BENEDIKT SZMRECSANYI

Trends in Linguistics

Morphosyntactic Persistence in Spoken English

A Corpus Study

MOUTON DE GRUYTER

Morphosyntactic Persistence in Spoken English

Trends in Linguistics

Studies and Monographs 177

Editors

Walter Bisang
(main editor for this volume)
Hans Henrich Hock
Werner Winter

Mouton de Gruyter
Berlin · New York

Morphosyntactic Persistence in Spoken English

A Corpus Study at the Intersection of Variationist Sociolinguistics, Psycholinguistics, and Discourse Analysis

by

Benedikt Szmrecsanyi

Mouton de Gruyter
Berlin · New York

Mouton de Gruyter (formerly Mouton, The Hague)
is a Division of Walter de Gruyter GmbH & Co. KG, Berlin.

♾ Printed on acid-free paper which falls within the guidelines
of the ANSI to ensure permanence and durability.

Library of Congress Cataloging-in-Publication Data

Szmrecsanyi, Benedikt, 1976–
Morphosyntactic persistence in spoken English : a corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis / by Benedikt Szmrecsanyi.
p. cm. – (Trends in linguistics. Studies and monographs ; 177)
Includes bibliographical references and index.
ISBN-13: 978-3-11-019012-0 (acid-free paper)
ISBN-10: 3-11-019012-5 (acid-free paper)
1. English language – Morphology. 2. English language – Syntax. 3. English language – Discourse analysis. 4. English language – Grammar. 5. English language – Variation. I. Title. II. Series.
PE1171.S96 2006
425'.9–dc22
2006007356

ISBN-13: 978-3-11-019012-0
ISBN-10: 3-11-019012-5
ISSN 1861-4302

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

Cover design: Christopher Schneider, Berlin.
Printed in Germany.

Acknowledgments

This book is a revised version of my doctoral dissertation, submitted to the University of Freiburg in fall 2004. A partial summary of some of the work reported in this book has appeared in *Corpus Linguistics and Linguistic Theory* as Szmrecsanyi (2005).

Needless to say, I owe large debts to many individuals and institutions that have helped me in various ways in this work. Thus, I thankfully acknowledge the generous research support from the German Research Foundation (DFG), for help in assembling part of this study's database and for later funding a related research project under grant no. Ko 1181/1–3. I am also grateful to the University of Freiburg, to its English Department, and – especially – to my research group, all of which provided me with an extremely supportive atmosphere in which to conduct my research and to write. Moreover, I wish to thank the Faculty of Philology of the University of Freiburg for a number of travel grants to present my work at scholarly conferences.

Second, many people have made valuable suggestions and comments relating to particular chapters or to the book as a whole. Along these lines, I owe particular gratitude to my supervisor, Bernd Kortmann, for his intellectual guidance, constant support and encouragement, and trust; to Lukas Pietsch, whose early explorations of syntactic priming in dialect data convinced me that the topic was rich enough to devote a PhD dissertation to it; to Nuria Hernandez, Daniela Kolbe, and Tamsin Sanderson, for stimulating discussions and for helping me establish intercoder reliability of some of my corpus annotation; to Christian Mair, for his encouraging feedback alongside with helpful suggestions concerning verbal complementation; to Manfred Krug, for reading a preliminary draft of this work *in toto*, and for providing me with a large number of helpful suggestions; to Stefan Th. Gries, for his expertise on particle placement and for a number of suggestions relating to both interpretation of my findings and statistical design; and to Martin J. Pickering, whose close reading of an earlier version of the manuscript has made this a better book and possibly saved me from embarassment. To all these individuals, colleagues, and friends, I am extremely grateful.

My deepest personal thanks go to my family: to Sigrid, not only for proofreading large parts of my writing, but also for her endurance of all the travail and tension inherent in a project like this; to my brothers, who never hesitated to provide me with some (much-needed, I suppose) diversion; and to my parents, who for longer than I can remember have always been there to support and encourage me in many important ways. This book is dedicated to them.

Of course, despite these many debts I have incurred, I alone remain responsible for the analyses, interpretations, and arguments in this book.

Freiburg, February 2006

Contents

List of tables

List of figures

List of abbreviations

Adj	Adjective
BNC	British National Corpus
BNCwri	Written section of the British National Corpus
CG	Context-governed spoken section of the British National Corpus
CSAE	Corpus of Spoken American English
CSPAE	Corpus of Spoken Professional American English
df	Degree(s) of freedom
DO	Direct object
DS	Demographically sampled spoken section of the British National Corpus
exp(*b*)	Odds ratio (logistic regression)
FRED	Freiburg English Dialect Corpus
ln	Natural logarithm
N	Noun
NP	Noun phrase
n.s.	Not (statistically) significant
Part	Particle
Perl	Practical Extraction and Report Language
POS	Part of speech (tagging)
PP	Prepositional phrase
R^2	Variance explained
Varbrul	Variable rules software package
V+ger.	Gerundial complementation
V	Verb
VIF	Variance Inflation Factor
V+inf.	Infinitival complementation
V+NP+Part	Noun phrase / particle order after transitive phrasal verbs
VP	Verb phrase
V+Part+NP	Particle / noun phrase order after transitive phrasal verbs

Chapter 1
Introduction

The present study is an extended, empirical argument that language users are creatures of habit whose behavior can be characterized as inertial; it seeks to show that this inertia is sufficiently patterned to be predicted by the analyst. Dwight Bolinger once noted that

> we have no way of telling the extent to which a sentence like *I went home* is a result of innovation, and the extent to which it is a result of repetition, countless speakers before us having already said it and transmitted it to us in toto. Is grammar something where speakers 'produce' (i.e. originate) constructions, or where they 'reach for' them, from a pre-established inventory ...? (Bolinger 1961: 381)

Indeed, as a corpus linguist dealing with naturalistic data, one regularly has the nagging suspicion that language users seem to 'reach for' at least as much as they 'produce.' Consider, for instance, (1):

(1) *I think it may be 10 years from now when people* ***start seeing*** *the long-term effects from it, and you* ***start having*** *problems with it.* (CSPAE Comm8a97)

In (1), there are two occurrences where the verb *to start* takes gerundial, as opposed to infinitival, complementation. Linguists have devoted considerable energy to answering the question why the verb *to start* sometimes takes infinitival complements, and sometimes gerundial complements (cf. chapter 8 for a review). Traditionally, a linguist interested in this kind of variation would look at both *start seeing the long-term effects* and *start having problems with it* in isolation and attempt to determine, for each of the two occurrences, semantic or other internal constraints that presumably caused the speaker to go for gerundial complementation in both cases; sociolinguists would also examine external variables such as sex, age, or class and see how these interfere with the observable variation. By contrast, the main argument of the present study is that an important reason why gerundial complementation is used in the second slot in (1) is because gerundial complementation has just been used immediately before. This comparatively simple explanation has received far less corpus-linguistic, empirical attention than it, I believe, de-serves. In what follows, I will refer to the underlying phenomenon – namely, that speech production is inertial – as *persistence* in language use.

