'negative' but 'adventurous' and 'appropriate'. In a nutshell, as these two statements by travellers put it:

- still, I was going to take the **bus** cause it was more adventurous
- The **bus** ride to V. was an experience I will never forget.

3.2 The keyword sleep

How and where they sleep seems to be as much of an in-group marker for travellers as taking the bus. People habitually complain, but nobody takes action against the problems concerned with sleeping. The complaints are always rather mild, never outraged; there is never a clear signal to other travellers not to go there, do this, or book that. The principal complaint about lack of, or poor quality sleep most often goes together with situations implying adventure. This fits with the group prosody for buses, apparently being part of the same behavioural pattern. Money is of course an issue, as for example in:

< miserable sleep followed, but money was saved >

More focused than the financial question, however, is the microcosm of the community of travellers with their habits and rituals. Foregrounding this 'in-group' experience, the contexts of `we` outnumber those of `I` by far.`I` occurs mainly as part of a group. Not spending money on comfort could be seen as another group marker.

The majority of collocates around forms of `sleep` concern **disruptions of and hindrances to sleep**, such as shown in the following semantic sets:

- Noise from other people or animals
  `<shouting, rustling, lovemaking, row, argue, music, rabbiting on>`
- Ground and beds
  `<(hard/stone) ground, hostel floor, grass, dormitory, outside>`
- Small animals as nuisance or danger
  `<cockroaches, ants, (tsetse) flies, mosquitoes >`
- Big animals felt to be a danger, mainly because they come too close to the sleeping place.
  `<circling, trampling, hippos, elephants, smell>`
- Natural forces
  `<gale, rain, storm, (howling) wind, water, soak, muddy, cold>`
- How people describe their sleep
  `<restless, not a wink of, grumpy, deprived, little, irritable from lack of, fitful>`
- Wanting to sleep
  `<much needed, need to catch up on, try(ing/tried) (desperately)) to, manag(ing/ed) to get some, not a chance of getting, stiff from, unable to, terrible, angry, rough>`

The semantic profile indicated by the above collocates can be further illustrated with the following concordance lines:

```
- Not a chance of being able to get a decent night's sleep. If I
   could have slapped them
- I'd had a terrible night's sleep due to the incessant ramblings of
   the watchmen.
- Such sleep as could be managed was punctuated by braying from the
   donkey
- procession of ants marched past over the sheet. Excellent, I had
   friends to sleep with.
- miniature scorpion and giant leech. I tossed and turned, finally
   drifting off to sleep
```

The keyword `sleep` is not as frequent as `bus`, although both are among the ten most frequent keywords in each subcorpus. Together with issues of provision with food, issues such as means of transport and choices of accommodation usually determine the comfort of travelling. Such comfort generally enables the pleasure one can gain from visiting and sightseeing. In the present data however, the collocates around both nodes support the same local schema: 'real' travellers gladly undergo the hardships of local transport and makeshift sleeping arrangements. This marks them as both sharing the travellers' creed of soaking up the local way of life and as conforming to self-chosen, but apparently conventionalised in-group-values. The linguistic evidence for this conclusion has been clearly demonstrated in the above analysis.

4. Diachronic example - the semantic change of a phrase

It is a trivial truth that words are frequent because they occur in frequent phrases (Sinclair 1999: 162, Stubbs 2004). And every language learner soon realizes that context around a word is needed to recognize meaning. This leads to phrases as units of meaning in language beyond the word boundary. Take as an example the meaning of the word `middle`*:* The explanation given by the Cobuild Dictionary (1995) is: "The middle of something is the part of it that is furthest from its edges, ends, or outside surface". This is certainly true. But how does the meaning of the following uses from the travel corpus fit in?

```
- as if to highlight the day's absurdities, we got stuck in the
   middle of the desert
- he was outraged at being called out in the middle of the night for
   basically nothing
```

It is hard to imagine that the first speaker meant the exact geographical mid-point in a clearly confined desert area. And when does 'night' start and when does it end? So, when exactly is the middle of a night? The layer of meaning 'furthest from its edges / ends' is doubtlessly still present, but rather as a basis for a more specific, pragmatic meaning. We will come back to this point and more examples shortly.

Many researchers have stressed that ordinary language use is to a large extent made up of more or less pre-constructed chunks (cf. e.g. Pawley and Syder 1983, Cowie 1988, Moon 1998, Hunston and Francis 2000). Such units are variably called extended lexical units, lexical items, phrases, clusters, and more. Sinclair (1991: 110 and 1998) showed that, although such stretches of words "might appear to be analysable into segments", they have to be seen as one choice, as a form-function unit that is habitually used and that expresses a conventionalized meaning within a language community.

So while 'middle' itself seems to have a clear, de-contextualized meaning, the phrase 'in the middle of (+article)' frequently expresses a rather specific pragmatic meaning, which will become clear from the examples discussed below. If there is consensus that our language use is largely made up of such more or less variable phrases, we naturally want to be able to find them in corpora. Depending on the way a corpus is annotated, there are two options:

First, as some concordance programs offer searches for n-grams, i.e. recurrent strings of uninterrupted word-forms stopping at sentence boundaries, we can check their frequency in a text and their preferred co-texts.

Second, if the words in the corpus are marked for grammatical categories, i.e. if the corpus is tagged, we can also look for strings of 'part of speech' (POS)-tags. In the British National Corpus, which is the largest corpus of contemporary written and spoken British English, Stubbs (2004) has shown that the prepositional phrase structure 'preposition-determiner-noun-*of*-determiner' is the most frequent 5-word string, realized in expressions such as `at the end of the`, `at the beginning of the`, `in the middle of the`. The frequency of this POS-5-gram can partly be explained on semantic grounds since (as realized in the noun slot) we often talk about wholes and parts, beginnings and ends, in order to express spatial and temporal structures in discourse.

The prepositional phrase `in the middle/midst of the/a` occurs 49 times in the subcorpus of weblogs (21st century). Out of these, in only 14 instances does the meaning of `middle` correspond to the description in the Cobuild Dictionary, given above. In 35 occurrences, the phrase rather has the pragmatic function of an intensifier, underlining the speaker's surprise or anger about a situation or about a situation's inappropriateness, as shown in these examples from the travel corpus:

C21: `in the middle of ART` (72% used in a pragmatic function)

- around 4pm and landed in Auckland in the middle of a rain storm. A
- main road was an oasis of a place in the middle of a hot, dusty dirt track.
- we were served breakfast in the middle of the desert.
- or make mad dashes to toilet blocks in the middle of the night.
- I am not being kicked out of here in the middle of the night and having to carry
- it's brutally expensive to email in the middle of the jungle. So,
- Out of nowhere appeared a hut in the middle of the ocean. Sticking

The string `the middle of` is actually redundant for the message and functions as an evaluative marker. The texts from the 21st century do not show any occurrences of `midst`. Apparently, young people have stopped using this form. Here are some more examples from the 20th century, for `middle` as well as `midst`:

C20: `in the middle of ART` (64% used with a pragmatic function)

- office buildings and executive flats in the middle of a vast urban nowhere,
- caravan parks standing in fields in the middle of a lonely, windbeaten nowhere,
- confusion, like someone wakened in the middle of the night by an emergency,
- on their white womanless island in the middle of the sea. As we
- this one in bed (in the middle of the day, remember); and this

C20: `in the midst of ART` (80% used with a pragmatic function)

```
- I had scant sense of being in the midst of a rich, proud city
   built of
- had the slightest sense that I was in the midst of a lot of
   granite, and it was
- pychiatrist, until he turned to me, in the midst of a detailed
   explication of the
- new development. It was like being in the midst of an ugly-
   building competition.
```

While the spatial and temporal meanings are both residually present in these contemporary examples, the pragmatic meaning gains force through time. An investigation of the diachronic part of the travel corpus shows this interesting change. Work in construction grammar and grammaticalization (Heine, Claudi and Hünnemeyer 1991, and Hopper and Traugott 1993) has demonstrated such processes of semantic weakening of elements in habitual phrases with ensuing pragmatic strengthening of the entire phrasal use. Typical examples are metaphorical processes of grammaticalization with terms for body parts[4] which over time come to be used first as locatives, then as temporals and finally with a more pragmatic function.

I analysed the use of the phrase 'preposition-article-*middle-of*-article' throughout all centuries in the travel corpus, from the 21st down to the 16th. A clear development can be seen:

Table 2: Pragmatic use of *middle* and *midst* from C21 to C16

Century	*middle*, pragmatic use in % of total	*midst*, pragmatic use in % of total
21	(35 of 49) = 72%	no occurrence
20	(11 of 17) = 64%	(4 of 5) = 80%
19	(3 of 11) = 27%	(7 of 17) = 41%
18	(2 of 14) = 14%	no occurrence
17	(3 of 26) = 12%	no occurrence
16	(0 of 6) = 0%	(3 of 11) = 27%

The actual numbers are quite small for using percentages. This only serves to make the proportions clear. In the 21st and 20th century the majority of the occurrences have a pragmatic function. The numbers for `midst` show that it is used more often in its evaluative function, increasingly so in more recent language data. This evaluative expression is an idiomatic form-meaning complex. Backwards from the 19th to the 16th century however, the use is increasingly literal. It mainly denotes a specific place or time, as illustrated in two examples from the 16th century below, where `middle` means 'furthest from the edges / ends'.

```
- they are reported to have their eyes in their shoulders, and their
   mouths in the middle of the breasts, and that a long train of
   hair groweth
```

```
- that of a buffalo, feet like those of an elephant, and a horn in
   the middle of the forehead, which is black and very thick
```

A comparison with the Early Modern English section of the Helsinki Corpus (1500-1710), shows that out of in total thirteen occurrences of the phrase 'preposition-article-*middle-of*-article', there are six uses indicating at least an approximate spatial or temporal reference. The other five examples indicate concrete spatial uses of the phrase. There is not a single use of the phrase in the pragmatic function described above in connection with the travel corpus.

HC, EModE: spatial / temporal approximation

```
- But this work may be done in the middle of the day, if the heat be
   not violent
- London buildings; there is in the middle of the town the Duke of
   Norfolks house
- stately building, placed by it selfe about the middle of the
   outside of a Town,
- the Fort in the middle of the City is circular; toward
- In the middle of the River we had a pleasant Prospect on both
   sides;
- in the middle of the Vale we repaired to the
- In the middle of the Munsel i.e. a whole Day's Journy the Butler
   alights
```

HC, EModE: concrete spatial use

```
- but you must first make Incision alongst wise, vpon the middle of
   the foresaid (^Escharre^): Then put in some small quantity (that
- (^cleave^) just through the midst, so as the (^bud^) may be
   directly in the middle of the one half; and then snip off a part
   of the (^leaf^),
- You may make the (^cross cut^  )in the middle of the downright
   (^score^) on the Stock, and lifting up the four
- outhouses very handsome; a coach yard and stables in the middle of
   which is large gate into the ground and built over with a high
- that a good space may be left in the middle of the Schoole, so as
   six men a breast may walk up and down
```

5. Conclusion - and possible applications

The present study has provided access to a local intertextual net where uses of particular expressions (such as those involving `bus` and `sleep`) are tightly linked. They derive their meaning partly through delimitation from uses in other data, in this case other travel data or the BNC, and thus form a local (group-based) schema in their representation of travellers' preferences and conventions.

Researchers of language and culture tend to agree that linguistic representations and cultural concepts are related in a non-trivial way (cf. research in the Whorfian tradition and language philosophy, as briefly discussed above, as well as work in critical discourse analysis). They do not, however, agree about the form and extent of this relationship, nor about the kind of research necessary to document it. Of course no direct link can be presupposed between language use, cognition and culture. The point made here, however, is that frequently used linguistic routines in a particular area of meaning are inseparably linked to both the cognitive schemata the language users have formed for this area of meaning/part of their experience, and to institutionalised cultural practices There are always alternative ways of expression, but if particular forms are habitually chosen, this points to a cognitive preference. A cultural basis for such shared preferences seems plausible.

The travel corpus is specific in its topic, that is, stories about travelling. However, the text types are diverse, from letters, reports, diaries and adventure novels to publications on an internet platform. This provides for a range of information on form-meaning relations. In response to questions of ambiguity or variation, Sinclair (2004: 281) speaks of a "guiding principle", that "each distinct meaning in language can be associated with a word pattern that is unique to it". Work in Construction Grammar has come to similar conclusions about the form-meaning relationship. Kay (2001: 19) states that "pragmatic information … can be directly associated with linguistic form in irreducible grammatical constructions – that is, constructions whose form cannot be produced by combining smaller units of the grammar according to general principles".

Again, the question we are addressing here is that of the size of a unit of meaning. This question has to take into account structures and word choices in their semantic co-text, often motivated by a pragmatic intention. By alerting language learners and teachers to the principles of collocation, language awareness will grow. A fairly easy start is to investigate sets of semantically related words (semantic preference) and conventionalized evaluation (discourse prosody), where these prosodies are often the pragmatic reason for making the choice at all.[5]

The travel corpus offers insights into the changing role of travelling in society. Judging from the popularity of travel literature throughout the centuries, travelling has always been an interesting experience, for the travellers themselves as well as for the readers of their writings. On the basis of diachronic linguistic data we can observe shifts in cultural practices; from travelling as a privilege for the aristocracy and the rich to travelling as a life-style of adventure and intercultural experience for backpackers, as could be seen in the data from C20 and 21 (for a related study showing such development see Gerbig and Shek 2007). We can also observe, at each historical stage, which role the different forms of mobility take in the value systems of a culture. The diachronic parts of the travel corpus are a repository of information about cultural historical events and changes in the English language and can thus be equally of interest to students of both culture and linguistics.

As most of the texts in the travel corpus are literary texts, it is a particularly suitable basis for work in the field of stylistics. In linguistic descriptions of 'English usage' we need to give more prominence to the place of literature in our communicative lives. The travel corpus offers a real potential for such interdisciplinary research (for literary linguistic analyses of a related kind see e.g. Stubbs 2005, Müller-Wood and Gerbig 2006).

There are many corpora available today**:** very large, general background corpora covering language uses in everyday situations as well as many smaller, specialized corpora covering particular topics or genres, like the one used in this study. It would be desirable of course to have these different corpora accessible in compatible forms. This would give both researchers and students the opportunity to search for exactly those uses they are interested in at a particular moment in their studies, so as to arrive at a more detailed picture of 'norm' and 'variation'. From the teacher's point of view, variations need to be introduced gradually, so that the most frequent, usually canonical, forms are prioritized in teaching (Sinclair 2004: 275), proceeding to sets of more specialized uses, at first at the receptive level, as the learner's competence increases. Local and specialized uses, such as those in parts of the travel corpus can then be pointed out to advanced learners. Furthermore, a corpus in the hands of advanced learners is the ideal resource for them to increase their phraseological awareness (using material like the small study of *middle* above). In terms of improving learners' competence, it is important to draw on authentic language use. This need has been extensively discussed (see e.g. Sinclair ed 2004); it has been demonstrated in comparative studies of large background corpora (such as the BNC) with EFL / ESL textbooks that the latter still misrepresent the distributions and patterns of use as found in actual language data (cf. e.g. Römer 2004, Conrad 2004).

The analyses in this paper can be viewed as one module of a possible ethnographic study to discover the meanings people construe, which circulate and become embedded in their daily experience. Corpus linguistics offers the possibility of documenting this relationship from the language side. As Allen (2000: 37) puts it: "Meaning … is always at one and the same time 'inside' and 'outside' the text". The textual basis, however, is the common stock from which we all draw, which is analysable and therefore accessible.

Notes

[1] I would like to thank Patricia Sift, Barry Morley and Ingo Bachmann for their cooperation in planning and compiling the corpus. I am further grateful to Patricia Sift for discussions about some of the data and Barry Morley for writing tailor-made pattern matching software.

[2] I would like to thank Naomi Hallan and two anonymous reviewers for helpful critical comments on an earlier draft.

[3] See the full list of texts from the 16th to the 21st century in the appendix. The most recent texts (21st century) are taken from the internet at "BootsnAll.com".

[4] Here is an example of 'middle' being used for a body part (Helsinki Corpus, Early Modern English section)

```
and a paire of olde broken slip shooes on his feet, a rope
about his middle instead of a girdle, and on his head an old
greasie cap
```

5 As these small examples indicate, from a corpus linguistic view, the autonomy of 'linguistic levels' is not fully tenable.

References

Allen, G. 2000. *Intertextuality*. London and New York: Routledge.

Conrad, S. 2004. "Corpus linguistics, language variation, and language teaching". In J. Sinclair (ed.) 67-85.

Cowie, A. 1988. "Stable and creative aspects of vocabulary use". In R. Carter and M. McCarthy (eds). *Vocabulary and Language Teaching,* 126-37. London: Longman.

Foucault, M. 1980. *Power/Knowledge*, ed. C. Gordon. London: Harvester.

Gerbig, A. 2003. *Korpus und Kultur: Korpuslinguistische Analysen zu Repräsentationen deutscher und britischer Politik in den Printmedien.* Unpublished manuscript.

Gerbig, A. and A. Shek 2007 "The phraseology of tourism: a central lexical field and its cultural construction". In P. Skandera (ed.) *Phraseology and Culture in English*, Amsterdam: De Gruyter, 303-322.

Hall. S. 1997 (ed.) *Representation: Cultural Representations and Signifying Practices*. London: Sage.

Halliday, M. 1992. "Language as system and language as instance: the corpus as a theoretical construct". In J. Svartvik (ed.) *Directions in Corpus Linguistics,* 61-77. Berlin: Mouton.

Halliday, M. 1993. "Quantitative studies and probabilities in grammar". In M. Hoey (ed.) *Data, Description, Discourse*, 1-25. London: HarperCollins.

Heine, B., U. Claudi and F. Hünnemeyer 1991. *Grammaticalization: A Conceptual Framework*. Chicago: The University of Chicago Press.

Hopper, P. and E. Traugott 1993. *Grammaticalization*. Cambridge: Cambridge University Press.

Hoey, M., M. Mahlberg, M. Stubbs, W. Teubert 2007. *Text, Discourse and Corpora: Theory and Analysis (Corpus and Discourse).* Continuum International Publishing Group Ltd.

Hunston, S. and G. Francis 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam / Philadelphia: Benjamins.

Kay, P. 2001. "Pragmatic Aspects of Grammatical Constructions". http://www.icsi.berkeley.edu/~kay/cg.prag.pdf

Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon.

Müller-Wood, A. and A. Gerbig 2006. "A Literary-Linguistic Reading of Graham Greene's *Brighton Rock*: Interdisciplinarity in Practice". In A. Gerbig and

A. Müller-Wood (eds). *How Globalization Affects the Teaching of English: Studying Culture Through Text*. Lampeter: Mellen.

Pawley, A. and Syder, F. 1983. "Two puzzles for linguistic theory". In J. C. Richards & R. W. Schmidt (eds). *Language and Communication*. Longman. 191-226.

Römer, U. 2004. "A corpus-driven approach to modal auxiliaries and their didactics". In J. Sinclair (ed.) 185-199.

Scott, M. 1997. WordSmith Tools Manual. Oxford: Oxford University Press.

Scott, M. and C. Tribble 2006. *Textual Patterns*. Amsterdam: Benjamins

Searle, J. 1995. *The Construction of Social Reality*. London: Penguin.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, J. 1998. "The lexical item". In E. Weigand (ed.) *Contrastive Lexical Semantics*. Benjamins. 1-24.

Sinclair, J. 1999. "A way with common words". In H. Hasselgård and S. Oksefjell (eds). *Out of Corpora*. Amsterdam: Rodopi. 157-79.

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

Sinclair, J. 2004. "New evidence, new priorities, new attitudes". In J. Sinclair (ed.) 271-299.

Stubbs, M. 2001. Words and Phrases: Corpus Studies of Lexical Semantics. Oxford: Blackwell.

Stubbs, M. 2004. "On very frequent phrases in English". http://www.uni-trier.de/uni/fb2/anglistik/Projekte/stubbs/icame-2004.htm

Stubbs, M. 2005. "Conrad in the computer: examples of quantitative stylistic methods". *Language and Literature*, 14, 1: 5-24.

Stubbs, M. 2007. On texts, corpora and models of language. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert. *Text, Discourse and Corpora*. London: Continuum.

Appendix

List of books included in the travel corpus, 16th to 20th century

16th century

Leland, John, *The Itinerary of Lohn Leland in or About the Years 1535-1543*

Torkington, Richard, *Ye Oldest Diarie of Englysshe Travell: Being the Hitherto Unpublished Narrative of the Pilgrimage of Sir Richard Torkington to Jerusalem in 1517*

Hakluyt, Richard, The Principal Navigations, Voyages, Traffiques and Discoveries of the English Nation, Vol. I - VI

17th century

Coverte, Robert, *A Trve and Almost Incredible Report of an Englishman, 1612.*

Taylor, John, *The Penniless Pilgrimage. All the Workes of John Taylor the Water Poet, 1630*

Chardin, John, *Travels in Persia 1673-1677*, Book Two.

Dampier, William, *A Voyage to New Holland*, London, 1702.

Dampier, William, *A Continuation of a Voyage to New Holland*, London, 1709.

Fiennes, Celia, *Through England on a Side Saddle in the Time of William and Mary. Being the Diary of Celia Fiennes*

Fryer, John, A New Account of East India and Persia, Being the Nine Years' Travels, 1672-1681.

18th century

Defoe, Daniel, *Tour through the Eastern Counties of England, 1722.*

Fielding, Henry, *The Journal of a Voyage to Lisbon*, 1755.

Johnson, Samuel, *A Journey to the Western Islands of Scotland*, 1775.

Cook, James, *A Voyage Towards the South Pole and Round the World*, Vol. I, London, 1777.

Piozzi, Hester, *Observations and Reflections Made in the Course of a Journey through France, Italy, and Germany*, 1789.

Smollett, Tobias, *Travels through France and Italy.*

Sterne, Lawrence, A Sentimental Journey through France and Italy.

19th century

Stevenson, Robert L., *Travels with a Donkey in the Cevennes*, New York, 1911.

Franklin, John, *The Journey to the Polar Sea*, London, 1823.

Dickens, Charles, *American Notes*, London, 1842.

Kinglake, Alexander, *Eothen*, London, 1844.

Burton, Richard F., *First Footsteps in East Africa*, London, 1856.

20th century

Hudson, William, *Afoot in England*, London, 1909.

Douglas, Norman, *Old Calabria*, 1915.

Douglas, Norman, *Alone*, London, 1921.

Chatwin, Bruce, *In Patagonia*, 1977.

Bryson, Bill, *Notes from A Small Island*, 1995.

Fowler, Beth, *Half Baked in Taiwan*, 2000.

An extended view of extended lexical units: tracking development and use

Naomi Hallan

Trier University

Abstract

Using corpus data derived from recordings of spontaneous conversations, the article examines the acquisition and use of various types of extended lexical units involving the path morpheme ***out****. Data from children acquiring British English is compared with material from the COLT corpus of teenage English and with adult conversational data from the British National Corpus and some of the changes in usage that occur during development are examined.*[1]

1. Introduction

The astronomical increase of computing power and data storage available to the ordinary researcher has led to a massive change in the way many scholars approach the study of language phenomena. It seems no longer necessary to deal in detail with the objections of the prevailing orthodoxy of the mid-20th century to the use of empirical data. Scholars such as Stubbs (1996, 2001), Sinclair (1991), Hunston and Francis (2000) and Moon (1998) have shown how the application of corpus linguistic methods can lead to new insights into the way language is used. The methods they describe are being applied ever more widely, and provide a valuable addition to the study of language acquisition, enabling the researcher to detect and quantify patterns which are easily overlooked when simply reading through transcriptions of child speech (c.f. Theakston et al. 2005 on the complexities of the acquisition of auxiliaries).

The investigation discussed in this article forms part of an on-going study, part of which has already been reported in Hallan (2001). As discussed in that article, the function of word-forms such as *on, over* or *out*, which are traditionally classed as prepositions, is in fact quite complex. Data on the acquisition of *over* and *on* showed that they are initially acquired not only or even primarily with prepositional function, such as *on there* or *over the road*, but rather in adverbial or particle functions as part of multi-word expressions, such as *over here*, *come on* or *fall over*. Following Bowerman (1996) I use the function-neutral term **path morpheme** to refer to the closed class of grammatical words under investigation. The term is of course not entirely neutral, since it assumes that these forms are fundamentally spatial terms of some sort, which may not be the whole story. In the case of *over*, for example, there is some evidence that the deixis involved could be as much interpersonal as spatial (Hallan 2001: 100). However, the form considered in this article is clearly spatial from its very first use.

Linguists have observed for some time that a great deal of language is produced and understood as relatively unanalysed multi-word building blocks (e.g. Pawley and Syder 1983, Langacker 1987, 2000, Sinclair 1991, Stubbs 2001, Bybee 1998). It has become clear that the meaning of such extended lexical units (ELUs) is not simply built up additively from that of the individual word-forms they contain. Many of these units are framework constructions with variable slots (Goldberg 1995, Stubbs 2004). I shall use here Stubbs' (2004) term **phrase-frame** to describe such units. As became clear in the earlier work in this project, some path morphemes are acquired from the beginning as parts of extended lexical units as well as, or even before, their acquisition as free-standing lexemes.

2. The meanings of out[2]

Different languages vary enormously in the way they encode spatial relations. An important typological distinction, first described by Talmy (e.g. Talmy 1985, 1991), is between **verb-frame** and **satellite-frame** languages. Verb-frame languages, such as Spanish or French, encode the path of motion in the verb itself (*sortir, entrer, monter, descendre*) and the manner of movement as an optional adjunct (*sortir en courant*). Satellite-frame languages, such as English, encode the path of motion using satellites — path morphemes — either as adverbs or prepositions (*go/come out, in, up, down*); the manner of movement is encoded in the verb (*run out, crawl in, jump up, roll down*). Basic verbs of motion and caused motion in English encode some sense of a direction of movement, but only in terms of changing distance from a reference point (*go, come, bring, take*).

2.1 Primary *out*

The path encoded by *out* is bound up with one of the earliest learned spatial concepts — that of a container and its contents (see Bowerman 1996 for an overview of work on conceptual development). In cognitive linguistics this notion is considered to be one of a number of pre-conceptual **image schemas** (e.g. Lakoff 1987: 271 ff.), arising naturally out of the configuration of our own bodies and underlying our interpretation of the physical world. The function of *out* is assumed to be the encoding of the path of something moving from the inside to the outside of a container of some sort.

As stated above, the encoding of a path is not necessary the only or even the first function in which young children encounter these word-forms (cf. Hallan 2001: 99, 104). However, in the case of *out*, the spatial function does seem to be primary, and has such force that the word can occur independently, in a quasi-verbal function (cf. Tomasello 2003: 87) with directive force, in the speech of the youngest learners:

(1) *GER: **out**. *GER: **out**. *MOT: not **out**. *MOT: although we are going to the swings.
(Wells age band 2, Gerald 1;6.6 wants to go outside)

The word-form is perceptually salient in continuous speech, as it often carries a stress, and is rendered more so by the adults, who produce the construction *out you* *, where the slot in the phrase-frame is filled by the change-of-state verb *get* or the basic motion verbs *come* or *go*.

(2) *MOT: **out** you get. (Mother lifts him out of the bath). *GER: I'm cold. *MOT: you what? (Mother dries him). GER: I want. *GER: I want to be wrapped in a towel.
(Wells age band 5, Gerald 2;3.5)

In theory it ought to be possible for people to say *on you get*, *down you go* or *over you come*; in practice these phrases are not found in the Wells corpus. The motion verbs *go* and *come* are found with *in, out* and *up*, while the change of state verb *get* occurs with these three and also with *off*. The reason for this would certainly repay investigation; in the present context however it is enough to note that the construction occurs with *out* and must contribute to the form's salience for the young language learner. Since it is accompanied by actions, as in (2), this presumably further reinforces learning.

2.2 Containers, frontiers and goals

However, it is important to look carefully at what is actually being learned. Being lifted out of the bath might be perceived as emerging from a container, and somehow comparable to removing toy cars from a box:

(3) *DAR: brm brm. *MOT: do you want the brm brm **out**?
(Wells age band 2, Darren, 1; 6.2)

It seems improbable however that children as young as 18 months should perceive an analogy between the removal of the toy car from the box and their own desired exit from the house into the garden (cf. example (1) above). The path encoded by this use of *out* seems to have more to do with the crossing of a frontier, with all the formalities such transitions can require:

(4) *ELL: goak [coat]. *MOT: you wants your coat on? *ELL: yeh. *MOT: it's windy **out**, El.
(Wells age band 3, Ellen 1; 9.0 wants to go outside so asks for her coat)
(5) *ELS: boots. *MOT: you don't need your boots on. *MOT: you can go **out** with your shoes on.
(Wells age band 4, Elspeth 2; 0.2)

(6) *ABI: (action: goes into the garden with no shoes on and steps in a puddle) *MOT: goodness sake you've come **out** in your tights. MOT: after I've just dressed you.
(Wells age band 2 Abigail 1; 5.28).

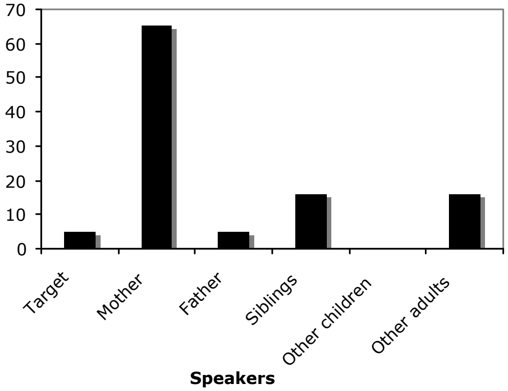


Figure 1: Use of *out* by different speakers in Wells age band 2

In the early age bands the children themselves say very little: only 5 of 107 uses of *out* in the recordings at 18 months (see figure 1). However they are hearing the word all the time, as parents, siblings, grandparents and visitors attempt to direct their movement and actions:

(7) (Betty has shut a door on another child, Joanne) *MOT: Betty let Joanne come **out** please because she's a pest on her own. *FAT: she in the bedroom? *MOT: yeh. *FAT: Betty come on. *FAT: let Joanne **out**. *BET: Joanne **out**.
(Wells age band 4, Betty, 2;0.3)

In addition, there seems to be a hierarchy of *out*-ness within the house, with bedroom and living room being the most *in*, and hall and kitchen referred to as *out*, although the whole house is still on the same side of the greater frontier:

(8) *MOT: Sarah? *MOT: go on out and play. *SAR: go on out and play? *MOT: that's right. *SAR: well it's raining. *MOT: go out in the hall then.
(Wells age band 6, Laura, 2;6.3 and her elder sister are annoying the grown-ups)

(9) *MOT: get on your bike then. *MOT: out in the kitchen. (Neville gets his bike in the kitchen - noises as bike sets off) *NEV: round corner. (rides his bike from kitchen towards sitting room) *MOT: coming in here are you?
(Wells age band 5, Neville, 2;3.0)

At the same time, the children's entourage regularly use *out* to refer to a different sort of path: going or being absent from home for a relatively short period (not more than a day) in pursuit of some goal:

(10) *ELS: I want to go **out**. *MOT: well we'll be going **out** soon. *MOT: we've got to go up to school shortly.
(Wells age band 6, Elspeth, 2; 6.6)

This use is more frequent than the others in the teenage and general corpora and will be discussed below. Although it can be related to the others we cannot assume that it is necessarily perceived as a derived meaning by the young children who are learning it. The children are confronted with three uses of *out* which they will doubtless integrate into a more generalised category as they grow older, but which are at least in the first years perceptibly distinct.

3. Functions of *out*

3.1 Quasi-verbs and adverbs

As we have seen, *out* encodes a variety of related paths and can do so as an independent adverb or even with a verb-like function. In addition to the dynamic path, *out* can encode a static locative, and in both functions it is frequently associated with *there, here,* or another locative expression (cf. example (11) below).

In both these adverbial functions it can be used with either an intransitive or a transitive verb:

(11) *BET: bubble. *MOT: no, have them **out** in the garden. *BET: bubbles. *MOT: because they're no good indoors, because they don't go up.
(Wells age band 3, Betty, 1; 9.4)

3.2 Phrasal verb particles

Other constructions involving *out* and a verb can be classed as phrasal verbs: unlike those described above, *out* functions as a particle rather than an independent adverb and the constructions convey meanings which can be described as metaphorical extensions of the path emerging from a container. Examples include *spread out*, *clear out, find out.* It is easy to imagine how the sense of these ELUs might have developed from the primary meaning of *out*, such derivations are a feature of the classic expositions of cognitive linguistics (for examples, see Hallan 2001, Langacker 2000). With verbs like *find out, write out,* or *work out*, where the particle has completive force rather than any locative sense (or at most a very tenuous one), it is doubtless possible to show how their meanings can be related to the central uses of *out*, but questionable whether language users really make the connection when actually uttering the forms. Since the phrasal verbs are present in the language they hear from the very beginning, it seems reasonable to suppose that children acquire the supposedly derived meanings in parallel to the 'primary' ones. Tomasello (1992: 172) found that the phrasal verb particle function was acquired later than the verbal/adverbial one, and this is borne out by the later production of such uses in the Wells corpus:

(12) *JON: I'll show you. *JON: you just stretch it **out** like that look. *MOT: let me see then.
(Wells age band 8, Jonathan 2; 11.29 helping his mother to make a paper butterfly)

The more complex processing of the placement of the direct object between the verb and particle could be a sufficient explanation for the later production of such constructions — the children show by their behaviour that they understand them much earlier.

3.3 Prepositions

Any investigation of the prepositional use of *out* is complicated by the parallel existence of the complex preposition *out of*, which shares most but not all of he functions of *out*. The choice of one form or the other is probably influenced by variety, register and sociolect.[3] But there is some meaning difference as well. There are no examples of this in the Wells corpus, but the following, from the BNC sub-corpus, will illustrate the difference:

(13) The only thing is, see what I'm scared of now, you go and put the mower in there and cart it **out the back** and the bits of moss dropping off the mower onto that patch! (KCN 6153)

(14) I said I want to get my teddy bears **out of the back**. (KBE 1020)

Clearly, in example (13) the simple prepositional phrase encodes the goal of the movement and, unsurprisingly, is the only one used for the corresponding static locative — after it has been moved, the mower is *out the back*. In (14) the compound prepositional phrase encodes the source of the movement, not the position of the direct object: the corresponding locative would be *in the back*, since it appears from the context that the teddy bears are in a back room rather than outside the building.

4. The corpora

The material used for this study comes from three corpora of spoken British English, covering acquisition by children up to five years old, teenage language and casual conversation among adults. The current lack of a corpus of spoken language for the primary school period is a real problem for a study of this type, but there is nevertheless a great deal to be learned even without the intermediate data such a corpus could provide.

4.1 The language acquisition data

The Wells corpus (Wells 1981, 1986) is made available through the CHILDES databank (MacWhinney 2000), a large and continuously growing collection of child language data contributed by scholars all over the world. The Wells corpus contains transcriptions of recordings of spontaneous speech from 32 children, 16 boys and 16 girls, born in the Bristol area in the second half of 1972. The children were fitted with a radio microphone on a harness and samples were recorded at random times during the day using a tape recorder in another part of the house. The children and their entourage were not aware whether or not they were being recorded at any given moment. Recordings were made during a full day every three months between ages 18 months and 3 years 6 months and again at almost 5 years, giving ten age bands in all. The material obtained covers the full range of everyday activities and contains not only the children's own productions but also anything said in their presence, whether addressed to them or not. Contextual information provided by parents enables a better interpretation of elliptical speech. The corpus contains about 395,000 words.

4.2 The teenage language

The Bergen Corpus of London Teenage English (COLT) (Stenström et al. 2002) was recorded in the late spring and early autumn of 1993 by teenagers recruited from schools in Barnet, Camden, Hackney, and Tower Hamlets, as well as from a boarding school in the Greater London Metropolitan Area, in Hertfordshire. The teenagers carried Walkman tape recorders with a lapel microphone, and taped their conversations in a variety of situations over three to five days, keeping a log of the situations and participants. The material was first of all transcribed by the Longman team preparing the British National Corpus, and then corrected and edited by the Bergen team before being annotated in various ways. For this study I have made use of the edited orthographic transcription. The whole corpus contains approximately half a million words, however 13 files (out of 337) contain extended passages of teacher monologue, so I removed these from the analysis, reducing the word count by some 31,000 words

The pupils were aged between 13 and 17 at the time of the recordings and were thus born between 4 and 8 years later than the children in the Wells corpus. In addition, they were all Londoners, and a number were from ethnic minorities, whereas the Wells study was carried out in Bristol, and deliberately excluded ethnic minority families. There is consequently a certain problem with the comparability of the two corpora. Unfortunately, as is so often the case with spoken language data, there are no more suitable datasets available — collecting and transcribing spoken language is difficult and time-consuming, and above all expensive, so researchers are very often obliged to make the best of what is already out there. In the case of the child data, the universality of the topics and situations in a household with a toddler makes for a reduction in the variability of the child-directed speech. The teenage data also shows a relatively restricted set of situations, with peer-group interactions predominating, and many of these being concerned with different types of relationships (Stenström et al. 2002: 28-29). This pre-occupation has a similar unifying effect, and it seems reasonable to assume that, had the recordings been made in Bristol, the talk would have been remarkably similar

4.3 The spoken BNC

The spoken section of the British National Corpus (2001) contains about 10 million words. However the material includes a wide variety of speech types, from more or less scripted monologues such as lectures, sermons or reports in meetings, to casual conversation. In order to restrict the material to something more directly comparable with the two other corpora, I made use of David Lee's Index to the BNC (Lee 2001, 2003). I was able to select the spoken conversations in Lee's 'demographic' domain, which are spontaneous conversations recorded by informants belonging to different social classes. Since the recordings used to compile COLT were also included in the BNC, in the initial transcription made by the Longman team, it seemed sensible to avoid simply comparing two versions of the same thing, by identifying and removing the BNC texts containing COLT material. The remaining texts contain 3741769 words, according to Lee's spreadsheet. This is clearly a much larger body of data than either of the two other corpora, which means that a detailed analysis of individual contexts is scarcely practical. In addition, the material contains speech not only from adults but also from young children and teenagers, as well as drawing on informants speaking a wide range of regional varieties of British English. Nevertheless it seems reasonable to suppose that the material averages out to a snapshot of spoken English as a whole and can thus serve as a basis for comparison when attempting to identify specific characteristics of the other datasets.

5. The data

The corpora were examined using concordancing software, either WordSmith Tools for the PC or *conc* for the Macintosh, and the results were displayed as a KWIC (Key Word In Context) concordance. It might be thought that much could be done by using part-of-speech tagged data to identify particular constructions. There are two reasons for not doing so, one general and one specific to this enquiry.

As a number of scholars have pointed out (e.g. Sinclair 1991, Tognini-Bonelli 2001, Römer 2005), the use of tagged corpora can pre-judge the issue one is trying to investigate. The categories ascribed by the parser to the different word-forms in the corpus are based on a pre-existing view of how the language is structured. Particularly in the case of spoken language, with its fluid structures and often elliptical style, a category assignment based on an often unconscious acceptance of quite traditional (and sometimes fundamentally prescriptive) ideas about grammar will not do justice to the data (cf. Hallan 2001: 91-2, 94-6).

In the case of this particular study, the boundaries between the functional categories for path morphemes are very fuzzy:

> The distinction between verb particles and prepositions is a problematic one in adult language, and, as usual, that means that it is even more problematic in early child language. (Tomasello 1992: 172)

The tags in the %mor tier of the Wells corpus (an interlinear tagging level which is provided in the CHILDES CHAT format, MacWhinney 2000), are a case in point: *out* is frequently, but not consistently, tagged as a preposition where the utterance clearly contains an adverb or particle. A similar problem exists with the tagging of the BNC (cf. Hallan 2001: 102).

Clearly the practicability of a direct categorization depends on the volume of material one has. So far as the Wells corpus is concerned, concordances of the separate age bands can easily be inspected and a functional categorization made on the basis of context as well as utterance structure. The individual texts can easily be examined using the CLAN software (MacWhinney 2000), and the contextual comments are a valuable aid. The COLT corpus has a similar volume of occurrences (see below) and thus a direct inspection is possible here too. In the case of the BNC sub-corpus, the volume becomes unmanageable in the short term. The concordances have therefore been restricted by searching for interesting constructions rather than the word-form *out* on its own.

5.1 Initial results

There were 1270 occurrences of *out* in the Wells corpus, between 110 and 139 in each of the different ages bands, and 1546 occurrences in the COLT corpus. This is equivalent to approximately 3215 and 3092 occurrences per million words respectively. The BNC sub-corpus had 12644 occurrences of *out,* giving a figure of approximately 3379 per million words.

Thus the frequency of *out* in the three corpora is not radically different. The important findings lie in the distributional differences across different constructions. As the analyses proceed, it will ultimately be possible to produce a **distributional profile or** usage **profile** of *out* for the different language samples. As an example, figure 1 shows the distributional profile for *out* in the first age band of the Wells corpus. The construction codes are explained below.

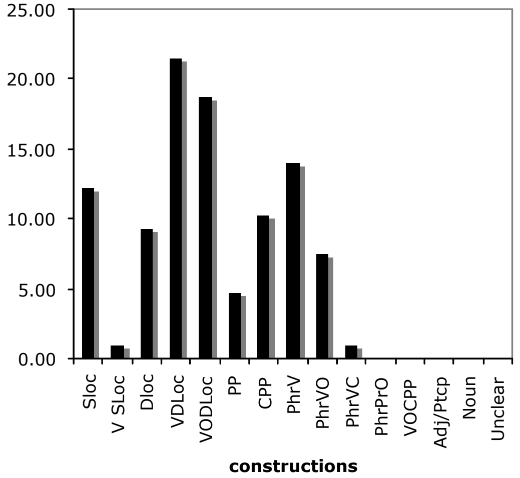


Figure 2: Uses of *out* in Wells band 2

Some possible constructions are not represented in this sample, which may be an indicator of the features of child-directed speech in this population (for an overview of issues related to CDS see Snow 1995).