1. Scope

I operationally define

(2) **persistence** as referring to the tendency that if speaker A faces a variable Z where he or she has the choice between two or more semantically equivalent variants (regardless of whether they are lexical, morphological, or syntactic in nature), speaker X's choice will be affected by

(α) previous exposure to the variable Z, such that use of a specific variant (either by speaker A or by another speaker B, to whose output speaker A has been exposed) in previous discourse will make it more likely, all other things being equal, that the same indexvariant variant will be used again by speaker A (henceforth: α-persistence; see example (3), next page); or by

(β) previous exposure to a linguistic pattern Z*, which is not variable in the same way as variable Z but parallel to one of variable Z's variants, such that use of the linguistic pattern Z* (either by speaker A herself or by another speaker B, to whose output speaker A has been exposed) in previous discourse will make it more likely, all other things being equal, that the variant of variable Z which is parallel to the linguistic pattern Z* will be used by speaker A (henceforth: β-persistence; see example (4), p. 4).

It is important to point out that persistence may be implicit and does not necessarily involve any conscious intent by the speaker – though it may, but this is impossible to determine on the basis of corpus data. The aforesaid points to a basic terminological and methodological dilemma that needs to be spelled out clearly from the outset: as we shall see in chapter 2, persistence can be either viewed as a primarily psycholinguistic phenomenon (it is then referred to as *production priming*), or as a primarily discourse-functional phenomenon. Yet, I avoid referring to discourse-analytic and especially psycholinguistic terminology *a priori*. This is because corpus study may simply not be the appropriate method to investigate psycholinguistic mechanisms such as production priming effects (see, e.g. Branigan et al. 1995 and Bock and Griffin 2000 on this point, and Gries 2005 for arguments against this view). Carefully designed experimental procedures, which can control for alternative accounts, are the data source of choice to present evidence for such effects. Trying to accomplish this would be a tall order for a corpus study: in corpus data, speakers' output may exhibit persistence for reasons of rhetoric, politeness (for instance, Tannen 1982, 1987, 1989), or thematic coherence, to aid the process of gap filling in creating and processing elliptic utterances (for instance, Matthews 1979), to open up question-answer pairs (for instance, Levelt and Kelter 1982), because speakers feel like intentionally repeating items from previous discourse, or because they were primed in preceding discourse. Because it is not easily (if at all) possible to disentangle the above motivations through corpus study in a waterproof fashion, the phenomenon will be referred to using the relatively neutral term *persistence*. In the remainder of this study, therefore, the term *persistence* will be used when referring to the observation of surface parallelism in corpus data. More specific terminology – for instance, *production priming* or *discourse management* – is going to be employed when we turn to exploring likely causes of the phenomenon.

1.1. α-persistence and β-persistence

My definition of persistence in (2) captures dependencies between two or more occurrences of the same variable (α-persistence). This is exemplified in (3), where three future marker slots are positioned in adjacency and presumably influence each other:

(3) JOE: *I mean, or (. . .) is there* ***gonna*** *be a separate, they're* ***gonna*** *have an account in Chicago, for the funds to pass through?*

Or is it ***gonna*** *be passthrough funds here at the bank?* (CSAE 0906)

Crucially, however, the definition also seeks to capture dependencies between a variable and a linguistic pattern which is not necessarily variable itself, but which shares one or more syntactic, morphological, or lexical properties with one of the variable's variants. This is what I call 'β-persistence.' For illustration, (4) exemplifies a context where persistence appears to obtain between a variable and a non-optional usage of a pattern which is parallel to one of the variable's variants. The interviewee employs an optional affirmative DO periphrasis (*We did have a trap*), thereby mimicking the interviewer's question structure with non-optional DO-support (*did you go in with a trap then?*).

(4) INTERVIEWER: *Then I suppose that –* ***did you go in*** *with a trap then in those days, or . . . ?*

CAVA_HJ*: Well, we* ***did have a trap****, but uh later on, Charlie had a, uh, an old delt motorbike.* (FRED CON002)

Hence, the present study will assume that persistence can be either an intervariable effect, or an effect between a variable and some other parallel though non-optional linguistic pattern in the variable's context. In the former case (α-persistence), both the 'prime' and the 'target', in psycholinguistic parlance, are optional in that the speaker could have twice used alternative ways of expressing herself. In the latter case (β-persistence), it is only the target that is optional and variable, while the prime is not (at least it is not variable in the same way as the target). Thus, in (4) above, only the affirmative DO periphrasis (*We did have a trap*) is optional; DO-support in the interviewer's question (*did you go in with a trap then?*) is, strictly speaking, mandatory. Let us look at another set of examples for further illustration:

(5) a. *donate money to the church*
b. *give money to the church*

(6) a. *send money to the church*
b. *give money to the church*

Persistence from (5a) to (5b) would fall under the scope of β-persistence because *donate* cannot normally take double object constructions, while (5b) is just another way of saying *give the church money*. This is why *donate* is not variable in the same way as *give*. By contrast, persistence from (6a) to (6b) would qualify as α-persistence because in (6), both *send* and *give* are choice contexts in that both verbs can occur in either subcategorization frame – in other words, *send the church money* and *give the church money* are fully fledged paraphrases of (6a) and (6b), respectively.

Crucially, therefore, the analytical distinction between α-persistence and β-persistence – which may be more critical to variationist sociolinguists than to psycholinguists – is a methodological consequence of, and relies heavily on, the Labovian notion of the *linguistic variable* (as does, indeed, the present study as a whole): a linguistic variable is a linguistic item which has clearly identifiable variants (i.e., alternative realizations) where one variant can be substituted by the other with no semantic change (cf. Labov 1966a, 1966b). That is to say, if both the prime and the target are variants of the same linguistic variable, we are dealing with α-persistence; else, the relationship between prime and target falls under the scope of β-persistence.

1.2. Persistence within and across turns

It is also postulated in (2) that persistence is not only an intra-turn phenomenon, but can also have scope across turns and across speakers:

(7) JIM: *Matt* ***'ll*** *find this out, and, I mean, we* ***'ll*** *get involved in it.* (CSAE 0906)

(8) LYNNE: *But you know,* ***they do it for a living****. You know, most people that you would get to trim your horse do it all the time. And I'm not that good, or I'm not very strong.*

LENORE*:* ***Did they train you?*** (CSAE 0408)

Consider (7), where persistence in future time reference (i.e., *Matt'll* and *we'll* instead of *Matt's going to* and *we're going to*) is observable in two subsequent variable sites within a speaker's turn, or (8), where persistence of generalized actives (as opposed to agentless passives) has scope across turns and across speakers. The parallelism in (8) is due to the fact that *they do it for a living* is another way of saying *it's done for a living*, and that *did they train* *you?* can be considered a syntactic variant of *were you trained?* (cf. Weiner and Labov 1983).

The psycholinguistic literature on priming effects certainly lends support to the assumption that persistence can have scope across turns (cf., for instance, Levelt and Kelter 1982). Moreover, repetition across turns and repetition of what another speaker says (*allo-repetition* in discourse-analytic terminology [cf. Tannen 1989: 54]; *cross-speaker priming* or *comprehension-to-production priming* in psycholinguistic parlance [cf. Branigan, Pickering, and Cleland 2000]) are widely observed in discourse and serve important functions.

2. Objectives

Why study persistence? Thanks to Labov (1969), the notion of "inherent variability" plays an important role in modern linguistics: speakers alternate between semantically roughly equivalent options of saying the same thing in a statistically regular way. While linguists have always known that speech production is inertial and repetitive, the phenomenon has as yet not been systematically exploited in variationist research designs, i.e. in research seeking to quantitatively pin down the determinants of linguistic variation.[1] Presumably, this is because the phenomenon has been thought to be too unpredictable and chaotic to serve as an explanatory variable. By contrast, one of the main claims in the present study is that persistence is actually sufficiently patterned and predictable to help us understand better the linguistic choices that speakers make. In the spirit of Labov (1969), then, this study will seek to show that when persistence is factored into variationist model building, it turns out that speakers' choices are even more regular and patterned than has hitherto been thought. This would play havoc with a basic assumption underlying empirical linguistic inquiry: namely, that an occurrence of a linguistic phenomenon can in theory be considered the result of a new throw of the dice, and that it can be investigated in isolation and out of the wider discourse context. I would like to submit that this assumption is likely to be flawed.

In a nutshell, the present study will investigate – primarily, but not exclusively, through multivariate analysis methods – the effect previous linguistic choices in discourse have on upcoming linguistic choices by conducting five case studies (comparison strategy choice, genitive choice, future marker choice, particle placement, and complementation strategy choice) on the basis of several naturalistic data sources. This investigation will have two overarching objectives: first, to suggest a variationist methodology to deal with the phenomenon; second, to demonstrate that consideration of the phenomenon can substantially increase the linguist's ability to account for linguistic variation, and to predict speakers' linguistic choices more accurately. More specifically, the following research questions will guide this book's analyses:

1. How important a factor is persistence in the linguistic choices that speakers make? How much does consideration of persistence-related factors help us to account for linguistic variation?

2. What is the relative empirical showing of α-persistence and of β-persistence?

3. Presumably, persistence itself is subject to several determinants – for instance, the more recently a given option was used, the more likely it is to be used again (compared to when it was used a long while ago). Which factors influence persistence, and in what way?

4. Is persistence different for different groups of speakers – for instance, are there differences between older and younger speakers, or between male and female speakers?

5. The alternations investigated in the present study differ in their nature: some are primarily syntactic, others are additionally characterized by lexical differences. Does the magnitude of persistence also depend on the nature of the alternation examined?

A word on the disciplinary orientation of the present study might be useful to the reader. This book is primarily concerned with extralinguistic or intralinguistic factors which impact speakers' linguistic choices. Yet, the present study also draws heavily on ideas and evidence developed by psycholinguists and discourse analysts. Still, the present study is closer to variationist research traditions than to either psycholinguistics or discourse analysis. On a more general level, it is my hope that the present study succeeds in contributing to a theory of how spoken language works.

3. The organization of the present study

This study is structured as follows. Chapter 2 will review previous research on persistence, repetitiveness, priming effects, and related phenomena. Chapter 3 makes explicit the methodological and empirical framework of the present study. Chapters 4 to 8 constitute the empirical core of the present study, investigating persistence on the basis of five well-known alternations in the grammar of English as case studies. Each of these chapters will be concluded by a plain-English interim summary. Chapters 9 and 10 are, in effect, two concluding chapters. Chapter 9 is a synopsis of the five previous chapters where the findings will be generalized and discussed in a wider context. Chapter 10, finally, summarizes the present study's main findings, discusses their greater relevance, and points out areas for further research.

Chapter 2
Previous research on persistence phenomena

Those familiar with the study of rhetoric, public oratory, and poetry may be acquainted with the stylistic feature called *parallelism*, employed in (1):

(1) *Quod rogat illa, timet; quod non rogat, optant* ... (Ovid, Ars Amatoria I, 485)

This feature, and the effects that can be achieved by using it, have been known to authors and speakers for millennia. (1) exploits persistence in that the syntactic structure of a clause is copied to an adjacent clause to achieve a mesmerizing, rhythmic effect. Be that as it may, persistence can serve many more functions and can have many more motivations besides making poetry more pleasant to listen to. In fact, persistence, automaticity, and repetitiveness appear to be ubiquitous no matter which population of speakers is being looked at: preschoolers, second language learners, and adolescents (Miller and Weinert 1998: 384), neuropsychological patients (for instance, aphasics; cf. Blanken, Dittmann, and Wallesch 1992), and also academics (see, e.g. Biber et al. 1999: 987–1036).[2] With the focus of this study being on persistence and repetitiveness in adult language production, this chapter will review previous research on this population in what follows. There are three major, and somewhat distinctive, perspectives from which such research has been carried out: psycholinguistics, discourse analysis, and (quantitative) corpus linguistics. The following review is structured accordingly.