5.2 Classification

For the Wells and COLT corpora, the individual examples were examined and classified according to the constructions they represent. The constructions identified were as follows:

1. **Sloc**: *out* used as a free-standing locative adverb or in combination with *there, here*, or another locative expression, with or without a form of BE;
2. **Vsloc**: a static verb other than BE, such as STAND or STOP used with *out* as a locative adverb;
3. **Dloc**: *out* used as a free-standing directional adverb;
4. **VDLoc**: an intransitive verb of motion (*come, go)* or change of state, or a manner-of-motion verb (jump, dance, climb) used with *out* as a directional adverb. This classification is restricted to clear examples of the physical movement of entities or substances from physical containers;
5. **VODLoc**: as 4., but with transitive verbs of caused motion (*take, bring*, etc) and a concrete direct object. Passives of such constructions are included here. *Throw X out* belongs here when it refers to the violent ejection of an entity from a location, but in 9. below when it means *dispose of/get rid of X*;
6. **PP**: a simple prepositional phrase, *out X*;
7. **CPP**: a compound prepositional phrase, *out of X*;
8. **PhrV**: a phrasal verb with *out* as the particle;

9. **PhrVO**: a transitive phrasal verb with a direct object. Passives of these were also included here;
10. **PhrVC**: a phrasal verb with a wh-complement;
11. **VOCPP**: a transitive verb and object followed by a compound prepositional phrase, *V X out of Y*. This construction is mostly found as a small group of fixed expressions and is discussed below (6.2.1);
12. **Adj/Ptcp**:a participial form (mostly past participle) of a phrasal verb used as an adjective, *stressed out*;
13. **Noun**: a compound noun formed from a phrasal verb, *drop-out*;
14. **Unclear** anything where a detailed examination of extended context failed to bring comprehension, often because the speaker was indistinct at a crucial point in the utterance. Where the verb and particle were clear but the object not, it was often possible nevertheless to classify the construction.

6. Some specific cases

Many corpus studies are concerned with the behaviour of specific word-forms and their collocates. In this analysis, rather than extracting frequent collocates and attempting to measure the strength of attraction between them (e.g. Gries and Stepanowitsch 2004), I have considered the behaviour of *out* in some specific phrase-frames and constructions.

6.1 *Going out*

It seems obvious that *out* will frequently be found next to a form of the lemma GO. What is interesting is how the collocation functions. Table 1 shows some frequencies for the different corpora.

As discussed in section 2.2 above, *GO out* not only encodes movement along a path, specifically one exiting from a container, but also the crossing of the frontier between "indoors" and "outdoors" (see examples 4-6 above), and the pursuit of social goals outside the home:

(15) No, if you were feeling okay and wanted to **go out** we'd **go out** on Saturday night somewhere. Let's **go out** for a nice meal somewhere. (KD3 426)

The construction *GO out with* can function as a subset of this usage, where the other participants in the social activity are made explicit. This phrasal-prepositional verb has another, idiomatic meaning, however, particularly (though not exclusively) for teenagers: *X GO out with Y* means that X is having a romantic or a sexual relationship with Y (cf. Stenström et al. 2002: 39):

(16) Sarah's going out with someone nice, Sarah. who's she **going out with**? someone in the R A F. not Mike? no, this guy called Bob (KD6 4191)

(17) Why are you and James cuddling up to each other, you **going out with** each other, or you just a friendly hug? (b133701)

Table 1: *GO out* in the different corpora

Frequency	Wells	COLT	BNC
Out	1270	1546	12644
GO *out*	186	260	1819
as percentage of total	14.65	16.82	14.39
GO *out with*	2	122	159
as percentage of total	0.16	7.89	1.26
as percentage of GO *out*	1.08	46.92	8.74
unclear	-	1	2
with something	2	-	22
as social/work outing	-	37	77
percentage of GO *out with*	-	30.33	48.42
percentage of GO *out*	-	14.23	4.23
percentage of all *out*	-	2.39	0.61
as romantic relationship	-	84	58
percentage of GO *out with*	-	68.85	36.48
percentage of GO *out*	-	32.31	3.19
percentage of all *out*	-	5.43	0.46

Table 1 shows the proportions of *GO out*, and of *GO out with* in their different meanings, for the three corpora. Although this is not shown in the table, the progressive forms predominate, as would be expected when referring to an on-going situation (although all tense and aspectual forms are present). Clearly, who is, or has been, or will be *going out with* whom occupies a disproportionate place in the conversation of teenagers compared with speakers in general: the percentage of all uses of *out* to be found in this phrase-frame with this meaning is almost 12 times as great in COLT as in the BNC. In the Wells corpus it is completely absent, despite the recordings containing a proportion of conversation between adults on purely adult topics. The only examples of the phrase frame in

Wells refer to *going out with something*, and are about wearing the right or wrong footwear for crossing the frontier to the outdoors (cf. examples 4-6 above).

It is well known that different text types or genres can be distinguished by the range and type of grammatical structures they contain (e.g. Biber et al. 1994). It is trivial that the topic(s) of discourse in a text can be identified from its lexis. What Table 1 appears to show is that speakers' pre-occupations can be tracked in their lexicogrammar. Stenström et al. (2002: 32-41) discuss the importance of 'Romance' and 'Sex talk' as topics of conversation, something that is confirmed by one of the teenagers in no uncertain terms:

(18) Oh God this whole school revolves around snogging people, **going out with** people, shagging people. It's just a nightmare. (b142303).

Not only is it a nightmare, but it substantially alters the proportion of phrasal prepositional verbs in the usage profile of *out* in COLT!

6.2 Prepositional phrases

As discussed in section 3.3 above, *out* forms part of prepositional phrases both on its own and in the compound preposition *out of.* An examination of the phrase-frame *out NP* in the spoken BNC, using the search facility at Mark Davies' VIEW web-site, suggests that the most common simple prepositional phrases will be those where NP is definite. I therefore compared *out the X* and *out of the X* in each of the three corpora. Table 2 shows the data for the Wells corpus.

Table 2: *Out the* and *out of the* in the Wells corpus

	out the	***out of the***
number	79	46
percentage of *out*	6.22	3.62
unclear	2	-
PhrVO	6	-
percentage of *out*	0.47	-
percentage of *out the*	7.59	-
PP	71	46
percentage of *out*	5.59	3.62
percentage of *out the/out of the*	89.87	100
Frequency of nouns in PP (% of PP)	16 nouns	11 nouns
way	24 (30.4)	21 (15.7)
window	8 (11.3)	6 (13.0)
garden	7(9.9)	-
front/back	7 (9.9)	-
door-	2	-
kitchen	-	3
fridge/freezer	1	2
Verbs preceding PP (% of PP)		
no verb: static	5 (7)	1
no verb: dynamic	14 (19.7)	4 (8.7)
get	12 (16.9)	26 (56.5)
go	13 (18.3)	-
come	4 (5.6)	-
look/watch/see	9 (12.7)	4 (8.7)
manner of (caused) motion	2	-

The table suggests that both types of prepositional phrases are in fact acquired and used as ELUs. There are two nouns that are used frequently with both prepositions: *way* and *window* (since the numbers are so small, the percentages are calculated only to give an idea of proportions). The frequent verbs are *get* and the basic motion verbs. The stand-alone prepositional phrases are almost all *get out (of) the way*, and the phrases with *window* are instructions to the child to direct its gaze towards events outside the house (apart from a sentence from a story being read out, where a cat *jumped out of the window*). A similar situation was found with *on* (Hallan 2001: 105-6): although there were many more nouns occurring in PPs with *on*, they appear to be common fixed expressions for referring to everyday situations.

The frequencies in COLT (table 3, below) show how the functions of the two PPs have changed. *Out the* is still found overwhelmingly with a handful of nouns: these are clearly the fixed expressions acquired in childhood, which have doubtless persisted largely unanalysed, especially since they are correspondingly infrequent with *out of the*. On the other hand the range of nouns found with *out of the* is much greater, suggesting that this is the construction of choice for talking about a range of more specialised situations. There were also some interesting idioms (see 6.2.1, below) and one occurrence of *out of the question*.

The verbs used with both constructions have changed too: in particular the proportion of manner-of-motion verbs has gone up at the expense of the basic motion verbs. In particular, words for different styles of *throwing* are very frequent with *out the*, and this is primarily due to a notable difference in behaviour between toddlers and teenagers: whereas young children mostly look out of windows, teenagers (especially boys) spend more time throwing things out. This is perhaps another expression of the ebullient energy of the teenage informants (cf. Stenström et al. 2002), although it might also be a result of the opportunities for experiment and amusement afforded by first floor classrooms!

In the BNC sub-corpus (table 4, below) the sequence *out the* is much more frequent than *out of the*. Surprisingly the number of prepositional phrases with *out the* is higher too. However the table makes clear that this is largely due to the familiar unanalysed fixed expressions we saw in the other corpora. The PPs with *out of the* include additional fixed expressions such as *out of the blue* and *out of the ordinary,* which are presumably acquired during schooling. The large number of static locatives for *out of the*, often referring to the origin of something, confirms the suggestion made above that *out of the* focuses on the beginning of the path of motion, while *out the* relates to its end. A closer examination of the other verbs used with the different preposition forms, and the type of nouns they govern, would doubtless provide more information about the meaning difference which enables the two to continue to exist and function alongside one another.

Table 3: *Out the* and *out of the* in COLT

	out the	***out of the***
number	83	50
percentage of *out*	5.10	3.1
unclear	1	1
other	-	5
PhrVO	30	-
percentage of *out*	1.84	-
percentage of *out the*	36.14	-
PP	52	44
percentage of *out*	3.36	2.84
percentage of *out the/out of the*	62.65	88.00
Frequency of nouns in PP (% of PP)	24 nouns	38 nouns
way	6 (11.5)	2 (4.5)
window	19 (36.5)	4 (9.1)
door	5 (9.6)	-
front/back	1	-
car/cab/taxi	1	3
cinema	2	-
vending machine	-	3
Verbs preceding PP (% of PP)		
no verb: static	-	8 (18.2)
no verb: dynamic	-	2
get	5 (9.6)	4 (9.1)
Go	-	2
Come	3 (5.8)	4 (9.1)
look/watch/see	8 (12.7)	1
Take	1	5 (1.4)
throw/chuck/lob	12 (23.1)	2
other manner of (caused) motion	12 (23.1)	10 (22.7)

Table 4: *out the* and *out of the* in the BNC sub-corpus

	out the	***out of the***
Number	713	386
percentage of *out*	5.64	3.05
Unclear	10	-
Other	2	-
PhrVO	220	3
percentage of *out*	1.73	-
percentage of *out the*	30.86	-
PP	481	383
percentage of *out*	3.8	3.0
percentage of *out the/out of the*	67.5	99.2
Frequency of nouns in PP (% of PP)	~130 nouns	~130 nouns
way/road	110 (22.9)	44 (11.5)
window	5 (1)	8 (2.1)
door	36 (7.5)	6 (1.6)
front/back	49 (10.2)	5 (1.3)
car/cab/taxi	19 (4)	10 (2.6)
house	13 (2.7)	11 (2.9)
fridge/freezer	9 (1.9)	7 (1.8)
Verbs preceding PP (% of PP)		
no verb: static	-	33 (8.6)
no verb: dynamic	25 (5.2)	5 (1.3)
get	66 (13.7)	113 (29.5)
go	40 (8.3)	9 (2.3)
come	36 (7.5)	36 (9.4)
look/watch/see	17 (3.5)	1
take	10 (2.1)	5 (1.3)
throw/chuck/lob	12 (2.5)	3
other manner of motion	~25	~50

6.2.1 Verbal and physical assault

In the course of examining the prepositional phrases, I observed a number of uses of a more complex construction, *X V N out of Y* (no. 11 in section 5.2 above). The vast majority of examples were from two phrase-frames:

(a) X take the mick/mickey/piss out of Y;
(b) X beat/knock/punch/blow the shit/crap/fuck/stuffing out of Y.

Stenström et al. (2002: 200-8) show how important ritual insult is as a part of teenagers' interactions, particularly among the boys. It is therefore unsurprising that (a), which is used to talk about what one has said or will say, or in some cases to defuse a situation, is so frequent in COLT.

(19) <laughing>I'm just taking the piss out of you Jock and it's working, for the first time in my life it is working. (b141801)

Phrase-frame (a) occurs 38 times in COLT, all but 4 with *piss*. In contrast, it only occurs 18 times in the BNC sub-corpus, 9 with *piss*, 4 with *mick* and 5 with *mickey*. Phrase-frame (b), used to talk about physical assault, is reassuringly infrequent in both corpora, occurring only 4 times in COLT and, even less frequently over all, only 5 times in the BNC sub-corpus. The BNC also has *frighten the life/living daylights out of X* (3 times) and *take it out of X* once. These are the same construction, and similarly negative in effect, but very infrequent. The use of phrase-frame (a) in COLT, almost 1 per cent of all occurrences of *out of* (312) in any construction, testifies to the importance of this sort of interaction in teenage talk.

7. Conclusion

The cases I have discussed here represent only a small part of the information which the systematic study of path morphemes is bringing to light. An important goal is to construct the usage profiles of *out* for children of different ages, for teenagers and for a large sample of English conversation. It is clear that even the interim results presented here can give new insights into the functions of a class of English words whose role is too often assumed to be unambiguously prepositional. The range of information, from construction frequency to sociolinguistic insights, to which the techniques of corpus linguistics give access makes the use of such empirical methods an essential part of linguistics.

Notes

1 I would like to thank Andrea Gerbig and Oliver Mason for inviting me to contribute to this volume and for their helpful comments, and above all Mike Stubbs for all his encouragement and patience.

2 In this section and section 3 I will mainly give examples from the language use of young children hearing and acquiring British English, and from that of their entourage. In the following sections I will also include data from the COLT and BNC sub-corpora. Examples from the Wells corpus are identified by the age band, based on the numbering of the Wells files, and the name of the child, with the child's age given in the format y; m.d. There is no age band 1. Most of the Wells formatting, such as intonation codes, has been removed, and more than one utterance has been placed on one line. The mother and father are designated by *MOT and *FAT, the target child and other participants by *XXX, where XXX is represents the first 3 letters of the name. Examples from COLT and the BNC sub-corpus are identified by the text reference. BNC data cited in this article have been extracted from the British National Corpus World Edition, distributed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

3 Biber et al. (1999) do not mention the prepositional use of simple *out*. It is not apparently something taught to second language learners: works such as Murphy (1994) or the *Longman Language Activator* (1993) contain no mention of it. The *Collins COBUILD English Dictionary* (1995) states that "[i]n American English and informal British English, *out* is often used instead of *out of*." The *Collins COBUILD Dictionary of Phrasal Verbs* (1989: 477), in the entry on *out* in the Particle Index, states that

> "[i]n some varieties of English such as American English, and also in non-standard British English, **out** can also be used as a preposition with verbs of movement. […] in standard British English you need to add the preposition 'of'."

Since "non-standard" is often a euphemism for "low-class", it is worth examining the distribution of prepositional *out* in the BNC sub-corpus, using the socioeconomic class information in Lee's (2003) index. 27% of the texts were recorded by AB (upper and upper-middle class) speakers, and produced 17% of the examples. The 33% of texts recorded by C1 (lower-middle class) speakers contained 29% of the examples. C2 (skilled working class) informants accounted for almost 23% of texts and 37% of examples, DE (unskilled working class and unemployed) and unclassified informants provided nearly 18% of texts and almost 17% of examples. These figures are scarcely conclusive, especially since the class indicators refer only to the person providing the recording; without identifying the individual speaker for each example it is difficult to claim that the whole of any text represents only the language use of a particular social group. One thing is certain, however: British people are all taught in school that using *out* rather than *out of* is simply "wrong". That so many of them do so, at least in casual conversation, would doubtless be regarded by many as yet another example of declining standards!

References

Biber, D., S. Conrad and R. Reppen 1994. 'Corpus-based approaches to issues in applied linguistics', *Applied Linguistics,* 15 (2), 169-189.

Biber, D., S. Johansson, G. Leech, S. Conrad, E. Finegan 1999. *Longman Grammar of Spoken and Written English*. Harlow, England: Pearson Education Limited.

Bowerman, M. 1996. 'The origins of children's spatial semantic categories: cognitive versus linguistic determinants', in: J.J. Gumperz and S Levinson (eds.), *Rethinking Linguistic Relativity*, (Studies in the Social and Cultural Foundations of Language No. 17), Cambridge, Cambridge University Press, 145-76.

The British National Corpus, version 2 (BNC World) 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Bybee, J 1998. 'The emergent lexicon', *CLS 34: The Panels*, Chicago: Chicago Linguistic Society, 421-35.

Davies, M. 2005. 'VIEW: Variation In English Words and Phrases' URL: http://view.byu.edu/

Gries, S. Th. and A. Stefanowitsch 2004. 'Co-varying collexemes in the into-causative', in: M. Achard and S. Kemmer (eds.), *Language, culture, and mind*, Stanford: CSLI Publications, 225-236.

Hallan, N. 2001. 'Paths to prepositions? A corpus-based study of the acquisition of a lexico-grammatical category', in: J. Bybee and P. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*, Amsterdam/Philadelphia: John Benjamins Publishing Company, 91-120.

Hunston, S. and G. Francis 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam/Philadelphia: Benjamins.

Lakoff, G. 1987. *Women, fire and dangerous things: What categories reveal about the mind).* Chicago/London: University of Chicago Press.

Langacker, R.W. 1987. *Foundations of cognitive grammar, 1: theoretical principles*. Stanford: Stanford University Press.

Langacker, R.W. 2000. 'A dynamic usage-based model', in: M. Barlow and S. Kemmer (eds.), *Usage-based models of language*, Stanford: CSLI Publications, 1-63.

Lee, D.Y.W. 2001. 'Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle', *Language Learning and Technology*, 5 (3): 37-72.

Lee, D.Y.W. 2003. *The BNC Index* (Excel spreadsheet available for download at http://clix.to/davidlee00).

MacWhinney, B. 2000. *The CHILDES project: Tools for analysing talk.* Mahwah, NJ: Lawrence Erlbaum Associates.

Moon, R 1998. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Clarendon Press.

Murphy, Raymond 1994. *English Grammar in Use: A self-study reference and practice book for intermediate students.* Cambridge: Cambridge university press.

Pawley, A. and F.H. Syder 1983. 'Two puzzles for linguistic theory: Nativelike selection and native fluency', in: *Language and Communication*, J.C. Richards and R.W. Schmidt (eds.), London: Longman, 191-226.

Römer, Ute, 2005. *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics.* (Studies in Corpus Linguistics 18). Amsterdam/Philadelphia: Benjamins.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University press.

Snow, C.E. 1995. 'Issues in the study of input: Finetuning, universality, individual and developmental differences, and necessary causes', in: P. Fletcher and B. MacWhinney (eds.), *The Handbook of child language*, Oxford: Blackwell, 180-193.

Stenström, A.-B, G. Andersen and I.K. Hasund 2002. *Trends in teenage talk: Corpus compilation, analysis and findings.* Amsterdam/Philadelphia: Benjamins.

Stubbs, M. 1996. *Text and corpus analysis: Computer-assisted studies of language and culture*. Oxford: Blackwell.

Stubbs, M. 2001. *Words and phrases: Corpus studies of lexical semantics.* Oxford: Blackwell.

Stubbs, M 2004. 'On very frequent phrases in English: distributions, functions and structures', plenary lecture given at ICAME 25, the 25th anniversary meeting of the International Computer Archive for Modern and Medieval English, in Verona, Italy, 19-23 May 2004.

Talmy, L. 1985. 'Lexicalization patterns: semantic structures in lexical forms', in: T. Shopen (ed.) *Language typology and syntactic description. Vol. III: Grammatical categories and the lexicon*, Cambridge: C.U.P., 57-149.

Talmy, L. 1991. 'Path to realization: Via aspect and result', *Proceedings of the 17th Annual Meeting of the Berkeley Linguistics Society*, Berkeley: Berkeley Linguistics Society. 480-520.

Theakston, A.L., E.V.M. Lieven, J.M. Pine and C.F. Rowland 2005. 'The acquisition of auxiliary syntax: BE and HAVE', *Cognitive Linguistics*, 16-1, 247-277.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work* (Studies in corpus linguistics 6). Amsterdam/Philadelphia: Benjamins.

Tomasello, M. 1992. *First verbs: A case study of early grammatical development.* Cambridge: Cambridge University Press.

Wells, G. 1981. *Learning through interaction: The study of language development.* Cambridge: Cambridge University Press.

Wells, G 1986. *The meaning makers: Children learning language and using language to learn.* Sevenoaks, England: Hodder and Stoughton.

I don't know – differences in patterns of collocation and semantic prosody in phrases of different lengths

Bettina Starcke

University of Trier

Abstract

Different realisations of a lemma have different collocations (cf. Sinclair 1991 and Sinclair in Moon 1987: 89). This is also true for the different longer variants of a short phrase.

Taking the most frequent 3-gram in the BNC, 'i do n't', as a basis or nucleus phrase for the analysis, 4- to 6-grams which include 'i do n't' are analysed for their collocations and semantic prosodies. The results reveal that there are distinct differences in the usage of the n-grams. While the 3-gram 'i do n't' and the 4-gram 'i do n't know' collocate with hedging expressions and markers of uncertainty, their literal meaning, not knowing something, frequently disappears as their core meaning. In contrast to that, the 5-gram 'i do n't know what' and the 6-gram 'i do n't know what you' are mostly used in their literal sense which is negating knowledge of something. Unlike the 3- and the 4-grams, the longer phrases also have a distinct semantic prosody, namely that of anger, aggression, despair and frustration. A second, briefer study of phrases containing 'the end of' toward the end of the article will support the hypothesis that phrases of different lengths but with a shared nucleus phrase have different collocational patterns and distinct semantic prosodies.

In the second part of this article, explanations for these differences in collocations and semantic prosodies are offered. It is suggested that the longer a phrase, the more predetermined is its use. This is because semantic, pragmatic and syntactical restraints increase with the length of a phrase. Finally, implications of these findings for corpus linguistics in general are discussed.

1. Introduction

Language is frequently understood as a system of words which are formed into strings on the basis of grammatical rules. Words form into phrases, sentences, paragraphs and texts. The main unit of meaning and therefore also linguistic analysis has traditionally been the word. Only fairly recently have phrases become a focus in the analysis of language when linguists have recognised that they too carry meaning in language.

Since this finding has been relatively recent, basic concepts of the analysis of language have so far mainly been applied to the study of words and not to the study of phrases and phrasal meaning. This is for example the case with the concepts of collocation and semantic prosody.

1.1 Collocation and phraseology

The concept of collocation is one of the most essential in corpus linguistics. Firth discusses it as early as 1951 and later on describes it by saying that "you shall know a word by the company it keeps" (1958: 179). Leech (1974: 20) takes up the idea and introduces the concept of

> collocative meaning [which] consists of the associations a word acquires on account of the meanings of words which tend to occur in its environment. (...) collocative meaning is simply an idiosyncratic property of individual words.

He then goes on to classify collocative meaning as an associative meaning and defines it as "What is communicated through association with words which tend to occur in the environment of another word" (26). Since it is an associative meaning, it is individual to the speakers of a language depending on whether they recognise this association or not.

The introduction of corpus linguistic techniques in the analysis of language has changed this perception of collocation as individual to the different speakers back to Firth's original notion. Collocation is now defined as the habitual co-occurrence of words as observable in a corpus. It is not intuitive anymore and the concept is firmly based on empirical evidence that was not available to Firth and Leech at the time when they wrote their definitions.

But not only are collocations intersubjective, empirical research has also revealed that different realisations of one lemma have different collocations. Sinclair (1991) first demonstrated that with an analysis of the different forms of YIELD in his discussion of how the lemma should be represented in a dictionary. What looked like a fairly easy concept – a lemma – turned out to be highly complex.

Partington (2004) confirms that different forms of one lemma, in his case study HAPPEN, occur in different contexts. While *happens* occurs in predominantly neutral contexts in his corpus of academic writings, *happened* mostly occurs in a negative context, and *happen* occurs twice as much in negative or neutral contexts as in positive ones. He also finds that this, as he calls it by referring to Hoey (2004), semantic priming is "realised within and through separate and typical phraseologies, characteristic syntactic patterns" (141).

While the differences in collocational patterning between different forms of lemmata have been discussed in depth, the application of this finding to phrases of various lengths has not been analysed. Phraseological research has frequently focussed on multi-word units as units of meaning (for example Sinclair 1996 and Stubbs 2005). Stubbs (2005) approaches this question by asking why the node *world* is among the most frequent words in the BNC. He finds that it is often part of a longer phrase such as *best in the world* which is extremely frequent in the language. It is therefore not the single word but the phrasal construction which is frequent. Consequently, phrases are seen to be the units of meaning in language and not single words. Words cannot be seen in isolation but have to be looked at in their contexts.

This ties in with one of Stubbs' earlier findings (2001), namely that there is a relationship between the semantic prosody of a word and its syntactic surroundings. For example ACCOST has a predominantly negative semantic prosody and consequently frequently occurs with passive constructions in which someone is accosted by somebody else. It seems unlikely that somebody would describe one's own behaviour as accosting someone.

It is again Partington (2004) who argues in a similar way. In his analysis of semantic prosody and semantic preferences in language, he finds that "certain prosodies/preferences are typically expressed using certain phraseologies" (144). This means that phrases instead of single words carry prosodic and connotative meaning, and adds further evidence for Sinclair's proposition (1996: 30) that "units of meaning are expected to be largely phrasal". Sinclair gives evidence for this claim by discussing the phrasal meanings of e.g. "naked eye" and "true feelings" which he both finds to be "the core of a compound lexical item" (21) with inherent components such as semantic prosody and semantic preference. The phrases must therefore have inherent meanings.

A second major area of phraseological research has been to look at phrases as characteristic features of text-types. Studies by for instance Stubbs & Barth (2003), Cortes (2002) and Biber & Conrad (1999) have revealed that different phrases are characteristic for different genres. The latter two articles discuss academic writing by undergraduate students and find that much of it is patterned. Stubbs & Barth (2003) show that the most frequent lexis and phrases differ between genres. In this article, I discuss how situation specific frequent phrases are.

1.2 Semantic prosody

The concept of semantic prosody describes the characteristics of a word in terms of its semantic context. This context has implications for the meaning of a word since the prosody becomes part of the word-meaning. The term and the concept of semantic prosody have been introduced by Louw (1993) who defines prosodies as "reflections of either pejorative or ameliorative [semantic] changes [over a period of time]" (169) and says that "prosodies based on frequent forms can bifurcate into 'good' and 'bad' " (171). This means that prosodies of words have developed diachronically and that words can adopt an either positive or negative connotation through their contexts.

The unit of analysis of semantic prosodies has mostly been the traditional unit of meaning – the word. For example Sinclair (1991) analyses the phrasal verb *set in* and discovers that it mostly occurs in a negative context. Looking at this finding in retrospect, we can now say that it has a negative semantic prosody.

2. The analysis

The analysis of semantic prosodies and collocations has, as shown above, mostly focussed on single words or phrasal verbs. But also phrases of different lengths which share a core part differ in their collocations and semantic prosodies. This will be demonstrated in the following analysis.

To do so, I will firstly analyse the semantic prosodies and connotations of four related phrases of different lengths containing *i do n't* on the basis of concordance lines. I will demonstrate that their prosodies are distinct and display a progression from an expression of uncertainty and of hedging to a distinctly unpleasant semantic prosody of aggression or defiance as a response to an aggression. Following this analysis, I will give some tentative explanations for the variation in semantic prosody and connotation with phrases of different lengths. I will show that semantic, pragmatic and syntactic constrains on the phrases increase with their lengths. My explanations will be supported by further analyses of phrases containing *the end of* which tie in with my original findings.

2.1 The terminology

By the term *nucleus*, I understand the core part of a phrase, for example *i do n't*. The *extension* is that part of the phrase which is added to the nucleus in longer realisations of it, for example *i do n't* <u>*know*</u> and *i do n't* <u>*know what you*</u>. While the nucleus is the core of the phrase and contains its core meaning, the extension determines the contextual meaning and the collocational patterning of the phrase. A phrase, that is an uninterrupted string of n words, is called n-gram.

2.2 The data

The data for the analysis is the British National Corpus (BNC) and some of its most frequent phrases of different lengths that share a nucleus. The phrases for my primary analysis are

- *i do n't*
- *i do n't know*
- *i do n't know what* and
- *i do n't know what you*

(the spacing between do and n't occurs in the original data from the BNC and will therefore be retained in the entire article, see the Appendix for the frequency lists of multi-word units in the BNC). The findings from this analysis will be supported by an analysis of a second, much briefer, analysis of the phrases

- *the end of*
- *the end of the*
- *at the end of the* and
- *at the end of the day.*

These phrases are also among the most frequent phrases of the BNC of their respective lengths.

The data for the analyses have been extracted from the BNC by disregarding the POS markup of the corpus. While the knowledge about parts of speech (POS) is interesting and important for much research, it is of no consequence for the present one. My aim is to look at the most frequent phrases in the language in absolute terms and independent of grammatical categories.

The phrases I look at are all among the 20 most frequent phrases of their respective lengths (see table 1). They share a nucleus 3-gram as their basis which provides adequate data for a comparison of semantic prosodies and collocations of phrases of different lengths.

I will proceed by firstly analysing concordance lines of the different phrases and by secondly comparing my findings for the phrases. The concordance lines have been extracted from the BNC by using the database *Phrases in English* (PIE, Fletcher 2003/04) which displays randomly selected concordance lines for every query. This is the textual evidence for my findings. Concordance lines given as evidence for my claims have been chosen so that no concordance line occurs as evidence for phrases of different lengths.

2.3 First findings

When looking at the 20 most frequent 3- to 6-grams of the BNC, one thing to notice is that only a limited number of phrases occurs on the lists (see the Appendix for the complete lists). Longer phrases frequently consist of one nucleus with various extensions of different lengths. This is for example the case with *the end of, the end of the, by the end of the, at the end of the day* and *by the end of the year*. While the nucleus remains constant, the extensions vary with the length of the phrase.

The nucleus *i do n't* occurs in the following realisations of the top 20 3- to 6-grams in the BNC:

Table 1: phrasal realisations of *i do n't*

Phrase	occurring mostly in written or spoken language	Position on the respective n-gram list
i do n't	spoken	1
i do n't know	spoken	1
i do n't think	spoken	4
i do n't want to	spoken	3
i do n't know what	spoken	4
i do n't think i	spoken	13
i do n't know how	spoken	20
i do n't think it 's	spoken	9
i do n't know what you	spoken	20

It is striking that all of the phrases above occur mostly in spoken language as represented by the BNC while most of the other phrases among the top 20 which are not listed here mostly occur in written language. The occurrence of the short forms *do n't* and *it 's* hints at their use in spoken language.

The frequent occurrence of *n't* within the most frequent phrases indicates that negations are particularly formulaic in spoken language while written language seems to be more formulaic in a wider range of aspects. Since, according to Watt (1960), there are no negatives in nature but only in the human mind, this finding suggests that expressing one's disappointment of positive expectations is frequent in spoken language. The fact that the negation mostly refers to mental processes (*know* and *think*) indicates that these are dominant in people's discourse.

The data selected for the primary analysis of this article are the various realisations of *i do n't* in connection with *know*. Findings for phrases which include *think* in their extensions will be discussed with the implications of the analysis.

2.4 The analysis

As already mentioned, the different realisations of *i do n't* plus *know* have distinct semantic prosodies and are increasingly restricted to being used in particular kinds of situations the longer the phrases are. They are not necessarily text-specific but rather specific to situations of usage. In addition, the phrase becomes increasingly delexicalised the longer it is. While the 3-gram is mostly used in its literal sense of negating something, speakers of the longer phrases frequently use them to express indignation, anger or frustration. The semantic and pragmatic contexts of the phrases differ between phrases of different lengths.

As a first step, I will analyse the connotations and semantic prosodies of each phrase. Secondly, I will compare my findings for the four phrases and will, thirdly, give some tentative explanations for the differences in usage found between the phrases. This explanation suggests that variation in semantic prosody and connotation between related phrases of different lengths is a pervasive phraseological phenomenon.

2.4.1 i do n't

The concordance lines of *i do n't* show that the phrase is mainly used in its literal sense. It occurs in constructions which literally describe something that the speaker does not, for instance, know, agree to or think. It frequently collocates with mental verbs, hedging expressions and expressions of uncertainty. This is exemplified by the following selection of concordance lines:

"I wanted to kill her. **I don't** know what stopped me.

"Can't do any harm. But **I don't** think it would do any good."

It was Keith who spoke first, looking red-faced and embarrassed." **I don't** want to sound awkward, but -- you said Free People were devious. Don't you think it might be a hoax?

Don't you have any feelings for what I'm goin' through?" "Waal, **I don't** exactly feel too good myself," he replied morosely.

Fiona Fullerton **I don't** think I was told a great deal about the facts of life. My mother was a Calvinistic Methodist, so you can draw your own conclusions.

And why are we telling them this if it's not eligible for consideration as a possible way of depriving us of an cash? **I don't** I mean probably it's just to be to the safe side. Mm.

Doesn't mention her mother, and Leo's whole thesis about was built on this single phrase in Who's Who. words, some future researched, that the printer here, missed that bit, you know, she should have said, and her mother's name, but her mother's got missed out on the proofs or somethin, **I don't**, this is the kind of thing that happens, of course. Leo's entire book will collapse, er, as, as, as perhaps it should.

The analysis of the concordance lines does not tell us anything unexpected. The phrase is used as a negation and frequently occurs in spoken language. What is intuitively perhaps not quite expected is the frequent collocation of the phrase with mental verbs and expressions which denote uncertainty in this particular context, such as *exactly* and *but*. They function as hedging expressions and the negation inherent in the sentence appears tentative. The occurrence of mental verbs as collocates of the phrase hints at mental processes being frequent topics in conversations.

The 3-gram frequently collocates with unpleasant notions, such as a project collapsing or wanting to kill somebody. It therefore has a semantic prosody of unpleasantness. This is not entirely surprising with a phrase containing a grammatical negation, but the fierceness of this unpleasantness could not be predicted from the grammatical negative alone.

2.4.2 i do n't know

On the first glance, there are again no surprises in the analysis of *i do n't know* following the analysis of the 3-gram. The phrase includes a mental verb, a semantic class which has been identified as a frequent collocate of the 3-gram, and the phrase collocates with hedging expressions so that it appears tentative in its negation. But unlike the 3-gram, *i do n't know* is frequently delexicalised in its use and fulfils the function of a filling phrase or a hedging expression itself. It is used to soften a statement and to hedge it when its literal meaning, not knowing something, can be inferred from the context. The explicit acknowledgement of not knowing is only a secondary message of the 4-gram. This is exemplified by the following concordance lines:

But I would have thought, I mean it **I don't know** if it was No, no

What actually happened there? Well **I don't know** really. All I know is the reports which the lads filed.

Oh what's the time then? Oh **I don't know**, I don't know Nearly eleven, this stop

When is it? **I don't know**, not very long. I think so I think we'd said we'd go.

Mm. **I don't know**, but I know he wanted it. about the turbo or summat.

It seems so recent. But **I don't know** whether that policy still prevails but Erm I I think does but er I don't think it er became quite the successful initiative that er that they hoped it would be.

What? **I don't know.** I didn't know that, I thought I'd put down what, some of the people one of the babysit for as one of my references because I babysit her kids at primary school age.

These concordance lines show that the phrase does not express a definite negative of a fact or process in all of its occurrences. It is instead used as a hedger or to conceal uncertainty. The collocation of the phrase with softening expressions softens the negation *not*. While the negation of a state or process is one function of the phrase, the second function is to express uncertainty and to weaken the message of the surrounding sentence(s).

2.4.3 i do n't know what

While there are similarities in the usage of the 3- and the 4-grams, *i do n't know what* shows distinct differences in its connotations and collocations. The 5-gram is mainly used in its literal sense of not knowing and the connotations of the phrase include anguish, frustration, anger or failing to understand something but thinking of its consequences as negative. One distinct connotation is that of gratefulness for an action or service which prevented something negative happening to the speaker. This is observable in the following concordance lines:

Make a copy for me while you're at it." "**I don't know what** it will do to her when she hears that they're the wrong bodies." "Does she have to be told?"

He thought he'd got a job for life when he got his old mate Humphrey in as master -- they were at school together, you know -- but all that's backfired pretty badly. **I don't know what** they fell out over, but it must have been serious.

Eat it while it's hot." "**I don't know what** I'd have done without you these last weeks, Carrie. You've been a God-send, and no mistake."

Now look here! **I don't know what** you're implying, but, for your information, I have no idea what those goons wanted. I've done nothing to put myself in a position where I have… hit men coming after me!"

Me sister never showed 'er legs in all 'er life, nor me, neither. **I don't know what** girls are comin' to. Ain't it shockin', mister?" she said to Joe.

Last night, Johanna's father broke down as he revealed his daughter had mysteriously taken down her 50 Christmas cards from her bedroom wall only hours before disappearing. "**I don't know**

what happened,″ said Robert, 40, a self-employed TV repairer and electrician. "The last time I saw her she was bubbly and full of life.

Oh be quiet Er **I don't know what** we'd do without you Paul I'm gonna, I'm gonna Potatoes are nearly ready

The concordance lines show a semantic prosody of conflict and problem which differs from that of the 3- and 4-grams. The semantic prosody for the shorter phrases was that of hesitation and uncertainty. While the 5-gram collocates with conflict, violence or aggressions do not seem to be involved.

2.4.4 i do n't know what you

Again, the collocations of the 6-gram slightly differ from those of the previously discussed n-grams. While the previous phrases had semantic prosodies and connotations that were mostly peaceful, the connotation of the 6-gram is often that of either aggression or defiance with the literal meaning of not knowing prevailing. Despair or frustration do not anymore collocate with the phrase. This is exemplified by the following concordance lines:

The question made her flinch. **"I don't know what you** mean.″ "I mean -- have you been avoiding me like the very plague simply because of who I am?″

Sweat from the washing-up misted her forehead and nose. **I don't know what you**'re talking about, Léonie lied: I don't know what's the matter with you. Thérèse clasped the biscuit tin in the crook of her arm.

"So you can't tell me much?″ **"I don't know what you**'re after,″ she said. "I don't reckon you know yourself.

He hadn't backed down on that. Her throat constricted and she swallowed hard before stammering, "I -- **I don't know what you** mean. I'm not acting.″

She'd never felt so weak, so helpless. **"I don't know what you** mean,″ Ruth said huskily. "Let me tell you.″

Court had decided to bluff. He said, **"I don't know what you**'re talking about. What's more, you now that I don't.

Surely we should try and see **I don't know what you**'re talking about, we of course we're interested in peace, we want peace. We

The semantic prosody of the phrase is, in a large number of occurrences, either explicitly or implicitly that of open and strong conflict. Usually it is the person in the weaker position, that is the person being accused of something, who uses the phrase. It therefore takes on the characteristics of a defence against verbal aggression. It is frequently used as a response to a statement, often an accusation or unwelcome question, and the statement either preceding or following the utterance frequently includes an aggression against the speaker of the phrase. The phrase itself functions as a defence against this aggression. Also a gender bias in the use of the phrase is observable: it seems to be used by women mostly. This can either hint at a gender bias in terms of who sender and receiver of aggression in society are or at a female communicative strategy.

2.5 An attempt at explanation and systematisation

The analysis above has shown that collocations and semantic prosodies of the phrases seem to fall into two groups. The 3- and the 4-grams *i do n't* and *i do n't know* frequently function as hedges and softeners in discourse. They express uncertainty but without having distinct semantic prosodies. The 5- and 6-grams *i do n't know what* and *i do n't know what you* frequently collocate with expressions of aggression, despair or frustration. Their semantic prosody is that of conflict. The two groups of n-grams seem to differ significantly in their functions.

But on closer analysis, a gradation rather than a division into two groups of collocations and semantic prosodies of the four phrases becomes visible. The gradation of usage ranges from a literal use of the 3-gram to an increasing restriction of context and semantic prosody of the phrases. The 3-gram is often used in its literal sense to negate something and frequently collocates with hedging and softening expressions. The 4-gram is already more restricted in its usage: negating knowledge of something becomes secondary to hedging a statement and to softening its propositional force. The phrase is partly delexicalized and its semantic prosody is more restricted than that of the 3-gram.

This development away from the literal meaning of the phrase is reversed with the 5-gram: *i do n't know what* is again used mainly in its literal sense of not knowing something – even though it is not always visible whether the speaker really does not know or whether he pretends not to – , but it is used in different kinds of situations than the 3- and 4-grams. While the latter frequently occur in contexts where the speaker aims at avoiding conflict, the 5-gram occurs in contexts where conflict has already arisen. The function of the phrase appears to be to deescalate conflict.

Also the 6-gram *i do n't know what you* is mainly used in its literal sense and it also collocates with expressions of conflict such as *Shut up, bloody, flinch* and *bluff.* Again, conflict is the dominant semantic prosody of the phrase. Compared to the 5-gram, there is an increase in fierceness in tone of the utterances in which the phrase occurs. Its frequent use as a defence against aggression distinguishes it from the 5-gram.

An increase in length of the phrase leads to an increasingly fixed semantic context in which the phrase occurs. While the 3- and the 4-grams do not have distinct semantic prosodies, the 5- and the 6-grams do. This hints at a relationship between the length of a phrase and possible restrictions in its usage with restrictions increasing with its length.

Also the functions of the phrases become more distinct with an increase in length. While the 3- and the 4-grams are mostly used in their literal senses as responses to queries, the 5- and 6-grams are mostly used as responses to verbal aggression. They are used as a defence against accusations and, by not giving in to these accusations, the phrases function as markers of hidden aggression or defiance on the part of their speakers. This becomes more pronounced the longer the phrases are. The functions of the shorter phrases are included in those of the longer ones even though they occur in increasingly restricted contexts. The longer phrases therefore have two pragmatic functions: They firstly function as a defence against aggression or as an expression of ignorance on being faced with aggression. Secondly, they fulfil a hedging function in so far as not the actual content of the utterance is of primary importance but rather the expression of innocence.

These findings entail that the longer phrases are, the more situation-specific they become. Therefore also the semantic and pragmatic contexts of the phrases become increasingly restricted and predetermined. This results in a distinctly negative semantic prosody of the 5- and the 6-grams.

This gradual change in connotation and the emergence of a semantic prosody with the longer phrases is due to the length of the phrases and the resulting increase in semantic, pragmatic and possibly syntactic restrictions in their use: the longer a phrase, the more specific its function. The shorter a phrase, the greater the possible variation of language co-occurring with the phrase. A wider range of semantic, pragmatic and possibly syntactic variation may occur with short phrases which can be used in a larger number of contexts. Conversely, also the language co-occurring with the phrases becomes increasingly restricted the longer the phrase is. Restrictions on the semantic context of the phrases increase with their lengths. This results in pragmatic restrictions which are mirrored in the semantic prosody of the phrases. This indicates that phrases do not become increasingly text specific the longer they are, but rather situation- or genre-specific.

3. Further evidence: the case of *the end of*

A second analysis of phrases provides further evidence for the claims made above. The data for this analysis are the following phrases:

- *the end of*
- *the end of the*
- *at the end of the* and
- *at the end of the day.*

The phrases all range among the top 15 phrases of the BNC of their respective lengths disregarding the POS markup of the corpus.

The nucleus *the end of* occurs in the following realisations among the top 20 n-grams of the BNC:

Table 2: phrasal realisations of *the end of*

Phrase	Position on the respective n-gram list
the end of	3
the end of the	2
at the end of the	1
at the end of the day	1
by the end of	15
by the end of the	2
by the end of the year	7

In the following section of this paper, I am going to discuss the four phrases containing the nucleus *the end of* either preceded by *at* or with *the* as part of the extension, but excluding those phrases in which *by* precedes the nucleus. The selected phrases are all among the three most frequent phrases of their respective lengths in the BNC.

3.1.1 the end of

The dominant use of the phrase is mostly to express its literal meaning – indicating a final state –, with the temporal dimension being dominant. This can be seen for example in the following sample of concordance lines:

Members often meet up to carry out homework tasks together, and contact each other between sessions. At **the end of** each course, members are invited to exchange telephone numbers and addresses, and small self-help groups often develop.

On one such incident I was in command of Venturous patrolling in the Straits of Dover **at the end of** a very busy Bank Holiday, during which we had followed a suspect yacht from just outside Calais, and handed her over to our special unit in Dover.

And there's the target for go to go for. It's not necessarily erm a target if you don't get it hard luck er that's **the end of** your s your time with us, it's just a target that we would like we know that we're gonna clear all our costs out of that first year.

Swayed by the prospects for economic revitalisation, governor Weld gave MASSMoCA another$688,000 and until **the end of** 1992 to raise $12 million as proof of private-sector support. The campaign fell $8 million short.

Non-literal usage is infrequent in realisations of the phrase that do not continue with the extensions discussed in the following.

3.1.2 the end of the

Also the 4-gram is mostly used to express its literal meaning, again with a predominantly temporal dimension. Evidence for this are the following concordance lines:

The EC Commission said that much the same in its submission to the Energy Committee: "At the present time the FBR is the only reactor type which could, if introduced early enough, extend the lifetime of our uranium resources to **the end of the** next century -- and beyond."

My preconceived ideas about this course, which was held at Manchester University, were completely erased by **the end of the** first evening.

In fact, during the Hundred Years War (1337 to 1453), the people of Bordeaux took the English side, and many vineyards were destroyed in revenge. It was not until **the end of the** 18th century that the first bottle of claret as we now know it was put down for ageing at the famous Chateau Lafite in 1797.

So the patterns of moral respectability, far from being a simple assimilation of the middle-class norm, were effects of specific class experiences and a growing sense of class identity. There were even signs, by **the end of the** nineteenth century, of increased intermarriage between the skilled worker and other strata of the working population, a sure indication of a diminishing sense of social distance.

While the temporal bias in the usage of the phrase is visible, it is so to a lesser degree than the bias of the 3-gram. A tendency toward a more metaphorical, that is delexicalized, usage of the phrase becomes visible.

3.1.3 at the end of the

The 5-gram has got three main tendencies of usage. It firstly and primarily indicates temporal finality and therefore carries its literal meaning. This tendency is less dominant than with the 3- and 4-grams though. It secondly collocates with spatial expressions. This is a move away from the mostly temporal dimension dominant with the shorter phrases.

The third tendency is a more metaphorical usage of the phrase as part of a fixed phrase. While most of the phrases in which this metaphorical usage is apparent are part of the 6-gram *at the end of the day*, the metaphorical usage also occurs when the phrase continues differently.

The following concordance lines give evidence for the three types of usage:

However, they have one disadvantage. **At the end of the** season their leaves are frequently dulled and disfigured by powdery mildew.

At the end of the Mass there is a touching prayer as men go out into the world linking the sacrament of the one bread which binds all men in God with the bread and ale of human meeting:

While HRP can be claimed by both sexes, it predictably applies more frequently to women. For more information, see "Pensions for Women" **at the end of the** chapter or obtain leaflet NP 27 Looking After Someone at Home? How to Protect Your Pension from your local Social Security office.

```
Gary Moore uses an idea similar to this at the end of the first
      verse of the track Story Of The Blues, from his "After Hours"
      album.
Another common argument is to point out that everything in the world
      must have a cause, but that at the end of the line there must
      be an uncaused or "first" cause.
```

The collocations between the 5-gram and temporal and spatial expressions are the most frequent in the sample of concordance lines from the BNC with the temporal expressions dominating.

3.1.4 at the end of the day

The dominant usage of the 6-gram differs considerably from that of the 3- to 5-grams. The phrase is predominantly used in a metaphorical context indicating the end of something. While the temporal dimension observable with the other phrases is still present, the main usage has shifted away from a concrete point in time toward an undefined period or occasion. Evidence for this are the following concordance lines:

```
When a bank creates a loan in a multi-bank system, the customers may
      write cheques in favour of customers of other banks and at the
      end of the day, all the banks have claims against each other.
If we are expecting a good level of practice from proprietors,
      whether private proprietors or statutory, then we have not to
      expect them to be out of pocket at the end of the day as a
      result.
It's not maybe as rare as you ought to have it but it really tastes
      nice. With disasters and all it's come out reasonably well at
      the end of the day.
The old prejudices still remain though, and at the end of the day
      far too few people really applauded this season's winner.
```

The phrase *at the end of the day* is non-compositional and its meaning cannot be deduced from the single words. Its idiomatic character is clearly reflected in the concordance lines above which show how the temporal core meaning of the nucleus phrase is still part of the 6-gram and its idiomatic meaning. The concordance lines also show a negative semantic prosody with negatively connotated words frequently co-occurring with the phrase.

3.2 Systematisation of the findings on *the end of* and connections to *i do n't*

The analysis of the four phrases above has shown a similar pattern to that of the phrases containing the nucleus *i do n't*. In both cases, the analysis of concordance lines has shown a pattern of progression of collocational meaning. Both sets of phrases progress from a literal meaning of the shorter phrases toward non-literalness with the longer phrases. The core meanings of the phrases remain part of the longer phrases.

The semantic prosody of the phrases containing *the end of* gradually develops from neutral – stating a literal end of something – toward a negative prosody. The 6-gram collocates with words and expressions such as *obsolete, impossible, too few* and *tired*. The phrase is used when describing future consequences of actions or processes, and it is frequently used to discuss insecurity, for example whether pupils will be able to find a job after leaving school or whether certain security measures are enough to protect a building. In these cases, the phrases express insecurity and hint at possible negative events. The semantic prosody is predominantly negative.

This negative semantic prosody stands in contrast to the neutral prosody of the shorter phrases. Both the 3- and the 4-grams mostly occur in their literal senses and no prosody, either positive or negative, is perceptible. Concordance lines of the 5-gram reveal instances of both neutral and negative semantic prosody. The 3- to 5-grams do not have distinct prosodies.

This finding shows that the phrases containing *the end of* are divided into two in terms of their semantic prosodies. The 3-and the 4-grams both have neutral prosodies, the 6-gram has a negative prosody. The 5-gram takes an intermediate position with no distinct prosody.

The distinction of the phrases into two groups in terms of their semantic prosodies resembles that of the phrases containing *i do n't*. Both groups of phrases have two groups of prosodies, namely a neutral one with the 3- and 4-grams and a negative one with the 5- and 6-grams in the one case and the 6-grams in the other. While the 5-grams do not have similar prosodies, they resemble each other in so far as they both differ from the shorter phrases and resemble the 6-grams. We can therefore talk about two groups of semantic prosodies.

4. Implications and conclusion

The findings from these analyses have several implications for corpus linguistic research:

First, frequent phrases of different lengths might derive from the same shorter phrases. This is further evidence for the dominance of patterns in language.

Second, phrases function as units of meaning in language. The analyses have revealed that phrases frequently have distinct semantic prosodies or fulfil distinct functions in discourse. A phrase might, for instance, function as a hedging expression.

Third, phrases of different lengths have different semantic prosodies or fulfil different functions even though they are realisations of the same shorter phrase. This means that a phrase does not have one meaning but that its meaning depends on its realisation, that is its length. This in turn has implications for phraseology in general since it attaches greater importance to the choice of phrase-length for an analysis than has been done so far. The choice of phrase length might predetermine the results obtained in an analysis and a different choice might have generated different results.

Fourth, the analyses have shown that not all phrases have semantic prosodies. In this paper, only the longer phrases, that is 5- and 6-grams, have distinct semantic prosodies. This is independent of whether the phrase is compositional as *i do n't know what you* or non-compositional as *at the end of the day*. This shows that non-compositional phrases pattern in the same way as compositional phrases.

Fifth, there appears to be a structural similarity between words and phrases since both pattern differently with different realisations of the original unit (word or phrase). While different realisations of a lemma have different collocations, phrases of different lengths have different semantic prosodies and fulfil different discourse functions. Similarities of prosody and discourse function seem to depend on the length of the phrase.

Acknowledgement

I am grateful to William Fletcher who provided me with lists of the most frequent 1- to 8-grams in the *BNC* disregarding the POS markup.

References

Biber, D. & S. Conrad 1999. 'Lexical bundles in conversation and academic prose', in: Hilde Hasselgard (ed.) *Out of corpora. Studies in honour of Stig Johansson.* Amsterdam: Rodopi. 181-90.

Cortes, V. 2002. 'Lexical bundles in Freshman composition', in: R. Reppen, S. M. Fitzmaurice, D. Biber (eds.) *Using corpora to explore linguistic variation.* Amsterdam, Philadelphia: John Benjamins. 131-145.

Firth, J.R. 1958. 'A synopsis of linguistic theory 1930-1955', *Studies in linguistic analysis*. 1-32

Firth, J.R. 1951. 'Modes of meaning'. in: J. R. Firth (ed.) *Papers in linguistics 1934-1951.* London: OUP. 1957. 190-215.

Fletcher, W. H. 2003/04. *Phrases in English.* database. http://pie.usna.edu/ 5 January 2007

Hoey, M. 2004. 'Lexical Priming and the Properties of Text', in: A. Partington, J. Morley, L. Haarman (eds.) *Corpora and discourse*. Bern: Peter Lang. 385-412.

Leech, G. 1974. *Semantics*. Harmondsworth: Penguin.

Louw, B. 1993. 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in: M. Baker, G. Francis, E. Tognini-Bonelli (eds.) *Text and technology. In honour of John Sinclair*. Philadelphia, Amsterdam: John Benjamins. 157-176.

Moon, R. 1987. 'The analysis of meaning', in: J. Sinclair (ed.) *Looking up. An account of the COBUILD project*. London, Glasgow: Collins ELT. 86-103.

Partington, A. 2004. ' "Utterly content in each other's company". Semantic prosody and semantic preference', *International journal of corpus linguistics*. 9 (1): 131-156.

Sinclair, J. 1996. 'The search for units of meaning', in: G. Corpas Pastor (ed.) *Las lenguas de Europa: Estudios de fraseología, fraseografía y traducción*. Granada: Comares. 2000. 7-37.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: OUP.

Stubbs, M. 2001. *Words and phrases*. Oxford: Blackwell.

Stubbs, M. 2005. 'The most natural thing in the world: quantitative data on multi-word sequences in English', Conference presentation at *Phraseology 2005, Louvain.*

Stubbs, M. & I. Barth 2003. 'Using recurrent phrases as text-type discriminators: a quantitative method and some findings', *Functions of language.* 10 (1): 65-108.

Watt, I. 1960. 'The first paragraph of *The Ambassadors*: an explanation', *Essays in criticism.* 10. 250-74.

Appendix

Top 20 3-grams in the BNC	Top 20 4-grams in the BNC
i do n't	i do n't know
one of the	the end of the
the end of	at the end of
part of the	for the first time
do n't know	on the other hand
some of the	between # and #
a number of	as a result of
there is a	the rest of the
a lot of	in the case of
# and #	one of the most
there was	# per_cent of the
it's a	the secretary of state
be able to	by the end of
it was a	from # to #
the fact that	do n't want to
you do n't	is one of the
to be a	to be able to
it's not	i do n't want

Top 20 5-grams in the BNC	Top 20 6-grams in the BNC
at the end of the	at the end of the day
by the end of the	on the other side of the
i do n't want to	ask the secretary of state for
i do n't know what	to ask the secretary of state
as a result of the	from the point of view of
in the middle of the	my hon. Friend the member of
the secretary of state for	by the end of the year
the other side of the	at the other end of the
at the time of the	i do n't think it 's
you do n't have to	in such a way as to
for the first time in	the department of trade and industry
at the top of the	from # per_cent to # per_cent
i do n't think i	in the second half of the
at the beginning of the	in the middle of the night
the end of the year	our next bulletin is at # p.m.
in the case of the	secretary of state for the environment
there are a number of	the secretary of state for the
on the other side of	at the end of the year
the end of the day	i do n't know what you
i do n't know how	

Stubbing your toe against a hard mass of facts: corpus data and the phraseology of STUB and TOE

Hans Lindquist

University of Växjö

Abstract

In this paper Fletcher's database Phrases in English is used to extract frequently recurring n-grams containing the verb 'stub' and the noun 'toe' from The British National Corpus. After analysing some of these n-grams, the paper focuses on the formulaic sequence 'stub one's toe' and investigates this in the BNC, The New York Times and The Independent. Additional searches are also made on the World Wide Web by means of WebCorp. It is found that the phrase is used with equal frequency in American and British English, but that the American use differs in that approximately half of the tokens are non-literal, while such use is relatively rare in British English. It is hypothesized that the non-literal use originated in American English and that it may be spreading to other varieties.

> Somewhere down there he stubbed himself against an ill-defined but hard mass of fact, and brought it up to the surface to examine it. (Michael Frayn: Towards the end of the morning, 1967; BNC G12 1612)

1. Introduction

Stubbs (1996, 2001) pioneered the use of corpora in the study of semantic and pragmatic meaning. In a number of more recent articles (Stubbs 2002, in press, forthcoming a, forthcoming b), he has suggested methods for retrieving n-grams from corpora in order to study frequent collocations and collocations with frequent words. These methods constitute innovative ways of bringing data, hard masses of fact, from the depths of large corpora to the surface where they can be examined and analysed.

The present study will use some of the methodology suggested in these papers to investigate the phraseological patterns, or, with Wray's (2002) terminology, the formulaic sequences which form around two lemmas, the fairly frequent body-part noun TOE and the rather infrequent verb STUB. My purpose is first to describe and analyse the semantic and pragmatic features of recurring formulaic sequences with these words, and, second, to discuss some theoretical and methodological consequences of this type of study.

The paper will have the following structure. After a brief section on the method and material there will be a few paragraphs on the history of the individual words *stub* and *toe*. Then the phraseological tendencies of these two words will be illustrated by a study of their occurrences in formulaic sequences in the British National Corpus, which leads up to an investigation of the specific phrase *to stub one's toe* in the BNC, *The New York Times*, *The Independent* and on the World Wide Web (by means of WebCorp). Finally there is a conclusion and a Coda.

2. Method and material

The method used has been called "from lexis to n-grams" by Stubbs (forthcoming a) and is described in some detail in Lindquist and Levin (forthcoming a and b). Basically it means starting with a particular word or lemma, or a set of words or lemmas, and investigating which recurring n-grams they occur in. Lists of n-grams in the British National Corpus with the search word in all possible positions can easily be extracted by means of William Fletcher's (2003/2004) database *Phrases in English*, which includes all n-grams between 2 and 8 words occurring 3 times or more in the BNC.

At the next stage, these recurring n-grams have to be manually analysed to judge which may be considered to be formulaic sequences and which may be just chance occurrences without interior structure and integrity. For instance, in the BNC, the most frequent 3-gram beginning with *toe* is *toe of his* (21 occurrences), but this is not a likely candidate for a formulaic sequence to be stored holistically in the brain. The second most frequent 3-gram beginning with *toe*, however, is *toe the line* (16) which is clearly a formulaic sequence.

In the present paper, the phraseologies of *stub* and *toe* were also studied in two sets of newspapers on CD-ROM: the *New York Times* and the *Independent* from 1990, 1995 and 2000. Here, searches for the words were made by means of the program Wordsmith to create concordances which were then sorted and analysed manually. Finally, searches were made on the Web through the mediation of WebCorp (2007). However, as has been shown by e.g. Mair (2006) and Lüdeling et al. (2007), mining the rich resources of the Web is complicated due to a number of technical obstacles caused by the conflict between the information the linguist wants to extract through for instance WebCorp and the information that search engines like Google provide. Furthermore, due to various technical limitations, WebCorp at present returns many fewer hits than direct searches through Google. It is not the aim of the present paper to specifically evaluate WebCorp or Google data in comparison with data retrieved from traditional, tidy corpora or text archives like newspaper CD-ROMs. The Web data has however been added as a complement to the findings based on the other corpora. As has been pointed out by Mair (2006: 370), "corpus linguists of the future will […] [be] working in a vast and expanding corpus-linguistic working environment in which one of the chief skills required will be to identify the resources which are relevant to the problem studied from a vast range of possibilities". Comparing data from different corpora will often be a necessity (cf. Lindquist and Levin 2000).

3. The verb STUB and the noun TOE

3.1 Stub

The first record of *stub* in the *OED* is from 967, meaning "A stump of a tree or, more rarely, of a shrub or smaller plant; the portion left fixed in the ground when a tree has been felled […]" (*OED* s.v. *stub*, n. 1. a.). As a transitive verb it occurs c1400 with the meaning "To dig up by the roots; to grub up (roots)" (*OED* s.v. *stub*, v^1, n. 1. a.) and by 1577 it could mean "To reduce to a stub or stump" (*OED* s.v. *stub*, v^1, n. 6. a.). The first citation in the OED for *stub one's toe* is from John R. Bartlett's *Dictionary of Americanisms* 1848: "'To stub one's toe', is to strike it against anything in walking or running; an expression often used by boys and others who go barefoot" (*OED* s.v. *stub*, v^1, n. 9. a.), and by 1927 another frequent present-day meaning is recorded: "To extinguish (a cigarette) by pressing the lighted end of the stub against a hard object. Freq. with *out*. Also *fig.*" (*OED* s.v. *stub*, v^1, n. 12). In terms of semantic prosody (cf. e.g. Sinclair 1991, 2004; Louw 1993; Partington 1998, 2004; Hunston and Thompson 1999; Levin and Lindquist 2007) it seems that in general *stub* has negative prosody, probably based on the original nominal meaning of something that is left when valuable material has been extracted, and strengthened by the associations to violent actions resulting in this extraction and later to various verbal metaphorical uses leading to similar results. For some positive examples, however, see the Coda below.

3.2 Toe

The earliest citation for the noun *toe* in *OED* is from c725, with the meaning "Each of the five digits of the human foot" (*OED* s.v. *toe* n. 1. a.) Through history it has occurred in a number of phrases with figurative meaning like *stepping on someone's toes*, *being on one's toes*, *toe to toe* (for a treatment of this particular phrase, see Lindquist and Levin forthcoming b), *a toe in the door*, *to dig in one's toes* and several others. Behind many of these figurative meanings seems to lie either the balancing and gripping function of the toes or their exposed position at the outer points of the longest extremities of the human body.

4. STUB and TOE in the British National Corpus

4.1 Stub

The verb *stub* occurs 100 times in the BNC, making it a fairly rare word with 1.03 occurrences per one million running words (compared for instance with *hit* which occurs 106 times per one million words). Of the 100 verbal *stub(s),* 69 were in the phrasal verb *stub out*, 62 of which referred to the physical putting out of cigars or cigarettes as in (1), 2 referred metonymically to giving up smoking as in (2), and one metaphorically to the crushing of an object as in (3). In one example, from a poem where pigs are uncharitably likened to women at a jumble sale, *stub out* seems to mean 'stick out' (4).

(1) She rolls off the bed and **stubs** the fag **out**. (A74)

(2) Now, though the office air is clean, the butt-crammed ashtray outside testifies that smoking is far from **stubbed out**. (A4K)

(3) The car was so low it looked as if a giant had tried to **stub** it **out** and it was clear that getting out of the bucket seat gave the Greek momentary altitude sickness. (FR3)

(4) Floppy hats, high-heeled trotters, massive hams, a double row of buttons done up neatly, salmon pink on beige – they squeal and **stub** their noses **out**, flushed and burning with the change of life ... (HRL)

In the remaining 31 examples, *stub* is used about a number of different entities. The most common is *toe*, with 14 tokens. Most of the instances have concrete meaning, as in (5), but there were also two where the meaning is figurative, as in (6).

(5) Distracted, Luce **stubbed** her **toe** against a piece of raised planking and tripped. (JY2)

(6) As a prominent figure in Rottweiler rescue, she's **stubbed** her **toe** on more unfair bullying and downright idiocy than most. (C8U)

Other things that were stubbed and somehow damaged or terminated included *foot*, *finger*, *toecap*, *himself*, *its shock absorbers*, *a blue trace* [on a screen], *his miling exploits*, *another English prejudice* and *their ego*. In (7), however, the verb seems to refer to a thrusting manner in which a goal was scored.

(7) After 70 minutes, Paul Robinson, a tall, long-necked forward, **stubbed** Scarborough's second from close range. (A2S)

The "cigarette meaning" also occurs with *out*-less *stub* (three examples with abstract meaning as in (8) and one with figurative meaning, as in (9)).

(8) A ringed hand held a thin cigar which – as if in impatient expectation of her arrival – he **stubbed** in a silver tray. (H82)

(9) Putting out the cigarette again, as though **stubbing** Alice out of existence, he said […] (EV1)

4.2 Toe

Toe occurs 1616 times in the BNC, which gives it a frequency of 16.55 per million words. In Adam Kilgarriff's frequency list based on the BNC ([1995] 1998) it has rank 4253, which can be compared with other body nouns like *foot* (frequency 21,339, rank 484) or *hand* (frequency 53,265, rank 176). Since *toe* is more than 15 times more frequent than *stub*, it is not practical to study all the concordance lines where it occurs with equal attention to detail. Instead, we will look at the most frequently recurring n-grams with the forms *toe* and *toes*. As mentioned in the method section, although the PIE program supplies n-grams up to 8-grams, very few linguistically significant n-grams of that length recur. At the 5-gram level, however, *the toe of his boot* occurs 10 times, *the toe of his shoe* 5 times and *the toe of her shoe* 3 times. This means that the most frequent long formulaic sequences with *toe* are ones where *toe* does not refer to a body part at all but rather to a part of a piece of footwear. In all, with various possessive pronouns, there were 15 *the toe of X's boot* and 9 *the toe of X's shoe*. In addition, there were occasional references to the toes of Doc Martens, wellingtons and flip-flops. Typical examples are (10) and (11).

(10) Benjamin tapped **the toe of his shoe** on the soft carpet. (HH5)

(11) She raced for the protection of mast and water butt, caught **the toe of her shoe** on a raised nail and went sprawling. (C85)

Another 5-gram with a frequency of 3 is *the toe of the club*, where *toe* refers to a protruding part of a golf club. Such technical meanings of *toe* are not uncommon. *A toe in the water* also occurs three times, but is better treated as the 4-gram *toe in the water*, which occurs 8 times. None of those examples refers to human toes dipped into real water; they are all used figuratively as in (12) – (14).

(12) It's always best to dip a **toe in the water** first, rather than plunging in with a programme of hopefully helpful ideas for the improvement of her life and comfort. (C8Y)

(13) […] Cognos is still at the "**toe in the water**" stage with the AS/400 market. (CSH)

(14) But Nordstrom's catalogue is merely a **toe in the water**. (CR8)

Similarly, the 5-grams *covered from head to toe* (4 instances), *dressed from head to toe* (4) and *clothed from head to toe* (3) are better treated under the 4-gram *from head to toe* (74 instances) and its variant *from top to toe* (19). While all instances of *from head to toe* refer to the human body being covered in clothes or other material or being injured or treated or scrutinized in its entirety, as in (15) – (17), *from top to toe* is also occasionally used about other concrete objects, as in (18), or about abstract entities, as in (19).

(15) The likes of Naomi Campbell and Linda Evangelista were clad **from head to toe** in leather, rubber, latex and PVC. (A7N)

(16) The King's legs were broken, there were injuries **from head to toe**. (BMN)

(17) Her glance raked Polly **from head to toe**. (H7W)

(18) But John and Veronica Saunders still make time to decorate their home **from top to toe**. (ED4)

(19) The overwhelming impression left after the debate is of a Tory Party split **from top to toe** over Europe, and a Prime Minister unable to heal the rift. (CEN)

Moving down to 3-grams, there are two noteworthy phrases: *toe to toe* with 6 tokens and *stubbed his toe* with 5. Counting all variants of the latter like *stubbing her toe* etc. the figure goes up to 14; these were treated above under *stub*. Five of the instances of *toe to toe* are used adverbially to refer to people standing opposite each other at close range, often literally with touching toes, and exchanging blows as in (20), while one token refers to the measuring of the distance from the toes of one foot to the toes of the other. Lindquist and Levin (forthcoming b) also found a number of cases where *toe to toe* was used non-literally, meaning 'in direct confrontation' in an abstract sense.

(20) […] we stood **toe to toe** and swapped blow for blow. (H0A)

4.3 Toes

As has been pointed out by Sinclair (2003: 167–172), singular and plural forms of nouns often occur in quite different contexts, and indeed the searches for *toe* and *toes* yielded totally different n-grams. One 6-gram with *toes* occurred 5 times: the line from the children's rhyme *head and shoulders, knees and toes*, which is repeated over and over again in one single text. On the 6-gram level we also find *to keep you on your toes* (3), but this is better treated under *on X's toes*, which has a frequency of 169 with five main meanings as shown in Table 1.

Table 1: *On X's toes* in the BNC

	Attention	Posture	Embellishment	Encroaching	Perceiving	Other	TOTAL
On his toes	12	21	1	2		2	38
On their toes	31	4	2				37
On your toes	9	9		2	1		21
On the toes		1	7	4	3	1	16
On her toes	3	11	1	1			16
On our toes	14			1			15
On its toes	9	3	1				13
On my toes	6	5		1			12
On yer toes		1					1
TOTAL	83	55	12	11	5	3	169

The five meanings 'attention', 'posture', 'embellishment', 'encroaching' and 'perceiving' are illustrated in (21) – (25).

(21) […] just as black people keep changing the name you are allowed to call them in order to keep whitey **on his toes**. (ECU)
(22) Culley swung at him, coming up **on his toes** for the blow. ((FS8)
(23) On his legs were hose striped with red and gold, while his feet were hidden in crimson velvet slippers with silver roses **on the toes**. (H9C)
(24) It claims it will stick to commercially-led Unix information not treading **on the toes** of existing non-commercial Unix networks […] (CTJ)
(25) […] come on look **on your toes** now, get up come on […] (KB8)

As Table 1 shows, the 'attention' meaning as in *keep sb on their toes* or *be on one's toes* is clearly the most common, followed by the 'posture' meaning as in *standing/running/dancing etc. on one's toes*. The 'embellishment' meaning, where something is applied or placed on the toes (usually the toes of a particular pair of shoes, but sometimes on someone's bare feet), comes next, and then the 'encroachment' meaning as in 'treading on someone's toes'. Finally there are a few cases referring to the fixing of one's gaze etc. on one's toes, and some unclear cases.

A few more things can be said about these findings. First of all, the distribution of meanings over forms is far from even. The singular *on his toes*, *on her toes* and, to some extent, *on my toes*, are mainly about posture. With this meaning there is also a gender difference: about men the reference is especially to fighting contexts like in (22) above, while about women it is to romantic contexts like in (26).

(26) She once again burst into tears and, crossing rapidly to George, threw her arm around his neck and stretching up **on her toes**, began to kiss him with a fervour which shocked him. (C98)

The 'attention' meaning predominates clearly with the plural *on their toes* and *on our toes*, and also with *on its toes*. (27) and (28) are typical examples. *On its toes* often refers to collectives that need to be on the alert, as in (29).

(27) "I also carry out random spot checks with a probe at various points during the week, just to keep everyone **on their toes**," he adds. (HC3)
(28) "A lead of 2–1 guarantees nothing, except the fact that we need to be **on our toes** against a side who will relish the big occasion." (CH3)
(29) The competitive situation keeps the governing party **on its toes** and sensitive to the public's view of policy.

It is easy to see how the non-literal, metonymic 'attention' and 'encroaching' meanings have developed from the physical acts of standing on one's toes, ready to move quickly, on the one hand, and accidentally or intentionally stepping on someone else's toes, on the other. The 'embellishment' and 'perception' instances are just cases of fully transparent compositional constructions.

5. Stubbing one's toe(s) in the BNC, in The New York Times and The Independent, and on the World Wide Web

In this section we will take a closer look at the construction *stub one's toe(s)* in a number of corpora. Table 2 shows the overall distribution. It is hard to make a comparison between the corpora since the exact number of words is only known for the BNC. The newspaper figures were arrived at partially by counting, partially by extrapolation (for the procedure, cf. Lindquist 2007) and as regards the World Wide Web, its constantly growing size can only be roughly estimated (Keller and Lepata 2003: 467 and Mair 2006: 365–366 independently arrive at the ballpark figure 100 billion words for the English part of the Web). The unexpectedly low figures for WebCorp (accessed 4 March 2007) are due to the limitation of the present prototype version of 200 accessed webpages (WebCorp 2001–2007: Guide). In addition, the reliability of the search engine figures is not to be trusted (cf. Lüdeling et al. 2007) and the WebCorp figures are only included for internal comparison.

Table 2: Frequency of the phrase *to stub one's toe(s)* against something

	Total number of words	*Stub one's toe*	*Stub one's toes*	Total
BNC	100 M	13	1	14
NYT	180 M	18	5	23
IND	115 M	20	0	20
WebCorp	?	53	21	74
Total		104	27	131

One conclusion that can be drawn from Table 2, in spite of the reservations above, is that phrases with *toe* in the singular are more frequent than phrases with the plural *toes*. This is largely a function of the fact that a majority of the subjects are singular. As pointed out by Moon (1998:95), "nothing systematic accounts for the way in which words denoting parts of the body inflect in some [fixed expressions and idioms], in accordance with the number of the grammatical subject or referend, but not in others." In the case of *stub one's toe*, however, the singular *toe* is normally used with singular subjects and the plural *toes* with plural subjects, even if there are exceptions, as in (30) and (31):

(30) Plus for someone like me who is always hiking through streams climbing boulders and tripping over everything in my path it's a real asset to discover I can't **stub my toes** in these sandals. (http://www.lonelyplanet.com/travel_services/trvl_gear/item/keen_youth_newport_sandal.htm)

(31) I grew up in a small town in Maryland, and Dr. Roth was our family doctor, he said. I remember calling him in the middle of the night. We called him if we **stubbed our toe**. (NYT2000)

Furthermore it seems that the plural form is less current in British English (BNC and IND) than in American (NYT) and mixed (WebCorp) varieties.

One way in which formulaic sequences increase their frequency is through achieving non-literal, extended meanings which can be used in more contexts. This has happened with *stub one's toe*. Even if one can argue that the borderline between literal and non-literal is not always crystal clear, and that some vestiges of the literal meaning often remain in the non-literal uses and can be reactivated in the mind of the language users, it is usually possible to ascertain from context with reasonable certainty if a particular token is used (mainly) literally or non-literally. In Table 3, the distribution between literal and non-literal meaning is given.

Table 3: Literal and non-literal meaning of the phrase *stub one's toe(s)*

	Literal	Non-literal	Unclear	Total
BNC	12	2	0	14
NYT	12	11	0	23
IND	18	2	0	20
WebCorp	65	6	3	74
Total	107	21	3	131

The American corpus (NYT*)* stands out from the two British corpora, BNC and *The Independent*, and the mixed web material. In American English, the meaning of *stub one's toe(s)* is evenly distributed between literal and non-literal, while in the other corpora there is a clear predominance of literal examples (ratios between 6:1 and 10:1). The non-literal examples can for instance refer to unsuccessful business encounters, as in (32), political debacles, as in (33), literary fiascos, as in (34), and, very frequently, defeats in sports, as in (35).

(32) P. & G. **stubbed its toe** pretty badly with Carrefour, the international retailer, said Mr. Flickinger, managing director of Reach Marketing in Westport, Conn. (NYT 2000)

(33) Among the democrats who regularly crop up on lists of potential Democratic candidates, a pair of northeasterners, Gov. Mario M. Cuomo of New York and Senator Bill Bradley of New Jersey, **stubbed their toes** with less than impressive victories. (NYT1990)

(34) Lanford Wilson, who won the Pulitzer Prize for the romantic "Talley's Folly" and who has written some of the most poetic and graceful plays in the contemporary theatre, **stubs his toe** over "Burn This." (NYT1990)

(35) "This is not last year's team," he said. "We're a young team. We **stubbed our toes**. We're going through some growing pains but it doesn't excuse our performance tonight." (NYT 1990)

This sports meaning can be seen as a possible link between the literal and the non-literal use, since the mishaps of football teams and athletes are to some extent physical in their nature, or strongly related to the physical, as when players actually hit their toes against the turf, the goalposts or their opponents, as in (36) (cf. the phrase *bite the dust* with similar meaning).

(36) Even worse to have someone **stub their toe** on your head – as happens to Louis Saha, who feels the force of Andre Bikey's boot.
(http://news.bbc.co.uk/sport2/low/football/fa_cup/6368735.stm)

6. Conclusions

The study of lexical phenomena like collocation and phraseology requires very large corpora. By means of *Phrases in English (PIE)* (Fletcher 2003/2004) and the accompanying interface it is possible not only to look for already well-known phrases found in dictionaries or retrieved from memory, but also for frequently recurring n-grams of which we are not always consciously aware but which may nevertheless be stored and retrieved holistically. This paper started out with such n-grams retrieved from the British National Corpus by means of *PIE*, discussed some of the most frequent types, and then zoomed in on one particular phrase, *stub one's toe(s)*, which was found to be used both literally and figuratively. This phrase was then further investigated in other sources of data (*The New York Times, The Independent*) and on the Web (through WebCorp).

Beginning with the individual verb *stub* in the BNC, it was found that it was used about cigarettes, cigars et cetera in 69% of the cases and about toes in 14% of the cases. Studying the much more frequent noun *toe*, it turned out that the frequent 3-gram *on* X's *toes* occurred with a number of specific meanings, some of which were literal and some figurative, and furthermore that different meanings co-varied with different pronouns and the definite article, so that e.g. *on his toes* was most likely to be about 'posture', while *on their toes* was most often used about 'attention' and *on the toes* was used referring to 'embellishment'. It is thus not enough to describe the meaning of the phrase or frame *on X's toes* as such, since its meaning is influenced by the word which is put into the X slot.

The phrase *stub one's toe(s)* occurred most frequently in the singular, which was seen mainly to be a trivial function of the number of the subject. Plural or collective subjects were rarer in the British corpora, and consequently there were fewer plural forms of *toe*.

As for the meaning of the phrase, it was literal in 84% of the tokens in the total material. In the purely American corpus, however, the meaning was non-literal in approximately half of the cases (11/23). The cases of non-literal or figurative meaning in the corpus referred to unlucky incidents in various areas of human public endeavour: business, politics, literature and especially sports.

The picture provided by these findings is that the phrase *to stub one's toe(s)*, which was first attested in American English, now occurs with equal frequency in British English, but that that the extended, figurative meaning of encountering an abstract obstacle of some sort is primarily American and not frequent in British English at all. This is an illustration of a phenomenon discussed by Buchstaller (2006) in relation to attitudes towards certain linguistic features when they are borrowed from one variety to another: the connotations are not always taken over wholesale. Similarly, the figurative meaning of this phrase does not seem to have been taken over by many British speakers (as yet).

7. Coda

The motto of this paper was taken from Michael Frayn's comic novel *Towards the end of the evening*. (The title of that novel is by the way an instance of the string "PREP *the* NOUN *of the*", which Stubbs (2002:232–235) has shown to be a very frequent pattern in English; in Stubbs's material, *towards the end of the* is the 11th-most frequent exponent of this structure). At this stage, one may of course ask oneself if *stub one's toe* (or *stub oneself*) is really a suitable metaphor for the pursuit of corpus linguists. After all, as was mentioned in the *OED* reference above, the semantic prosody of the phrase is almost always negative. However, there are also some more positive contexts, like in (37) and (38):

(37) The young French explorers literally **stubbed their toes** on treasures buried in the sand: obelisks and sphinxes; ruined temples at Karnak, Dendera, and Luxor; broken colossal funeral statues in the Valley of Kings. (Dora B. Wiener, 'With Bonaparte in Egypt'. *Isis* 2000, 91:755)

(38) Otters quickly became scarce, but then three guys standing in a Yukon creek **stubbed their toes** on a thimble's worth of gold.
(http://www.where.ca/alaskayukon/article_feature~listing_id~36.htm)

It is therefore possible to end this paper on a happier note: Wading through sand dunes of words, standing waist-deep in swirling data, the corpus linguist will ever so often stub his or her toe(s) on unexpected treasures.

References

Buchstaller, I. 2006. 'Social stereotypes, personality and regional perception displaced: Attitudes towards the "new" quotatives in the U.K. ', *Journal of Sociolinguistics*, 10 (3): 362–381.

Fletcher, W. 2003/4. *PIE: Phrases in English*. http://pie.usna.edu.

Frayn, M. [1967] 2005. *Towards the end of the morning*. London: Faber and Faber.

Hunston, S. and G. Thompson (eds.) 1999. *Evaluation in text. Authorial stance and the construction of discourse*. Oxford: Oxford University Press.

Keller, F. and M. Lapata 2003. 'Using the Web to obtain frequencies for unseen bigrams', *Computational Linguistics*, 29 (3): 459–484.

Kilgarriff, A. [1995] 1998. *BNC database and word frequency lists*. http://www.kilgarriff.co.uk/bnc-readme.html

Levin, M. and H. Lindquist 2007. 'Sticking one's nose in the data. Evaluation in phraseological sequences with *nose*', *ICAME Journal* 31: 63–86.

Lindquist, H. 2007. 'Viewpoint *-wise*: The spread and development of a new type of adverb in American and British English', to appear in *Journal of English Linguistics* 35 (2): 132–156.

Lindquist, H. and M. Levin. 2000. 'Apples and oranges: On comparing data from different corpora', in C. Mair and M. Hundt (eds.) *Corpus linguistics and linguistic theory*. Amsterdam: Rodopi. 201–213.

Lindquist, H. and M. Levin forthcoming a. 'FOOT and MOUTH. The phrasal patterns of two frequent nouns', to appear in S. Granger and F. Meunier (eds.) *Phraseology: An interdisciplinary perspective*. Amsterdam: Benjamins.

Lindquist, H. and M. Levin forthcoming b. 'The syntactic properties of recurrent phrases with body part nouns: The N_1 *to* N_1 pattern'. To appear in U. Römer and R. Schulze (eds.) *Exploring the Lexis-Grammar Interface*. Amsterdam: Benjamins.

Louw, B. 1993. 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies.', in: M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and technology: In honour of John Sinclair*. Amsterdam: Benjamins. 157–176.

Lüdeling, A., S. Evert and M. Baroni 2007. 'Using web data for linguistic purposes' in: M. Hundt, N. Nesselhauf and C. Biewer (eds.) *Corpus linguistics and the web*. Amsterdam: Rodopi. 7–24.

Mair, C. 2006. 'Tracking ongoing grammatical change and recent diversification in present-day standard English: The complementary role of small and large corpora' in: A. Renouf and A. Kehoe (eds.) *The changing face of corpus linguistics*. Amsterdam: Rodopi. 355–376.

Moon, R. 1998. *Fixed expressions and idioms in English. A corpus-based approach*. Oxford: Clarendon Press.

Partington, A. 1998. *Patterns and meanings. Using corpora for English language research and teaching*. Amsterdam: Benjamins.

Partington, A. 2004. '"Utterly content in each other's company". Semantic prosody and semantic preference.' *International Journal of Corpus Linguistics*, 9 (1): 131–156.

Sinclair, J. 1991. *Corpus Concordance Collocation*. Oxford: Oxford University Press.

Sinclair, J. 2003. *Reading concordances. An Introduction*. London: Longman.

Sinclair, J. 2004. *Trust the text: language, corpus and discourse*. London: Routledge.

Stubbs, M. 1996. *Text and corpus analysis*. London: Blackwell.

Stubbs, M. 2001. *Words and phrases. Corpus studies in lexical semantics*. London: Blackwell.

Stubbs, M. 2002. 'Two quantitative methods of studying phraseology in English', *International Journal of Corpus Linguistics*. 7 (2): 215–44.

Stubbs, M. 2007. 'An example of frequent English phraseology: Distributions structures and functions', in: R. Facchinetti (ed.) *Corpus linguistics 25 years on*. Amsterdam: Rodopi. 89-105.

Stubbs, M. forthcoming a. 'Quantitative data on multi-word sequences in English: The case of the word "world"', to appear in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds.) *Text, discourse and corpora*. London: Continuum.

Stubbs, M. forthcoming b. 'Quantitative data on multi-word sequences in English: The case of prepositional phrases'. Lecture given at the Berlin-Brandenburgische Akademie der Wissenschaften, 3 November 2006.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Stringing together a sentence: linearity and the lexis-syntax interface

Oliver Mason

University of Birmingham

Abstract

Following existing approaches to linear grammar we explore the application of automatically identified multi-word units to the analysis of sentence structure. After looking at several sample sentences we then move on to a discussion of routine use vs. creativity in language.

The proposed new phraseological grammar does away with both syntactic and functional categories and reduces syntax to a by-product of a linearising thought in the form of phraseological units of meaning.

1. Introduction

Phraseology is concerned with the study of units above the level of the single word, which seem to become increasingly important with the widening application of empirical principles to the field of linguistics. The single word, while a convenient starting point, is not a suitable entity when describing either sentence structure or aspects of meaning. A word has no meaning in isolation, and even its syntactic environment is usually idiosyncratic when we consider actual use rather than theoretical possibilities.

So far, multi-word units (MWUs) have been identified as units of meaning (eg Danielsson 2001), as it is only in conjunction with other words that we are able to decide which aspect of its meaning potential has been realised in a particular instance. The phrasal environment of a lexical item thus serves as a shorthand description of its use, if we consider the definition of meaning as use. Stubbs (2001) gives the examples of *surgery* and *bank*, which despite having several distinct meanings in isolation cannot ever be confused when used in authentic sentences. It is, of course, possible to deliberately invent sentences where their use is ambiguous, but the important issue here is that this is not what speakers do in real life.

However, MWUs are not only important in semantics, where they displace the single lexical item as the central element, shifting the focus from lexical meaning to phrasal meaning, but also in syntax, where they compete with abstract descriptions ultimately based on phrase structure grammar (Chomsky 1957). As Stubbs (1993) observes, grammarians are traditionally interested in structures only, and view the lexical items as mere instantiations of the grammatical categories which they belong to. More recent approaches (eg Sinclair 1991, Francis 1991, Brazil 1995, Hunston and Francis 2000, Sinclair and Mauranen

2006), on the other hand, have demonstrated that lexical items are more than that, and that instead there is a correlation between grammatical structures and the words which occur in them. As is the case with everything in the description of language, this correlation is not an absolute, but rather expresses strong tendencies reinforced by everyday usage.

Some of these alternative approaches furthermore view sentence structure not as hierarchical (as in analyses derived from phrase structure grammars) but instead as linear. Such a linear sequence of units (*elements* in the terminology of Brazil (1995), *patterns* in Hunston and Francis (2000), and *chunks* in Sinclair and Mauranen (2006)) is constructed mainly according to the principle of prospection, where one unit places constraints upon the range of possible successor units.