1. Psycholinguistic approaches

One fundamental characteristic of native speakers' knowledge of their language is productivity. In an ideal word, this productivity should translate into native speakers' capacity to generate and process an infinite number of grammatical sentences. Chomsky and his followers have argued that this capacity is constrained by somewhat trivial performance factors such as memory limitations, the human tendency to make mistakes, and fatigue (Chomsky 1965). One important – and not all that trivial – constraint is arguably missing from this list: researchers of the human speech production system have known for some time now that there is a number of empirically robust psycholinguistic phenomena, so-called production priming effects, in which utterances "of a linguistic form (e.g. a word or sentence) . . . tweak the production system in a way that may be reflected in changes in the production of subsequent forms" (Bock 1990: 1225). Hence, processing and production of material is manipulated by having been exposed in prior context to something related. The effect of this is that similar forms and patterns tend to persist in speech production, and ultimately that production is less creative than in could be. "Related," along these lines, can mean remarkably many things. This section will review research on the following priming phenomena:[3]

Lexical priming. Processing of lexical material is facilitated if related material has just been activated in the mental lexicon.

Morphological priming. Processing of a word is aided by just having processed a morphologically related word.

Form Priming. Processing of a word is aided by just having processed a word with a similar phonological form.

Syntactic priming, type I. Processing of a word is aided by the compatibility of that word with the prior syntactic context.

Syntactic priming, type II. Processing of a sentence is aided by just having processed a sentence with the same syntactic structure; using a particular syntactic structure is more likely given previous exposure to that structure. This phenomenon is also sometimes referred to as *syntactic persistence* or *structural priming*.

Priming phenomena are often (though not universally) assumed to be due to spreading activation levels in a network of memory which is presumably organized in terms of lexical, morphological, phonological, or syntactic similarity. When a word, morpheme, phonological form, or syntactic structure is recognized, some site in the network is activated, and this activation may subsequently spread to nodes of related patterns or tokens (see, for instance, Tanenhaus, Flanigan, and Seidenberg 1980: 519). An alternative account is that priming is a form of procedural or implicit learning:

> Structural priming can arise within a system that is organized for learning how to produce sequences of words . . . structural priming is a dynamic vestige of

> the process of learning to perform language. We call this process *learning to talk*, in the completely literal sense of *talk*. It is not learning language but learning to produce it. In this sense, learning to talk involves learning procedures – cognitive skills – for efficiently formulating and producing utterances. What structural priming suggests is that these procedures may undergo fine-tuning in every episode of adult language production. Similarly, structural priming in language comprehension . . . might be interpreted as *learning to understand*. (Bock and Griffin 2000: 188–189; emphases original)

The literature on priming phenomena is extensive, so the review below will necessarily be somewhat eclectic.

1.1. Lexical priming

Levelt and Kelter (1982) conducted a number of experiments to investigate what they called the "correspondence effect" with regard to word repeats. More specifically, Levelt and Kelter were interested in contexts such as (2).

(2) a. *Aan wie laat Paul zijn viool zien?*
'To whom lets Paul his violin see?'
b. *Wie laat Paul zijn viool zien?*
'Whom lets Paul see his violin?' (Levelt and Kelter 1982: 78)

These were the kind of questions that Levelt and Kelter's informants were asked after having been shown corresponding pictures. There was a significant tendency for informants to use the (optional) preposition in their answer if the question contained the preposition, as in (2a), and to not use a preposition when the question did not contain a preposition, as in (2b). Levelt and Kelter also manipulated the task and inserted interfering materials to study how memory might be involved in the effect. They moreover designed an experiment to test whether some degree of correspondence between a question and an answer is even perceived as 'natural' by informants. Levelt and Kelter (1982: 103) concluded that subjects tend to repeat lexical items from previous talk, both their own and other parties' and that question-answer sequences appear more natural to subjects when the sequences agree in prepositional form. Crucially, then, "a previous element of speech which is available in the speaker's working memory can, by its mere presence, affect the formulation process and reproduce itself during the speaker's turn" (Levelt and Kelter 1982: 103).

Because of the lack of closed-class priming reported in Bock (1989) (see below, section 1.4.2), it might be argued that the findings of Levelt and Kelter (1982) are just syntactic in nature. Brennan and Clark (1996) and Wheeldon and Monsell (1992), however, reported experiments where repetition is clearly lexical. Wheeldon and Monsell (1992) investigated repetition priming and found that recent exposure to a name (for instance, in response to a definition or after reading the name) substantially facilitates the subsequent naming of a pictured object. Crucially, the effect persists after as many one hundred intervening trials, and prior production of a mere homophone of the object's name does not have the same priming effect; this is why the phenomenon cannot be due to phonological relatedness but must be lexical in nature.

Brennan and Clark (1996), on the other hand, sought to elucidate why and how conversationalists, when repeatedly referring to the same object, come to use the same terms eventually. In experiment 1, for example, Brennan and Clark (1996) had subjects go through different card sets with pictures of common objects – a shoe (loafer), a dog (retriever), and so on – on each card. In simple terms, the difference between the card sets was that some sets only contained one basic-level item (for instance, one dog only), while other sets contained both a retriever and a Scottish terrier. Brennan and Clark tested how subjects, in interaction with other subjects in the experiment, would refer to particular objects (say, a retriever) after successive trials involving several sets of cards – as a *dog* or a *retriever*? This choice is presumably governed by several factors, among them recency and frequency of use. In conclusion, Brennan and Clark argue that conversationalists establish conceptual pacts when referring to objects, and that these pacts result in what Brennan and Clark call 'lexical entrainment': "the repeated use of the same or closely related terms in referring to an object on successive occasions", even when conversational parties have the option to user simpler references (Brennan and Clark 1996: 1491).

In summary, there is solid experimental evidence that it is easier to reactivate recently activated lexical representations in the mental lexicon than to activate new lexical representations from scratch, and that lexical repetition and, indeed, lexical priming constitutes an important interactional mechanism in dialogue.