1.1 Open Choice *vs* Idiom

Sinclair (1991) discusses two principles of grammatical descriptions, connected to the Saussurian notions of *syntagma* and *paradigma*: the open-choice principle treats each position in an utterance as a (complex) choice, basically like a slot that is filled by an appropriate item (hence it is also referred to as 'slot-and-filler model'). The idiom principle, on the other hand, states that the user has at their disposal a set of larger units, so that they do not select individual lexical items (as they would do following the open-choice principle) but instead larger chunks. He argues that neither principle is sufficient to describe language, but that the idiom principle is the more important of the two, which should be used by default for describing texts. Only when a phenomenon cannot be accounted for by the idiom principle should we fall back on the open-choice principle.

This view fits in well with a model that uses MWUs as their basic units, as these would represent the larger chunks that make up utterances instead of single lexical items. As we will see below, Sinclair was right in stating that both principles are required for a more comprehensive description.

2. Multi word units and Phraseology

We now look at multi word units, which go beyond the single lexical item. While they can of course be described intuitively, there are two principal ways of automatically identifying them through computer algorithms. In this section we will explain those algorithms, as they form the basis for extracting MWUs that we later apply to our grammatical description.

2.1 Chains

When looking at computer-identified phrases in the past, most work has been concerned with *n*-grams, where word sequences of a particular length are extracted from a text. Values for *n* are typically in the range of 2 to about 8 (as on Fletcher's *Phrases In English* site). This is a great step forward from early studies which were mainly limited to bigrams and trigrams; this step has been facilitated by advances in computing power and storage. The problem with *n*-grams is that as their number quickly increases for larger values of *n*, their respective frequencies quickly diminish, leading to large sparse matrices that are difficult to process.

Despite not being grammatically well-formed, these *n*-grams are usually referred to as phrases, which is less awkward than the more general term 'multi word units'. Stubbs and Barth (2003) use the term *chain* to describe an *n*-gram, a usage we will adopt here as well.

In our own work we collect all *n*-grams with *n* ranging from 2 to 7. In order to filter out those that are not 'interesting' we assign to these chains a weighting, based on its frequency of occurrence and length: since short chains are usually more frequent while longer ones occur less often, using frequency alone would favour short chains. But longer chains are more specific (and thus interesting), provided they are also used often enough, so we take length into account as well. The use of a weighting function has been inspired by Kita et al. (1994) who applied it to a similar problem.

2.2 Frames

Another way to derive multi word units (MWUs) is described by Danielsson (2001), who aims to identify units of meaning, starting with the premise that single words do not carry meaning unless embedded in context. Danielsson uses collocation to extract larger units. Renouf and Sinclair (1991) start off with a gapped sequence of high frequency words, eg *as* — *as*, and investigate typical fillers for the gaps in their templates. Mason (2006) reverses this procedure by attaching words to a central node word which have a higher frequency than the node word itself; this is based on the idea that context words (low frequency) alternate with function words (high frequency). This procedure can also be combined with *n*-gram extraction to produce a good range of MWUs for a particular node word.

There are two different uses of MWUs apparent: first, one can extract all MWUs from a text or corpus to look at the properties or distribution of MWUs within the text (eg Stubbs and Barth, 2003), and second, one can look at the MWUs involving a particular target word or phrase (eg Starcke, this volume, and Mason 2006). In this paper we will be pursuing the second kind of analysis.

We combine the frames with the chains mentioned in the previous section, and thus can collect for any particular word a good range of multi-word units. As a starting point we take MWUs extracted from the written part of the British National Corpus (BNC). These lists of MWUs are calculated for individual words, and are filtered so that all items with a frequency of less than 1% of the highest frequency MWU are discarded.

Before we apply those MWUs to the description of sentence structures we will further investigate the transition from grammar to phraseology.

3. Linear Grammar

Linguists generally describe the structure of language in terms of sentences, and those sentences are commonly assigned a hierarchical structure, represented as an upside-down tree. Lexical items are of minor importance only; they get replaced by syntactic categories in the first step, and from then on there are only nouns, verbs, adjectives and a few other elements which make up the sentence. It is obvious that the actual words do not play any significant role in this model, and thus phraseology is completely irrelevant to it.

However, the study of collocations and work in phraseology has shown that word class labels cannot adequately replace lexical items, as they are too general and cannot describe the combinatorial idiosyncrasies of words. For an example analysis of *of* see Sinclair (1991). Sinclair shows that even the most frequent member of the class 'preposition' does not at all behave as it would be expected to. Other studies have shown that even the various inflected forms of a lemma have little in common when looked at in more detail (eg Stubbs (2001) on CONSUME and SEEK). Francis (1993) describes the importance of phraseology for the co-selection between lexical items and syntactic structures.

There are two related issues here: first, the use of hierarchical structures for describing sentences, which necessitates abstract labels such as NP and VP for phrasal units, and second, the use of abstract labels itself which leads to overgeneralisation, as elements belonging to the same syntactic category usually show divergent behaviour in authentic language.

In this section we will look at three alternative approaches, which postulate a linear structure for utterances. The grammar devised by Brazil (1995) focuses on spoken utterances, but there is in principle no reason why it should not also be applied to written data. Brazil uses category labels, but is prepared to relinquish them should they turn out to be superfluous, ie he is not fixated on traditional grammatical terminology. Hunston and Francis (2000) in their pattern grammar use a mixture of category labels (for both phrases and individual items) and lexical items, in order to achieve a higher level of precision. Sinclair and Mauranen (2006) in their linear unit grammar, on the other hand, do without word class labels, and classify elements of utterances according to their role in discourse.

3.1 A Grammar of Speech

Brazil (1995) sets out to describe (spoken) utterances in a linear way, deliberately focusing on the process rather than the product of language. In the absence of any alternatives he tentatively adopts traditional categories such as nominal and verbal elements, but he restricts himself to a purely linear structure. As an underlying formalism he chooses a finite-state model, which seems to be appropriate for his purposes.

His central concern is communicative function rather than supposed grammaticality, and so he identifies target states in the speech chain which fulfill a communicative purpose. Various paths through the chain can either reach a target state (which leads to a completed utterance) or an intermediate state (which requires further elements for completion). Although it has not been adopted into mainstream linguistics, this incremental approach works well without the need for a constituent structure.

One reason why a finite state model works, even though Chomsky (1957) states that it could not, is that authentic language is more restricted when it comes to those features which a finite state model would have difficulties with. While in theory there can be an infinite number of central embeddings in a sentence, this simply does not occur in practice, especially not in spoken discourse. Taking this into account thus allows us to use a simpler and less powerful apparatus for language description.

3.2 Pattern Grammar

Hunston and Francis (2000) describe a grammar based on syntactic patterns centred around lexical items, predominantly verbs, but also nouns and adjectives. The syntactic behaviour of these words can be captured in a finite set of such patterns, which are specific to the word in question (but does not have to be unique across the vocabulary). In traditional terminology the closest equivalent would be that subcategorisation information is contained in those patterns, eg 'V n' describes a verb that takes one object while 'V *on* n' describes a verb that typically takes a prepositional phrase with *on* as its complement.

Pattern grammar, however, goes well beyond merely describing subcategorisation: it can be used to model the structure of sentences as well. Similar to Brazil's approach the description is a linear one, where a sentence is seen as a sequence of patterns of its lexical items which are realised. These patterns can either be end-on-end, or they can overlap. The latter phenomenon is called 'pattern flow', as one pattern 'flows' into the next, as illustrated in table 1: here the pattern ***there* V n** of the verb *are* flows into the pattern **N that** of the noun *signs*, which fulfills the role of the 'n' in the first pattern.

Table 1: Example of 'pattern flow'

and	there	are	signs	that	the	…
	there	V	n			
			N	that		

3.3 Linear Unit Grammar

While both Brazil and Hunston/Francis make use of traditional category labels, Sinclair and Mauranen (2006) abandon them completely. They also employ different units of analysis, neither words nor phrases, but chunks, groups of lexical items which by intuition belong together. They deliberately do not offer a definition of 'chunk', but instead prefer it to remain as a pre-theoretical term. Each chunk in a text gets assigned a functional label, classifying it either as message oriented or (text-)organisation oriented. These two classes also have a number of sub-classes, which are used eg to mark incomplete elements which require completion (similar to Brazil's 'suspension'), or fragments (false starts and repetitions in spoken utterances).

After the chunks have been classified, they suggest several steps for further processing, which involves removing the organisational elements and incomplete fragments, and results in a 'cleaned-up' version of the original utterance. This revised utterance is more suitable for analysis with a traditional grammar, as it will more closely resemble the kind of well-formed utterance that such grammars have been designed to handle.

Their method of segmentation (using intuition) remains unsatisfactory, though they suggest that there is sufficient overlap between different analysts to assume its validity. It would also fit in with an individualistic view of language, where we would be dealing with the internalised knowledge of language of an individual: here we can easily accept that different speakers have different internalised grammars, and different segmentations could be part of that.

However, if we use the MWU algorithm described above, then we would have an objective means of deriving chunks, which is also capable of modeling the linguistic experience of an individual, as it will be based on a particular corpus. Different corpora will yield different sets of MWUs, which is consistent with the differing language experience of different users.

4. Phraseology and Grammar

As we have seen in the above section, it is perfectly possible to do away with hierarchical structure when describing utterances. Of the three approaches mentioned here, two explicitly commend themselves for the description of spoken language, namely Brazil's Grammar of Speech and the linear unit grammar of Sinclair and Mauranen. However, there is no reason why written language should not also be describable by a linear approach, as it essentially is a linear process by which it is created. The only difference seems to be the possibility of additional editing once a sentence has been written down, and the influence of the conventions of writing systems and written genres.

In the remainder of this article we will present the outline of a grammar based on phraseology. It will adhere to similar principles as the three linear grammars, and it will be based on corpus data rather than intuitive judgments.

The basic premise goes back to Stubbs (1996:41) who states as the seventh principle of neo-Firthian linguistics: "Much language use is routine." Essentially, re-using bits of language has several benefits: it reduces the strain on the speaker to create ever new and previously unheard constructions; it makes it easier for the listener to recognise patterns in the language stream; it can help establish larger chunks and contexts for disambiguating words used with multiple senses and meanings.

Essentially we assume that pretty much everything has been said before, though that is of course an over-simplification. There are indeed new and creative constructions, but they are the exception rather than the rule. Most of language will consist of chunks that have occurred before, just as we tend to re-use words and only occasionally introduce new coinages. But it is not only the words themselves that we re-use, it is also their contexts, as they are inseparable. And their contexts are effectively multi word units.

Stubbs (1996) further notes that the study of language is basically part of the social sciences, as language transmits (and creates) culture. But society is not an independent entity, but a collection of individuals, each of them acting independently (though according to behavioural rules which can be observed). The same applies to any language, which exists only through the internalised grammars of their speakers, which of course are all independent and therefor distinct. This provides support for the use of corpus data, and it also highlights the fact that any grammatical description of a language will only ever be an approximation, as there cannot exist a single grammar describing a (natural) language.

Combining these points, we will try to identify fragments of a sentence to be analysed in our corpus, but we will also not worry about gaps in the description: these can either be caused by a lack of evidence (ie the fragment in question has not been perceived before and is thus not part of the speaker's linguistic experience), or it could be a genuine instance of creative usage. While these two potential causes look pretty much identical, there is an important difference between them: a creative use will also not be found in any other corpus, whereas a case of lacking evidence might have occurred there.

The identification of the fragments is as follows: for each word token in the sentence we look up all the MWUs found for that item in the reference corpus. We then tabulate all the MWUs that can be matched with the syntactic context. Eventually it is only necessary to keep track of the longest match, but for the time being it will be interesting to also find shorter (overlapping) fragments, as these might point us to issues of interest in the sentence.

5. Examples and Explanations

We will now look at several examples of the MWU analysis procedure applied to sentences. All the sentences have been chosen at random from a variety of (written) sources. To keep processing issues simple, all words have been converted to lower case and punctuation has been removed. Where a MWU is given, words in capitals are the node words of that particular MWU, ie the word that has been used as the starting point in the extraction procedure.

5.1 *The long dark tea-time of the soul*

The first paragraph we will be looking at is from Douglas Adams' "The long dark tea-time of the soul":

> (1) at the top of the stairs was a minute landing which opened on one side into a bathroom so small that it would best be used by standing outside and sticking into it whichever limb you wanted to wash . (2) the door to it was kept ajar by a length of green hosepipe which trailed from the cold tap of the wash-basin out of the bathroom across the landing and into the only other room here at the top of the house . (3) it was an attic room with a severely pitched roof which offered only a few spots where a person of anything approaching average height could stand up .

The beginning of the first sentence is very well covered by MWUs, as shown in table 2.

Table 2: Fragment of sentence 1 of the Adams text

at	**the**	**top**	**of**	**the**	**stairs**	**was**	**…**
AT	the	top	of				
AT	the	top	of	the			
AT	the	top					
AT	the	top	of	the	stairs		
	THE	top	of	the			
	the	top	of	THE			
	THE	top	of				
	the	top	of	THE	stairs		
	THE	top	of	the	stairs		
		TOP	of	the			
		TOP	of	the	stairs		
at	the	top	of	the	STAIRS	was	

Here we have full coverage of the fragment *at the top of the stairs was* in a single MWU, but also several partial overlaps, which would allow for minor variations. The overlap continues with a MWU *was a MINUTE*; the following *landing* is not covered, but then the gap is closed with *which OPENED on*, and coverage continues with *on one side*, which is identified as a MWU for each of its three elements independently. Then there is no overlap with *into a* BATHROOM, which again does not overlap with the next part covered by the analysis, shown in table 3 (duplicate MWUs with different node words have been omitted).

Table 3: Fragment of sentence 1 of the Adams text

…	**so**	**small**	**that**	**it**	**would**	**best**	**be**	**used**	**by**	**…**
	SO	small	that							
		SMALL	that	it						
			THAT	it	would					
					would	BEST	be			
							BE	used	by	

Here we have a phrase that is less obviously repeated than *at the top of the stairs*, and we can easily see how it is composed of smaller units 'flowing' into one another. The remaining parts of the sentence contain one further stretch covered by MWUs, namely *YOU wanted to* and *wanted to WASH*. In total, ¾ of the sentence can be described through sequences of automatically derived MWUs.

In the second sentence we have a similar pattern, multiple MWUs matching and overlapping. In the following representation we have put in brackets words which are not covered, and in places where there is no MWU flow/overlap we have placed a vertical bar:

> The door to | it was kept | (ajar) | by a length of | (green hosepipe which trailed) | from the cold tap | (of the wash-basin) | out of the bathroom | across the landing and into the only other | (room) | here at the top of the house

The coverage rate is even higher than in the first sentence, at 79%. We can identify some kind of larger constituents, which might be an approximation to Sinclair and Mauranen's chunks, though they certainly violate the elements posited by a traditional analysis. One could furthermore argue that the second chunk, *it was kept*, is a mere accident, a mismatch from a usage where *it* is the subject of a clause rather than part of a postmodifier, and human analysts would separate it into *the door to it* | *was kept* instead. However, an automated analysis is never going to reproduce exactly what human beings would do, and we would need to look at more data before being able to evaluate the quality of the results.

One interesting aspect is the first gap, where *ajar* was not found as part of an MWU. Here we can go back to pattern grammar, where a pattern for *keep* is **V n adj/prep**, which would need to be transformed here due to the passive voice used. The adjective *ajar* thus seems to indicate a choice point, a case where the idiom principle (Sinclair 1991) cannot account for the variation, as there is a point of lexical choice more suited for a slot-and-filler model. But by combining the two, linear MWU analysis and pattern grammar, we can achieve a complete analysis of this stretch of text.

We can also see that the *at the top of the* part is even a direct repetition from the previous sentence, so it is not surprising that we get a similar result.

The analysis of the final sentence in this small sample is not that different from the previous one:

> It was an attic room with a severly | (pitched roof which) | offered only a few spots | where a person of | (anything approaching average height) | could stand up.

The rate of coverage of these sentences is between 74% and 79%, which is remarkable in that we have retrieved our MWUs from a general reference corpus only. In section 6 below we will discuss the issue of routine usage versus creativity in more detail.

5.2 *Words and Phrases*

Next we will look at a sample of academic prose, taken from Stubbs (2001). To ease the analysis we will not tabulate the complete set of MWUs identified in the text, but instead use the compressed format introduced in the previous section. Coverage ranges from 54% (final sentence) to 73% (second sentence), so it is less than the fiction sample discussed above:

> **1** a brief summary of the argument so far is the slogan | (meaning is use) . **2** (words) | do not have | (fixed) | meanings which are recorded | once and for all | (in dictionaries) . **3** (they acquire or change meaning) | according to the social and linguistic | contexts in which they are used . **4** (understanding) | language in use | depends on a balance between | (inference and convention) . **5** (here) | are more detailed | (examples which use textual) | data to show that our | (communicative competence relies on) | knowledge of what is expected | (or typical) .

Regarding the first sentence we can observe that the first part is relatively well covered, whereas the second part (which is a quoted statement introduced by the first part) is not. In terms of the informational structure of this sentence Stubbs makes use of well-established prefabricated units to introduce what is new, namely the brief summary. We thus have an introductory phrase as the 'given' or 'theme', followed by the 'creative' formulation of the 'new' or 'rheme'.

In the second sentence the first word, *words*, does not occur in this usage as an MWU; it is more frequently used in a non-technical sense, such as *in other words*. The adjective *fixed* can be viewed similar to *ajar* in the Adams example in that it is a point of high variation and can best be described following the open-choice principle. The final part, *in dictionaries*, is probably not covered as *dictionaries* is a relatively specialised word too unusual for a general reference corpus such as the BNC. It only occurs 348 times in the written part used for this study, compared with 929 instances of the singular form *dictionary*.

The next sentence (3) again begins with a sequence which is not covered by MWUs, but after the first five words coverage is rather unexpectedly good.

In the fourth sentence we can see that we again have good coverage of the middle part which relies more on core words (see Carter 2004). In the BNC, *inference* is not typically used in this way, but mainly in MWUs such as *the inference that.*

We also have two MWUs in an end-on-end position, which tends to indicate a higher level boundary, and especially the second one (*depends on a balance between*) seems to be a highly re-usable stock unit. Typically at such boundaries we find more variability in lexical choices, which reflects in a break point when it comes to finding re-current sequences. Conceptually this is related to the procedure Harris (1955) used to derive morphemes from an unsegmented stream of phonemes. His distributionalist approach looked at points of maximal variation in a sequence, such as how many different phonemes can follow a given sequence. A local maximum would indicate that there was a morpheme boundary. In this sentence, then, *language in use* would be a (higher level) unit, as would *depends on a balance between*, since there is no 'bridge' that links *use* and *depends*, possibly due to too much variability.

Stubbs' final sentence has the least coverage of the five sentences investigated here. We can explain this again through the technical vocabulary (eg *communicative competence*). It seems obvious that we would require a more academic corpus in order to achieve a higher degree of coverage, as only the more general parts of the sentences can be found in the set of MWUs. Interestingly, the sentence-initial *here* mainly occurs as an MWU in *here are some*, *here are the*, and *here are a few*; the singular *here is more* also occurs in the MWUs extracted from the BNC.

5.3 Conference Call for Papers

The next example is a single sentence taken from a call for papers of a (linguistics) conference. This is a fairly standardised kind of text which does not leave much room for creativity, and hence it is unsurprising that coverage of the sentence is very good:

> The papers presented at the conference will be available in | proceedings on the first day.

As with the previous sentences we have looked at, the MWUs overlap and link up to form a longer sequence, similar to what Hunston and Francis (2000) describe as 'pattern flow', and Gledhill (2000) as 'collocational cascade'. Looking more closely at some example MWUs, we could say that in *papers presented at the* the word *papers* prospects the following items *presented at the*, whereas in *at the conference* the initial *at* prospects *the conference*, and so forth. Prospection is an important organising principle in language use, as it restricts the expectations of possible elements in the remainder of the utterance. As Sinclair and Mauranen (2006) state, prospection is not fixed and prescriptive, but flexible, based on the frequency of phraseological patterns, and can be used to great effect when violated (in a similar vein as discourse prosodies can, see Louw (1993)).

Interestingly, *conference* then flows into *will be*, which could be classed as an instance of 'colligation' in the sense used by Firth and later Hoey (2005): the word *conference* tends to occur frequently with expressions of futurity, in this case *will be*.

There is only one point in this sentence where the flow of MWUs is interrupted, between *in* and *proceedings*. Here we can hypothesise the existence of a higher-level unit boundary, which is not crossed by the MWU chunks.

Even though we speak of *at* prospecting *the conference*, we need to be careful about the scope of such statements: they only apply to the analysis of an utterance, not its creation. While we could undoubtedly generate natural-sounding utterances by randomly stringing together overlapping MWUs, we would ignore the semantic aspect and the utterances would not be comparable to authentic ones. But in the analysis we presuppose that the utterance we are looking at is meaningful, so that the semantic dimension is implicit. There are certainly many more (in fact, 565 in total) MWUs that begin with *at*, but out of these that particular one has been chosen.

Especially in fairly standard situations (such as giving information about conferences), we do not need to be creative. On the contrary, going back to routine usages we make it easier for the recipients to understand what we are saying, as it involves less effort to process something that one has already encountered before. New or creative sentences, on the other hand, require more decoding effort, as they cannot be matched against previously experienced utterances.

We could thus assume that the degree of MWU coverage changes according to the text type: texts which are easy to read ought to be described better using MWU chunks than highly creative ones or those which are more difficult to read. This obviously has to take into account other considerations, such as topic: since we model the speaker's linguistic experience through a general reference corpus (the BNC), texts which make use of specialised vocabulary will clearly have less coverage. But from a theoretical point of view this poses no problem, as the BNC is only an approximation in the first place. If we were to analyse an academic article, then we would get a higher coverage if we used a corpus of academic language for the retrieval of MWUs. This is consistent with the notion of a separate speech community, that of academics, which have separate shared linguistic experiences from other communities.

5.4 Summary

We have examined several (authentic) sentences to see how far we get in terms of accounting for their composition using multi-word units extracted from a corpus. It is clear that there is considerable variation in the degree of coverage, which can usually be explained by the nature of specific constructions (eg adjectives/adverbs which allow for slot-type variability) or the use of technical vocabulary that would not occur frequently enough in a general reference corpus. We have also noticed, that 'stock' sentences are covered to a larger extent than more 'creative' ones, an issue we will look at in the following section.

6. Routine and Creativity

In section 4 above we mentioned Stubbs' principle that language use often is a matter of routine, and that re-use rather than outright creativity dominates our language experience. This is more so in spoken dialogue, but should also be found in written language, even though there is more scope for creativity through the added time available for careful composition and editing. As our approach to the description of sentences involves the identification of repeated sequences in the form of MWUs, we tap right into the re-use aspect of language production.

One possible application of our procedure which is a by-product of our attempt at a grammatical description is therefore to determine the degree of re-use of a sentence: if a sentence make use of a lot of 'existing' language, then we should find more MWUs that match, and our coverage of that sentence should be greater than that of a sentence whose creation involved more creativity.

We have already seen in the previous section that a stock phrase like the one taken from the call for papers is almost completely covered, whereas with the Stubbs data we could find certain introductory phrases which are routine whereas the statement introduced involves less re-use. Other instances that were notable were places where there is a lot of slot-like variation, as with the adjectives in both the Adams and the Stubbs sample.

In this section we will be looking at several other extracts that we have not already described in great detail, as we are only interested in quantitative analysis of their coverage ratio. Before we look at the results, here a brief description of the samples, presented in the compressed format, with sentence numbers printed in bold:
adams: (see above)
stubbs: (see above)
conference: (see above)
clock: a section from a children's book, *My grandmother's clock* by McCaughrean and Lambert

> ***1** (in my grandmother 's house) | there is a grandfather clock | but it does not go **2** (the hands on its big face never move) **3** (once) | i opened the door in the front of the clock to find out why | and there was nothing inside | (but one umbrella) | a walking stick and a picture of | (king zog)*

independent: a section from the Independent newspaper, also analysed by Sinclair and Mauranen (2006)

> ***1** (mr kennedy now) | declares that it must be bold in its thinking | and ready to | (plan long-term) **2** (sounding nice) | is no longer enough | (he argued) **3** from now on the liberal democrats have | to present themselves as a party that | (wants power) | and knows what it wants to do if it gets it **4** with that in mind he | (announced two reviews) | one to take a broad | (look at policy) | the other to look at | (tax) | policy as well as a number of internal | (reviews) | into the party | ('s structure and communications)*

clarkson: an extract from *The world according to Clarkson* by Jeremy Clarkson (non-fiction)

> ***1*** *last week the queen of england | (very) | kindly agreed to break off from her | (waving duties) | and lend a hand with a television programme | (i'm) | making about the | (victoria cross)* ***2*** *and so on | (wednesday) | i slipped into a whistle and went to | (buckingham palace) | to see some | (prototype medals she'd) | found in a cupboard* ***3*** *(sadly) | i never met | (my new researcher) | but i did have a | (snout around) | the state rooms | which provided a rare insight into the life of the royals*

bbc: an extract from a BBC news story taken from the BBC website

> ***1*** *(a bickering new york couple) | have had a dividing wall | (constructed inside) | their home as part of an | (acrimonious divorce)* ***2*** *(chana and simon taub both 57 have endured) | two years of | (divorce negotiations) | but neither is prepared to give up their | (brooklyn home)* ***3*** *(now a white partition) | wall has been built | through the heart of the house to keep the | (pair apart)* ***4*** *(mr taub) | asked a judge to allow him to erect the | (partition when the couple 's divorce stalled over financial details)* ***5*** *(the taubs') | divorce has been rumbling | through the new york | (divorce courts) | for two years* ***6*** *(but despite owning another home - just) | two doors away | (- the unhappily married couple) | have decided to carry on living under the same roof*

blink: seven sentences from *Blink* by Malcolm Gladwell (non-fiction)

> ***1*** *(the videotape of bill and sue 's discussion seems) | at least at first to be a random sample of a very ordinary | kind of conversation that | (couples) | have all the time* ***2*** *no one gets | (angry)* ***3*** *there are no | (scenes no breakdowns no epiphanies)* ***4*** *i'm just not a dog | (person is how bill starts things) | off in a perfectly reasonable tone of voice* ***5*** *(he complains) | a little bit but about the dog | (not about susan)* ***6*** *(she complains too) | but there are also | (moments when they simply) | forget that they are supposed to be arguing* ***7*** *when the subject of whether the | (dog smells comes up for example bill and sue banter) | back and forth | (happily both) | with half a smile on their lips*

sunken: six sentences from the *New Scientist*, also analysed by Hunston and Francis (2000)

> ***1*** *as a rule | (books proclaiming) | the solution of a mystery | (deal) | with something that | (isn't mysterious) | or fail to deliver* ***2*** *(the sunken kingdom falls into both categories* ***3*** *(plato 's atlantis) | is a mystery | only to those who care* ***4*** *and the solution | offered in a readable and | (well-argued fashion) | is not conclusive* ***5*** *(peter james distinguishes between believers chasing the real atlantis and sceptics) | who are hostile to the very idea* ***6*** *but i fear he | (is omitting) | a far larger | (category -) | those who find | (this) | a waste of time*

bryson: a sentence by Bill Bryson, analysed in Hoey (2005)

> ***1*** *(in winter hammerfest is a thirty-hour ride by bus from oslo though) | why anyone would want to go there | (in winter) | is a question | (worth considering)*

hoey: Hoey's rephrasing of the 'bryson' sentence, deliberately avoiding typical lexical relationships for illustrative purposes

> ***1*** *(through winter rides between oslo and hammerfest use thirty hours) | up in a | (bus though why travellers would select to ride there then might be pondered)*

All these extracts have been analysed as described in the previous section. From a genre/register point of view they are not very wide-ranging, but there is some scope for variation; all are written, two are fiction ('adams', 'clock'), two non-fiction ('blink' and 'bryson'), several are newspaper/magazine ('bbc', 'independent', 'sunken'), one is a reprint of a newspaper column ('clarkson'), one academic prose ('stubbs'), and one is a made-up paraphrase ('hoey').

If we express the rate of re-use as a percentage of words of a text covered by MWUs, we get a wide range from 11% for Hoey's non-idiomatic reformulation of the Bryson sentence up to 100% for the formulaic call for papers extract.

Table 4: MWU coverage per text (in %)

text	%	text	%
Conference	100.0	sunken	64.7
Clarkson	79.2	bbc	59.3
Independent	78.8	blink	53.5
Adams	76.1	bryson	38.0
Stubbs	74.3	hoey	11.0
Clock	67.2		

Perhaps unexpected is that some of the non-fiction samples tend to score lower than the fiction ones; this suggests that coverage cannot simply be equated with routine/lack of creativity (in a literary sense). It also shows features of a readability measure, which makes sense if we consider that routine usages are easier to decode.

It is interesting to note that the lowest coverage is for a piece of made-up text, which Hoey deliberately created to sound clumsy and un-natural for the purpose of comparing it with the original. Our method of analysis seems to be able to pick up the non-naturalness, as almost none of the sentence's lexical items are used in their expected contexts.

Hoey argues that from a traditional point of view there is nothing wrong with his invented sentence, as it is perfectly grammatical; yet it is obviously odd. His point is that it is the collocational and colligational patterns that are broken which make the sentence un-idiomatic. However, if we look at it from the point of view of phraseology, then we could argue that it is the multi-word units (ie the basic elements of the idiom principle) which are ignored, that untypical words are combined individually and taken out of their context (according to the open-choice principle). From that angle, collocation (and colligation) are just epiphenomena of multi-word units: simply because lexical items are used in a fixed set of contexts, all the words that are also contained in these contexts occur near to them more often than would be expected. The collocation algorithms simply pick this up and identify those context words as significant collocates. Reversing this process, Danielsson (2001) does in fact use collocations to construct her units of meaning.

To summarise, this brief exploration of the routine re-use of language fragments and creativity indicates that there is a link between the degree of creativity as identified through multi-word units and certain properties of the sentence: a high degree of re-use points to stock phrases; a low degree reveals non-natural language; and anything in-between is 'normal'. Obviously, a lot more further analyses are required in this area to establish valid benchmarks. This would involve longer texts from a wider range of genres, and also cross-comparisons with corpora other than the BNC for MWU extraction. It would also be interesting to explore where poetry fits into this range.

7. Summary and Discussion

In this paper we investigated a phraseological approach to grammar. Based on the notion of a multi-word unit, a re-current combination of words, we noted that parts of sentences can be viewed as (often overlapping) sequences of MWUs. This is a first step towards a 'proper' grammatical description; so far we are not using grammatical categories, but neither does similar work by Sinclair and Mauranen (2006).

What is the point of this description, when it does not provide labels of either syntactic or functional nature? The answer is that it is a lexicalised approach to grammar, where the function realised by a particular segment is the meaning instantiated by the MWU. The expression of meanings of some sort is the fundamental reason for using language, whereas syntax is merely a by-product of the linearisation of thought.

This, however, is not the full story. There are gaps in the descriptions, some of which we have attempted to patch up in the discussion of the samples above. While the idiom principle goes a long way when composing structures, there are instances where there are slots or wildcards which allow a greater degree of choice. This choice cannot be captured with MWUs, as they rely on repetition, but for the purpose of accounting for the full structure we could combine the MWU approach with a set of local grammars, which would result in a hybrid model, combining both the idiom principle and the slot-and-filler one, only with constraints put on the possible items than can occur in a slot.

Another related problem that we are facing is that our definition of MWUs might be too restrictive. There can be a lot of variability in the actual ordering of words, with certain elements inserted or omitted, but none of that is currently captured by our MWU identification algorithm. Eventually we will want to be able to relate similar MWUs to each other, rather than treating them as separate sequences.

So, what we have presented here is only the first step towards establishing phraseology as a legitimate way of describing sentence structure. There is still a lot of work to be done, but at the same time we have already established the main principles, combining work in grammar with that in phraseology.

References

Brazil, D. 1995. *A Grammar of Speech*. Oxford: OUP.

Carter, R. 2004. *Language and creativity*. London: Routledge.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

Danielsson, P. 2001. *The automatic identification of meaningful units in language*. PhD dissertation, Göteborg University.

Francis, G. 1991. Nominal group heads and clause structure. *Word*, **42**, p. 144—156.

Francis, G. 1993. A corpus-driven approach to grammar. In M. Baker, G. Francis, and E. Tognini-Bonelli (eds) *Text and technology: in honour of John Sinclair*, p. 137—156. Amsterdam: Benjamins.

Gledhill, C. 2000. *Collocations in Science Writing*. Tübingen: Narr.

Harris, Z. S. 1955. From phonemes to morphemes. *Language*, **31**(2), p. 190—222.

Hoey, M. 2005. *Lexical Priming: A new theory of words and language*. London: Routledge.

Hunston, S. and G. Francis 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

Kita, K., Y. Kato, T. Omoto and Y. Yano 1994. Automatically extracting collocations from corpora for language teaching. In T. McEnery and A. Wilson (eds) *Corpora in Language Education and Research: A Selection of Papers from TALC94*. Lancaster: Department of Linguistics.

Mason, O. 2006. *The automatic extraction of linguistic information from text corpora*. PhD Dissertation, University of Birmingham.

Louw, B. 1993. Irony in the text or insincerity in the writer?—the diagnostic potential of semantic prosodies. In M. Baker, G. Francis, and E. Tognini-Bonelli (eds) *Text and technology: in honour of John Sinclair*, p. 157—176. Amsterdam: Benjamins.

Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J. M. and A. Renouf 1991. Collocational Frameworks of English. In K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics: Studies in honour of Jan Svartvik*, p. 128—144. London: Longman.

Sinclair, J. M. and A. Mauranen 2006. *Linear Unit Grammar*. Amsterdam: Benjamins.

Stubbs, M. 1993. British traditions in text analysis—from Firth to Sinclair. In M. Baker, G. Francis, and E. Tognini-Bonelli (eds) *Text and technology: in honour of John Sinclair*, p. 1—33. Amsterdam: Benjamins.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs, M. 2001. *Words and Phrases*. Oxford: Blackwell.

Stubbs, M. and I. Barth 2003. Using recurrent phrases as text-type discriminators. *Functions of Language*, **10**(1), p.61—104.

'Sailing the islands or watching from the dock': the treacherous simplicity of a metaphor. How we handle 'new (electronic) hypertext' *versus* 'old (printed) text'

Wolfram Bublitz

University of Augsburg

Abstract

This paper looks at the validity of two tightly interrelated linguistic dogmas. They state that the dyadic nature of human communication is an indispensable precondition for negotiating meaning, which is understood as a dyadic, transitive and reciprocal act requiring two interactants. It will be shown that since the advent of the new electronic media, both dogmas have been subject to a process of gradual erosion. Some forms of computer mediated communication have altered our understanding of participation as a dyadic and focussed concept. Furthermore, despite their amazing possibilities and extraordinary interactive potential (which, however, is at least partly counteracted by the extremely high degree of fragmentarization), interacting with new electronic media does not per se guarantee easier understanding, i.e. an easier access to the world 'behind the screen' than when interacting with 'old' printed media. It is argued that the user's situation is not essentially different from the familiar situation of the reader who is trying to understand printed text. [*]

1. Introduction: even eternal truths are not what they used to be

Outside grammar, there are not very many 'eternal truths' in the science of linguistics. Arguably, that duality (i.e. dyadic orientation) is a basic feature of human communication, is one of them, and that meaning is always negotiated meaning, is another. These two dogmas are tightly interrelated. They cohere because negotiating meaning is dyadic by nature in that it is a transitive and reciprocal act requiring two interactants. Hence, the dyadic character of human communication is an indispensable precondition for semiosis, i.e. the act or process of meaning-making (in the Peircean sense).

As is sometimes the case with everlasting truths, however, an unforeseeable change of their conditional fundaments can lead to their erosion. As I will argue in my paper, this appears to have happened with duality as a dogmatic feature of human communication. Since the advent of the new electronic media, it has been subject to a process of gradual erosion and is no longer unrestrictedly valid for both 'old' (spoken and written) and 'new' electronic media. Some forms of computer mediated communication (CMC) in particular have altered our understanding of participation as a dyadic and focussed concept, and have also made negotiating meaning and thus understanding more difficult. The latter may come as a surprise because the possibilities of the electronically administered new media with their literally infinite number of audio-visual data are widely regarded as an asset rather than as an impediment to composition and thought. But, as we will see, the interactive potential of CMC is (at least partly) counteracted by the high degree of fragmentarization (with all its consequences). Thus, despite its extraordinary possibilities, interacting with this new medium does not *per se* guarantee easier understanding, i.e. an easier access to the world 'behind the screen' than when interacting with old media. In actual fact, the user's situation is not *essentially* different from the familiar situation of the reader who, when reading a book, a handbook or a newspaper, is trying to understand, i.e. to create his or her own inner world.

2. Communication is not as dyadic as is generally assumed

Among the long-established dogmas that spring to mind when studying how communication works is the following: Human communication is most obviously characterised by its speaker/writer – hearer/reader symmetry, i.e. its dual or dyadic orientation. To communicate means for someone to communicate *with* someone else; it is a reciprocal act.[1] As a fundamental principle, this time-honoured dictum has seldom been queried in its entirety, though every now and then in some of its aspects (as I will show presently). A succinct description was provided by Wilhelm von Humboldt. In an article about *dual* as a grammatical number (besides singular and plural), he reflected in a more general way on *duality* as a universal communicative principle:

> Besonders entscheidend für die Sprache ist es, daß die Zweiheit in ihr eine wichtigere Stelle, als irgendwo sonst, einnimmt. Alles Sprechen ruht auf der Wechselrede, in der, auch unter Mehreren, der Redende die Angeredeten immer sich als Einheit gegenüberstellt. [...] Es liegt aber in dem ursprünglichen Wesen der Sprache ein unabänderlicher Dualismus, und die Möglichkeit des Sprechens selbst wird durch Anrede und Erwiderung bedingt. (1827/1969: 138)

Duality is the most obvious defining feature of two-party talk as the archetypical kind of spoken face-to-face communication in a homogeneous and focussed social setting. Human verbal communication is by nature dialogic. At closer inspection, however, neither the prototypical speaker nor the prototypical hearer are monolithic concepts but fusions of various conceptual roles. To take a simple example from the production side of verbal interchange: The *speaker*, the *author* and the *source* of a piece of text can be *three* different persons (e.g., a government spokesman reading out a secretary's account of a cabinet minister's ideas to a journalist), *two* different persons (the secretary reading out her own account of a cabinet minister's ideas to a journalist), or just *one* person (the cabinet minister telling the journalist herself her ideas). Or, focussing on the reception side, that the *hearer* of an utterance is not necessarily its *addressee* (i.e.

the interlocutor who is expected and entitled to reply) is a blatant truth that has by now been accepted even in speech act and inference theoretical circles.[2]

To account for such diversity, various theoretical frameworks have been developed, in which the monolithic, unitary concepts of the producing and the receiving participant are being deconstructed.[3] Among such frameworks we find Goffman's (1981) well-known distinction between "production format" and "participation framework" or, to use Levinson's (1988: 169 ff) rather more comprehensible pair of terms "production roles" and "reception roles"; the following outline of categories is taken (and slightly adapted) from Levinson (1988: 169):

Production roles
1. *animator*: 'the sounding box'
2. *author*: 'the agent who scripts the lines'
3. *principal:* 'the party to whose position the words attest'

Reception roles
A: *ratified*
   1. *addressed recipient*: 'the one to whom the speaker addresses his visual attention and to whom, incidentally, he expects to turn over his speaking role'
   2. *unaddressed recipient*: 'the rest of the 'official hearers', who may or may not be listening'

B: *unratified*
   1. *over-hearers:* 'inadvertent, non-official listeners' or '*bystanders*'
   2. *eavesdroppers*: 'engineered, non-official followers of talk'

Even though the two juxtaposed composite poles have been deconstructed into several components, which can be adopted by one interlocutor or distributed among several interlocutors, each can still be regarded as an integrated unit or "Einheit" (in Humboldt's account), which is not unanalyseable and monolithic but flexible and multi-facetted, allowing for various degrees of internal variation. Thus, without losing its dyadic character, Goffman's participation model can be adapted to cover one-many-talk or many-one-talk, or other complex forms of interaction such as chaired panel discussions open to the public. And it can also easily be expanded by implementing an even more refined sub-categorization of its constituent roles. The analysis of the following two examples, for instance, calls for the introduction of the role of *intended* recipient, who is neither *directly addressed* nor simply *listening* (i.e. *unaddressed* in Goffman's account).

(Four US high school students, one middle school student and the recording teacher, Joan, are working in the school's writing lab.)

(1) Sue: oh you're you're from the Middle School
Cyndy: yep
Sue: I was wondering
Kim: oh yeah I didn't introduce you Cyndy this is Sue Sue this is Cyndy

(2) Sue: well you guys
Don: see my my introduction's like (using fingers to indicate two inches) just like that long and that's it
Sue: wait are we being recorded
Kim: keep it I don't know I mean I don't know if that's too long
Don: yeah that's really short
Mary: she's taking a class in linguistics and she's not really looking for what we're saying but how we say it (…)
Sue: ha
Mary: how they speak
Sue: okay
Joan: I promise you no one here will hear about this except for me
(Joan Wallace, *Mixed Sex Discourse*, USA 1994; private data; names of students changed; adapted)

In (1), the exchange is opened by Sue, who, as *animator*, *author* and *principal* addresses Cyndy (by asking her). Cyndy is *addressed recipient* and re-addresses Sue by answering her question. Kim, on the other hand, is not *addressed recipient* but *intended recipient* because she reacts to Sue's implicit reproachful request by addressing both her interlocutors. In (2), two exchanges overlap. Don and Kim address each other, are *animator*, *author* and *principal* as well as *addressed recipient*, respectively and subsequently. With both her initial utterances, Sue addresses everyone present. But only Mary accepts the role of *addressed recipient* and re-addresses Sue by answering her question "Wait are we being recorded?". For the length of this and the ensuing exchange, the other three persons present assume the role of *unaddressed recipient*. However, following Sue's "Okay", Joan, somewhat belatedly, reacts to Sue's original question (which she indirectly confirms) and Mary's utterances, thus switching from *unaddressed* to *intended recipient*. (She cannot be *addressed recipient*, though, because Mary refers to her in the third person singular.)

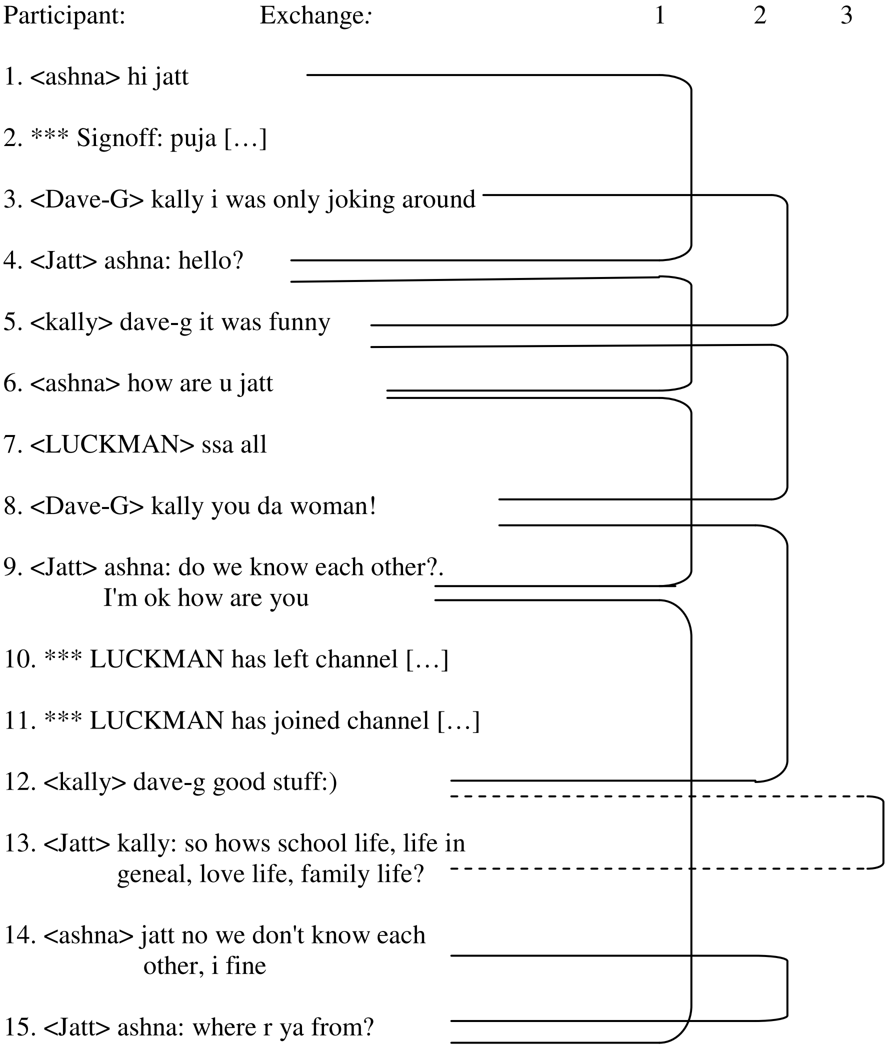
Goffman's participation model has been criticized for a number of deficiencies and limitations, which predominantly concern individual problems of definition and terminology or aspects of the assumed social setting,[4] but not for its general dyadic construction. Despite the intricate patterns of categories, co-categories and sub-categories at both the producing and the receiving side, it is still a dyadic model of communication, which upholds the underlying presumption that human interaction is principally and distinctly dyadic.

But blurred, fuzzy and generally unclear participant structure can even be found in 'old' printed texts, e.g. in handbooks or encyclopaedias with their wealth of references, cross-references, inserts, self-contained texts and strong iconic orientation, which is manifest in images, drawings, graphics, layout, etc; they reflect some aspects of modern electronic media, such as fragmentarization (constituting a kind of *reverse remediation* in Bolter's (2001) sense).[5] And it is obviously even more problematic for forms of new electronic media, which are typically and in varying degrees multimodal, fragmented and interactive (cf. below).

Hence, with the advent of the new electronic media the erosion of the presumption that communication is and must be dyadic has reached a new dimension. This becomes immediately evident when looking at two forms of electronic communication, computer mediated chats and Wiki-media communications.

Computer mediated chats differ from traditional conversation, as discussed by Goffman, in several ways.[6] Crucial differences are that their structure is not necessarily dyadic and that they are not "focused social interactions" occurring in "particular physical spaces" and involving "easily identifiable participants with clearly defined roles and relationships" (Jones 2004: 23), who monitor each other's actions, attitudes and presence. The clear distinction between participant roles is blurred in such online chats based on multi-functional technological gear supported by various kinds of instant-messaging software (like ICQ) (cf. Jones 2004). In these hybrid forms of human communication (which share features of both speaking and writing), ongoing interaction is basically multilateral, a multilogue, rather than bilateral and a dialogue, with no clearly discernable participation roles of the kind described by Goffman and other analysts.

To illustrate this point, I borrow one of Herring's (1999) examples of *overlap in CMC*, because "overlap in CMC is […] problematic. On the one hand, temporal overlap in display of turns is not an option in one-way CMC, since one-way systems force messages into a strict linear order. On the other hand, overlap of *exchanges* is rampant in computer-mediated environments. In dyadic communication, users are unable to tell whether their interlocutor is in the process of responding or not. They may become impatient and send a second message before a response to the first has been received, resulting in incomplete or interleaved exchange sequences […]. In group communication, unrelated messages from other participants often intervene between an initiating message and its response […]." (1999: 4) I have slightly adapted her example (which is taken from CM group communication "on a public IRC channel") and added the lines:



Participant: Exchange*:* 1 2 3

1. <ashna> hi jatt

2. *** Signoff: puja […]

3. <Dave-G> kally i was only joking around

4. <Jatt> ashna: hello?

5. <kally> dave-g it was funny

6. <ashna> how are u jatt

7. <LUCKMAN> ssa all

8. <Dave-G> kally you da woman!

9. <Jatt> ashna: do we know each other?.
I'm ok how are you

10. *** LUCKMAN has left channel […]

11. *** LUCKMAN has joined channel […]

12. <kally> dave-g good stuff:)

13. <Jatt> kally: so hows school life, life in
geneal, love life, family life?

14. <ashna> jatt no we don't know each
other, i fine

15. <Jatt> ashna: where r ya from?

Here is part of her analysis:

> Two extended dyadic exchanges are interleaved in this sample of chat, one between Ashna and Jatt, and the second between Dave-G and Kally. To complicate matters further, in line 13, Jatt initiates a third exchange by addressing a question to Kally. [...] The perspective [as represented by the lines] is anaphoric – the participant lower in the diagram is considered to be responding "backwards" (or in this case, upwards) to a previous participant in each case. Dotted lines indicate interactions in which the message either initiates a new exchange with an already active participant (as in 1 and 13) or responds to a turn not

> included in the example (as in 3). [The lines show] clearly that exchanges overlap, rather than taking place in sequence – turns from one exchange regularly "interrupt" another. (Herring 1999: 4 f)

The basically multilateral and multilinear character shapes, *mutatis mutandis*, also written communication in *Wiki- or Web 2.0-based media formats*, which are generally believed to support online collaboration among large numbers of users. Hence, the meaning-making and even text-building actions can no longer be assigned to individual but only to 'multiple authors', who, in exceptional cases, may still be identifiable as individuals but are usually unidentifiable members of collective networks.

Furthermore and going one step further, in computer based semiosis, even the recipient side is involved in that the roles of author and user regularly coincide. This leads to the central question of who is *doing* the semiosis in these cases. If the distinctions between participant roles blur or even vanish, if users cannot be distinguished from authors and users become their own authors, how is the essentially *collaborative* action of creating meaning, i.e. of understanding achieved? And, as meaning is always *negotiated* meaning, who is negotiating with whom (on what evidence or data input)?

To answer these questions, let me first explain what I mean by the two key concepts *understanding* and *negotiating*.

3. Negotiating is best seen as a metaphor

In agreement with hermeneutic, interpretive and usage based approaches within semantics and pragmatics, I adhere to a theory of comprehension which views understanding as a cooperative activity, resting on speakers' and hearers' immediate and writers' and readers' delayed collaboration, rather than on each person's autonomous and strictly individual action.[7]

And while it is a platitude to state that understanding as an act and cognitive process is a private affair in the sense that it happens in a single person's mind, it is likewise true that a person's mind is not autonomous and isolated in the sense that it is totally cut off from its *Umwelt*, including (the output of) other people's minds. There is a constant *interactive* exchange of information between an interactant's 'inside world' and the 'outside world'; understanding could not happen otherwise. Only from a neurocognitive (and, incidentally, also analytical) point of view are one interactant's understanding and meaning private, idiosyncratic, unique and fully distinct from another's. From a semantic-pragmatic point of view, however, which takes 'inter-action' as its focal point, they are compatible and concordant to a degree that we can talk of 'collective acts of comprehension' and 'shared meanings'.

It has been argued that understanding is not done (as an act) *by* someone (by interactively gathering and accumulating and arranging information), but that it *happens to* someone (in a quasi autopoietic manner).[8] While I reject the underlying highly mechanistic and de-humanized view, I readily concede, of course, that the amount of cognitive effort required when understanding may vary considerably. After all, much understanding is routine, with little or no hermeneutic distance to be bridged. Understanding means reducing distances and overcoming differences! Among them, relevant for our topic are the *linguistic difference* (hearers have to ‚translate' speakers' language into their own and vice versa), the *historical difference* (which separates quite literally the time of writing and the time of reading of a text; or, when related to the comprehension process itself, the difference between each of the emergent and successive states of understanding, which are always only provisional, reflecting the status quo, to be adapted and modified later), the *representational* (or *rhetorical*) *difference* (e.g. between familiar formal means of presentation and unfamiliar audio-visual, iconic and related signs and formats as used in CMC), and the *episodic and 'semantic' difference* (referring to interactants' memories and systems of knowledge, as described by Tulving 1972 and 1983). In various forms of CMC, some of these distances turn out to be difficult for users to bridge when negotiating meaning, as I will argue presently.

In communication, meaning is always jointly acquired and shared meaning, which is communicatively valid. We are not free agents when it comes to cooperate in order to understand; cooperation in understanding is an anthropological constant. And this is where negotiating comes in, because comprehension rests on the negotiation of *self* with *other*.

Unlike *self*, which I wish to believe is still a human individual (and not an electronic 'mind'), both *negotiating* and *other* can be taken literally or figuratively.

Literally, negotiating is a bilateral, reciprocal, dialectical action between the understanding person, the *self*, and his or her interlocutor, the *other*. What is to be understood is not given, a priori existing, static meaning but emergent, dynamic meaning, which is manifest in text or discourse. When negotiating, the interactants therefore refer to the piece of text or discourse in question, but they also need to and do rely on other sources of directly relevant information such as non-linguistic signs (pictures, sounds, kinetic signs) or the situational and socio-cultural setting. Negotiating in the literal sense hinges on speech acts such as suggesting (e.g. readings of a word or clause) and accepting or rejecting, querying and explaining, doubting and affirming, which themselves involve acts of supplementing or completing as in the following examples:[9]

(3) C you didn't have capital gains but of course you did háve ∂: .
a death du*ties*
C *dèath* duties
(Svartvik & Quirk 1980: 347)

(4) C because there was some pecùliar - ∂: convention about hyphens which just
B seemed quite *àrbitrary*
C *it was* absolutely illògical - - (Svartvik & Quirk 1980: 134)

In (3) and (4), by completing their interlocutors' utterances, *a* and *B* offer their understanding of what is meant as part of a negotiating process. In both cases, the offer is successful, i.e. accepted (also as part of the negotiating give-and-take) by the other. This is different in the following examples, where negotiating takes a different direction because the offer is 'wrong' and not accepted.

(5) A this is the one I could most live wìth . *the cardinals*
B _*the stàtues*_
A well the càrdinal áctually
(Svartvik & Quirk 1980: 203)

(6) A the amount you get from - wèll . firstly the […]
B _amount you get from the sun doesn't còunt_
A well no it dòes it's quite impòrtant

(Svartvik & Quirk 1980: 598)

In (7) and (8), supplementing other's talk is a means of negotiating shared stance as well as shared meaning.

(7) C course it was Pòrt réally that kept them wàrm . in the eighteenth cèntury
a _and enormous quantities of food ._
C yes

(Svartvik & Quirk 1980: 340)

(8) A very óirish with a màss of great - ∂ sort of grey háir
C _and a Catholic of course_
A and a Catholic presúmably - *∂:m*
C _*lots of* chìldren ._
A three grown-up chíldren - all márried [...]

(Svartvik & Quirk 1980: 748)

In these transcripts of spoken face-to-face conversation, *the other* is the interlocutor who is present. In written communication, the author or writer is either known or unknown, inferable or not, available (contactable) or not, more or less displaced, depending on the kind of written communication (e.g. a personal letter or a novel versus an unsigned newspaper editorial or advertisement copy writing). If the reader has no access to the writer, he or she has to resort to assumptions about the writer, i.e. has to *create the other*. Which leads us away from the literal and into the realm of the figurative.

So far, we have assumed that *the other* is the negotiating self's human collaborator. But since meaning is frequently negotiated not with the human collaborator directly (as in the examples above) but, in a figurative sense, with the verbal, non-verbal, situational and other available data, *the other* can actually stand for the totality of such evidence. Accordingly, *the other as the totality of data on which the self draws when negotiating meaning* stands metonymically for *the other as human collaborator*. Getting back to our dogma of the dyadic nature of communication, while the understanding pole, i.e. *the self*, is still a human individual and *Einheit* in new electronic media, the producing pole, i.e. *the other*, typically dissolves into a heterogeneous array of data.

The true explanatory power of *negotiation* unfolds itself only when *negotiating* is used as a metaphorical expression with the underlying concept NEGOTIATING IS MAKING ASSUMPTIONS. In other words, negotiating meaning is *making assumptions about* what is meant by the author of an utterance, a text, a picture, a sound etc *on the basis of* the relevant and available data. Negotiating being a distinctly empathetic act, this means that the self also draws on the (known or assumed) collaborator's inferred and construed linguistic and world knowledge, episodic and conceptual memory, cultural background and emotional frames of mind, as part of the cognitive scaffold which supports the emerging meaning.

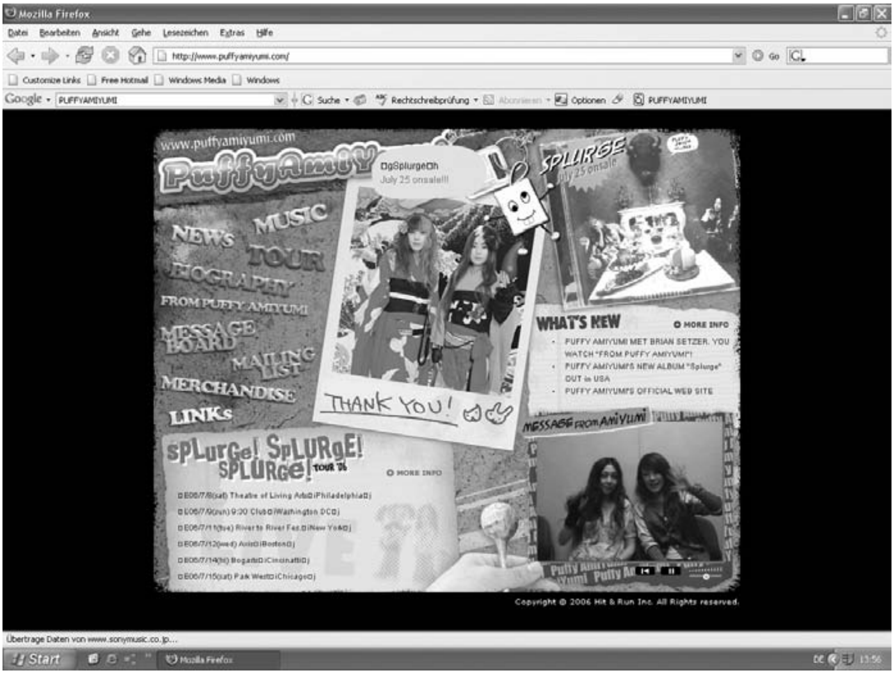
To demonstrate how negotiating of meaning and text actually works in CMC and what problems the negotiating (i.e. meaning understanding and text building) user is confronted with, we have to cast a brief look at the concept of (hyper-)text first and recapitulate its characteristic features.

4. Hypertext and the privilege of the eye

Disregarding *e-documents*, which are nodes that are simply electronic versions of conventional written texts and thus of no interest for our topic, the notion of *text* can be used when studying CMC, even though spoken and written language in CMC is obviously not the sole and, sometimes, not even the most important medium of information.

In accordance with the hermeneutic orientation of this paper, a *text* is seen not as a document that is 'there' for the reader or user 'to find', i.e. not as input to understanding, but as the output of the reader's or user's interpretation. Each user creates his or her own text by relating perceivable data to the cognitive framework of his or her mind. The emerging text is understood as coherent and meaningful, with a topic, a purpose, a function, embedded in and dependent on a situation, a socio-cultural environment, a set of other texts.

A chief difference between traditional printed media and new electronic media is that for the latter's user to decide what (i.e. which data) constitutes a text, can often be a serious challenge. It can be difficult to decide, which of the perceivable assortments of signs displayed on the site can be taken to count as text. There may be no clear demarcation lines between sequences of words (which may look like a text, i.e. *Fließtext* or continuous text) and surrounding bits of audio-visual data appearing in a kaleidoscopic wealth of signs, which range from pictures, graphics, pop-ups, bars, frames, links, films, sounds to layout and color.


Mozilla Firefox
Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
http://www.puffyamiyumi.com/
Go
Customize Links Free Hotmail Windows Media Windows
Google PUFFYAMIYUMI Suche Rechtschreibprüfung Abonnieren Optionen PUFFYAMIYUMI
www.puffyamiyumi.com
PuffyAmiYumi
NEWS
MUSIC
TOUR
BIOGRAPHY
FROM PUFFY AMIYUMI
MESSAGE BOARD
MAILING LIST
MERCHANDISE
LINKs
□gSplurge□h
July 25 onsale!!!
SPLURGE
July 25 onsale
THANK YOU!
WHAT'S NEW
MORE INFO
MESSAGE FROM AMIYUMI
sPLurGe! SpLURgE! SPLURGe! TOUR '06
MORE INFO
Puffy AmiYumi
Copyright © 2006 Hit & Run Inc. All Rights reserved.
Übertrage Daten von www.sonymusic.co.jp...
Start
Mozilla Firefox
DE 13:56

- PUFFY AMIYUMI MET BRIAN SETZER. YOU WATCH "FROM PUFFY AMIYUMI"!
- PUFFY AMIYUMI'S NEW ALBUM "Splurge" OUT in USA
- PUFFY AMIYUMI'S OFFICIAL WEB SITE

□E06/7/8(sat) Theatre of Living Arts□iPhiladelphia□j
□E06/7/9(sun) 9:30 Club□iWashington DC□j
□E06/7/11(tue) River to River Fes.□iNew York□j
□E06/7/12(wed) Axis□iBoston□j
□E06/7/14(fri) Bogarts□iCincinatti□j
□E06/7/15(sat) Park West□iChicago□j

(http://www.puffyamiyumi.com)

Figure 1: Kaleidoscopic wealth of signs in a website of a Japanese pop duo

One can actually *see* the difference between 'old' printed media and 'new' electronic media as a result of the often cited *iconic turn*, first noted by Mitchell (1994) and Boehm (1994), who called attention to the powerful (and still growing) contribution of pictures and other iconic signs (including metaphor) to semiosis, i.e. to our semiotic approach to and interpretation of reality.[10]

Considering the wealth of electronic possibilities of creating information in CMC, it is reasonable to argue for a broader reading of *text* which takes into account both symbolic (mostly linguistic) and iconic signs. To this end, the notion of *hypertext* has been coined. It refers to a much wider concept than *text*; indeed, it incorporates text as one of its components. I adopt Slatin's definition of *hypertext* as "an assemblage of texts, images, and sounds – nodes – connected by electronic links so as to form a system, whose existence is contingent upon the computer. The user/reader moves from node to node either by following established links or by creating new ones." (Slatin 1991: 56)[11]

From this definition, we can deduce five constitutive features of hypertext: (a) *computer mediation*, (b) *multilinearity*, (c) *multimediality*, (d) *fragmentation* and (e) *interactivity*.

(a) Computer-mediation: Hypertexts consist of digitalized, electronically mediated bits of information. Anything that can be written, spoken, drawn, filmed, etc can be turned into digital signs to be put on the Web.[12] This in itself is not a feature of hypertext that impedes meaning- and text-making negotiation.

(b) Multilinearity: It is generally distinguished between *medial linearity* and *conceptual linearity*. Medial linearity depends on and is conditioned by the *medium* in varying degrees. It is weaker in a newspaper, where the reader can easily deviate from the order in which the items are printed; it is stronger in a video or audio tape, where the medium restraints the viewer's or listener's choice of order. Conceptual linearity is given by the *author*, who is merely suggesting in which order the written material can be perceived; the binding factor is stronger in a work of fiction (but cf. below on hyperfiction) than in a travel guide or a dictionary. Both types of linearity, medial and conceptual, can also be transferred, *mutatis mutandis*, to multilinearity. Of course, the internet and thus hypertexts are particularly well suited for conceptual multilinearity, which in itself is an asset rather than an impediment to meaning- and text-making negotiation.

(c) Multimediality, also often called multimodality: The term *medium* is polysemous.[13] It can either refer to the *hardware*, the material devices that carry and transmit information, such as newspapers, books, radios, tv-sets, MP3-players, computers as well as mobile devices such as pdas, cell phones and, though less material and more virtual, the internet. Or it can refer to the *representation format* such as the spoken words (utterances, discourse), the written texts, pictures, illustrations, drawings, graphics, layout, typography, film, melody. The alternative term *mode* as in *multimodality* is usually applied to the latter reading because it generally refers to the ways in which information materializes. Multimodality (-mediality) in CMC is a scalar feature, reaching from monomodality (-mediality), e.g. in text-only CMC, to multi-modality (-mediality), which may comprise textual, visual and audio modes. The internet and hypertexts are multimodal (-medial) because of their rich diversity of representation formats. On occasion, it may be difficult for users to find their way through the wealth of kaleidoscopic data in order to negotiate meaning and textuality.

(d) Fragmentation[14] is among the most noticeable features of internet sites. Every node, picture, pop-up, hyperlink etc is a fragmentary informational and thus communicative unit, and as such a challenge for the user to ascribe meaning and textuality to. Like multilinearity and multimodality, interactivity is based on the fragmentary assemblage of different text clusters. In hypertexts, e.g., texts are interactively aligned across different nodes (internodal), while multimodality is mirrored in the fragmentary combination of text units within one and the same node (intranodal). Furthermore, it is generally distinguished between fragmentation across modes (e.g. textual and pictorial node) and fragmentation within one mode. The latter is a characteristic feature of chat communication.

Here, comments are broken down into smaller bits of incomplete information (or "messages", Beißwenger 2005), which have to be linked up by users with the help of cohesive means such as types of address, backchannels, turn-taking signals, cross-turn references.

(e) Interactivity is a scalar feature. At one end of a cline, we find self-contained asynchronous e-documents lacking in interactive potential, since they do not allow for online feedback or manipulation by the user. At the other end, there are chat environments (instant messenger, ICQ, IRC, etc) and video-conferences, which have a high degree of interactivity because they allow for synchronous and near simultaneous feedback and manipulation. The user can obviously choose from a much wider range of options to participate in the production, alignment and negotiation of content. Between these two extremes, there are intermediary types of interactivity, among them purely physical acts (in which the user connects self-contained 'text' units by, e.g., simply clicking hyperlinks) and purely cognitive acts (in which the user relates textual and audio-visual nodes to each other across the screen).

There is one more constitutive and quite salient feature of hypertext, the *(hyper-) link*. Links relate structured pieces of information, i.e. nodes, in an electronic and non-linear way.[15] They can be used to interfere with existing data or to create new data, e.g. when submitting to weblogs or using search engines. Links are mechanistic *instructions* and as such a significant and apparent component of negotiation. They substitute for identifiable authors (giving instructions). As such, they can even be described as a residual feature of a dyadic orientation of CMC.

5. The loss of the 'other participant' or the erosion of the dyadic principle

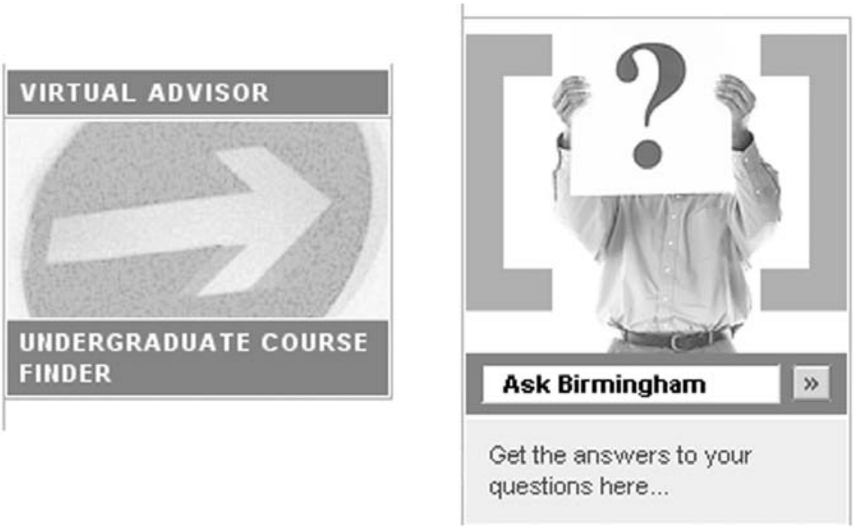
These properties of hypertext explain why different users can (and regularly do) create their own distinctive hypertexts which are different from other users'. Using the link function, they can easily and freely (i.e. multilinearly) navigate through a broad spectrum of modally different informational fragments that are easy to handle, i.e. to shift around, to replace, to (re-)arrange and to manipulate. In doing so, users create meaning by negotiating with 'the other' in the figurative sense explained above. The decisive factor, which promotes and supports negotiation in this way, is, of course, the interactive nature of the medium.

However, on closer inspection it is a very restricted kind of negotiation, whose outcome somehow thwarts the promising possibilities of the new media. Their insufficiencies when creating meaning and building own hypertexts are readily addressed, let me pick out a few.

Users can, of course, re-arrange the nodes, i.e. the fragments of information they are confronted with in CMC, even though they are not of their own but (frequently) of some anonymous author's making. Given fragments and newly created fragments, however, may not always be compatible as to their topic, their evaluative load, their register and function. Other-authored given fragments may, therefore, resist manipulation and thus impede the free creation of a new hypertext.

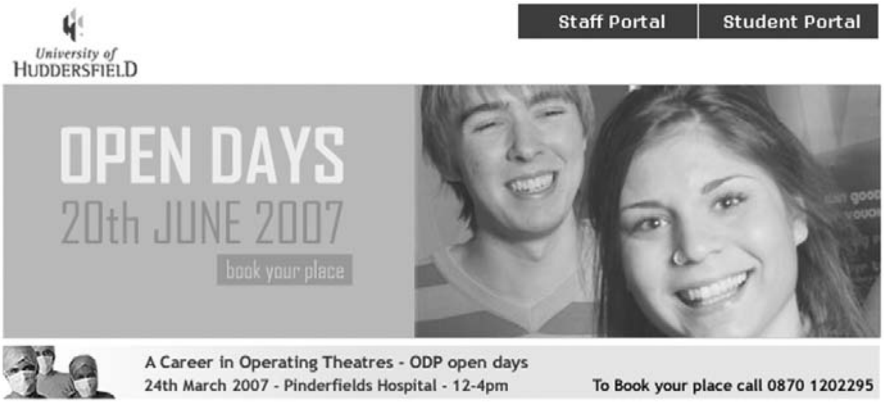
By the same token, bringing together textual and pictorial nodes which were not adjacent and related before can cause incoherence. The difficulty of making them cohere is due to a lack of 'traditional' and familiar cohesive means. In hypertext, the user is faced with a large number of nodes, which are frequently self-contained. Accordingly, all the relevant and expected means of linkage are often missing such as inter-node gambits and discourse markers, referring expressions other than definite descriptions and generally known proper names, tag questions etc. Users are familiar with the problems of scanning the screen for signs and signals that help to relate textual and audio-visual nodes to each other. While we are accustomed to the linguistic, non-linguistic and cognitive ways of establishing cohesion in non-electronic spoken or written communication, users have to learn what (other) means and strategies are used in CMC to relate current items or nodes to preceding or prospective other items or nodes. This is at least partly due to the interplay of various modes of presentation; the reiteration of pictorial elements, e.g., can support semantic connectivity proposed by the verbal structures in the text. Of course, refined and sophisticated types of hyperlinks have been developed to serve as cohesive devices. They are of a different, i.e. mostly non-linguistic nature and are thus parasitic on the multimodal form of hypertext (cf. Hoffmann 2006).

Hoffmann (2007), e.g., points out that hypertext users often miss "important evidence in hyperlink anchors needed for the appraisal of hyperlink trails. In this respect, all too often hyperlinks are insufficiently marked by their authors. Most forget that visualizing the content of target nodes is essential for determining which hypertext path one will follow. Likewise, informing the user about possible link trajectories is a central motive in maintaining the cohesive foundation of hypertexts. For these reasons, more and more hypertext authors use extensive, multimodal means (audio-visual signs) to provide users with information about the "hidden content" of their hyperlinks." And he takes his examples from the websites of several universities, which by now "have caught on to the new iconic possibilities of web design", making ample use of "images or photos in their online presentations":


VIRTUAL ADVISOR
UNDERGRADUATE COURSE FINDER
Ask Birmingham
»
Get the answers to your questions here...

(http://www.bham.ac.uk/)

Figure 2: Multimodal link anchor in university website

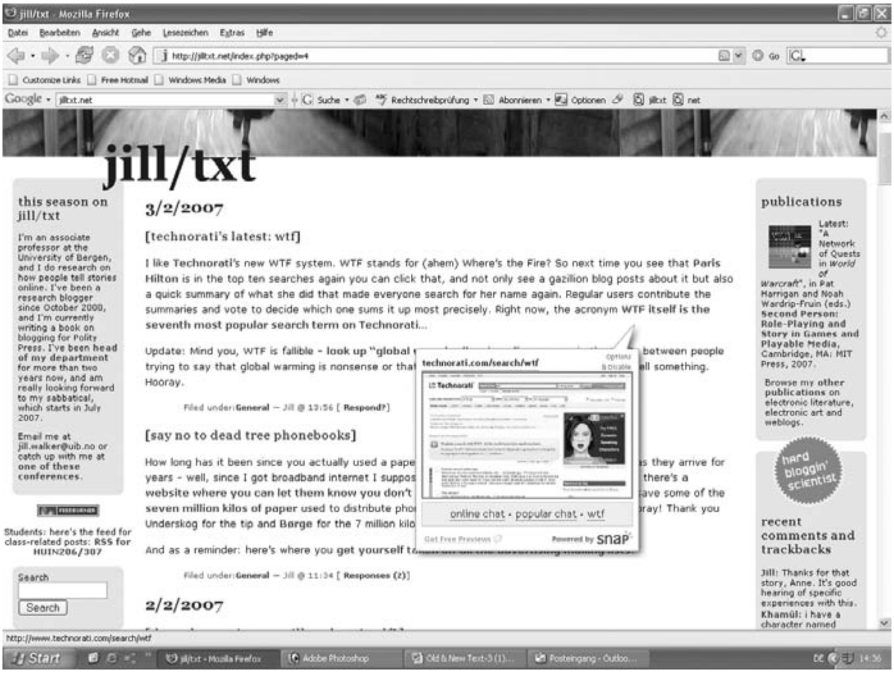

University of HUDDERSFIELD
Staff Portal
Student Portal
OPEN DAYS
20th JUNE 2007
book your place
A Career in Operating Theatres - ODP open days
24th March 2007 - Pinderfields Hospital - 12-4pm
To Book your place call 0870 1202295

(http://www.hud.ac.uk/)

Figure 3: Multimodal link anchor in university website

In general, it seems safe to claim that cohesion relies strongly on the particular mode used: speech, writing or, indeed, CMC.[16]

There is another reason why coherence building in hypertext is a challenge. Unlike readers of 'old' printed text, CMC users cannot operate on a default assumption of coherence. They cannot assume as a matter of course that what they will read or see next (when activating a link) *is* coherent (cf. Bublitz 2005a, 2006). However, work to downsize the problem is in progress. As Hoffmann (2007) points out, an interesting feature, recently introduced by the company Snap.com© and appearing on some weblogs "may have the potential to bridge coherence breaks between websites. Once a freeware program is installed successfully on the hard drive of a webpage owner, users can direct their mouse cursor over a hyperlink, and an additional window will appear instantly delivering an appropriate preview of the respective target area. The preview picture includes a search engine which can be used for looking up concrete words or phrases within the future website." Here is an example (courtesy of Hoffmann):



this season on jill/txt

I'm an associate professor at the University of Bergen, and I do research on how people tell stories online. I've been a research blogger since October 2000, and I'm currently writing a book on blogging for Polity Press. I've been head of my department for more than two years now, and am really looking forward to my sabbatical, which starts in July 2007.

Email me at jill.walker@uib.no or catch up with me at one of these conferences.

Students: here's the feed for class-related posts: RSS for HUIN206/307

3/2/2007

[technorati's latest: wtf]

I like Technorati's new WTF system. WTF stands for (ahem) Where's the Fire? So next time you see that Paris Hilton is in the top ten searches again you can click that, and not only see a gazillion blog posts about it but also a quick summary of what she did that made everyone search for her name again. Regular users contribute the summaries and vote to decide which one sums it up most precisely. Right now, the acronym WTF itself is the seventh most popular search term on Technorati...

Update: Mind you, WTF is fallible - look up "global between people trying to say that global warming is nonsense or that ll something. Hooray.

Filed under:General — Jill @ 13:56 [Respond?]

[say no to dead tree phonebooks]

How long has it been since you actually used a pape s they arrive for years - well, since I got broadband internet I suppos there's a website where you can let them know you don't ave some of the seven million kilos of paper used to distribute pho ray! Thank you Underskog for the tip and Børge for the 7 million kilo

And as a reminder: here's where you get yourself t

Filed under:General — Jill @ 11:34 [Responses (2)]

2/2/2007

publications

Latest: "A Network of Quests in *World of Warcraft*", in Pat Harrigan and Noah Wardrip-Fruin (eds.) Second Person: Role-Playing and Story in Games and Playable Media, Cambridge, MA: MIT Press, 2007.

Browse my other publications on electronic literature, electronic art and weblogs.

recent comments and trackbacks

Jill: Thanks for that story, Anne. It's good hearing of specific experiences with this. Khamûl: i have a character named

(http://jilltxt.net)

Figure. 4: Snap.com© applet used in Jill Walker's weblog

As Hoffmann (2007) points out, "it is highly probable that these simple applications could at least provide partial or preliminary solutions to the cognitive overload which stems from forward-looking planning strategies of hyperreading" (cf. also Bublitz 2005a: 321 f).

A further difference between 'old' printed text and hypertext nodes is that the latters' propositional meanings are often much more readily understood than interpersonal meanings. This is not only due to a lack of relevant means of emotive prosody, empathetic orientation, evaluative judgement, ideological stance etc. As any piece of text is created by a particular person (or several persons) and addressed to an intended reader, browser, user, it carries and displays subjectivity to a greater or lesser extent. But unlike with printed text, say a book, the probability is rather high that hypertext users who do not belong to the circle of intended addressees, visit nodes whose interpersonal, subjective impact they do not understand. The same holds for meaning 'between-the-lines', insinuations and allusions.

Generally, there is a much higher demand on the empathetic, knowledge inferring skills of the hypertext user than on those of the traditional reader of printed text. Pronounced empathetic proficiency is necessary for users to establish and develop *common ground*, which is an essential prerequisite for comprehension (and as such aimed at in any negotiating process) (cf. Bublitz 2006). The insufficiencies related so far clearly indicate that the establishment and maintenance of common ground is much more difficult and sometimes hardly possible in CMC.

6. The user as sailor or as watcher

Where do we stand? I have argued that the defining properties of hypertext can seriously interfere with the users' effort to ascribe meaning, coherence and textuality to the kaleidoscopic flux of online textual and audio-visual fragments of information, and to build and thus create own new hypertexts. Comprehension can become a challenge, because it rests on empathetic acts of negotiating, which require a real or assumed other interactant. The lack of a human other forces the user to negotiate with, i.e. to make assumptions about the assumed author's linguistic and world knowledge, episodic and conceptual memory, cultural background and emotional frames of mind, in order to establish a common ground as a prerequisite for understanding. In computer mediated communication, the simple dyadic set-up of prototypical speaker-hearer or writer-reader communication has been replaced by a set-up with the human self as user as one collaborator and the internet with its wealth of data as the other collaborator.

That semiosis in CMC can be more demanding than in printed texts or spoken discourse, is, on the face of it, not obvious at all. But actually, the user has to rely on his or her own interpretive skills, knowledge and experience to a much greater extent than the overwhelming wealth of electronically based devices and mediated data has us expect. Somewhat ironically, such a literally unlimited pool of audio-visual data offered by the internet can be an impediment rather than an asset for understanding.

Things may be different when we move from negotiating meaning to negotiating text, i.e. building own hypertext. Here, both the coincidence of the roles of user and author and the multitude and diversity of electronically provided means are definitely an asset. It is certainly regarded as an asset in *hyperfiction* where the user as reader is strongly invited to act as author by actively interfering with the composition of a story. That hyperfiction gains by the user turning author is nicely captured in a revealing metaphor (taken from the opening directions to a piece of hypertext fiction):

> The difference between reading hyperfiction and reading traditional printed fiction may be the difference between sailing the islands and standing on the dock watching the sea. (Guyer & Petry 1991)

There are obviously two sides to this metaphorical coin, which focusses on the recipient's (i.e. the reader's or user-reader's) stance. Underlying the metaphorical concept READING IS SAILING are the two presumptions that reading is passively happening to a reader, i.e. is a state rather than an act, whereas sailing is a dynamic activity by the sailor. By adding the specification "the islands" the verb acquires even greater power because the valency of "sailing" changes from transitivity ('sailing a boat') to complex-transitivity ('sailing a boat through the islands'). By pairing the two concepts, reading is turned into an undertaking or even a venture, which demands from the reader to constantly make decisions about the course to be taken, i.e. about what to do next, and to actually do it. The presupposed inactivity or passivity as a semantic feature of reading is, of course, confirmed in the second metaphorical concept READING IS WATCHING and further emphasized by the supplement "standing on the dock". The message of the two creators of the metaphors is clear: Printed texts are just there, for the reader to take them or leave them, to understand them or not, while hyperfiction is of the reader's own building and thus read and understood in the act of creating. And there is no doubt that in their eyes, the latter stance is preferable to the former.

The obvious next step is to transfer the metaphors of 'the user-reader as sailor' *versus* 'the reader as watcher' from building hyperfiction to the more general realm of building non-fictional hypertext and also to the question of how understanding is managed in hypertext and printed text. I believe that only the 'sailing'-side of the metaphor is apt. CMC users can indeed devise and control hypertext in an author-like manner. They 'sail the islands' in the sense that they actively handle the data provided by the internet in various ways. Briefly recapitulating, when building their own hypertexts, users

- employ prospective, forward-looking planning strategies, much like *authors* do when composing printed text;
- interact with the data provided by assembling fragments of information, i.e. nodes in an unforeseen, unforeseeable and ad hoc way (i.e. in a manner reminiscent of making a collage);[17]
- exploit the multilinear and multimodal nature of hypertext to create unusual but nonetheless meaningful sequences of (textual and audio-visual) bits of information, thus moving from exclusively monomodal to multimodal semiosis.

I contend, though, that the other side of the metaphorical coin does not hold. Even though the metaphor refers to 'reading', what is actually meant is the act of 'understanding while reading'. Neither reading nor understanding can be likened to standing (a state of immobility) and passively watching. As I have argued before, understanding is the act of ascribing meaning (coherence, function etc) to (textual or other sign-based) informational input. And that applies regardless of whether the understanding individual is looking at a book or a computer screen, at monomodal and steady print or multimodal and moving texts and pictures, is listening or reading, or, indeed, watching.

The authors of the metaphor insinuate that *watching* is a state rather than an act or process. In their view, watching appears as a one-dimensional mode of perception which regards the perceiver as passively receiving visual sensations together with their meanings, which are somehow inseparably attached to them. But such view does not hold. It is hard to imagine how 'watching the sea' (or 'the islands') does not involve recognition and thus conceptualisation of the perceived image *as* the sea. And conceptualising, i.e. understanding a linguistic object involves more than merely matching it with its mental image listed in the lexicon. Depending on the nature of the object, i.e. the linguistic expression, conceptualising can be a highly complex act of meaning construal, and even more so when strings of expressions, i.e. texts are involved as is normally the case when reading.[18]

What our metaphor does not cover is that watching can involve a higher or a lower degree of cognitive effort, can be difficult or easy, depending on the object watched. Applying this to understanding, it can be more difficult in some forms of media, notably in CMC, and less difficult in others.

7. Conclusion: The loneliness of the user

This paper has started out with a recapitulation of two of those dogmas that have widely been taken for granted in linguistics. Bringing them together, they state that the dyadic nature of human communication is an indispensable precondition for negotiating meaning, which is a dyadic, transitive and reciprocal act requiring two interactants. The subsequent attempt to find proof in the new electronic media for the universal validity of the two dogmas was not entirely successful. They are no longer unrestrictedly valid for both 'old' (spoken and written) and 'new' electronic media. Some forms of CMC in particular have altered our understanding of participation as a dyadic and focussed concept, and have also made negotiating meaning and thus understanding more difficult.

It is the latter fact that may come as a bit of a surprise because the development of the electronically administered new media with their infinitely large quantity of audio-visual data was certainly not intended as an impediment but as an asset for handling its possibilities and understanding its content. But the extraordinary interactive potential of CMC is (at least partly) counteracted by the extremely high degree of fragmentarization (with all its consequences as related above). Wiki- and related CM media are thus to some extent deconstructive media.

However, this, I would like to argue, is no cause and certainly no need for despair and lament. The multilinear, multimodal, fragmentary, intricate and occasionally perplexing way of presenting information is by no means confined to CMC, but is a totally familiar phenomenon to all of us. Even ordinary printed text and discourse do not always transmit their messages in a linear, orderly, explicit and straightforward way, which forces the hearer or reader to ascribe order to disorder, to create his or her own linearity, to make fragments of information cohere that are not cohesively connected, to infer the implicit from the explicit, the additional from the given information, and all that in an associative and occasionally roundabout way. Furthermore, the way users go about ascribing meaning to a vast array of fragments, is suggestive of the associative way human minds work when understanding. Contemplating can run on different levels of modality simultaneously, can be extremely fragmentary (like the outside-world) and quite unfocussed. Minds slip sideways. This metaphor holds both in virtual reality, just think of Molly Bloom's mind slipping sideways in her soliloquy (an often quoted long stream of consciousness passage in James Joyce's *Ulysses*) and in the real world, where it can be applied to users when moving on an unguided, self-constructed tour through the internet.