1.2. Morphological priming

Morphological priming is the phenomenon in which "responses to a target word (e.g. *counted*) can be facilitated when it is preceded by a morphologically related prime word (e.g. *counting*)" (Drews 1996: 629). Kempley and Morton (1982) are credited as being the first to describe the effect. They performed an experiment in which they first primed subjects for a relatively long period, and then tested for the effect on recognition of spoken words presented in noise during the test phase. Kempley and Morton made three observations: (i) there was no priming effect for "physically" related words, hence *deflecting* produced clearly during the priming phase did not facilitate identification of *reflecting* in the test phase; (ii) for words with a regular inflectional relationship, there was a clear facilitation effect. Hence *reflected* spoken clearly during the priming phase facilitated identification of *reflecting* during the test phase; (iii) there was a complete lack of facilitation for irregularly related words such as *sing–sang*, *man–men*, etc. Kempley and Morton (1982: 441) conclude that "there is a morphological/structural level of analysis which is pre-semantic, at which these long-term effects take place." Another well-known study on morphological priming is Boyce, Browman, and Goldstein (1987). Prior to this study, it had not been entirely clear whether the strong priming effect between affixual variants, such as *reflecting–reflected*, was really due to regular morphological relatedness rather than to phonological identity of the stems, which is not given in pairs such as *sing–sang*. To investigate this matter, Boyce, Browman, and Goldstein studied morphological priming in Welsh, a language in which initial consonants in words undergo systematic mutations as a function of their syntactic context. These mutations are regular, but considerably more complex than English affixation. Crucially, phonological identity of the stems in morphologically related words is not given. Replicating the method used by Kempley and Morton (1982) with Welsh-speaking subjects, Boyce, Browman, and Goldstein show that in spite of the lack of phonological identity, Welsh "mutation is similar to affixing in English in that mutated variants prime each other" and that "abstract morphological categories, rather than identity of phonological form, are required to organize the Welsh lexicon" (Boyce, Browman, and Goldstein 1987: 419).

Therefore, morphologically related words seem to have common lexical representations. Thus a prime that is related to a target through regular morphological processes facilitates recognition, and possibly production of the latter. Morphological priming has been studied extensively in the auditory, but not in the production domain (cf. Drews 1996: 633). All the same, a similar effect in the domain of production is to be expected.

1.3. Form priming

When subjects are presented with a target word, to which a response is required (preceded by a prime), the processing of the target word is aided when the prime word and the target word are in some way related in form, as in *plank–blank* (Zwitserlood 1996: 589). Tanenhaus, Flanigan, and Seidenberg (1980) is the pioneering study on form priming. Tanenhaus, Flanigan, and Seidenberg used a so-called Stroop or color naming paradigm, in which subjects are first presented, auditorily or visually, with a prime word, and then with a target word which is printed in some color. The subjects' task is to ignore the target word itself and to just name the color in which it is printed. However, subjects cannot inhibit processing the target word, which interferes with naming the color. The logic of the task is that priming effects have been shown to additionally interfere with naming the color, meaning that subjects have a hard time *not* processing the target word and concentrating on naming the color when they have been exposed to a related word before. Thus, in Stroop paradigms, priming effects manifest as longer latencies in color naming. Tanenhaus, Flanigan, and Seidenberg (1980) used this experimental design to test for both orthographic and phonological priming. To disentangle phonological and orthographical relatedness experimentally, primes and targets sometimes were only orthographically similar (as in, e.g., *freak–break*) and sometimes only phonologically similar (as in, e.g., *light–cite*). To test for orthographic priming, Tanenhaus, Flanigan, and Seidenberg exposed subjects to the prime words visually, and found that this exposure increased color naming latencies only for orthographically (i.e., not phonologically) related targets. As for phonological priming, Tanenhaus, Flanigan, and Seidenberg (1980) presented the prime words auditorily, which led to increased color naming latencies only in the case of phonological relatedness.

Thus, Tanenhaus, Flanigan, and Seidenberg (1980) found evidence for both phonological and orthographic priming, and that phonological and form relatedness between words is a factor that aids the processing of linguistic patterns.

1.4. Syntactic priming

Two related, though analytically distinct phenomena have been referred to as *syntactic priming* in the psycholinguistic literature (cf. Nicol 1996: 675). I will refer to these as *syntactic priming, type I* and *syntactic priming, type II*, respectively. Because it is the latter type that is especially relevant to the present study, it will receive a more detailed survey in what follows.

1.4.1. Syntactic priming, type I

This variety of syntactic priming is concerned with "the facilitation in the processing of a word due to the compatibility of that word with preceding syntactic context" (Nicol 1996: 675). Tyler and Marslen-Wilson (1977) are credited with having been the first to describe the phenomenon. Their actual research question was whether syntactic processing is (un)affected by the semantic context. To this end, they designed an experiment where they presented subjects with sentences consisting of a context clause which was supposed to semantically disambiguate a following phrase fragment ambiguous in deep structure, as in (3a) and (3b):

(3) a. *If you walk too near the runway, landing planes ...*
b. *If you've been trained as a pilot, landing planes ...* (Tyler and Marslen-Wilson 1977: 684)

At the offset of the final word in the ambiguous phrase, a probe word was shown to subjects. The probe word was either compatible with the semantic context, or not; for instance, *are* is compatible with the deep structure indicated in (3a), *is* is compatible with the deep structure indicated in (3b). The subjects' task then was to repeat the probe word as rapidly as possible while the latency to name the probe word was recorded. Tyler and Marslen-Wilson's hypothesis was that if syntactic processing is unaffected by semantic context, latencies should be the same regardless of whether the probe word is compatible with the preceding clause or not. However, if subjects' syntactic representation of the fragment clause is affected by the meaning of the preceding clause, Tyler and Marslen-Wilson expected latencies to be longer for incompatible probe words. Latencies were indeed differential, which Tyler and Marslen-Wilson (1977: 689) argue is evidence that syntactic analysis is not completely unaffected by semantics. This is

> because one syntactic structure is more appropriate than another in view of the semantic constraints exerted by the preceding clause, the listener develops expectations about the syntactic structure of the remainder of the clause. Therefore, when the probe word consists of an inappropriate verb form, his naming latency is longer than it is when he sees a probe word which meets his structural expectations. (Tyler and Marslen-Wilson 1977: 689)

So, the above syntactic priming effect is the tendency that when subjects encounter a lexical item that somehow 'fits' into the preceding syntactic context, they can process this item faster than they could otherwise. This asymmetry is referred to as syntactic priming. Along rather similar lines, Wright and Garret (1984) investigated syntactic factors influencing word recognition. Subjects were presented visually with incomplete sentences which ended in a target word. Subjects then had to decide whether the target word fits into the larger context, and their reaction time was recorded. The target word was syntactically either compatible with the preceding context, as in (4a), or incompatible, as in (4b):

(4) a. *If your bicycle is stolen, you must* ***formulate*** ...
b. **If your bicycle is stolen, you must* **batteries** ... (Wright and Garret 1984: 32; emphases original)

The results obtained by Wright and Garret (1984: 39) show that "some process within subjects slows responses whenever the target word is 'odd' with respect to the preceding context."