Despite its extraordinary possibilities, interacting with 'new' electronic media does not *per se* guarantee easier understanding, i.e. an easier access to the world 'behind the screen' than when interacting with 'old' printed media. The user's situation is not *essentially* different from the familiar situation of the reader who is trying to understand printed text. Like these readers, who have no one to negotiate meaning with, we as users are on our own. When building fictional or non-fictional text, we may be invited by the medium to sail or to surf with the others, i.e. the community of authors, but when it comes to understanding, we are on our own and as lonely as the 'old' readers sitting in their libraries surrounded by thousands of books with no human interlocutors helping them to create their own inner worlds.

Notes

[*] I wish to thank Jenny Arendholz, Volker Eisenlauer and especially Christian Hoffmann for a number of valuable comments on an earlier version of this paper.

[1] To reach a complete characterization of human communication, as, e.g., upheld by Bühler (1965) and going back to Plato, we need to add that communication is '*about* something' and, of course, that it involves *language* as its *organon*.

[2] For a critique of the concept *hearer* in speech act theory and inference theory cf. Clark and Carlson (1982) and Levinson (1988).

[3] Cf. for a useful overview Levinson (1988); he presents and reviews a) the traditional account (which is based on the grammatical distinction between 1st, 2nd and 3rd person, and relies heavily on the criterion of whether a participant is present or absent), b) Shannon & Weaver's (1949) communication model, and c) Goffman's (1981) theory (cf. below) before introducing his own elaborate proposal (1988: 170 ff).

[4] For an overview cf. Levinson (1988: 169 ff).

[5] For *remediation* cf. Bolter (2001), and Eisenlauer and Hoffmann (submitted).

[6] Cf. Beißwenger (2005), Beißwenger (ed.) (2001), Hess-Lüttich and Wilde, and Arendholz (2006).

[7] Elsewhere (Bublitz 2006), I have juxtaposed and explained in more detail the 'collaborative' or 'cooperative' and the 'autonomous' views of comprehension, adopting Clark's (1992) terminology.

[8] I.e., that creating meaning is autopoietic (literally 'self-creating'), to put it differently. I take and adapt the term *autopoiesis* from the system theoretician Niklas Luhmann (1984), who defines a society as a social system of communication.

[9] In a wider sense, any *metalingual* use of language can be taken to be a means of negotiating meaning; for an overview of such use cf. Hübler and Bublitz (2007). The following examples from the *London Lund Corpus* have been adapted; the remaining conventions refer to intonation, pauses (. = brief pause, - = unit pause) and simultaneous talking (*… *).

[10] Mitchell, who uses the term "pictorial turn" and Boehm, who talks of "ikonische Wendung" (iconic turn) do, of course, take up Richard Rorty's (1967) famous topos of the *(linguistic) turn*.

[11] For an overview of hypertext definitions cf. Bublitz (2005a), Hoffmann (2006), Huber (2002), Jucker (2002), Kuhlen (1991), Storrer (2000), (2002).

12 Some authors do not count this criterion among the defining criteria of hypertext; as a consequence, they also apply the term hypertext to some kinds of printed media such as encyclopediae and handbooks, cf. e.g. Ansel Suter (1995), Bucher (1998), Kaplan (1995: 13).

13 Cf. Esser (2004), Kress and van Leeuwen (2001).

14 Also called "modularity", cf. Jucker (2003).

15 Nodes can be visual or aural, i.e. they can be read as text or seen as a visual image and even heard.

16 This is why the term 'e-cohesion' is an apt description for electronically mediated forms of cohesion in CMC.

17 That the user can literally (and not merely metaphorically) turn author is, of course, primarily due to the interactive nature of CMC; after all, hypertext has aptly been called "a medium for composition" and "not just […] a presentational device" (Slatin 1991a: 153).

18 Cf. for a similar account of *(to) see* both in a literal and a metaphorical sense (as a metaphor of understanding) Bublitz (2005b).

References

Ansel Suter, B. 1995. *Hyperlinguistics. Hypertext Lernumgebungen im Akademischen Kontext: Eine Fallstudie.* Diss. Zürich: University of Zürich.

Arendholz, J. 2006. *Kommunikative Unfälle in Chat-Gesprächen. Wie und warum Online-Kommunikation misslingen kann.* MA-thesis. Augsburg: University of Augsburg.

Beißwenger, M. 2005. 'Interaktionsmanagement in Chat und Diskurs', in: M. Beißwenger and A. Storrer (eds.) *Chat-Kommunikation in Beruf, Bildung und Medien: Konzepte - Werkzeuge - Anwendungsfelder.* Stuttgart: ibidem, 63-87. www.michael-beisswenger.de/biblio/interaktionsmanagement.pdf

Beißwenger, M. (ed.) 2001. *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation.* Stuttgart: ibidem.

Berners-Lee, T. *FAQ*. http://www.w3.org/People/Berners-Lee/FAQ.html

Boehm, G. 1994. 'Die Wiederkehr der Bilder', in: G. Boehm (ed.), *Was ist ein Bild?* München: Fink. 11-38.

Bolter, J. D. 2001. *Writing space. Computers, hypertext, and the remediation of print.* New Jersey: Mahwah.

Bublitz, W. 2005a. 'The user as 'cyberego': text, hypertext and coherence', in: L. Moessner and C. M. Schmidt (eds.), *Anglistentag 2004 Aachen, Proceedings*. Trier: WVT. 311-324.

Bublitz, W. 2005b. 'Seeing as a metaphor of understanding: the visible and the invisible', in: A. Schuth, K. Horner and J. J. Weber (eds.), *Life in language. Studies in honour of Wolfgang Kühlwein*. Trier, WVT. 135-149.

Bublitz, W. 2006. '*It utterly boggles the mind*: knowledge, common ground and coherence', in: H. Pishwa (ed.), *Cognitive aspects of language and memory*. Berlin: Mouton de Gruyter. 359-386.

Bucher, H.-J. 1998. 'Vom Textdesign zum Hypertext. Gedruckte und elektronische Zeitungen als nicht-lineare Medien', in: W. Holly (ed.), *Medium im Wandel*. Opladen: Westdeutscher Verlag. 63-102.

Bühler, K. 1965. *Sprachtheorie. Die Darstellungsfunktion der Sprache*. Stuttgart: Gustav Fischer. (First publ. in Jena, 1934).

Clark, H. H. and T. Carlson 1982. 'Hearers and speech acts', *Language* 58: 332-373.

Clark, H. H. and E. F. Schaefer 1992. 'Contributing to discourse', in: Clark: 144-175.

Clark, H. H. 1992. *Arenas of language use*. Chicago: UP.

Eisenlauer, V. and Chr. Hoffmann submitted, 'The metapragmatics of remediated text design', *Information Design Journal.* Amsterdam: Benjamins.

Esser, J. 2005. 'Hypertext and taxonomies of text-types', in: L. Moessner and C. M. Schmidt (eds.), *Anglistentag 2004 Aachen, Proceedings*. Trier: WVT. 297-309.

Fritz, G. 1999. 'Coherence in hypertext', in: W. Bublitz, U. Lenk and E. Ventola (eds.), *Coherence in spoken and written discourse: how to create it and how to describe it.* Amsterdam: Benjamins. 221-232.

Goffman, E. 1981. *Forms of talk.* Oxford: Blackwell.

Graesser, A. C. and L. F. Clark 1985. *Structures and procedures of implicit knowledge.* Norwood, NJ: Ablex.

Guyer, C. and M. Petry 1991. *Izme Pass*. Disk included in the magazine *Writing on the Edge*. Reprinted in Coover, Robert *The End of Books*. www.tnellen.com/ted/endofbooks.html.

Herring, S. C. 1999. 'Interactional coherence in CMC', *Journal of Computer-Mediated Communication*, 4(4). http://jcmc.indiana.edu/vol4/issue4/-herring.html:

Herring, S. C. (ed.) 1996. *Computer-mediated communication. Linguistic, social and cross-cultural perspectives*. Amsterdam: Benjamins.

Hess-Lüttich, E. W. B. and E. Wilde. Undated. 'Der Chat als Textsorte und/oder als Dialogsorte?' http://www.linguistik-online.de/13_01/hessLuettich-Wilde.pdf

Hoffmann, Chr. 2006. *Lost in hyperspace? Kohäsion und Kohärenz in Hypertexten.* MA-thesis. Augsburg: University of Augsburg.

Hoffmann, Chr. 2007. *The conceptual origins of hypertext.* (Unpubl. ms.) Augsburg: University of Augsburg, Dpt. of English Linguistics.

Huber, O. 2002. *Hyper-Text-Linguistik: TAH: ein textlinguistisches Analysemodell für Hypertexte.* Diss. München: University of München. www.huberoliver.de/index.htm

Hübler, A. and W. Bublitz. 2007. 'Introducing metapragmatics in use', in: W. Bublitz and A. Hübler (eds.). *Metapragmatics in use*. Amsterdam: Benjamins. 1-26.

Humboldt, W. von. 1827/^{3}1969. 'Ueber den Dualis', in: *Wilhelm von Humboldt. Werke in 5 Bänden.* Bd. III: *Schriften zur Sprachphilosophie.* Darmstadt: Wissenschaftliche Buchgesellschaft. 113-143. [First read in Berlin, Akademie der Wissenschaften, 26 April 1827].

Jones, R. H. 2004. 'The problem of context in computer-mediated communication', in: P. LeVine and R. Scollon (eds.), *Discourse and Technology. Multimodal Discourse Analysis*. Washington, D.C.: Georgetown UP. 20-33.

Jucker, A. H. 2002. 'Hypertextlinguistics: textuality and typology of hypertexts', in: A. Fischer et al. (eds.), *Text types and corpora. Studies in honour of Udo Fries*. Tübingen: Narr. 29-51.

Jucker, A. H. 2003. 'Mass media communication at the beginning of the twenty-first century: dimensions of change', *Journal of Historical Pragmatics* 4.1: 129-148.

Kaplan, N. 1995. 'E-literacies', *Computer-Mediated Communication Magazine* 2 (3), March 1.

Kress, G. and T. van Leeuwen. 1996. *Reading images. The grammar of visual design*. London: Routledge.

Kress, G. and T. van Leeuwen. 2001. *Multimodal discourse: the modes and media of contemporary communication*. London: Arnold.

Kuhlen, R. 1991. *Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank*. Berlin: Springer.

Lee, B. P. H. 2001. 'Mutual knowledge, background knowledge and shared beliefs: their roles in establishing common ground', *Journal of Pragmatics* 33: 21-44.

Levinson, S. C. 1988. 'Putting linguistics on a proper footing: explorations in Goffman's concepts of participation', in: P. Drew and A. Wootton (eds.), *Erving Goffman. Exploring the interaction order*. Boston: Northeastern UP. 161-227.

Luhmann, N. 1984. *Soziale Systeme. Grundriss einer allgemeinen Theorie*. Frankfurt am Main: Suhrkamp.

Mitchell, W.J.T. 1994. *Picture Theory: essays on verbal and visual representation*. Chicago: UP.

Rorty, R. 1967. *The linguistic turn: recent essays in philosophical method.* Chicago: UP.

Shannon, C.E and W. Weaver. 1949. *A mathematical model of communication*. Urbana, IL: University of Illinois Press.

Slatin, J. 1991a. 'Reading hypertext: order and coherence in a new medium', in: P. Delany and G. Landow (eds.), *Hypermedia and literary studies*. Cambridge, MA: MIT Press. 153-167.

Slatin, J. 1991a. 'Composing hypertext: a discussion for writing teachers', in: E. Berk and J. Devlin (eds.), *Hypertext / hypermedia handbook*. New York: Intertext Publication. 55-64.

Storrer, A. 2000. 'Was ist hyper am Hypertext?', in: W. Kallmeyer (ed.), *Sprache und neue Medien. Jahrbuch des Instituts für deutsche Sprache 1999.* Berlin: de Gruyter. http://www.hrz.uni-dortmund.de/~hytex/storrer/-papers/hyper.pdf

Storrer, A. 2002. 'Coherence in text and hypertext', *Document Design* 3: 156-168. (www.hytex.info).

Svartvik, J. and R. Quirk (eds.) 1980. *A corpus of English conversation*. Lund: Gleerup.

Tulving, E. 1972. 'Episodic and semantic memory', in: E. Tulving and W. Donaldson (eds.), *Organization of memory.* New York: Academic Press. 381-403.

Tulving, E. 1983. *Elements of episodic memory*. Oxford: OUP.

Linking the verbal and visual: new directions for corpus linguistics

Ronald Carter and Svenja Adolphs

School of English Studies, University of Nottingham

Abstract

This paper discusses an ongoing research project to investigate the compilation of a small corpus and the development of appropriate software tools that enable a more multi-modal approach to language data. The research draws on recent experience developed in the development of spoken corpora to explore alignments of the verbal and the visual and, as a starting point, does so with particular reference to gestures in communication and the role of head nods in particular. Issues of appropriate data capture and description are discussed alongside questions about the nature of language necessarily raised by language research that goes beyond the textual.

1. Introduction

Advances in the field of corpus linguistics over the past two decades have made it possible to develop computerised multi-million word databases of spoken and written language alongside powerful software tools to analyse this data quantitatively and qualitatively, a development that has contributed to pioneering research in many areas of communication studies and language description. However, while the analysis of large-scale text corpora can provide insights into language patterning and can help establish linguistic profiles of particular social contexts, it is limited to the textual dimension of communication. Communication processes are multi-modal in nature and there is now a distinct need for the development of corpora that enable the user to carry out analyses of both the speech and gestures of the participants in a conversation, and of how the verbal and non-verbal complement one another. In other words, corpus linguistics and discourse analysis might begin to be more closely aligned and descriptions made of rich contexts of language use of the kind advocated and illustrated by Michael Stubbs throughout his career.

1.1 Multi-Modal Communication

Recent work in multi-modal communication has seen advances in both theory and practice. The theoretical starting point for much significant work has been systemic-functional linguistics. Systemic linguistics is a theory that focuses on meaning, choice and probability in language and on the significance of language as a social phenomenon, underlining how particular choices of word, grammar and structure encode different meanings in different contexts of language in use.

Foundational work in multi-modal communication such as Kress and van Leeuwen (1996) has illustrated how choices of image can align with verbal choices and this work has been extended in recent years to embrace the multi-modal analyses of word, image and sound within different language varieties, including cartoons, comics, film, information leaflets, maps, advertisements (including TV advertisements), web pages and classroom textbooks (e.g. Baldry and Thibault, 2004, 2006). The emphasis has been on how choices of one image or camera angle or colour tone can cumulatively encode particular meanings. The almost exclusive focus has been on written text.

A particular challenge for current research is therefore to integrate the computer-enabled power of corpus linguistic methods, the theories and practices of multi-modal linguistic research and, with particular reference to the analysis of spoken discourse, the non-verbal signals of human gestures and bodily communication. In other words, one key aim is to provide computerised analyses of patterns of verbal and non-verbal meaning in ways that allow new understandings of textuality to emerge.

1.2 What is a Gesture?

Human communication functions within a variety of direct and indirect 'semiotic channels' (Brown, 1986: 409) which interact with, complement and 'counteract' each other (Maynard, 1987: 590). The occurrence of such channels is affected by modes of communication that differ widely according to their form, function and context-of-use (see foundational work by Argyle, 1969 and Ekman and Friesen 1969, 1976) and more recent studies by Wilcox 2004 and Gu, 2006). However, most studies have been undertaken within a research paradigm of psychology and in experimental rather than naturalistic conditions.

To date, experimental studies of the multi-modal nature of discourse have in general been designed to answer one or both of the following questions (Kendon, 1994: 177):

1 If recipients are offered utterances which include gestures and if they are permitted to see these gestures, do they interpret these utterances differently than when they are not permitted to see them? (examples of such studies include (Dobrogaev, 1929, reported in Kendon, 1980; Rogers, 1978; Riseborough, 1981).

2 If recipients are asked to make judgements about the gestures of others in the absence of speech to which they were related, do they make such judgements in a consistent way, and, if they do, do these judgements show that they have some understanding of the utterance of which they were a part?

Studies of gesture and the multi-modal nature of communication have focused upon gaze, (see Griffin, 2004 and Beattie & Shovelton, 1999, 2002) hand movements (see Rimé & Schiaratura, 1991 and Thompson & Massaro, 1986), head movements and other related gestures. In these studies the focus tends to be on language use in experimental conditions and does not embrace spontaneous, natural conversation. In addition, such studies tend to be more concerned with the gesture in relation to the basic content of talk, and do not explicitly explore the links between specific forms of language and accompanying gestures.

Current gesture detection and recognition systems developed in computer science within a tradition of automated vision recognition (see Nixon and Aguado, 2002; Kapoor and Pickard, 2001)) often focus on precise, intentional gestures. This is particularly true of hand gestures, where applications in sign-language recognition and human-computer interaction mean that specific gestures are made that are designed to be clearly distinguished by the observer. Gestures made in authentic, face-to-face conversation, by contrast, are much fuzzier, their form and meaning open to a greater degree of interpretation – a shake of the head can, for example, indicate disagreement, disbelief, or confusion, creating particular challenges for automated analysis of conversational gesture. Gestures are unlikely to be uniquely identifiable and interpretation will need to take into account other cues, such as the current role of the gesturer (speaker/listener) and the co-text of the conversation (i.e., what occurs before and after a sequence of gesture and talk). Furthermore, intentional gestures arise in a more constrained set of situations than conversational gestures. As a result, image sequences are usually acquired from a small, and known, set of viewpoints. Most intentional gesture recognition systems assume that a lone participant is in clear view, facing the camera from a short distance away. Many also assume the background to be uniform and fixed. Real conversational gestures arise in a wide variety of situations and involve dynamic activities from a variety of viewpoints and distances and include multiple participants, cluttered backgrounds and other moving people and objects.

However, for a corpus of gestures to be developed a record of the image is required and current computer technology provides one of the best available means of capturing such images digitally. The next sections report on a corpus-based project to investigate such a phenomenon with a focus on naturally occurring interactive two-party discourse.

2. Headtalk: an outline

HeadTalk is the first step in a project based at the University of Nottingham, involving interdisciplinary research between applied linguists and computer scientists, (in particular experts in vision recognition). The project aims to combine both linguistic expertise and new computational techniques and applications to provide the knowledge, research tools and procedures for exploring the behaviour of some salient gestures in naturalistic conversation. An initial focus on head nods was selected on account of their significance in communication.

The *Headtalk* project team has collected to date (January, 2007) five hours of video data, all based on face-to-face conversational episodes involving native English speaking academics and students based at the University of Nottingham.

The participants were filmed face-on, in close proximity to the cameras in order to create high quality, high resolution images, but were filmed in such a way as to minimise the interference and invasiveness of the recording equipment, to make the participants feel at ease and comfortable in the environment and to allow for (relatively) natural, authentic communication. This data can be properly described as 'multi-modal', as the transcribed recordings provide three different modes of discourse, offering three separate streams of data for analysis: the audio, the visual and the textual.

Utilising computer vision technology

The project utilises research in computer vision technology to allow the research team to detect, recognise and extract descriptions of head nod movements. For the detection and extraction of these movements, a variety of techniques were tested on significant samples of the data. After numerous evaluations, a head tracker was developed which can be placed upon the face of image data. Successions of movements can then be monitored and matched to the basic up-down sequence of a head nod in order to define where the movements occur, with the head tracker tracking movement in the videos. The headtracker allows multiple targets to be tracked in parallel, producing a description of the motion of each and showing intermediate results as they are obtained. (For further description of the tracker used (Cvision) see the Acknowledgements to this paper).

Developing linguistic categories

Head nods are vital for conversational maintenance and management (McClave, 2000) and often function as a form of 'back-channel' (Yngve, 1970), that is, a 'mechanism used for feedback' in discourse (Allwood et al, 1992), involving a strategy which involves a form of 'minimal response', a way for the listener to communicate that they have heard and perhaps understood a speaker's message, while allowing the speaker to continue talking. Although there has been research into and analysis of verbal back-channels, for example minimal responses or 'vocalisations' such as *mmm* and *yeah*, (Gardner (1998, 2002) integrated explorations of the verbal and visual components of head nod behaviour of this nature are limited. Preliminary linguistic analyses and classifications of each stream of data, (i.e. the transcribed text of the talk, as well as the audio and the video) was undertaken to determine patterns that may occur both within and across each data stream. The findings were then compared with the computational image analyses to define basic parameters for this particular gesture.

Coding back-channels

One of the key areas of concern of this project is how the head nods should be encoded. In terms of verbal realisations of back-channels most existing schemes focus upon grouping these in terms of their functions in discourse. This is a useful point of categorisation as every back-channel has a function in discourse, even if it may be unconscious to the interlocutor. Indeed, a wealth of research exists which agrees that 'back-channels have more than one macro function' (O'Keeffe and Adolphs forthcoming) as defined below (see also Schegloff, 1982; Maynard, 1987, 1990). As a guide to the key functions, the framework provided by O'Keeffe and Adolphs, has been adopted in the Headtalk project:

- **Continuers**: Maintaining the flow of discourse (see Schegloff, 1982)
- **Convergence tokens**: Marking agreement and disagreement
- **Engaged response tokens**: High level of engagement, with the participant responding on an affective level to the interlocutor.
- **Information receipt tokens**: Marking points of the conversation where adequate information has been received.

While this basic categorisation can be a useful starting point in analysing verbal realisations of back-channels, the question of how verbal and visual realisations interact within and across such categories has remained largely under-explored. For example, a back-channel such as *yeah* or *right* or *I see* or *mm* can be accompanied by a continuum of possibilities ranging from minimal head gesture to an emphatic nod of significant duration. And duration can also comprise several smaller nods within the same unit and still be linked to the same verbal token. Much depends on how an interlocutor is responding, whether he/she is simply maintaining contact or is signaling something altogether more engaged and involved. It is not just verbal form or duration that are significant but such factors as pitch and intensity govern how the form is interpreted and coded in relation to its verbal counterpart. The relationship is a complex and elusive one and a definitive coding scheme is still very much in development and will be extended beyond this phase of the project.

3. Methods

3.1 Record

For ease of transferability and consistency the data involved only native English language speakers taking part in 45-60 minute PhD supervision sessions. This meant factors such as intra and cross-cultural differences, which can potentially influence the way in which individuals gesture or signal feedback, were minimised.

For the recording of the video participants were required to face each other, with 4 cameras angled towards them and two microphones situated on the floor between them. These images are displayed in a split screen and have been positioned to ensure that they provide the highest quality images possible (see figures 1 and 2).

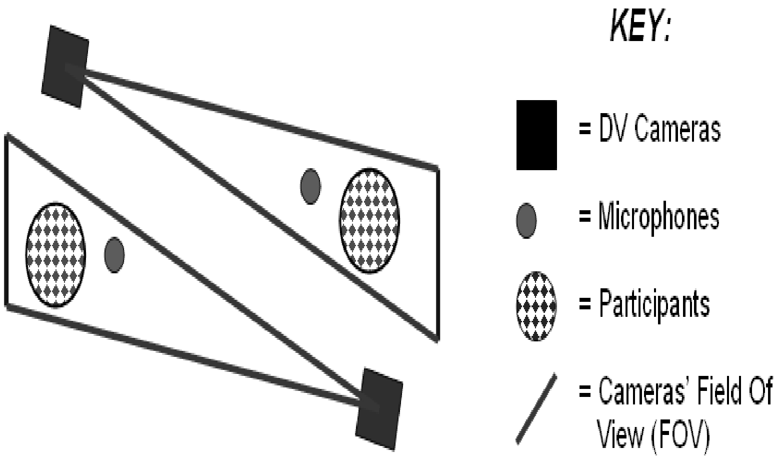


Figure 1: Setup for video recording

In order to keep the images as large as possible, the following screen split for capture was decided upon (figure 2):

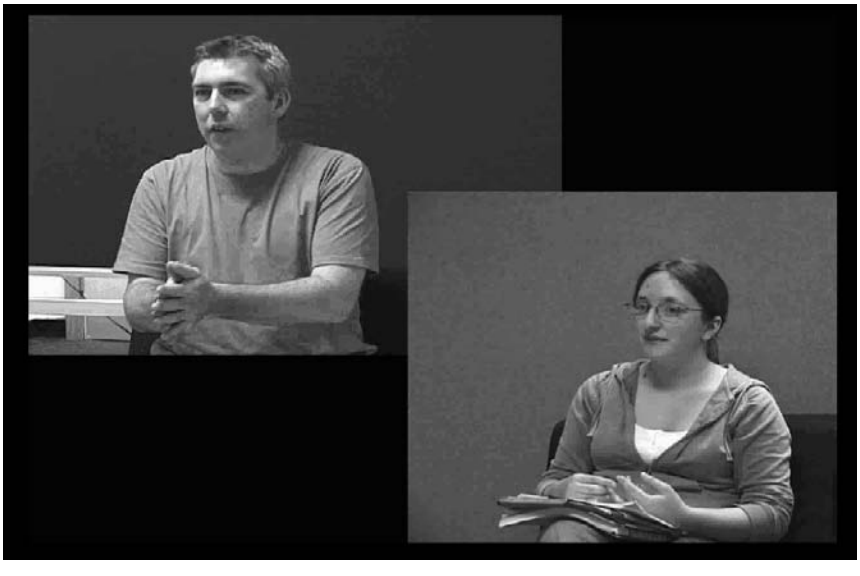


Figure 2: Screen shot example of data collected with the modified recording set-up.

Transcription of the data also created challenges as no universal, standardised transcription conventions exist for this type of data. Therefore, for continuity, the conventions used in the CANCODE (Cambridge and Nottingham Corpus of Discourse in English) corpus based at Nottingham University (http://www.-cambridge.org/elt/corpus/corpora_cancode.htm were reapplied here for the purely verbal component with Transana used to allow time stamping and annotation of synchronised video and audio data streams (http://www.transana.org/).

3.2 Coding

Following the data collection and rendering of video to display both participants as shown in figure 3, the next step is to develop a coding scheme for verbal and visual signals of active listenership in the data. The main aims of the development of the coding scheme are as follows:

- Defining and classifying verbal back-channel behaviour
- Defining and classifying non-verbal back-channel behaviour
- Combining verbal and non-verbal classifications and highlighting the potential for exploring patterns and relationships between the two

In relation to the coding phase methodological approaches for each of these processes needed to be closely explored. In order to create a new coding scheme, preliminary linguistic and gestural analyses and classifications of the data were undertaken. The findings were cross-compared in order to define basic parameters for gesture-in-talk for use as a corpus coding scheme.

The basic linguistic functions that were used in the analysis of the transcript are those outlined above (O'Keeffe and Adolphs, forthcoming). The analysis of head-nods, on the other hand, is based on classifications established with the use of computer-vision techniques. Five broad types of head-nods were identified in our training data:

Type A: small (low amplitude) nods with short duration
Type B: small (low amplitude), multiple nods with a longer duration than type 1
Type C: intense (high amplitude) nods with a short duration
Type D: intense and multiple nods with a longer duration than type 3
Type E: multiple nods, comprising of a combination of types 1 and 3, with a longer duration than types 1 and 3.

Using the functional categories, as well as the head nod classifications above, we carried out a preliminary analysis of a 10 minute stretch of video extracted from a longer MA supervision session. The participants in the session are a male supervisor and a female student, both of whom are British. The extract was taken from the middle section of the supervision, between 15.00 and 25.00 minutes. The data was transcribed and annotated (see below). The overall word-count of the transcript is 2156 words, of which 1401 words were uttered by the supervisor. For the purpose of illustrating the coding scheme we will focus here only on the description of back-channels used by the supervisor.

The supervisor uses 40 verbal back-channels in total, of which 18 are accompanied by a nod and 22 are purely verbal. In addition the supervisor uses 24 nods which are not accompanied by a verbal signal. Thus, the supervisor uses 42 head nods, 18 of which are accompanied by a verbal signal.

Focus on verbal back-channels

So far, linguistic research has focused mainly on the classification of verbal back-channels as outlined above. When we consider the discourse functions of the 40 verbalised back-channels used by speaker 1, the following breakdown emerges:
Continuer: 11
Convergence Token: 9
Information Receipt Token: 14
Engaged Response: 6

Focus on head-nods

In order to analyses the interface between verbal and visual, we have, as a second step classified the head-nods of the supervisor according to the different criteria (amplitude and duration) that led to the five head-nod types outlined above. Our analysis of the different types of head-nods used by the supervisor generates the following results:
Type A: 13
Type B: 13
Type C: 12
Type D: 2
Type E: 2

Integrating back-channel function and head-nods

We are particularly interested in this analysis to see whether any patterns emerge in those instances where a verbal back-channel is accompanied by a nod. This is the case in 18 of the back-channels used by the supervisor. In terms of linguistic functions and head-nod types the 18 instance are categorised as shown below:
Continuer: 4
Convergence Token: 4
Information Receipt Token: 3
Engaged Response: 7
And:
Type A: 6
Type B: 4
Type C: 6
Type D: 1
Type E: 1

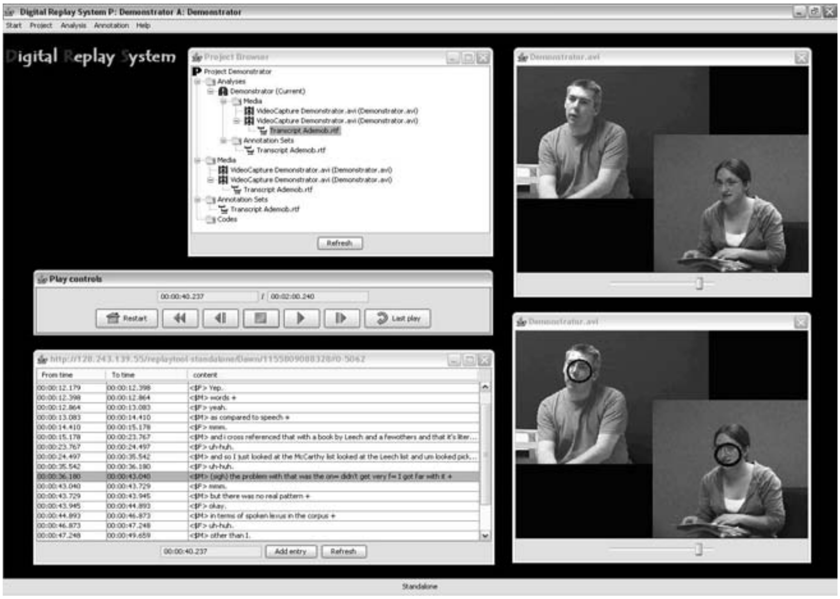
An analysis of back-channel functions as coded with the use of the linguistic coding scheme in relation to the type of nods that co-occur with the different functions highlights a number of interesting trends. Half of the small nods of short duration (type A) co-occurred with the information receipt function, while half of the small nods of longer duration (type B) co-occurred with the function of a convergence token. All of the type C nods (i.e. short and intense nods) used by the supervisor are accompanied by a verbal signal that has been classified as carrying either the continuer or convergence function. Overall, it is important to take a discourse level perspective to this kind of analysis, as preliminary inspection of the data suggests that some of the functions of head-nods can be aligned with the place at which they occur, i.e. where they are placed vis-à-vis the main speaker's utterance.

These are preliminary results and more data needs to be analysed to see whether there is any stable relationship between head-nods and linguistic signal. However, this very brief illustration of the different coding schemes highlights the need for an integrated analysis of verbal and visual, as the functions of back-channels are modified through the use of head-movement, and it remains to be established whether this modification is one of degree or of kind. One of the main challenges of multi-modal corpus analysis and representation is that corpus linguistics has traditionally focused on discrete items, such as individual words or grammatical categories. The complexities of gesture and movement, on the other hand, mean that they might not be able to be studied alongside traditional corpus linguistic units of analysis in a straightforward manner. Baldry and Thibault (2006: 181) point out that it is 'critically important [..] that corpus-based approaches to text engage with the level of discourse analysis and discourse-level meaning relations on various scalar levels of textual organisation'. While the integration of scalar levels and discrete categories is likely to cause problems in the development of an integrated framework, it also promises to lead to a much richer description of patterns in social interactions.

3.3 Coding the Data: An emerging replay tool

As we have seen, the primary challenge for developing support for analysis of multi-modal corpora is one of developing tools that provide an *integrated* approach to the representation of data. In general, there is a need to create tools that support the 'marking up' or identification of multi-modal patterns and the subsequent codification of recognizable patterns. Coding schemes for marking up textual records and verbal aspects of talk already proliferate. However, there is a paucity of such schemes for handling non-verbal elements: gestures, facial expressions, gaze, head and body movement, posture etc.

Existing tools do not generally support the extraction of linguistic patterns and thus fail, for example, to enable links between different types of listenership and accompanying head movements to be established. There is a need to develop new tools from the ground up to support linguistic analysis and, as an initial step towards this, and by means of developing concrete requirements for technical support, we have sought to exploit an emerging Digital Replay System (DRS) that has been developed to support ethnographic inquiry (Crabtree et al. 2005; French et al, 2006). The Digital Replay System provides some limited mechanisms of representation and below we consider both their potential and limits as a basis for articulating future requirements.


Digital Replay System P: Demonstrator A: Demonstrator
Start Project Analysis Annotation Help
igital eplay ystem
Project Browser
Project Demonstrator
Analyses
Demonstrator (Current)
Media
VideoCapture Demonstrator.avi (Demonstrator.avi)
VideoCapture Demonstrator.avi (Demonstrator.avi)
Transcript Ademob.rtf
Annotation Sets
Transcript Ademob.rtf
Media
VideoCapture Demonstrator.avi (Demonstrator.avi)
VideoCapture Demonstrator.avi (Demonstrator.avi)
Transcript Ademob.rtf
Annotation Sets
Transcript Ademob.rtf
Codes
Refresh
Play controls
00:00:40.237 / 00:02:00.240
Restart
Last play
http://128.243.139.55/replaytool-standalone/Dawn/1155809088328/0-5062
00:00:40.237
Add entry
Refresh
Demonstrator.avi
Demonstrator.avi
Standalone

From time	To time	content
00:00:12.179	00:00:12.398	<$F> Yep.
00:00:12.398	00:00:12.864	<$M> words +
00:00:12.864	00:00:13.083	<$F> yeah.
00:00:13.083	00:00:14.410	<$M> as compared to speech +
00:00:14.410	00:00:15.178	<$F> mmm.
00:00:15.178	00:00:23.767	<$M> and i cross referenced that with a book by Leech and a fewothers and that it's liter...
00:00:23.767	00:00:24.497	<$F> uh-huh.
00:00:24.497	00:00:35.542	<$M> and so I just looked at the McCarthy list looked at the Leech list and um looked pick...
00:00:35.542	00:00:36.180	<$F> uh-huh.
00:00:36.180	00:00:43.040	<$M> (sigh) the problem with that was the one didn't get very f= I got far with it +
00:00:43.040	00:00:43.729	<$F> mmm.
00:00:43.729	00:00:43.945	<$M> but there was no real pattern +
00:00:43.945	00:00:44.893	<$F> okay.
00:00:44.893	00:00:46.873	<$M> in terms of spoken lexus in the corpus +
00:00:46.873	00:00:47.248	<$F> uh-huh.
00:00:47.248	00:00:49.659	<$M> other than I.

Figure 3: Digital Replay System

The Digital Replay System allows video data to be imported and a digital record to be created that ties sequences of video to a transcribed text log, accompanied, where appropriate, by samples of data that are also tracked by the adopted C-vision recognition system (indicated in blue circles in Figure 3). The text log is linked by time to the video from which the transcript is derived so that the text log plays alongside the video. Further annotations can be added to the log to show where gestures – head nods in this case – occur and these annotations are also tied to the video. An index of annotations is produced and each can be used to go to that part of the log and video at which they occur. The annotation mechanism provides an initial means of marking up multi-modal data and of maintaining the coherence between spoken language and accompanying gestural elements.

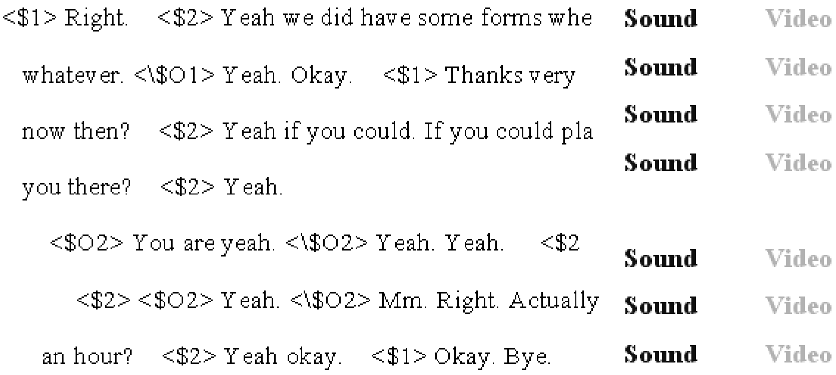
The data in the Digital Replay System is presented as a continuous, linear sequence of communication. Yet within any sequence a substantial number of utterances and gestures made by speaker and hearer overlap. This means that the representation of gestural patterns can appear somewhat disjointed, as there is no way at present to represent overlaps. The result of this is that it appears that head nods last only for a specific time and only occur between two verbalisations, which is inaccurate and misleading, as a nod may start after one verbalisation and continue for a long period into the next. Head nods are, in short, variable. They are not fixed in length and, given the limitations of the current incarnation of the Digital Replay System, are difficult to code as events occurring over time and not at particular moments in time. There is, then, a need to develop ways in which their occurrence across utterances can be time stamped and marked out. One way in which this might be achieved is to represent the utterances and gestural patterns made by parties to a conversation in individual transcripts, so to develop a 'textured text' consisting of separate layers. However, this is not necessarily an ideal solution since head nods need to be represented in relation to the behaviour of both the speaker *and* the listener. Here it is important to identify the defining parameters of the visual aspect of the particular gestural episode and align this with the verbal realisation that may or may not coincide with it.

While there is necessarily a level of interpretation in transcribing spoken discourse, the textual element of the head nod episode is relatively easy to establish and patterns of common back-channels can be extracted from existing multi-million word corpora. These types of patterns have been linked to particular functions in the area of corpus linguistics. For example, minimal verbalisations such as "mhm" have been linked to a continuer function while certain multi-word units such as "that's right" have been linked to an agreement function. Yet, the accompanying head movement, as well as the intonation pattern, can change the function of the back-channel realisation, which in turn will affect the surrounding discourse. It is therefore important to be able to establish some way of recognising the visual elements in a *principled way* so that these can be studied in relation to the verbal elements without adding a burdensome layer of interpretative intervention to the initial representation of the data. There is, thus, a need to marry vision recognition tools to machine learning techniques to reduce the overhead of interpretive work and have these tools and techniques work *across utterances* to adequately represent the verbal-visual character of multi-modal back-channels.

There is thus a need to *marry visual coding schemes to verbal coding schemes*, which may then be exploited by machine learning techniques to codify recognizable multi-modal patterns. In terms of using corpus linguistic techniques to analyse patterns in language it becomes even more important that recognizable patterns are consistently coded with reference to an agreed and replicable coding scheme. If this is not applied throughout the corpus, any searches for patterns will inevitably fail as they rely on the recurrence of consistent representations of linguistic phenomena. Furthermore, coding schemes need to be developed in such a way that that they can be shared across different research communities with different community cultures and different representational and analytical needs. In such circumstances, as analytical categories are re-classified in the light of new audio-visual evidence and as new insights from different research communities emerge, coding schemes need to be maintained as dynamically as possible.

3.4 Representation

The final concern of the corpus development is related to the way in which the multiple streams of coded data are physically re-presented. Current corpora utilise concordance tools. At the click of a button, appropriate citations of speaker information, context of use and evidence of the specific conversation in which each instance occurs, are easily available. With a multi-media corpus it is more difficult to exhibit all features of the talk simultaneously. If all characteristics of the specific instances where a word, phrase or coded gestures (in the video) occur are displayed, the corpus would involve multiple windows of data with, for example, 1000 instances of a head nod with an associated audio track of a *mmm* verbalised back-channel and the textual rendering of such (as seen in figure 4). This would make the corpus confusing and impractical.



<$1> Right. <$2> Yeah we did have some forms whe

whatever. <\$O1> Yeah. Okay. <$1> Thanks very

now then? <$2> Yeah if you could. If you could pla

you there? <$2> Yeah.

<$O2> You are yeah. <\$O2> Yeah. Yeah. <$2

<$2> <$O2> Yeah. <\$O2> Mm. Right. Actually

an hour? <$2> Yeah okay. <$1> Okay. Bye.

Figure 4: Representing concordances of multi-modal linguistic corpora

The basic solution to this problem is, as with current corpora, to present the data in a 'textured' way and integrate relevant information, such as the codes and further annotations, layering it behind main frames that display the key search features in a similar way to current textual concordances. This would involve marking up the transcript data with relevant information, such as codes for different gestures or indeed with information on the function of each gestures as well as corresponding speaker and time stamps, whilst linking it to other frames of information.

When, for example, the code ***+NOD+*** is selected in the corpus, the user will have access to the video and accompanying audio. Indeed such features may be relatively straightforward when just marking up single gestures (this has been the basic method used so far in our explorations of the supervision data), but, if one were to mark up additional features, this would become even more complex, especially when 'reading' concordances of multiple sources of data. So with searches of the visual and audio information it is difficult to 'read' multiple tracks of such data simultaneously, as is the case with current corpora and text. Our aim is to create a balance between the amount of texture in the corpus, i.e. the complexity and amount of information held in the corpus, and its ease of use. This is still very much under development.

4. Future Research Priorities

There are a number of lines of research arising from this project that require investigation in the future. These include technical issues such as the development of a recognition system to operate over the tracking data, and issues of scope, such as the analysis of other gestures and the analysis of coupled gestures and linguistic accompanying signals, such as hand movements performed in parallel with head gestures. *Headtalk* has allowed us to gain a better understanding of how we may describe and represent multi-modal language data but has also generated a set of additional pertinent research questions in the process. In addition to those outlined above, these also include theoretical questions of how gesture and language integrate and whether they can be described within a single framework. Major theoretical questions in this connection include consideration of the extent to which gestures may be said to be a language in the sense understood of language as a verbal medium. For example:

- Do gestures have rules and if so, how are the boundaries drawn?
- Do gestures have a syntax, that is, are they syntagmatically and/or paradigmatically organised. Or do they not conform to such structuring?
- If the relationship between language and image can be modally connected, as argued by theorists within a systemic linguistic tradition, and if images can be interpreted according to paradigms of choice, is the same true for gestures and for the relationship between human gestures and language?
- Is a system that is different to a linguistic system and are different underlying theories needed to account for the sheer multiplicity of different gestures?
- Many possible instantiations of headnods have been reported in this paper. What happens when researchers begin to try to explain the many possible meanings of hand gestures and their different cultural manifestations?
- What about 'body language' in the sense of movements encoded interactionally by proxemics?

Another important priority for future research in this area is the development of tools and methods to address ethical issues; for example, to anonymise video data while still being able to extract the salient features that are the focus of the analysis. Pixellating faces or using shadow representations of heads and bodies can blur distinctions between gestures and language forms and, when taken to its logical conclusion, anonymisation should also include replacing voices with voice-overs and with other speakers. Ethical considerations of re-using and sharing contextually-sensitive video data as part of a multi-modal corpus resource need to be addressed further in consultation with end users, informants, researchers and ethics advisors. The issues are especially acute when tools are shared or are developed to be web-enabled.

The *Headtalk* project complements core strands of work to be carried out by the e-Social Science Node at the University of Nottingham (see http://www.-ncess.ac.uk/research/sgp/headtalk/) As an extension to *HeadTalk*, the Digital

Record project, hosted in the e-Social Science Node (see http://www.ncess.-ac.uk/research/nodes/Digital/Record) allows for conversational gesture recognition and mark-up of a wider range of different gestures, from hand movements to gaze and facial expressions. This will enable researchers to start to 'complete the picture' of communication, to allow them to think about and explore communication from a variety of different perspectives, something for which *Headtalk* has endeavoured to provide the ground.

5. Conclusion

Natural language is a focus of a diverse range of disciplines and the continued explication of its real world, real time organization is of broad interest. The impetus towards multi-modal corpora recognizes that natural language is an embodied phenomenon and that a deeper understanding of the relationship between talk and bodily actions, particular gestures, is required if we are to develop more coherent understandings of the collaborative organization of communication (see also Saferstein, 2004).

Core requirements towards meeting this goal include the development of machine-based techniques that enable all visual and verbal patterns to be aligned and enable common multi-modal patterns to be recognized. There is also a pressing need to integrate visual and verbal coding schemes and to develop techniques whereby these analytic schemes can be exploited in machine learning environments to codify recognizable multi-modal patterns in large corpora of data. In order to achieve these developments we need to gain a better understanding of the particular requirements for recording, representing and replaying each of the different modes, and the research presented in this paper outlines some of the issues associated with this process.

The aim of this paper has been to begin to explore approaches that allow researchers simultaneously to review and analyse video, audio and textual records of naturally occurring communication. Such tools have the potential to provide a major resource for researchers in the field of applied linguistics and communication studies, film studies and drama in performance as well as in the field of face-to-face and remote human interaction. The development of research in this domain can also subsequently be extended to include pedagogic applications in the analysis of cross-cultural communication for modern foreign language learning as well as in professional discourse analysis, thus reinforcing the essentially interdisciplinary potential of applied research of which Michael Stubbs' work has been an exemplary instance.

Acknowledgements

The research on which this article is based is funded by the UK Economic and Social Research Council (ESRC), e-Social Science Research Node *DReSS* (www.ncess.ac.uk/nodes/digitalrecord), and the ESRC e-Social Science small grants project *HeadTalk* (Grant N$^{\text{o.}}$ RES-149-25-1016). Thanks are also due to Dawn Knight for providing copy for research reports which have in part at least formed the basis of this paper. All research reports so far are available via the National Centre for e-Social Science http://www.ncess.ac.uk/research/sgp/-headtalk/.

The HeadTalk demonstrator, Cvision, is an interactive program which allows users to apply the visual tracking algorithm developed within the project to selected targets in an input video clip. Cvision takes as its input an avi format video file and produces a text file giving the estimated position of each target in each frame of that video. This may be imported into the current version of the DreSS 2 tool DRS. An output video may also be produced, if desired. This shows the results of tracking overlaid on the input video images and is a useful debugging and interpretation tool. Cvision allows multiple targets to be tracked in parallel, producing a description of the motion of each and showing intermediate results as they are obtained. Cvision is written in C++ and provided as a Windows .exe file via http://www.ncess.ac.uk/research/sgp/headtalk/. User documentation is also provided, and incorporates full details of the algorithm employed.

References

Allwood, J., J. Nivre & E. Ahlsen 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics 9*: 1-26.

Argyle, M. 1969. *Social Interaction*. London: Methuen.

Baldry, A. and P. Thibault 2004. *Multimodal Transcription and Text Analysis* London: Equinox.

Baldry, A. and P. Thibault 2006. Multimodal corpus linguistics, in *System and Corpus* (eds Thompson and Hunston) pp. 164-183, London: Equinox.

Beattie, G. & H. Shovelton 2002. What properties of talk are associated with the generation of spontaneous iconic hand gestures? *British Journal of Social Psychology 41*, 403-417.

Beattie, G.W. & H. Shovelton 1999. Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica 123*, 1/2: 1 – 30.

Berger, K.W. & G.K. Popelka 1971. Extra-facial gestures in relation to speech-reading. *Journal of Communication Disorders 3*, 302 - 308.

Baldry, A. and P. Thibault 2004. *Multimodal Transcription and Text Analysis* London: Equinox.

Baldry, A. and P. Thibault 2006. "Multimodal corpus linguistics", in *System and Corpus* (eds Thompson and Hunston) pp.164-183, London: Equinox .

Brown, R. 1986. *Social Psychology* (2nd ed.). New York: Free Press.

Brundell, P. & D. Knight 2005. *Current Research and Tools to Support Data Intensive Analysis for Digital Records in e-Social Science.* Unpublished report. University of Nottingham.

Crabtree, A., A. French, C. Greenhalgh, S. Benford, K. Cheverst, D. Fitton, M. Rouncefield and C. Graham 2005. "Developing digital records", DReSS Working Paper 2, e-Social Science Research Node: University of Nottingham.

Ekman, P. & W. Friesen 1969. The repertoire of nonverbal behaviour: Categories, origins, usage and coding. *Semiotica 1*: 49-98.

Ekman, P. & W. Friesen 1976. Measuring facial movement. *Journal of Nonverbal Behaviour* 1, 1.

French, A., C. Greenhalgh, A. Crabtree, M. Wright, P. Brundell, A. Hampshire and T. Rodden 2006. Software Replay Tools for Time-based Social Science Data. Paper delivered at the *2nd annual international e-Social Science conference*, June 2006, University of Manchester.

Gardner, R. 1998. Between speaking and listening: The vocalisation of understandings. *Applied Linguistics, 19,* 204-224.

Gardner, R. 2002. *When Listeners talk: Response tokens and listener stance*. Amsterdam: John Benjamins.

Goldin-Meadow, S. 1999. The role of gesture in communication and thinking. *Trends in cognitive sciences 3*, 11: 419-429.

Griffin, Z.M. 2004. The eyes are right when the mouth is wrong. *Psychological Science 15*, 12: 814.

Gu, Y. 2006. Multimodal text analysis: A corpus linguistic approach to situated discourse. *Text and Talk 26*, 2: 127 – 167.

Kapoor, A. & R.W. Picard 2001. A real-time head nod and shake detector. *ACM International Conference Proceedings Series.* 1-5.

Kendon, A. 1980. Gesticulation and speech: Two aspects of the process of utterance. In M.R. Key (Ed.). *The Relationship of Verbal and Nonverbal communication.* The Hague: Mouton. pp. 207-227.

Kendon, A. 1994. Do gestures communicate? A review. *Research on Language and Social Interaction 27*, 3: 175-200.

Knight, D., S. Bayoumi, S. Mills, A. Crabtree, S. Adolphs, T. Pridmore & R. Carter 2006. Beyond the Text: Construction and Analysis of Multi-Modal Linguistic Corpora. Paper delivered at the *2nd annual international e-Social Science conference*, June 2006, University of Manchester.

Kress, G and T. van Leeuwen 1996. *Reading Images* Routledge, London.

Lock, A. (ed.) *Action, gesture and symbol: the emergence of language*, London: Academic Press.

Maynard, S.K. 1987. International functions of a nonverbal sign head movement in Japanese dyadic casual conversation. *Journal of Pragmatics 11*, 589-606.

Maynard, S.K. 1990. Conversation management in contrast: listener response in Japanese and American English. *Journal of Pragmatics 14*, 397-412.

McClave, E.Z. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics 37*, 7: 855-878.

Nixon, M. and A. Aguado 2002. *Feature Extraction and Image Processing*, Oxford: Elsevier.

O'Keeffe, A. & S. Adolphs forthcoming. Using a corpus to look at variational pragmatics: Response tokens in British and Irish discourse. in K.P. Schneider and A. Barron, ed., *Variational Pragmatics*. Amsterdam, Netherlands: John Benjamins.

Rimé, B. & L. Schiaratura 1991. Gesture and speech. In Feldman, R. & Rimé, B. (eds.). *Fundamentals of nonverbal behaviour*. Cambridge: Cambridge University Press. pp. 239-281.

Riseborough, M.G. 1981. Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication. *Journal of Nonverbal behaviour 5*, 172 - 183.

Rogers, W.T. 1978. The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research, 5,* 54-62.

Saferstein, B. 2004**.** Digital technology and methodological adaptation: Text on video as a resource for analytical reflexivity. *Journal of Applied Linguistics 1* (2): 197–223.

Schegloff, E. 1982. Discourse as interactional achievement: some uses of "uh huh" and other things that come between sentences. In D. Tannen (Ed.), *Analyzing discourse, text, and talk*. Washington, DC: Georgetown University Press. 71-93.

Thompson, L.A. & D.W. Massaro 1986. Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology, 42,* 144-168.

Wilcox, S. 2004. Language from gesture. *Behavioral and Brain Sciences 27*, 4: 525-526.

Yngve, V. 1970. On getting a word in edgewise. In *Papers from the 6th Regional Meeting, Chicago Linguistic Society.* Chicago: Chicago Linguistic Society.

The novel features of text. Corpus analysis and stylistics

Henry G. Widdowson

'These are only hints and guesses,
Hints followed by guesses.'
(T.S.Eliot: *The Dry Salvages*)

Abstract

This paper takes up the problematic stylistic issue that Michael Stubbs addresses in his study of Conrad's Heart of Darkness of the relationship between the analysis of a literary work and its interpretation. Inspired by his example, and applying his 'quantitative stylistic methods', I go in search of textual patterns and connections in the text of Conrad's novel other than those he has noted, and consider what possible significance they might have. The findings I come up with reveal features of the text that I would not otherwise have consciously noticed. Whether these simply serve as explicit confirmation of a subliminal literary awareness, or prompt new interpretative possibilities, is an open question. But there is no direct correlation between textual findings and literary effects. The precision of the analysis of the text does not lead to any greater precision in the interpretation of the novel, but on the contrary leads to a heightened recognition of the necessary variability and elusiveness of literary significance.

Of the many insightful enquiries that Michael Stubbs has undertaken into the significance of corpus analysis for our understanding of language, that which, for me at least, is the most intriguing concerns its contribution to literary stylistics. Stylistics claims to provide linguistic substantiation for the interpretation of literary texts. Since corpus analysis is *par excellence* a means of revealing textual features in precise detail, it seems reasonable to suppose that it must be relevant to the stylistic enterprise.

One advantage of corpus analysis by computer is that it can be so comprehensive in its coverage of the textual facts: it can yield a quantitative account of the recurrence and co-occurrence of all the words in a text. It is, however, precisely because it provides such detailed information that it brings into particular prominence the criticism that Stanley Fish levelled at stylistics in general, long before corpora and computers came on the scene (Fish 1973/1996). As Stubbs points out in a recent article (Stubbs 2005), Fish charges stylistics of being circular and arbitrary in that it presupposes relevance in advance. The analysis either selects literary features that are deemed to be significant and then adduces linguistic features to substantiate their significance, or it selects linguistic features and then claims that they are of literary significance. In the pre-corpus period, stylistics is particularly vulnerable to the first charge: generally speaking, what directed the selection of linguistic features was some impressionistic sense of literary significance. It worked from the literature to the language. With corpus analysis, however, we have the possibility of working in the other direction. Now that we have the linguistic facts of texts available to us in such comprehensive detail, we are in a position to make inferences from them about their literary significance. We can at least be certain about the linguistic facts. The problem of relevance, however, remains, and indeed becomes more difficult precisely because we have so much linguistic information to deal with. How do we decide what to select as significant?

This problem is both explicitly addressed and exemplified in the article already referred to, in which Stubbs applies methods of quantitative stylistics to Conrad's well-known short novel '*Heart of Darkness*'. This article is a fascinating exercise in corpus analysis which reveals textual facts which are likely to be unknown even to those readers who know the novel well. They certainly came as a revelation to me.

But the article is of particular interest because the analysis raises more general issues about text interpretation – about linguistic facts and literary significance, about the limits of analysis, about the Fish dilemma. What I intend in this contribution to this volume in honour of Michael Stubbs (henceforth MS) is to take up some of the points that he makes in his article in order to reflect on these wider issues.

The aim of his article is to apply a computer program to the text as corpus data and to demonstrate how the software can 'identify textual features which are of literary significance, including features which critics seem not to have noticed' (Stubbs 2005: 6). As I have already observed, it is unlikely that the literary critic will have noticed many textual features of the kind that computer software will reveal, and one can acknowledge that the value of corpus analysis is that it can provide textual substantiation to impressionistic interpretation. And indeed this particular analysis provides convincing linguistic evidence to support what literary critics have identified as the motif of dark indeterminacy that runs thematically through the novel. Thus, for example, MS points out that the computer reveals a high incidence of words denoting perceptual unclarity: *darkness*, *mist*, *shadow*, *gloom* and so on, and of expressions of vagueness: *seem*, *some*, *something*, and *like*. He notes that there is a repeated occurrence of adjectives with a negative prefix like *impossible*, *uneasy*, *unexpected*, *impenetrable* and so on.

Inspired by this kind of analysis, one finds oneself scrutinising wordlist and concordance for other findings which might be revealing. Use of the Wordsmith Tools software (Scott 1997, Scott & Tribble 2006), enables me to note, for example, that though these negatively prefixed adjectives occur frequently, adjectives generally seem to be in short supply in the text. Only two (*sombre*, and *black*) appear in the first 50 keywords (using BNC World as a reference corpus), and the most frequently occurring are simple, descriptively spare, monosyllabic (*long, great, black, white, old*). The description of river and forest is almost colourless (of the colours that one might expect to figure in such a description, *green* occurs only 5 times, *brown* only 4). One might conclude that it is a rather featureless world that Marlow describes, a monochrome world in black and white, a kind of abstraction.

All of these textual features can be said to substantiate the general impression that in Conrad's novel there is a pervasive presence of something essentially negative and indeterminate. The very texture of the style is, we might say, a representation of reality that can only be perceived, in the words of the apostle Paul, 'through a glass darkly'. The textual facts, then, can be adduced as evidence that 'This is a novel about the fallibility and distortions of human knowledge' (Stubbs 2005:12).

But, of course, only some textual facts are adduced as evidence, and their selection has been prompted by an impressionistic literary presumption that this is indeed what the novel is about. We have yet to contend with Fish. For the computer software will also reveal a whole host of textual features that the literary critic, or anybody else for that matter, would also fail to notice but which do not seem to be noteworthy. MS recognizes that textual features, however selected, 'still require a literary interpretation'. But then it cannot be the case that 'the software can identify textual features which are of literary significance'. This is because literary significance can only be assigned to *Heart of Darkness* as a **novel**, not as a **text**.

In this article, MS almost always refers to *Heart of Darkness* as a book or a text, hardly ever as a novel. But for the literary critic, of course, as for the normal reader, that is what it is. It is not just a book. Even less is it a text: a text is something you analyse, not something you read. MS, always admirably cautious in making claims for his analysis, acknowledges that 'textual frequency is not the same as salience, and does not necessarily correspond to what readers notice and remember in a text.' (Stubbs 2005:11) But the point is that readers do not process texts qua texts at all, and what they notice and remember are not textual features as such but their discursive realization in newspaper articles, manuals, leaflets, letters. And novels. The corpus analyst necessarily deals with *Heart of Darkness* as a text, a linguistic object. But the literary critic deals with it as a novel, a discourse, a particular genre of verbal art. So they are naturally inclined to notice different things: features of the text on the one hand, aspects of the novel on the other.

To return now to the point made earlier about the two possible directions of enquiry in stylistics, it seems obvious that we need to identify literary features first. In the present case, we need to consider not which linguistic features can be analysed out of the text, but which features seem to be significant in realizing different aspects of the novel. To take one simple example: the title. As part of the text, it will be included in the data to be analysed. But as a title, an aspect of the novel, it has a distinctive literary function which its textual features realize. There are two things one might note about it. First it has no determiner (*Heart,* not ***The*** *Heart of Darkness*), and second it is ambiguous ('heart consisting of darkness', cf *heart of gold*, vs 'at the heart, ie the centre of darkness'). What is the significance, if any, of these linguistic features? One might suggest that the theme of indeterminacy and uncertainty is already keyed in as a theme by the very title of the novel.

A text consists of words which combine with each other in various ways to form different kinds of linguistic pattern which the computer, of course, can identify. A novel consists of characters and events which combine in various ways to form narrative patterns which the computer cannot identify. But for the textual patterns to have literary significance, it has to be shown how they correspond or key in with the narrative patterns of the novel.

MS does talk about *Heart of Darkness* as a novel before proceeding to the main business of analysing it as a text by providing an account of its narrative structure. This, he says, is 'embedded in different frames' as follows:

1. The book starts with an unnamed narrator on a boat on the Thames
    2. Marlow becomes the narrator, and talks about the Thames in Roman times.
        3. Marlow tells of his visit to a European city
            4. Marlow tells the story which takes up most of the book….. .
        5. Marlow tells of his visit to Kurtz's fiancée back in the European city
    6. [There is nothing corresponding to frame 2, but some vocabulary from frame 2 is repeated in frame 7]
7. The book ends with a paragraph from the unnamed narrator back on the Thames.

(Stubbs 2005: 8)

Whatever criteria are used for identifying these frames, they are apparently not textual. Indeed the only textual feature that is mentioned here, vocabulary, is explicitly excluded. For if it were a factor then repetition of vocabulary would presumably give some textual grounds for putting frame 7 in correspondence with frame 2. Interestingly, later in his analysis, MS does, however, suggest that this framework is marked by textual features, pointing out that the phrase 'waterway leading to the uttermost ends of the earth' is repeated in the first and last frames. He also mentions that the phrase 'the pose of a….. Buddha' that occurs in the last frame is also a repetition, but in this case the first occurrence appears not in the first frame but the second, after Marlow has already assumed the role of the narrator. Such textual features, then, do not seem to be a reliable indication of the narrative structure.

I shall return to these verbal repetitions presently. For the moment, the point I want to make is that the framework that is proposed is an analysis of narrative - an aspect of the novel, not of the text. It can only be based therefore on a literary view on what is significant. For MS, who does the narrating is one significant factor: when Marlow takes over the narration we shift into a different frame. If only this factor were to be considered, then the novel would consist only of frames 1 and 7 with Marlow's story in the middle. But a quite different factor is introduced to distinguish the frames in between, namely the shift in the setting of Marlow's story. This gives us frames 3 and 5 in the European city with Africa in between. So in fact we have two separate kinds of narrative framework, and the attempt to fuse them into one leads to the postulation of the non-existent frame 6, which imposes a symmetry on the novel that appears to have no warrant in textual structure.

The question arises as to why these two aspects of the novel, narrator and setting, should be taken as the only significant determinants of narrative structure. One is led to wonder whether, if this is indeed the only structure that can be discerned, it is, perhaps, not in itself of much significance anyway. Certainly the analysis does not reveal what the significance might be: it concentrates on the content of frame 4, which covers a good 75% of the text, which, on this account, has no narrative structure worth mentioning at all.

In the case of the narrative framework, we have an aspect of the novel described without substantiation from textual features that a computer analysis might provide. But there are other kinds of narrative pattern that **are** shown to be textually realized, and here we return to the repeated phrases cited earlier. In his discussion of the way certain words are distributed in the text, MS indicates a number of other instances of intra-textual repetition, where the words repeated are not necessarily frequent in the text at all. Thus, for example, he points out the description of the city that Marlow returns to is in several ways a lexical reprise of its first description, and again that the words *voice* and *idol* are used in reference to both Marlow and Kurtz. Such facts 'start to say something about the structure of the whole text' (Stubbs 2005:12). Just what that something might be is left to the reader to ponder on. And in my case, these observations provoked a good deal of pondering. Though it is not clear to me what such facts tell us about the structure of the text, they set in train all manner of speculative reflection about the possible literary significance of intra-textual repetition of this kind.

So, with Wordsmith Tools at the ready, I set off in quest of other instances of such repetition. MS provides us with the example of recurring words that provide a textual link between the two descriptions of the city: 'high houses', 'narrow and deserted street', 'doors ponderously ajar' in the first description, 'a ponderous door', 'between tall houses' in the second. Further enquiry about the distribution of these words reveals that they also figure in the description of the jungle. A stretch of the river is described as 'narrow, straight, with high sides', there are 'high walls' of trees, the 'high stillness of primeval forest'. The word *deserted* only occurs three times in the text, and its other two occurrences are in descriptions of the African scene:

> And the village was deserted, the huts gaped black, rotting…
> the waterway ran on, deserted, into the gloom…

Directing my computer in quest of other repetitions, I discovered that the word *ponderous* occurs only twice in the text. Though its second occurrence in 'a ponderous door' may echo 'doors standing ponderously ajar' and so serve to link Marlow's two city visits, it also echoes its first occurrence where it appears in a description of the progress of Marlow's boat upriver 'between the high walls of our winding way, reverberating in hollow claps the ponderous beat of the stern-wheel'.

What is one to make of these inter-textual lexical connections? What do they hint at? What literary significance can we guess they might have?

If, as MS suggests, the use of *idol* and *voice* to describe both Marlow and Kurtz indicates their similarity, then the use of these other examples of repetition can presumably be said to have the same associative effect. If things that are described in the same terms take on a similarity, then just as Marlow assumes the likeness of Kurtz, so the city assumes the likeness of the river and takes on its darkness. This, we may suggest, is supported by other distributional facts that the computer reveals. As MS points out, 'the words *heart*, *dark* and *darkness* occur throughout the book, but increase in frequency at the very end'. This is the textual hint. What possible literary significance might be assigned to it? On, then, to the guesswork. And for this we need to set the computer aside for a moment, and look at the text for ourselves and consider its novel effects.

The end of the book is where Marlow visits Kurtz's fiancée – the 'Intended' - and he takes both the darkness and Kurtz's last whisper with him. At her door, as 'the dusk was falling', he 'seemed to hear the whispered cry, 'The Horror! The Horror!' She comes to meet him in the darkening room, 'dressed in black', with 'a pale visage', 'dark eyes'. On her appearance, 'The room seemed to have grown darker'. And as she speaks of the noble qualities of Kurtz, 'with every word spoken the room was growing darker'. Marlow listened. 'The darkness deepened'. As she talks, Marlow seems again to hear 'the whisper of a voice speaking from beyond the threshold of an eternal darkness'. The repeated words are like a sound track, a lexical leitmotif, that brings the room and the people in it into association with the African river that has been described in the same terms. And this is then confirmed by a quite explicit connection: a gesture of the Intended reminds Marlow of another woman – the 'wild and gorgeous apparition of a woman' who appears so dramatically and ominously out of the jungle earlier in the narrative:

> '…I shall see her, too, a tragic and familiar Shade, resembling in this gesture another one, tragic also, and bedecked with powerless charms, stretching bare brown arms over the glitter of the infernal stream, the stream of darkness.'

The two women become one and the realities of primitive savagery and apparent civilisation are fused by the presence of a common darkness.

This is a darkness of deception and delusion as well. As MS points out, the word *know* is, like *dark* and *darkness*, also evenly distributed through the text, often in negative form, but there is a cluster of positive instances at the end, mainly spoken by Kurtz's Intended. She knows that Kurtz was good and noble, and her belief is described as

> 'a great and saving illusion that shone with an unearthly glow in the darkness, in the triumphant darkness.'

And here we come to what (to me at least) is the thematic climax of the novel. Kurtz's last whispered words that have so haunted Marlow come back again:

> 'The dusk was repeating them in a persistent whisper all around us, in a whisper that seemed to swell menacingly like the first whisper of a rising wind. "The horror! The horror".'

The threefold repetition of the word *whisper*, insistently evoking the dark reality of Kurtz's world, is then immediately followed by another threefold repetition, and one that is an emphatic assertion of the counter-reality of the Intended's belief:

> 'Don't you understand I loved him – I loved him – I loved him'.

For a moment, this reality prevails, to such an extent that Marlow is drawn into it himself and tells his lie to sustain it:

> 'The last word he pronounced was – your name'.

Her reaction is first to give a light sigh, and then in the tense and darkened room she makes Marlow's heart stand still with 'an exulting and terrible cry, a cry of inconceivable triumph and unspeakable pain'. This cry is a dramatic and climactic moment in the novel. But it is also a textual echo. Where has the reader heard a cry before? This, of course, is where the computer comes in. It reveals that the word occurs 6 other times in the text. The first two occur in the phrase: 'a cry, a very loud cry, as of infinite desolation', which bears some formal resemblance to the Intended's cry, and breaks the stillness in a similarly sudden and startling way – but this time in the heart of darkness itself. As indeed does the third occurrence: 'a cry arose whose shrillness pierced the still air.' The other occurrences of the word relate to Kurtz – his 'cry that was no more than a breath', his 'whispered cry'. But the cry of the Intended is not a whispered but 'exulting', not one 'of infinite desolation' but 'of inconceivable triumph'. And here there is another echo, surely.

Back to the computer. And I find that the word *triumph* occurs only twice in the book. Its only other occurrence appears just before Marlow arrives at the Intended's door, when he describes the death of Kurtz in terms of 'a conquering darkness,' and as 'a moment of triumph for the wilderness'. It is not now the darkness that is triumphant, and in contrast to all the vague and menacing indeterminacy that pervades the book, we have a straightforward assertion of absolute certainty, which Marlow repeats.

> 'I knew it - I was sure!' She knew. She was sure.

Just how Marlow is supposed to actually say these words we cannot know – in a tone of irony, incredulousness? But they serve to confirm this other reality which he cannot deny. He cannot tell her the truth – to do so would have been to condemn her to the other darker world – 'It would have been too dark – too dark altogether.'

'..too dark altogether.' Here I am teased by another verbal echo, faint but persistent. What does this phrase remind me of? I check the concordance for *altogether*, and there it is: 'too dark altogether', 'too beautiful altogether'. And this latter phrase takes me to an earlier scene in the book: Marlow's visit to another woman – his 'excellent aunt'. This aunt is a minor figure, and as far as her role in the story is concerned, seemingly superfluous. Why then is she there at all? Marlow describes the visit:

> I found her triumphant. I had a cup of tea – the last decent cup of tea for many days – and in a room that most soothingly looked just as you would expect a lady's drawing-room to look, we had a quiet chat by the fireside.

Another lady's drawing room, but triumph here is associated with domestic normality – the cup of tea, the chat by the fireside. But it turns out that the aunt has the same kind of idealistic vision as does the Intended, and thinks of Marlow in the same way as the Intended thinks of Kurtz – as a kind of 'emissary' or 'apostle' with a mission, 'weaning those ignorant millions from their horrid ways'. Marlow comments:

> 'It's queer how out of touch with truth women are. They live in a world of their own, and there has never been anything like it, and never can be.'

And he adds:

> It is too beautiful altogether.

He expresses the same sentiment later, and in similar words, when, in his narrative, he anticipates the lie he will tell to the Intended:

> They – the women I mean – are out of it – should be out of it. We must help them to stay in that beautiful world of their own, lest ours gets worse. Oh, she had to be out of it.

The aunt and the Intended are thus brought into association, both inhabitants of an illusory world of conventional ideals that has to be sustained by deception. A reality 'too beautiful altogether' in contrast with the other reality which is 'too dark altogether' – the darkness of some pervasive moral corruption that Marlowe senses but cannot clearly discern, and which, as MS notes is reflected in the vagueness of his language.

But there are times when his language is not vague at all, and here we come to another aspect of the book as novel which the analysis of the text does not itself reveal. The intra-textual patterns I have been tracing can be taken as being indicative of the underlying theme of *Heart of Darkness*, and to lend support to MS view that: 'This is a novel about the fallibility and distortions of human knowledge'. What a novel is about is something, it would seem, that a quantitative analysis of text is particularly well suited to identify: its theme is reflected by the frequencies of linguistic features and their distribution, as MS demonstrates so convincingly. But there is, of course, more to a novel than what it is about. Its theme only becomes significant by the manner of its representation, by the way it is activated by events and characters. What seems to me most striking about the intra-textual patterns that I have noted is the way they give dramatic expression to the theme in the representation of the women characters and the events in which they figure.

The aunt and the Intended are associated in that they have in common the same idealistic view of the world, the same reality of conventional values. But the contexts of their appearance are in striking contrast: one a cheery drawing room, a quiet chat by the fireside, an atmosphere of relaxed normality, the other a sombre and sepulchral room, the atmosphere charged with intense feeling, and dialogue as different from a casual chat as it is possible to imagine. Neither woman is described in any detail. In fact the aunt is not described at all. She is 'excellent' and that's all. We have no indication about what she looks like. The Intended is hardly less sparely described – just one or two simple monosyllabic adjectives: 'fair hair' 'pale visage,' 'dark eyes'. Why this absence of descriptive detail, one might wonder. In the case of the aunt, one might suggest that since she is a minor character no description is called for, but then other minor characters are described in some detail. One of the women in the office where Marlow goes to get his job, for example, is described in very particular terms – warts and all indeed:

> Her flat cloth slippers were propped up on a foot-warmer, and a cat reposed on her lap. She wore a starched white affair on her head, had a wart on one cheek, and silver-rimmed spectacles hung on the tip of her nose.

The aunt and the Intended are by comparison featureless. As such, they seem to function more as thematic symbols than as individual characters: the aunt as representing conventional normality, defined by the typicality of her drawing room, and the Intended as representing too the darkness of delusion and deception in which she is embroiled.

But there are other figures that do not blend in with the thematic background, but are starkly foregrounded against it. The most striking instance of this is 'the wild and gorgeous apparition of a woman' that, as pointed out earlier, Marlow explicitly associates with the Intended. This is how the woman is described:

> She walked with measured steps, draped in striped and fringed cloths, treading the earth proudly, with a slight jingle and flash of barbarous ornaments. She carried her head high; her hair was done in the shape of a helmet; she had brass leggings to the knee, brass wire gauntlets to the elbow, a crimson spot on her tawny cheek, innumerable necklaces of glass beads on her neck; bizarre things, charms, gifts of witch-men, that hung about her, glittered and trembled at every step.

There is nothing vague or indistinct about this description. The vivid impression the woman's appearance makes on Marlow is etched on his mind in exact verbal detail, and the description stands out as particular and precise because the words in it are infrequent. It is this that makes the woman stand out against 'the gloomy border of the forest' 'the immense wilderness'. But at the same time she is also 'like the wilderness itself, with an air of brooding over an inscrutable purpose'. The clarity of the perception registered in this descriptive precision contrasts with the vaguely expressed sense of strangeness and foreboding and accentuates it. As MS notes, much of the language of the text, the recurrence of words to do with darkness, uncertainty, negation and so on reflects the underlying theme of the novel. In a way they serve as a backdrop, a mise-en-scene. But it is the events and characters, figures against this ground, that activate the theme and give it dramatic force, create the literary significance that make the text into a novel. And these are often described in language that a quantitative analysis of the text would not register as remarkable.

The observations made by MS in his article, particularly those that point out the recurrence of certain words and phrases, often themselves infrequent, in different parts of the text, have set me off looking for similar intra-textual links, and using them as hints to possible literary significance. Hints followed by guesses. And one bit of literary guesswork has set me in quest of other hints – in frequency lists, in concordances – looking for possible bits of textual evidence to support a particular literary interpretation. Hints and guesses. It is a fascinating exercise.

But not one that Stanley Fish would be likely to approve of. For it is, of course, open to the charge of circularity. This does not, however, make it invalid as a process of exploring significance which, prompted by the MS analysis, I have been pursuing here. On the contrary, circularity of a kind, is an essential feature of this process, for, to quote T.S. Eliot again:

> …the end of all our exploring
> Will be to arrive where we started
> And know the place for the first time.
>
> (*Little Gidding*)

The Fish objection only applies to the positivist claim that stylistics establishes a correlation between linguistic features and literary effects so that one can be read off from the other. But this is not the claim that stylisticians generally make. What they are principally concerned with is not correlations but correspondences, with ways in which textual features can be adduced to give warrant to different literary interpretations, not to ratify one of them as definitive. Stylistic analysis does not seek to foreclose on a particular interpretation but to open up alternative possibilities. It does not claim to discover meanings which are inscribed in a text and which may have eluded literary critics, but to provide the means for exploring one's own reactions to the text. Herein lies its educational value – for it offers an alternative to the traditional teaching of literature. Rather than being the passive recipients of the second hand interpretations of literary critics, students can be enabled (empowered even) to take the initiative and engage actively and directly with literary texts themselves. (Widdowson 1975,1992)

> Stylistics, then, is all about hints and guesses. As Verdonk puts it:
> '..it can serve not only to substantiate an impressionistic sense of meaning, but also to suggest the possibilities of reading other interpretations into a text'
>
> (Verdonk 2002: 78)

In his review of that book, MS calls this a 'weak' defence of stylistics and takes its author to task for not rising to the Fish challenge to 'defend a stronger position' (Stubbs 2004: 129). In this article on *Heart of Darkness* MS does rise to the Fish challenge, but interestingly does so by in effect arguing for the validity of the so-called 'weak' position himself, justifying his stylistic analysis, very much along Verdonk lines, by concluding that:

> '..observational data can provide more systematic evidence for unavoidable subjective interpretation'. (Stubbs 2005: 22)

As a text *Heart of Darkness* consists of observational data that can be analysed by computer. As a novel, however, it can only be subjectively interpreted. This means that what counts as evidence for interpretation can never be objectively determined, and any claim that it can (the 'strong' position) is mistaken. Hints and guesses are all we can reasonably expect. But the point about the computer is that it can provide so many hints for us to guess the significance of. This is what makes Michael Stubbs' article so stimulating – his own textual findings set the reader off in quest of others.

What, for me at least, is revealing about its 'quantitative stylistic methods', is that the results of the analysis are so different from its effects: the very precision of the findings provoke very imprecise speculation about their significance. The more you pin down and quantify features of the text, the more aware you become that features of the novel cannot be pinned down and quantified. They remain elusive, subjective, and variable, and cannot be reduced to textual terms. But, as MS says, it is not the claim of stylistics that they can be, or should be. His kind of analysis does not tell us what *Heart of Darkness* means, but what it might mean to different readers. And herein lies its value, and particularly its educational value: it demonstrates ways in which textual features can be explored, and how such exploration can open up possibilities of novel interpretation.

References

Fish, S. 1973. 'What is stylistics and why are they saying such terrible things about it?' In S. Chatman (ed.) *Approaches to Poetics*. Columbia University Press. Reprinted in Weber, J.J. 1996. *The Stylistics Reader*. London: Arnold. 94-116.

Scott, M. 1997. *WordSmith Tools* (Software). Oxford: Oxford University Press.

Scott, M. & C .Tribble. 2006. *Textual Patterns*. Amsterdam: John Benjamins.

Stubbs, M. 2004. Review of Verdonk, 2002. In *Applied Linguistics* Vol 25/1 126-129

Stubbs, M. 2005. 'Conrad in the computer: examples of quantitative stylistic methods'. In *Language & Literature*. Vol 14/1. 5-24.

Verdonk, P. 2002. *Stylistics*. Oxford Introductions to Language Study. Oxford:Oxford University Press.

Widdowson, H.G. 1975. *Stylistics and the teaching of literature*. London: Longman.

Widdowson, H.G. 1992. *Practical Stylistics*. Oxford: Oxford University Press.

Hocus pocus or God's truth: the dual identity of Michael Stubbs

Guy Cook

Open University

Abstract

Text analysis is in an anomalous position: hovering on the borders between the sciences on the one hand, and the arts on the other. As science it seeks to be descriptive rather than prescriptive, replicable by other analysts, expounding objective facts about language use. As an art it evaluates and prescribes, imposing the writer's views upon the external world, saying as much about the analyst as the analysed. This chapter explores the position of Michael Stubbs in relation to this dichotomy, suggesting that, while he advocates objectivity, and has made an outstanding contribution to linguistic description, his achievement - almost despite himself - is also to be an evaluator and interpreter. Like a good literary critic, he is worth reading not only for what he tells us about the external world (which is a great deal) but also for his own unique ideas.

> "Pure induction will never get you from empirical observations to interesting generalizations. You have to know where to look for interesting things. As Grice (1958:173) puts it: 'you cannot ask [....] what something is unless (in a sense) you already know what it is'. However, this is true only 'in a sense', since the aim is to say systematically and explicitly what something is: and that is where empirical, observational analysis can contribute. It is not possible (or desirable) to avoid subjectivity, but observational data can provide more systematic evidence for unavoidable subjective interpretation." (Stubbs 2005)

1. Hocus pocus or God's truth

In the early 1950s, when Michael Stubbs was just starting school in Glasgow, the lexicographer and semanticist Fred Householder, reviewed a book called 'Methods in Structural Linguistics' by Zellig Harris and evoked a distinction between two positions in linguistics.

> "On the metaphysics of linguistics there are two extreme positions, which may be termed (and have been) the 'God's truth' position and the 'hocus pocus' position. The theory of the God's truth linguists [...] is that language 'has' a structure and the job of the linguist is (a) to find out what the structure is, and (b) to describe it [...]. The hocus pocus linguist believes that a language (better, a corpus, since we describe

> only the corpus we know) is a mass of incoherent formless data, and the job of the linguist is somehow to arrange and organize this mass, imposing on it some structure [...]." (Householder 1952)

Householder describes both positions as extremes. He assumed perhaps what we would now call a negative discourse prosody for both phrases,[1] and that nobody would want to be identified as either. He criticises Harris for being too much of a "God's truth" linguist, but implies that a "hocus pocus" position would be just as flawed. The good linguist should be somewhere between the two poles and not at either end.

Householder was thinking of the division of his time between armchair structuralist linguists using examples drawn by intuition from their own minds, and empirical anthropological linguists going out and studying language behaviour. Given how linguistics has developed since, there are quite a few complications in applying Householder's distinction to linguists today. For one thing, there has been a revolution in corpus linguistics, in which that Glasgow schoolboy went on to play a leading part. The modern linguist is no longer limited to "only the corpus [the linguist] know[s]", but has access to millions of words beyond their own immediate experience. Nor are their corpora "a mass of incoherent formless data"; linguists like Michael Stubbs seek and find in them patterns and connections undreamed of in Householder's time. So it is now the corpus linguist who believes "that language 'has' a structure and the job of the linguist is (a) to find out what the structure is, and (b) to describe it".

Stubbs himself has paid attention to this dichotomy, though in different terms. Considering the ideas of Saussure, he writes:

> "In a famous and influential statement, Saussure (1916) argued that 'far from the object of study preceding from the point of view, it is rather the point of view that creates the object'. Due to advances in technology, new observational methods have made it possible to collect new types of data and to study patterns which had previously been invisible, but the point of view does not create the patterns. What we see certainly varies according to point of view, and it follows that any view is partial, but it does not follow that what we observe has been created by the point of view or by the observational tools." (Stubbs 2002a:220)

Saussure in other words was too hocus pocus, and has moreover been overtaken by events. But Saussure's work, though contested and discussed, is far from dismissed by Stubbs. It is a recurrent point of reference in his work. Saussure's ideas, my intuition suggests, are ones with which he has a love/hate relationship.

In this chapter, I shall use this distinction between 'God's truth' and 'hocus pocus' liberally and with poetic licence, interpreting it beyond the context of the time it was written to mean simply that there are two opposite tendencies in linguistics. In both, language is seen as ordered, but in the first case (God's truth) the order is an objective one, out there to be discovered, while in the second case (hocus pocus) it is something subjectively imposed upon the language by the analyst. Like Householder though I shall treat this as an idealisation, and charitably assume that no-one is so foolish as to be actually at one extreme or the other. We are all somewhere on a line between the two poles, pulled now in one direction, now in the other.

The distinction of God's truth and hocus pocus is a light-hearted echo of the many "unavoidable dualisms" (Stubbs 2000) "which trouble linguistics" and "other disciplines"

> "dualisms of subject and object, internal and external, agency and structure, process and product, *parole* and *langue*, language use and language system, pragmatics and semantics, communication and language, creativity and rules, intended action and unintended consequences" (Stubbs 1996: 56-57).

These are all dichotomies with which he has resolutely engaged, even agonised over. What I want to do in this chapter is assess where the work of Michael Stubbs belongs in relation to some of these dualisms, aware that in some hands, the weapon of corpus linguistics is indeed wielded as though it were God's truth, while for some of its critics it comes close to being hocus pocus: distorting the living actuality of language by freezing and dissecting it. And my conclusion will be that Stubbs is not where he appears or claims to be on that line! "What we see" as Stubbs writes of Saussure "certainly varies according to point of view, and it follows that any view is partial". In the partial view to be presented here, I shall suggest that on the continuum from hocus-pocus to God's truth, Stubbs is, like a quantum particle, actually in two places at once, depending on the observer.

I shall acknowledge the influence of Stubbs' ideas, but seek to develop two points. The first is the inevitable role of evaluation in language analysis. The second is the issue of where exactly "patterns which had previously been invisible" are. Are they in the mind or only in the corpus? If they are in the mind, are they conscious or subconscious, and how might we access them? If they are both in the mind and in the corpus, how might we relate the two?

2. Arts or Sciences

Let us begin with a related dichotomy, partly epistemological and partly institutional, between the arts and humanities on the one hand and the sciences on the other. In recent years, the tendency has been for linguistics, however institutionally placed, to emphasise its scientific credentials. "Mere" scholarly disquisition based upon reading and reflection has fallen out of favour, and for many journals, research assessments and examinations all valid conclusions must be based upon experimental or observational data. Rigorous analysis, quantitative measures, testable hypotheses, replicability, reliability, validity, are the order of the day.

This allegiance has large implications for how we write and what we write about. If we are wholly scientific, then our aim should be only to discover *object*ive facts about language as though it were an *object* out there in the world, untainted by our own intuitions, beliefs and values. Our aim is not to mould the data to our own *subject*ive vision of the world in the hocus pocus way, but like good natural scientists to provide a description and explanation of what exists independently - to try to approach God's truth. The facts we discover should therefore be the same, whoever the investigator. They should be exactly replicable by another researcher following the same rigorous procedures. The aim is for the personality of the investigator to melt away. The good researcher is one who becomes a clone of all other good researchers, emulating the proper scope of science as description and explanation of the material world, but not evaluation. Like the botanist describing a flower we should say everything possible about it, but not whether we think it is a pretty flower or the right one for the garden.