In all, the crucial dependent variable here is *always* reaction time and latency (cf. Nicol 1996: 676) and *not* persistence in subjects' output, which is why this priming effect is not as interesting to the present study as some other priming effects.

1.4.2. Syntactic priming, type II

This type of syntactic priming focuses on "the facilitative effect on the processing of a given sentence of having just processed a sentence with the same or similar syntactic structure" (Nicol 1996: 675). Bock (1986) is the seminal study in this field. She investigated syntactic priming in the choice of active/passive constructions, as in (5), and prepositional/double object constructions, as in (6):

(5) a. *Mary saw John.*
b. *John was seen by Mary.*

(6) a. *Mary gave John the present.*
b. *Mary gave the present to John.*

In the experiments Bock set up, subjects had to read out a priming sentence containing one of the above (a) or (b) constructions. Subsequently, they were presented with an unrelated event in a picture which they had to describe (see Figure 1 for some example pictures). The point is that the pictures could be described using an (i) active or passive construction, or (ii) a prepositional or double object construction. Bock found that the structural properties of the priming sentence significantly influenced subjects' subsequent description of the pictures in that "speakers tend to repeat the syntactic forms of sentences in subsequent utterances that are minimally related in lexical, conceptual, or discourse content" (Bock 1986: 378). This means that those subjects that had to read out sentences such as *John was seen by Mary* were substantially more likely to describe the left picture in Figure 1 as *the church was struck by lightning* than were subjects who received an alternative priming sentence (such as *Mary saw John*). The experiments were camouflaged as recognition memory tests that minimized subjects' attention to their speech, which allowed Bock (1986) to rule out influence of stylistic or other preferences. Bock (1986: 379) frames the interpretation of her findings in terms of activation processes:

> An utterance takes the grammatical form that it does because the procedures controlling its syntax are more activated than the procedures responsible for an alternative form, with the higher level of activation being an automatic consequence of the prior production of the same form. (Bock 1986: 379)

It follows from this explanation that factors such as frequency or recency of use of a syntactic construction figure prominently in explaining its overall distribution. This distribution is – if the activation-based account is correct – ultimately a function of the cognitive mechanisms that are involved in generating the construction (Bock 1986: 380–381).

Bock (1986) has sparked a great deal of research activity centering on syntactic persistence or priming. All the same, a number of questions concerning the phenomenon remain. For one thing, it is still not entirely clear at which stage of language production the effect is actuated (Wheeldon and Smith 2003: 432). A second controversy concerns the duration of syntactic priming: is the effect rather short-lived (as claimed by Levelt and Kelter

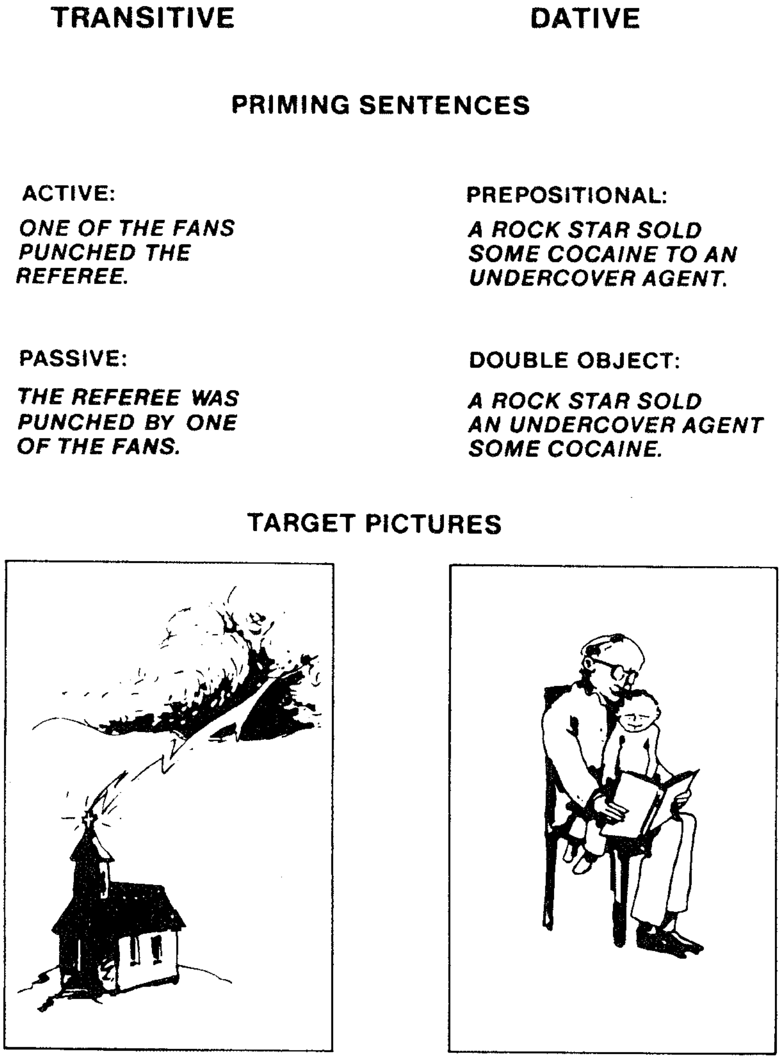

TRANSITIVE
DATIVE
PRIMING SENTENCES
ACTIVE:
ONE OF THE FANS PUNCHED THE REFEREE.
PREPOSITIONAL:
A ROCK STAR SOLD SOME COCAINE TO AN UNDERCOVER AGENT.
PASSIVE:
THE REFEREE WAS PUNCHED BY ONE OF THE FANS.
DOUBLE OBJECT:
A ROCK STAR SOLD AN UNDERCOVER AGENT SOME COCAINE.
TARGET PICTURES

Figure 1. Examples of transitive (left) and dative (right) priming sentences used by Bock (1986). Only one of the two alternative priming sentence forms was presented in each priming trial, followed by a target picture, which subjects then had to describe (from Bock 1986: 361)[4]

1982; Branigan, Pickering, and Cleland 1999; Wheeldon and Smith 2003, among others), or is it, after all, long-lived (Saffran and Martin 1997; Boyland and Andersen 1998; Bock and Griffin 2000; Branigan et al. 2000; Chang et al. 2000; Gries 2005)? Levelt and Kelter (1982) claimed that the likelihood of a prime-target match declines significantly within one single intervening clause. Branigan, Pickering, and Cleland found that in written sentence completion, "reliable priming occurred only when the target immediately followed the prime" (Branigan, Pickering, and Cleland 1999: 638). Similarly, Wheeldon and Smith (2003) reported that in their experimental set-up, priming effects did not survive even one intervening unrelated trial; and Pickering et al. concluded that "current evidence is inconclusive about how long syntactic information remains activated" (Pickering et al. 2000: 205).

Yet priming may be long-lived, after all. Bock and Kroch (1989) found that priming can persist over 12 intervening experimental trials, and Bock and Griffin reported priming effects after 10 intervening trials, a finding which, they argue, is "more compatible with a learning account than a transient memory account" (Bock and Griffin 2000: 177). Branigan et al., conducting a spoken sentence completion experiment, concluded that "syntactic priming in spoken sentence completion is not a very short-lived phenomenon" (Branigan et al. 2000: 1301) and that the rapid decay in Branigan, Pickering, and Cleland (1999) is specific to written sentence completion. Boyland and Andersen (1998) argued that priming can persist over a 20-minute interval, and Gries (2005) demonstrated that priming is rather long-lasting in corpus data. Saffran and Martin (1997) obtained priming effects over an interval of no less than a week in a study of structural priming in aphasic patients. It is true that these quite different durations are due to different methodologies that possibly trigger not quite identical effects with somewhat different loci (Wheeldon and Smith 2003: 431). Crucially, however, the issue of the longevity of the effect – controversial as it may be – has important theoretical consequences: if syntactic priming is a short-term memory or activation effect, it should be short-lived; if it is, in fact, long-lived, activation cannot be the whole story (Chang et al. 2000: 219).

Workers in the field have also sought to investigate whether and to what extent syntactic priming operates from production to production and from comprehension to production. Bock (1986) was clear evidence for production-to-production priming: subjects themselves had to read out the priming sentences before they were presented with the picture description task. As for comprehension-to-production priming, Branigan et al. (1995) have evaluated a range of experimental evidence for bidirectional priming between comprehension and production. More recently, Potter and Lombardi (1998) reported that reading as well as merely perceiving a sentence can prime its syntactic structure. Branigan, Pickering, and Cleland (2000) investigated if and to what degree speakers coordinate syntactic structures in dialogue. To this purpose, they employed a so-called 'confederate-scripting technique':

> Pairs of speakers took it in turns to describe pictures to each other. One speaker was a confederate of the experimenter and produced scripted descriptions that systematically varied in syntactic structure. The syntactic structure of the confederate's description affected the syntactic structure of the other speaker's subsequent description. (Branigan, Pickering, and Cleland 2000: B13)

In other words, speakers tended to mirror syntactic structures used by the other conversational party, which the authors took as evidence that syntactic priming is the result of residual activation at the lemma stratum (cf. Pickering and Branigan 1998), which is accessed during both comprehension and production. Using the same technique, Cleland and Pickering (2003) also showed comprehension-to-production priming, and Gries (2005), in his corpus-based investigation of syntactic priming in particle placement and the dative alternation, reported both production-to-production priming and comprehension-to-production priming.

Another issue concerns the question whether and to what extent morphosyntactic and lexical characteristics of the prime and the target can manipulate the strength of the priming effect. Pickering and Branigan (1998) investigated this matter utilizing five written completion task experiments with double object and prepositional dative constructions. They showed that on the one hand, priming is stronger when both the prime and the target consist of the same verb lemma than when different verb lemmas are used (although even then, the priming effect does not dissipate), an effect that was replicated by Branigan, Pickering, and Cleland (2000) in the domain of comprehension-to-production priming. On the other hand, Pickering and Branigan (1998) did not obtain differential effect sizes when they varied tense, aspect, or number of the involved verbs. Pickering and Branigan concluded that "combinatorial information is phrasal in nature, is associated with the verb's lemma rather than a particular form of the verb, and is shared between different lemmas" (Pickering and Branigan 1998: 633). In much the same vein, Cleland and Pickering (2003), using a confederate-scripting technique (see above) to investigate syntactic priming effects of noun phrases in dialogue, explored lexical representation – more precisely, how the formulation of complex expressions is conditioned on the structure of lexical entries. In a nutshell, Cleland and Pickering (2003) found (i) that repeating the head word of the target in the prime enhanced the syntactic priming effect, (ii) that the priming effect was also enhanced when the head words were semantically related, and (iii) that syntactic priming was not enhanced if the head words were merely phonologically related. Cleland and Pickering (2003) thus dovetails with Pickering and Branigan (1998): verb lemma or head word matches enhance priming, similarities in phonological or morphosyntactic form appear not to. These experimental findings notwithstanding, Gries (2005) submits that in the dative alternation and in particle placement, matching verb forms as well as matching verb lemmas might enhance the priming effect.

Along these lines, it is worth noting that much as Pickering and Branigan (1998) failed to obtain effects of varying morphology and Cleland and Pickering (2003) were unable to report effects of phonological relatedness of head words, Bock (1989) found that manipulation of closed-class items in priming sentences did not have an effect on the strength of syntactic priming. As in Bock (1986), subjects received priming sentences, such as (7), after which they were presented with events in pictures that they had to describe.

(7) a. *A cheerleader offered a seat to her friend.*
    b. *A cheerleader offered her friend a seat.*

(8) a. *A cheerleader saved a seat for her friend.*
    b. *A cheerleader saved her friend a seat.* (Bock 1989: Table 1)

As was to be expected given Bock (1986), subjects were more likely to describe a picture as, e.g., *the girl is handing a paintbrush to the boy* (instead of *the girl is handing the boy a paintbrush*) after having received (7a) instead of (7b). However, when Bock manipulated the prepositions in the priming sentences while leaving the overall syntactic structure unaltered – as in (8) – there was no discernible effect. This means that regardless of whether subjects received (7a) or (8a), they were equally likely to produce, e.g., a prepositional *to*-dative, as in *the girl is handing a paintbrush to the boy*. Bock concluded that "closed-class words are not inherent in the structural skeletons of sentences" (Bock 1989: 181).

In sum, there is substantial evidence that repetitions of syntactic structure often arise from the reiteration of mental processes responsible for building syntactic structure during language production, and that “this activity is to some degree dissociated from message content” (Bock 1990: 1228). Syntactic priming is an empirically robust phenomenon that has been shown for many languages and many constructions. Some of the issues that are currently subject to empirical discussion include: At which stage of language production is the effect generated? What is the time course of effect? How does the magnitude of the effect depend on its directionality (production-to-production and comprehension-to-production)? Do morphosyntactic and lexical characteristics of the prime and the target influence the size of the effect?