Taken at face value, the writing of Michael Stubbs would seem to be very much part of this movement, taking linguistics away from the arts and humanities, and towards the social and even natural sciences. (He often, for example, emphasises an aspiration to provide data which can be checked.) But the matter is more complicated. Curiously, he seems to avoid the word "science" in describing his methodology. A search of the eleven papers available electronically on his website (from which I created a mini-corpus for the purposes of this chapter), reveals only fifteen uses of "science"/ "sciences"/ "scientific" / "scientist(s)". Of these, nine occur because he is discussing scientific vocabulary. A further five occur in discussing trends in the "social sciences", which at times he seems to see as not scientific enough![2] The remaining two occurrences are because he is drawing an analogy - of which he is fond - between scientific observational instruments and corpus software. So he seems, quite studiously, to avoid describing corpus linguistics as scientific.[3] And there seems to be good reason for this. His own writing (e.g. 2002a: 232-238) is given over to discussion of subtler distinctions, elaborating in particular upon the nature of brute, social and subjective facts, as discussed by thinkers such as Popper and Searle. Thus however alienated he may feel from the contradictions of post-modern relativism, he is very much the social scientist, aware that linguistics is studying social rather than the brute facts of concern to the natural sciences. Like many natural scientists, however, he too is concerned with rationality, rigour and evidence - but in a different way.

Nevertheless, despite this avoidance of the term "scientific" and an informed take on the object of linguistic enquiry as different from that of the natural sciences, we might reasonably say that the methods advocated by Stubbs share more with those of science than of the arts. Verifiable replicable facts about language are what he seeks. And the imposing personality of the analyst is downplayed. His critique of Critical Discourse Analysis (CDA) (Stubbs 1997) for example takes the movement to task on exactly this ground of a lack of rational rigour, and for the imposition of the analyst's prior beliefs in analysis. At one point for example, he criticises CDA for being

> "reminiscent of the moral crusade against the vulgarising mass media and increasingly mechanized and capitalist society which was carried out by F.R. Leavis and his colleagues in *Scrutiny* in the 1930s."

This is not to say though that he is out of sympathy with Leavis' literary critical moralising,[4] any more than he is out of sympathy with the aims and opinions of CDA. (He expressly says that he is not.) It is rather that he finds such an approach inappropriate for linguistics. What he seeks (notwithstanding the usual philosophical caveats) is to be as objective as possible.

In the arts and humanities on the other hand, in literary criticism for example, there is also a legitimate but unscientific imposition by the writer upon their data, an assertion of ideas coming from inside as much as outside. (That is not to say that the *object* of study does not impose constraints upon what is said: the critic should not get their facts wrong.) Thus the literary critic (quite appropriately for the discipline) interprets what they find in the light of their own aesthetic or moral values, moulding it to their own system of thought. Thus we read a literary critic like F.R. Leavis (to take Stubbs' own example) not only to find out about the literature he critiques, but also to find out about his unique personality and view of the world. We are learning not only about the object of study, let us say for example the novels of D.H. Lawrence (Leavis 1955), but also the subject of the analyst. We enjoy, or get irritated by, the writer's company, and learn from his ideas, both matters of fact and matters of opinion. Rather as we would in reading the novels of D.H.Lawrence himself.

And in terms of style, the good literary critic should, like a good creative writer, be idiosyncratic and distinctive, with their own quirks of style - which Leavis certainly had. The good scientist (and linguist pace Stubbs) though, will aim to write about the facts s/he has discovered as lucidly and elegantly and possible, but to do so by removing all personal touches, idiosyncrasies and embellishments. Michael Stubbs' own lean and lucid style would be a very good model to follow. In earlier times, he was drawn to the story Cat in the Rain by Ernest Hemingway (Stubbs1983:194-219), a writer whose style though powerful, is famously clear, terse and unpretentious. Stylistically, Stubbs is in many ways the Ernest Hemingway of linguistics - though paradoxically, as with Hemingway, Stubbs' apparently depersonalised style is instantly recognisable and highly personable.

A further paradox with making any neat division between hocus pocus and God's truth is that hocus-pocus ideas, once uttered - especially if eloquently expressed - become part of the external world which the God's-truth scholar seeks to describe. Perhaps the arch example is Freud, who persuades through his skill as a storyteller (Fish 1986/1987), and whose concepts, however unscientifically arrived at, became part of people's experience, even if they were not so before. But I suspect that Stubbs, like Popper (who also included Marx in this category), would see such creations as "pseudo science" (Popper 1963). A converse paradox is that God's-truth thinkers, once they have found order, try to impose it upon the "formless mass" of their hocus pocus opponents.

3. Literary Criticism or Stylistics

Let us pursue the dichotomy of literary critical and linguistic analysis of texts a little further, as it is one with which Michael Stubbs is often concerned. Writing of this distinction in 1975, H.G. Widdowson characterises these two activities as follows:

> "the ultimate purpose of literary criticism is to interpret and evaluate literary writings as works of art and (...) the primary concern of the critic is to explicate the individual message of the writer in terms which make its significance clear to others" (Widdowson 1975:5)

while

> "the linguist, on the other hand, is primarily concerned with the codes themselves and particular messages are of interest in so far as they exemplify how the codes are constructed. Given a piece of literature, a poem for example, the linguist will be interested in finding out how it exemplifies the language system, and if it contains curiosities of usage, how these curiosities might be accounted for in grammatical terms." (ibid.)

The literary critic in other words is concerned with doing something to the object of study: interpreting and evaluating. That is to say, reading something into it. The linguist on the other hand is concerned with "finding out" something from it. It is the hocus pocus and God's truth distinction all over again - and formulated in this way, Michael Stubbs is clearly for the latter.

Widdowson on the other hand sees "the purpose of stylistics" (a subject of the book from which these quotations above are taken) as

> "to link the two approaches by extending the linguist's literary intuitions and the critic's linguistic observations and making their relation explicit." (ibid.)

He refers presumably to a linking of the two perspectives in the service of interpretation, rather than of the other literary critical activity he mentioned: evaluation.

4. Stylistics or Quantitative Stylistics

Stubbs seems to accept the distinction between literary criticism and linguistics, but, unlike Widdowson, to be far from happy with the position of stylistics between the two. He refers to it as leading "an uneasy half life, never fully accepted, for many related reasons, by either linguists or literary critics" (Stubbs 2005). His solution is to distance stylistics further than does Widdowson from "the linguist's literary intuitions", to shift it more towards the linguistic than the literary critical, to make it in others words more scientific, less - if you will - hocus pocus. He has done this by positing a further dualism, opposing traditional stylistics on the one hand to corpus or "quantitative stylistics" (Stubbs 2005) on the other. He does this partly on the practical ground that conventional stylistics struggles with longer works (it is "unworkable for novels") and partly on the more theoretical ground that a linguistic analysis of an individual text or extract cannot reveal facts of the same interest and reliability as an analysis setting those findings "against a background of what is normal and expected in general language use" - comparing in other words the language of (a) particular literary text(s) with a reference corpus. Conventional stylistics is in his view at fault, like CDA, for making arbitrary choices of which features to analyse, and then making assertions about those features without objective comparative evidence.

Thus where Widdowson had a trinary distinction in which stylistics played a mediating role between two extremes

literary criticism   stylistics   linguistic description

Stubbs sets up a new trinary distinction

(literary criticism)   conventional stylistics   quantitative stylistics

but appears to have dropped literary criticism as a candidate for the kind of linguistic work he is interested in - hence my bracketing of it here. In doing this he also introduces, by default (because the conventional stylistician can only handle short texts, and the quantitative stylistician can handle large ones), a further distinction, between

intensive analysis
extensive analysis.

But there is an important distinction to be made here. This concerns whether Stubbs is advocating *replacing* the intensive analyses of conventional stylistics with the extensive ones of quantitative stylistics, or whether he see the two as complimentary. The evidence points to the latter. He writes:

> "the most powerful interpretation emerges if comparisons of texts across corpora are combined with the analysis of the organization of individual texts" (Stubbs 1996:34)

Furthermore, in his many discussions of dualisms, he professes increasingly a desire to move beyond or even reconcile them. He is not apparently for picking one side and discarding the other. In *Words and Phrases* (Stubbs 2002a:228) writing again of Saussure, he cites Hodge and Kress's (1988:17) critique of Saussure for constantly dividing the subject matter of linguistics into binary oppositions and then dropping one half of the resulting opposition. (*Parole* is dropped in favour of *langue*, then a diachronic study of *langue* in favour of a synchronic one, then a syntagmatic analysis of *synchronic* langue in favour of a paradigmatic analysis). And he approves of Hodge and Kress's disapproval. Where his own dichotomy between conventional and quantitative stylistics is concerned, he is apparently not following this pattern. He is adding not subtracting, enriching not impoverishing.

Over recent years, Michael Stubbs has used this augmented stylistic technique to carry out a number of analyses of relatively short novels and stories,[5] notably James Joyce's *Eveline* (Stubbs 2002a: 123-144), Joseph Conrad's *Heart of Darkness* (Stubbs 2005) and Henry James's *Turn of the Screw* (Stubbs 2007). All of these analyses have yielded significant literary insights: an achievement of great note as these are among the most discussed authors and texts in literary criticism and conventional stylistics. Indeed, that is his express reason for having chosen these works. *Eveline*, for example, is chosen on the grounds that it is:

> "well known, and widely available to readers who want to check my analysis [and] has been the subject of many literary critical and stylistic analyses [and] we can therefore compare the computer's results with the interpretations of trained critics" (Stubbs 2002a:125)

This is not a case of throwing up interesting insights out of nothing in other words, but of adding interesting insights to a very well ploughed field. And it has also inspired others to follow his lead: see for example O'Halloran (2007a) on *Eveline*.

But what of literary criticism, whose business is, according to Widdowson, "to interpret and evaluate literary writings as works of art"? Is it completely beyond the pale? Something with which Stubbs is simply not concerned? The answer appears to be both yes and no.

On the one hand, Stubbs has argued very effectively for a quantitative dimension to support interpretation, which can be seen as augmenting literary criticism. Thus just as conventional stylistics added to literary critical interpretation by linking it to linguistic analysis, so quantitative stylistics has added the insights of corpus analysis. It is a cumulative process, of a kind of which Michael Stubbs seems to approve.

On the other hand, however, he is apparently not concerned in his own work with the other aspect of literary critical activity: evaluation. As he writes in a different context (discussing truth conditions):

> "Truth and falsity are also problematic with respect to evaluative utterances. If someone says *That's super!*, then that may tell us something about the speaker, but little about the world." (Stubbs 2002a:9)

But as to other logicians, one might ask the following question, especially considering that in Stubbs' express view (2002a:232-235), linguistics is not concerned only with the physical world. What conception of "world" excludes speakers and their utterances from membership? There are social and psychological as well as brute facts.

******

Part Two:

I feel somewhat out of place in this *festschrift* for I am persistently and unrepentantly guilty (as will be clear from this chapter itself) of all the crimes of which Michael Stubbs is so determined a scourge: arbitrary selection of texts and features within them, and the mixing of evaluation and description. I have also been a critic of applications of corpus linguistics in language teaching (Cook 1998, 2001b), and the subject of some counter attack (Carter 1998 and Hunston 2002: 192-197).

Where Stubbs situates my own work on the hocus-pocus spectrum is clear from his remarks on my book *The Discourse of Advertising* (Cook 1992[6]). It is one of three books chosen in chapter one of his *Text and Corpus Analysis* (Stubbs 1996: 14-21) to demonstrate shortcomings of three respective methods of analysis. And of it he writes:

> "The method is simply that of confident personal literary judgment (...) he picks out his own favourites (which is what all literary critics do), concentrating on memorable or famous examples, but does not analyse the majority which merely provide useful information and/or are just banal."

In these lines, a number of activities are singled out as inappropriate for stylistics:

- confident personal literary judgment
- picking out favourites
- concentrating on the memorable or famous
- ignoring the majority
- ignoring the banal

What I want to do in this section is to argue for the inevitability, and even desirability, of all of these activities in certain types of analysis, particularly literary analysis. I do not do this in a spirit of refuting Stubbs innovations in literary analysis, but rather of suggesting, in his own spirit of reconciling rather than reducing dualisms, a way of adding to and enriching both findings and methods. (I shall return to the relation of Stubbs to those with whom he disagrees in my conclusion.)

I should like to argue next that this second component of literary criticism, evaluation, cannot be so completely sidelined. It is part and parcel of language analysis, and - problematic though it is for linguists - it cannot be left to literary criticism. I do not mean by this only that linguists should take stock of the importance of evaluative judgments by language users and their effect on language use and usage,[7] but rather that evaluation is an inevitable part of the process of linguistic analysis itself, i.e. something linguists do as well as the other language users they study.

Stubbs' response, I suspect, would be that while he, like everyone else, has his own "favourites" among literary texts, such preferences are "simply" not part of the process with which he is professionally concerned. In this respect one half of the literary critical enterprise (evaluation) is indeed left behind, while the other (interpretation) is embraced and improved. This is related to the issues of hocus-pocus and God's truth continuum in the sense that evaluation, even more than interpretation, is of its nature an imposition upon the data, rather than something arising from it.

5. Evaluation or calculation: discourse prosodies

There are many examples of "confident personal .... judgment" in corpus linguistic analysis. Take for example the identification of negative and positive discourse prosodies (Sinclair 1991:112, Louw 1993): one of corpus linguistics outstanding contributions to the understanding of word meaning. There can be no doubt that the many studies of discourse prosody - including very significant ones by Michael Stubbs himself (1995, 1996, 2001a, 2002a) - provide invaluable insights into the relation between linguistic choices and their effects, of tremendous usefulness to discourse analysis in general and literary stylistics in particular. But the point I want to make is that only a part of this process of establishing a discourse prosody is automatic and objective. Thus the first stage is initiated by the software's statistical information about the collocates of a chosen word. But the second stage - saying whether those collocates are 'negative' or 'positive' - is guided by the analyst's "confident personal .... judgment". That is not to say they are wrong in their conclusions, but only that the established method for reaching such conclusions is partly subjective and evaluative. Thus in a recent cross-linguistic study, for example, Xiao and McEnery (2006) conducted an extensive analysis of the degree to which English and Chinese conventional translation word equivalents also have similar discourse prosodies. (The study is itself an illustration of the power of corpus linguistics to contribute to a range of sub fields, in this case translation and lexicography, and of the influence of Michael Stubbs to whom Xiao and McEnery frequently refer.) They list the following collocates of *BRING about* and *CAUSE*[8]:

> ***improvements***, *revolution, order, increase, death, downfall, war, government, situation, action, improvement, policy, reduction, result and state.*

The underlined words are described as "*obviously* negative, while the other collocates are either positive or neutral" (my italics). This is a standard enough procedure, of a kind followed routinely in corpus-linguistic analysis. My point is not to question the value of Xiao and McEnery's study, or dispute that these terms are indeed negative, but rather to make explicit the method by which corpus linguists reach such conclusion. Though I agree with such judgments in practice, I want to emphasise that there is no empirical basis for saying that a particular word - "death" for example - is negative. It is something which the corpus linguist feels intuitively to be "obviously" the case.

In addition, whether prosody is negative or positive will vary with to some degree with individual or cultural context. Even "death" is not negative in all contexts: for a suicide bomber, believer in human sacrifice, or Spartan warrior. Less far-fetchedly, the designation of "revolution" as "positive or neutral" is surely highly disputable, especially in a Chinese context, where it could be seen as positive by some people and markedly negative by others. Discourse prosody in other words is context and reader dependent. Indeed, the phrase "God's truth" is a superb example of a phrase whose prosody will be negative in some contexts and positive in others. A Google search reveals that it is used in two contradictory ways: positively by the fundamentally religious, and negatively by others as a synonym for bigotry.

Of course one could counter these arguments in two ways. Firstly one could survey opinion, specifying if necessary for which populations a given word is negative or positive. But this seems to involve the very circularity of which Michael Stubbs has been so critical elsewhere, as it is using people's intuition to access facts which are supposed to be unavailable to intuition! (Though the standard social-science assumption is that inter-coder agreement is itself a kind of objectivity.)

Alternatively, staying within the heuristic framework of corpus generated data, one could seek out the discourse prosody of the common collocates of the word in question, for example "death". But this would then, if we are to be rigorous, necessitate a further search for the collocates of these collocates, taking us into a game of everlasting deferral, more suited to a Derridean deconstructionist analysis that the new empiricism of corpus linguistics.

We should not be surprised though by this need in linguistics to combine empirical observation of external objective factors with internal subjective judgment for it nicely mirrors the ontological status of language, as discussed extensively by Stubbs (e.g. 2002a:226-242). Language is simultaneously both within us and without us, a mental and physical object. On the one hand it is an internal subjective fact which can be perceived internally in our minds, and on the other observable countable physical traces in the world (in the shape of marks on paper and screens and sound-wave vibrations) whose reality as language is only created by the perception of a human subject. Marks on paper are only words if they are perceived to be so by someone who reads that language! It is this which accounts for the ambiguous status of linguistics as a discipline and explains perhaps Michael Stubbs' careful avoidance of the term 'science'.

6. Verse or poetry: what would be a corpus of poems?

A similar point about the combination of objective fact and subjective intuition pertains to the assembly of the corpora from which the objective facts emerge. This is not a problem for finite specialist corpora, but it is one for general corpora. Thus one might have for example a corpus of *The Guardian* newspaper in 2003 which is complete - every word in every copy of the newspaper in that period - and searches of such a corpus would indeed yield hard indisputable facts about the language of *The Guardian* in this period. But when we come to the notion of a general corpus of English (or any other language) then we encounter serious problems, if we are to rule out the role of "confident personal ... judgment". Thus the written component of the British National Corpus for example - often taken as a standard - when broken down into components is based upon quite arbitrary choices about proportions of different text types,[9] making one wonder how the conclusions drawn from it about British English in general might be different if the proportion of say novels in it were higher or if the proportion of academic prose were lower. Again as with my remarks on discourse prosody it is important to be clear what exactly it is that I am criticising. I am not saying the selection of texts for the BNC is a bad one, nor am I offering an alternative. The point is that any selection in a general corpus must be arbitrary. As Stubbs himself says, acknowledging exactly the problem I have outlined:

> "the concept of a representative sample of the English language makes little sense.... A sample can be representative only if the population to be sampled is homogeneous, and this is possible only in special cases, say with a specialized sub-genre corpus (such as editorials from quality newspapers or research articles on biochemistry). Every time we enlarge a corpus, we increase the heterogeneity of the data, and there will always be text-types which we have not sampled, or which are arguably underrepresented. Unfortunately, the same problem arises with the concept of a balanced corpus: who is to say what percentage of the corpus should consist of weather forecasts, lonely hearts ads, business reports, the lyrics of pop songs, or whatever?" (Stubbs 2002a:223)

So the problem is recognised, and has been much discussed since the early days of computerised corpus linguistics (Francis 1979). Yet to quote Stubbs' own criticism of a methodological weaknesses in CDA:

> "the fact that this is noted from time to time by practitioners does not get CDA out of this particular Catch 22" (Stubbs 1997).
> So the same must presumably apply to corpus linguistics!

This however is not the point about corpus assembly I wish to argue, though it is an introduction to it. Stubbs and others may well be right in seeing increasingly large corpora, and the cross checking of findings from different corpora (Stubbs 2002a:223-224) as the best practical - if fallible - solution to the problem, and he has certainly obtained significant results by pursuing it. For the sake of argument, I am happy to take this particular objection (which I have raised elsewhere) as *passé*. But I would like to pursue further a particular problem about the construction of literary corpora, as I believe it illustrates my point about the necessity of using evaluation as a component in analysis. It is an instance of the inevitability of being at times hocus pocus, and abandoning the search for God's truth.

If we are studying literary discourse, then an example of a finite corpus (equivalent to *The Guardian* in 2003) would be the published works of a single author, or of one particular work such as *Heart of Darkness*, *Eveline* or *The Turn of the Screw* - as in Stubbs' analyses alluded to above. This finite corpus can be compared with a general corpus as Stubbs has done. But what if we want to compare such a finite literary corpus with literature in general? Then I believe we encounter a problem which only the literary critic can solve.

This is most easily illustrated with the case of poetry (a genre which significantly does not figure prominently in any of the standard general corpora such as the BNC or COBUILD).[10] Suppose that one had a finite corpus, let us say the published poems of W.B.Yeats, and wanted to compare the language used in it with the language used in poetry *in general* (or some more manageable but still general category such as nineteenth and twentieth century poetry *in general*). How could one construct the necessary reference corpus: that is to say, one of "poetry" *in general*? There would really be two options, reflecting different types of definition of the word 'poem' itself.

Let us consider how poetry is defined outside the academic linguistic or literary critical world, by a dictionary which is not corpus informed. Here is *Collins Concise Dictionary Plus* definition of the word poem:

> "**poem** *n.* **1.** a composition in verse, usually characterised by words chosen for their sound and suggestive power as well as for their sense, and using such techniques as metre, rhyme and alliteration. **2.** a literary composition that is not in verse but exhibits the intensity of imagination and language common to it: *a prose poem.* **3.** anything resembling a poem in beauty, effect etc. [C16: from L. *poēma*, from Gk, var. of *poiēma* something created, from *poiein* to make]"

This is, to say the least, a bit of a muddle. The first sense is in part a mechanical, technical and undiscriminating definition, suggesting that anything, however dreadful, can qualify as a poem, provided it has metre, rhyme and alliteration. Mixed in with this, however, is the notion that the words are chosen for their "sound and suggestive power". This seems to beg a host of unanswerable questions about authors' intentions, and whether other genres lack these qualities. The second definition repeats and even complicates the vagueness of the first. But the third definition, though tautological (a poem is "anything resembling a poem"!) introduces a quite different notion from either the mechanical technical criteria or the good intentions of the poet. It talks instead about the effect of poetry.

Nevertheless, despite its muddle and internal contradictions, this dictionary definition does capture a dualism which is acute in the case of poem, but true of other genres too, and poses a serious problem for corpus linguistics in that it cannot be solved by its own methods alone, but must appeal to others for help.

According to one sense of the word 'poem' you can proceed without evaluation of examples. Just collect everything that claims to be a poem and/or fulfils certain formal criteria - it is set out in lines on the page, or whatever. For certain purposes that is perfectly legitimate. Let me give you an example. In a recent research project, Brigitte Nerlich and colleagues examined the social and psychological effects of the Foot and Mouth epidemic in Britain in 2001, arguing against the media and governmental characterisation of the problem as an economic and health issue. They discovered that the traumatic experience of the epidemic in which 5 million farm animals were slaughtered and disposed often in sight of their owners and their families had occasioned an outpouring of poetry (in this first sense) from most unlikely sources: vets, government health inspectors, farmers, and schoolchildren. And she and her team collected and analysed a corpus of these poems (Nerlich and Döring 2005). The poems in this corpus are not necessarily good ones from a literary critical point of view, but it does not matter for these purposes.

But for other purposes it does matter. Of course one could compare the poems of W.B.Yeats with such a corpus, but the findings would be very different from those if one compared his poetry with other "good poetry". A mechanical definition of poetry (rhyme, rhythm, lineation and so on) is not enough to capture what is fully meant by this term. For "poetry", as the clumsy dictionary definition suggests, is an evaluative as well as merely descriptive term. To identify something as "poetry" is to applaud it for having achieved a certain kind of effect. So a corpus of poems in this second sense does need to concentrate on "the famous and memorable" and ignore the "the majority or the banal".

How is corpus linguistics to deal with such subjective and evaluative criteria? I can think of a number of answers, but they all seem to involve a degree of circularity and internal contradiction. To say that one would choose only published poetry simply dodges this issue, as it shifts the burden of evaluation away from the analyst and on to the publisher, and to an extent the public - insofar as what is published is a response to demand. To say that one would take poems from the "literary canon" (as evidenced for example in university literature syllabuses) is even less satisfactory, for that would ultimately rely upon the subjective evaluations of literary critics, whose faulty diagnoses and "confident personal literary judgment" we are supposed to be leaving behind.

One other solution might be to test findings on public. One of the effects cited in the clumsy dictionary definition, for example, is "beauty". Now this effect, as the truism observes, is a subjective one, "in the eye of the beholder". (Though within a certain discourse community, there could be quite a degree of consensus about what is beautiful language. Considering Shakespeare's language to be beautiful is not a minority view!) So one could objectify (and thus quantify) if not the beauty of a piece of language itself, at least the extent of its perception as beautiful among a particular population.

7. Evaluation and saliency

> "I cannot prove that a jury did interpret a summing-up in a particular way: I cannot look inside their minds. But I can attempt to show that patterns of language are likely to be interpreted in a certain way, because that is how they are likely to be interpreted in everyday life. The linguist has to try and show how a reasonable person, doing his or her best to understand, is likely to interpret language." (Stubbs 1996:102)

Evaluation is to a degree a conscious and explicit process: something which the reader actively does. There is then the possibility of correlating corpus findings foregrounding some particular linguistic feature with the reaction of actual readers, as a way - albeit partial and fraught with problems - of "looking inside their minds". Let us turn to this possibility next.

Now one apparent way out of these difficulties might be, as we are dealing with matters of effect and evaluation, to test out how readers do evaluate and react to certain texts and to certain features within them. Which poems do they find "memorable"? Which features within them do they find "beautiful" and so forth? In fact, just such a procedure of reader research is advocated by Stubbs (1997) in his critique of CDA, as a way out of its usual reliance upon the intuitions and personal responses of the analyst. This is one of his main criticisms of CDA when he laments the failure (since remedied in some CDA work[11]) to actually make any enquiries of real readers. What he seems to be advocated then are two alternatives to relying upon the analyst's intuition. The first is an appeal to corpus findings to ascertain how patterns in a single text relate to those in the language as a whole (or as near to a whole as one can model). The second is to try to ascertain the effects of linguistic choices, not by guesswork with reference to oneself , but by asking actual readers for their response. Stubbs himself, we might note, had made pioneering use of reader responses in his stylistic analysis of the Hemingway story *Cat in the Rain* (Stubbs 1983: 194-217) asking teachers, rather than himself, to summarise the story as a way of accessing what features were salient for them.

I myself now strongly favour such an approach, partly in response to having taken on board (despite my points here) some of Stubbs' criticisms. In a series of research projects examining the discourse of public debates over food issues, I have combined corpus analysis with interviews with text writers[12] and with focus groups in which readers discuss their responses to sample texts (Cook 2004; Cook, Robbins and Pieri 2006; Cook, Pieri and Robbins 2004). In this way, I and my colleagues have some evidence of which frequently occurring linguistic choices were made consciously by the text writer, and which were noticed by the text reader. Using a corpus from four British newspapers[13] of all articles published in the first six months of 2003 that mention genetic modification in any way.[14] We have been able to show, for example, that a frequent use of modals such as "could", "would" and "may" in hard-news reports about developments in genetic-modification technology is actually noticed without prompting by readers.[15] Thus our corpus analysis showed such constructions as the following to be typical - and there were many many more than these few (emphases added throughout):

> BANANAS **could become** extinct because a vicious disease is wiping the fruit out.
>
> The fungus, Sigatoka, is devastating plants in Africa. (...............) Emile Frison is top banana at the International Network for the Improvement of Banana and Plantain. He told New Scientist magazine that the only hope of saving edible bananas **may be** to create controversial GM versions -a new Frankenstein food. That **would involve** taking a gene from a disease-resistant, non-edible banana and injecting it into the threatened fruit. (*Sun* 16 January 2003)
>
> Decaffeinated brews **could soon be** cheaper and tastier after Japanese scientists grew GM coffee plants with 70 per cent less caffeine than normal. (*Sun* 19 June 2003)
>
> GM variety **could end** danger to children
>
> THE threat of peanut allergies **could soon be** wiped out by genetic engineering, scientists have revealed. (*Daily Mail* 17 February 2003)
>
> GENE therapy **could help** men undergoing surgery for prostate cancer to enjoy normal sex lives, scientists said yesterday. (*Daily Mail* 29 April 2003)
>
> (...............) Experts in Texas genetically modified ricin and found it killed off tumour cells in mice.
>
> Potentially lethal side effects were cut fivefold. Doctors who used ricin in tests on lymphoma patients believe it **could help** develop new drugs known as magic bullets.
>
> A hoard of ricin, extracted from castor beans, was found in London in January.
>
> (Sun, March 10 2003)

Such modals, we hypothesised, seek to confuse actual with hypothetical developments, and, if they are missed by the casual reader, are likely to create a more benign image of GM technology than it actually deserves.[16] However this hypothesis that readers do not notice this detail was not borne out by our findings. We showed the following text (chosen with the help of corpus analysis as typical) to our focus groups (recruited to represent different types of interest in food issues)

> GENETICALLY modified crops **could** help endangered birds such as lapwings and skylarks thrive again in Britain, says a study. It follows the development of herbicide resistant GM crops and a herbicide which can kill weeds much later in the year. (......) The study was funded by GM giant Monsanto. (*Daily Mail*, 15 January 2003)

We received the following spontaneous commentary (and again these are only examples):

> It *could* help endangered birds. They haven't proved that it does. (Farmers)
> It just says 'could' * Yes I just spotted that * Which is very ambiguous, isn't it? (Charity workers)
> Well another thing that's in there. It says it *could* help. It doesn't say it *will.* So therefore it's not necessarily scientifically proven either (Parents)
> It's not saying modified crops will help endangered birds i.e. lapwings, it just says 'such as' and it says it *can* kill weeds it doesn't say it *does* kill weeds. So the terminology's very vague. (Undergraduates)

There are many problems in pursuing such a methodology. It risks for example equating writers' and readers' own reports of their intentions and reactions with their actual intentions and reactions. It can rely too much upon the intensive reading and discussion generated in focus groups and interviews, when actual reading of newspapers is likely to be much more casual and less attentive (O'Halloran 2003). Nevertheless, for all its shortcomings, some such procedure is perhaps the best we have available if we are to distinguish between aspects of texts which are evident to the linguist (with or without the help of corpus analysis) and those which are salient and effective for the reader.

But for corpus linguistics, in addition to all the standard reservations about elicited data, there is a danger of a further contradiction. Stubbs (1996:21) makes the claim that "the deep patterning" revealed by corpus analysis is "beyond human memory and observation". Two questions come to mind. The first is to wonder why, if the patterns of language evidenced by corpora are not consciously available to one person, what the point is of asking anyone. The second is to wonder why, if an analyst's intuitions are to be discarded, they are better replaced with the intuitions of others.

What is needed is an integration of a corpus approach dealing with the records of language behaviour, and a cognitive approach dealing with how the mind produces or responds to that behaviour. The patterns of language revealed by corpus analysis need to be matched up with its acquisition and representation in the mind. There are certainly interesting attempts to do this (Wray 2000, Hoey 2005), though they tend to focus predominantly on subconscious rather than conscious interpretation and evaluation. Despite such pioneering studies, the lack of such integration is at present the main weakness of corpus linguistics, but also perhaps its main way forward. Interestingly, and significantly, Stubbs too seems to think so too:

> "When Chomskyan linguistics took a decisive step away from studying behaviour and its products, to studying the cognitive system which underlies behaviour, this led to the discovery of many interesting facts about language. Equally, when corpus linguistics took a decisive step towards the study of patterns across large text collections, this also led to the discovery of many new facts. The approaches are often seen as being in opposition, and the dualisms are perpetuated, but the long-term aim must be to integrate the insights from different approaches." (Stubbs 2002a:242)

8. CODA The hocus pocus Stubbs

The discussion above has sought to pursue a number of issues arising from Michael Stubbs' work on corpus linguistics. To some it may seem inappropriately over-critical for a *festschrift*, but I hope that, on the contrary, it follows Stubbs' own lead. Though critical of my work and disapproving of my methods, he has persistently engaged in dialogue with me: in print, by letter, by email, in conversation. I claim no special status for this. The evidence is that he does this with many people. He shuns the academic fashion for ignoring those with whom he disagrees, and allowing the discipline of linguistics to be safely compartmentalised. Indeed, he seems unable to leave the arguments of those with whom he disagrees alone, but thrives on engagement. This is as much the case with those who are narrow and ill-informed (Borsley and Ingham 2002, answered by Stubbs 2002b) as with those whose criticisms are of substance and depth (Widdowson 2000 answered in Stubbs 2001b) or powerful voices from the past, such as Saussure. He sharpens his own ideas by these encounters, eschewing the easier but more popular option of nailing his colours only to one methodological mast, and battening down the academic hatches in order to sleep more easily at night, untroubled by contradictory voices. And he is uncannily well informed, with a disconcerting habit of predicting and countering objections to his method and approach, even before they are uttered. It is difficult to catch him out. Many of my criticisms of corpus linguistics above are supported by quotations from Stubbs himself. Encounters with him have certainly made me think, and considerably changed my opinions over the years - for despite my quibbling, I do recognise the extraordinary richness of his way of analysing language.

So where is he on the God's-truth hocus-pocus continuum? He is by no mean a linguist for whom the corpus is a sausage machine, mincing up the living language and delivering it in manageable chunks. He has also exerted his own influence upon it. He seeks, like Saussure, to structure and understand the world he discovers, and to persuade others who do not agree. Thus although he may see himself as approaching God's truth, he is actually also hocus pocus in the best sense. He has exerted his own influence upon linguistics and shaped its development. And like a good literary critic, he is worth reading not only for what he tells us about the external world (which is a great deal) but also for his own unique ideas. I have always read his work with great interest, as much as to find out how he sees language, as to know what that language is "really" like.

Acknowledgement

My thanks to Kieran O'Halloran for his comments on earlier drafts of this chapter.

Notes

[1] I use Stubbs' own current term 'discourse prosody' rather than 'semantic prosody'. (Stubbs 2002a)

[2] He makes some wry criticisms of the influence of post-modern relativism. See for example the conclusion to Stubbs 2000: "That's logic... nevertheless, some facts are based on publicly-accessible empirical evidence. The post-modernists among you may argue that I don't realise the implications of my own text. But I can reply that, in order to study intertextuality, we need both historical and corpus methods. That's rhetoric...."

[3] He refers for example to "books such as Kuhn (1970) on paradigms of thought" (Stubbs 1997) where most people, including Kuhn himself would refer to Kuhn's ideas as referring to paradigms of science rather than thought. In my wider less reliable reading of his work outside what is available electronically, I know of only one exception, where he does seems to equate the proper practice of linguistics with good science in a critique of Chomsky (Stubbs 1996:29).

[4] Though his choice of the loaded phrase "moral crusade" does suggest so!

[5] Though still too long for the conventional stylistician to handle. Stubbs reports that Eveline is "a little over 1800 words"; Heart of Darkness "less than 40,000 words"; Turn of the Screw "a short text of only 42,880 words".

[6] Now in a second edition (Cook 2001a). The book's main aim is to show (through conventional stylistics analysis of adverts in comparison with literary works) that those linguistic features commonly identified with 'literariness' are as prevalent in ads as in literature.

[7] As argued by Deborah Cameron (1995:1-33), and implicit in Stubbs' own work on debates over standard English in schools (e.g. Stubbs 1976, 1980, 1986, 1995).

[8] A word whose discourse prosody is also analysed by Stubbs (2002a:45-49).

[9] For example, the written component of the BNC is roughly 60% books, 25% periodicals, 10% other published material, 10% unpublished written material, and 5% material written to be spoken. Broken down in another way, it is 75% informative writings, and 25% "imaginative" writing (BNC website http://163.1.0.36/corpus/creating.xml). Although these proportions are described as "enough to justify the claim that it characterises modern British English" (Burnage and Baguley) the basis for them appears to be intuition and subjective judgment.

10 Perhaps for the reason that it is too idiosyncratic to contribute to statements about normality. As Hunston and Francis remark: "[O]ne of the outcomes of using large quantities of data is that some of it may be discarded, in the sense that instances of word-play or language that is strange because it is being used in strange circumstances, are deliberately ignored in terms of the general description of the language. (Hunston & Francis 2000:17).

11 See for example Wodak et al. 1999

12 A similar supplementation of corpus analysis through interviews with writers can be found in Harwood 2007.

13 *The Sun*, *The Daily Mail*, *The Guardian* and *The Times*.

14 These were identified automatically by simply searching for every use of terms such as "GM", "Genetic modification" "genetically modified" etc. Thus the search was objective, even if our search times were inevitably not.

15 For a definition and corpus study of "hard news" see O'Halloran 2007b

16 See Cook 2004 for extensive elaboration and critique of the arguments.

References

Borsley, R. and R. Ingham 2002. 'Grow your own linguistics? On some applied linguists' views of the subject.' *LINGUA International Review of General Linguistics* 112: 1-7.

Burnage, G. and G. Baguley. 'The British National Corpus' http://163.1.0.36/archive/papers/gblibs.html (accessed February 10 2007)

Cameron, D. 1995. *Verbal Hygiene*. London: Routledge.

Carter, R. 1998. 'Reply to Guy Cook.' *English Language Teaching Journal.* 52.1:57-63

Carter, R. 1998a. 'Orders of reality: CANCODE, communication, and culture.' *English Language Teaching Journal.* 52(1): 43-56

Collins Concise Dictionary Plus 1989. London and Glasgow: Collins.

Cook, G. 1992. (first edition) *The Discourse of Advertising*. London: Routledge.

Cook, G. 1998. 'The Uses of Reality: A Reply to Ronald Carter.' *English Language Teaching Journal.* 52 (1): 57-64.

Cook, G. 2001a. (second edition) *The Discourse of Advertising*. London: Routledge.

Cook, G. 2001b '"The philosopher pulled the lower jaw of the hen": ludicrous invented sentences in language teaching.' *Applied Linguistics*. 22 (3): 366-387.

Cook, G. 2004. *Genetically Modified Language*. London: Routledge.

Cook, G., P.T. Robbins and E. Pieri 2006. '"Words of Mass Destruction": British newspaper coverage of the GM food debate, and expert and non-expert reactions.' Public Understanding of Science. 15 (1): 5-29

Cook, Guy, E. Pieri and P.T. Robbins 2004. ' "The scientists think and the public feels": expert perceptions of the discourse of GM food' *Discourse and Society* 15(4): 433-449

Fish, S. 1986. 'Withholding the Missing Portion: Power, Meaning and Persuasion in Freud's "The Wolf Man".' *Times Literary Supplement*, August 29: 935-938. (Longer version in F. Meltzer (ed.) (1987) *The Trial(s) of Psychoanalysis*. Chicago: University of Chicago Press. 183-209.)

Francis, W.N. 1979. 'Problems of assembling and computerising large corpora' in H. Bergenholtz and B.Schaeder (eds.) *Empirische Textwissenschaft.* Berlin: Scriptor. 110-23.

Grice, H. P. 1958. 'Postwar Oxford philosophy', in *Studies in the Way of Words*. 1989. Cambridge, MA: Harvard University Press. 171-80.

Harwood, N. 2007. 'Political scientists on the functions of personal pronouns in their writing: an interview-based study of "I" and "we".' *Text & Talk* 27(1): 27-54.

Hodge, R. and G. Kress 1988. *Social Semiotics*. Oxford: Polity.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language.* London: Routledge.

Householder, F.W. Jr. 1952. 'Review of Harris, Zellig S. (1951), *Methods in Structural Linguistics.*' *International Journal of American Linguistics*. 28:260-268.

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

Hunston, S. and G. Francis 2000. *Patterns of Grammar: a corpus - driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Kuhn, Thomas 1970. *The Structure of Scientific Revolutions*. Chicago: Chicago University Press.

Leavis, F.R. 1955. *D.H. Lawrence: Novelist*. London: Chatto & Windus; Toronto: Clarke, Irwin.

Nerlich, B. and M. Döring 2005. 'Poetic Justice? Rural policy clashes with rural poetry in the 2001 outbreak of foot and mouth disease in the UK.' *Journal of Rural Studies* 21:165-180.

Popper, K.R. 1963. *Conjectures and Refutations.* London: Routledge & Kegan Paul

Louw, B. 1993. 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies.' In M. Baker, G. Francis & E. Tognini-Bonelli (eds.) *Text and Technology* Amsterdam: Benjamins. 157-176.

O'Halloran, K. 2003. *Critical Discourse Analysis and Language Cognition.* Edinburgh: Edinburgh University Press.

O'Halloran, K. in press 2007. 'The subconscious in James Joyce's *Eveline*: a corpus stylistic analysis which chews on the "Fish hook"', *Language and Literature*

O'Halloran, K. in press 2007a. 'Critical discourse analysis and the corpus-informed interpretation of metaphor at the register level' *Applied Linguistics* 28 (1)

Saussure, F. de 1916. Cours de Linguistique Générale. Paris: Payot.

Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. 1976. *Language, Schools and Classrooms.* London: Methuen. (2nd edition 1983)

Stubbs, M. 1980. *Language and Literacy.* London: Routledge.

Stubbs, M. 1983. *Discourse Analysis.* Oxford: Blackwell.

Stubbs, M. 1986. *Educational Linguistics* Oxford: Blackwell.

Stubbs, M. 1995. 'Educational language planning in England and Wales: multi-cultural rhetoric and assimilationist assumptions.' In O. García & C. Baker, eds. *Policy and Practice in Bilingual Education.* Clevedon: Multilingual Matters. 25-39.

Stubbs, M 1995. 'Collocations and cultural connotations of common words.' *Linguistics and Education*. 7(4): 379-90.

Stubbs, M. 1996. *Text and Corpus Analysis.* Oxford: Blackwell.

Stubbs, M. 1997. 'Whorf's children: critical comments on critical discourse analysis.' In A. Wray and A. Ryan (eds.) *Evolving Models of Language*. Clevedon: Multilingual Matters. 100-16.

Stubbs, M. 1998. 'German loanwords and cultural stereotypes.' *English Today*, 53 (14): 19-26.

Stubbs, M. 2000. 'Society, education and language: the last 2000 (and the next 20?) years of language teaching.' In H. Trappes-Lomax ed. *Continuity and Change in Applied Linguistics*. Clevedon: Multilingual Matters.

Stubbs, M. 2001a. 'On inference theories and code theories: corpus evidence for semantic schemas.' *Text* 21(3): 437-65.

Stubbs, M. 2001b. 'Texts, corpora and problems of interpretation.' *Applied Linguistics* 22 (2): 149-72.

Stubbs, M. 2002a. *Words and Phrases: Corpus Studies in Lexical Semantics* Oxford: Blackwell.

Stubbs, M. 2002b. 'On text and corpus analysis: A reply to Borsley and Ingham.' LINGUA International Review of General Linguistics 112:7-13.

Stubbs, M. 2005. 'Conrad in the computer: examples of quantitative stylistic methods.' Language and Literature, 14, 1: 5-24 .

Stubbs, M. 2007. 'The Turn of the linguists: text, narrative, analysis.' Unpublished manuscript.

Widdowson, H.G. 1975. *Stylistics and the Teaching of Literature*. London: Longman.

Widdowson, H.G. 2000. 'On the limitations of linguistics applied.' *Applied Linguistics* 21(1): 3-25.

Wodak, R., R. de Cillia, M. Reisigl and K. Liebhart 1999. *The Discursive Construction of National Identity* (Transl. A. Hirsch and R. Mitten) Edinburgh: Edinburgh University Press.

Wray, A. 2000. Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics* 21:463-490.

Xiao, R. and T. McEnery 2006. Collocation, Semantic Prosody and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics* 27 (1):103-130.

BMN.06