2. Discourse-analytic and conversation-analytic approaches

What about the interplay between priming/repetition and discourse-functional goals? For the most part, psycholinguists have been silent on this issue, relying as they have on monological, decontextualized language fragments to study language production and comprehension. Pickering and Garrod (2004: 169–170) give the following two reasons for this abstinence: for one thing, studying dialogue experimentally is methodologically not trivial (as Pickering and Garrod 2004: 169 put it, “how can the experimenter stop subjects from saying whatever they want?”). Secondly, psycholinguistics has derived most of its theoretical apparatus from generative linguistics, which of course is notorious for ignoring performance phenomena such as conversation. Only recently, then, have experimental psycholinguists begun to investigate the psychological mechanics of dialogue. Garrod and Pickering (2004) and Pickering and Garrod (2004), for instance, have argued that massive priming on all levels creates ‘alignment’ of representations between interlocutors in dialogue. This means that

> what actually occurs in dialogue is lots of lexical, syntactic, and semantic activation of various tokens at each level, and activation of particular links between the levels. This leads to a great deal of alignment, and hence the production of routines. It also means that the production of a word or utterance in dialogue is only distantly related to the production of a word or utterance in isolation. (Garrod and Pickering 2004: 183)

We will now turn to a review of what the rich literature in the discourse analytic and conversation-analytical tradition (in the spirit of, for instance, Sacks, Schegloff, and Jefferson 1974) has to say about the function(s) of repetition in discourse. Workers in this field have long known that conversationalists, under certain circumstances, use parallel patterns to achieve certain effects. So obvious and pervasive is this tendency in talk that it often has been noted just in passing. Harvey Sacks has suggested that analysts should check if the variant chosen by the speaker is coordinated with things in its neighborhood when trying to explain the speaker's choice, because repetition creates rhythmic patterns (Sacks 1971). Jefferson (1972: 303) defines a 'repeat' as "an object that has as its product-item a prior occurrence of the same thing, which performs some operation upon that product-item." According to Halliday and Hasan (1976), exact repetition can serve as a cohesive tie. Ochs (1979) finds that one of the characteristics of unplanned spoken discourse is the use of parallelism: phonemes ("sound touch-offs"), lexical items ("lexical touch-offs"), and syntactic constructions are regularly repeated. Duranti and Ochs (1979: 396) view repetition as one of "two major ways in which referents are tied to the prior discourse" in conversation among Italians. Polanyi (1979) describes the systematic use of repetition to manage narrative flow in conversational story telling. Interactional sociolinguists have reported that speakers in social interaction often modify their speech to accommodate listeners ('accommodation theory'; cf. Giles 1980). Levin (1982) argues that *anadiplosis* – beginning a clause or phrase with the word in which the preceding clause or phrase ended – is an often-used figure of speech operating on linguistic form in discourse. Schiffrin shows that non-adjacent self-paraphrase can have multiple functions (Schiffrin 1982) and that local coherence in discourse is the "outcome of joint efforts by interactants to integrate knowing, meaning, saying and doing" (Schiffrin 1987: 29), and one way to "integrate saying" is repetition or persistence of linguistic structure. Johnstone (1984) notes that self-paraphrase in conversation is a paratactic modificational strategy in that frequent juxtaposition of two items makes them more similar. Shepherd (1985) proposes that repetitiveness is particularly pervasive in creoles. According to Abbi (1985), repetition of all parts of speech is frequent in South Asian languages and used to mark emphasis, intensity, and plurality.

Schenkein (1980) is an in-depth, qualitative analysis of taped interactions. As a starting point for his study, Schenkein presents a transcription of an interaction between a bank robber and his colleague and lookout:

(9) ROBBER: . . . ***You've got to hear*** *and witness it* ***to realize how bad it is.***
LOOKOUT: ***You have got to experience*** *exactly the same position as me, mate,* ***to understand how I feel****.*
(Schenkein 1980: 22; emphases mine)

In this passage, as well as in the whole interaction, significant parroting of structure takes place. The robber says "*You've got to hear . . . how bad it is*," and his colleague reiterates "*You have got to experience . . . to understand how I feel*." Schenkein (1980: 26) argues that "using materials from prior talk in current talk is enormously common in conversational interaction," regardless of whether the repeated material is topical, inflectional, structural, or thematic. In addition, Schenkein finds that in natural-occurring conversation, even action sequences (such as QUESTION/ANSWER/ANSWER-REPEAT) tend to be repeated more often than one would suspect.

Deborah Tannen has devoted two papers published in *Language* (Tannen 1982, 1987) and part of a book (Tannen 1989) to the question of why repetition is so pervasive in conversation, and what effects conversationalists can possibly achieve by being repetitive. In Tannen (1982), she studies how speakers or writers employ oral strategies – "those aspects of discourse which make maximal use of context, by which maximal meaning and connective tissue are implied rather than stated" (Tannen 1982: 3) – and literal strategies, "those [strategies, BS] by which maximal background information and connective tissue are made explicit" (Tannen 1982: 3). Comparing spoken narratives to written narratives, Tannen's focus is on how the distribution of oral/written features differs in spoken and written narratives. One of Tannen's findings is how pervasive parallel structures and repetitions are in spoken narratives. Consider (10a), which is drawn from a spoken narrative, and (10b), which is drawn from a written narrative where the writer was asked to retell the spoken narrative:

(10) a. *And he knows Spanish, and he knows French, and he knows English, and he knows German.*
b. *He knows at least four languages fluently – Spanish, French, English, and something else.* (Tannen 1982: 14)

This is not an isolated example; almost always, Tannen found parallel structures (" . . . *he knows* . . . *he knows* . . . *he knows* . . . *he knows* . . . ") to be more frequent in oral narratives than in the corresponding written narratives. From one view, it is certainly true that writers have more, if not unlimited, planning time at their disposal, hence they do not need to rely on repetition to aid memory. However, Tannen's key explanation for this skewed distribution is that persistence often caters to the communicative goal of inter-personal involvement between communicator and audience in spoken discourse:

> what seems most significant is that syntactic parallelism establishes a mesmerizing rhythm which sweeps the hearer along; hence it is perfectly geared to knowing through involvement ..., which underlies both oral performance and conversation. (Tannen 1982: 7)

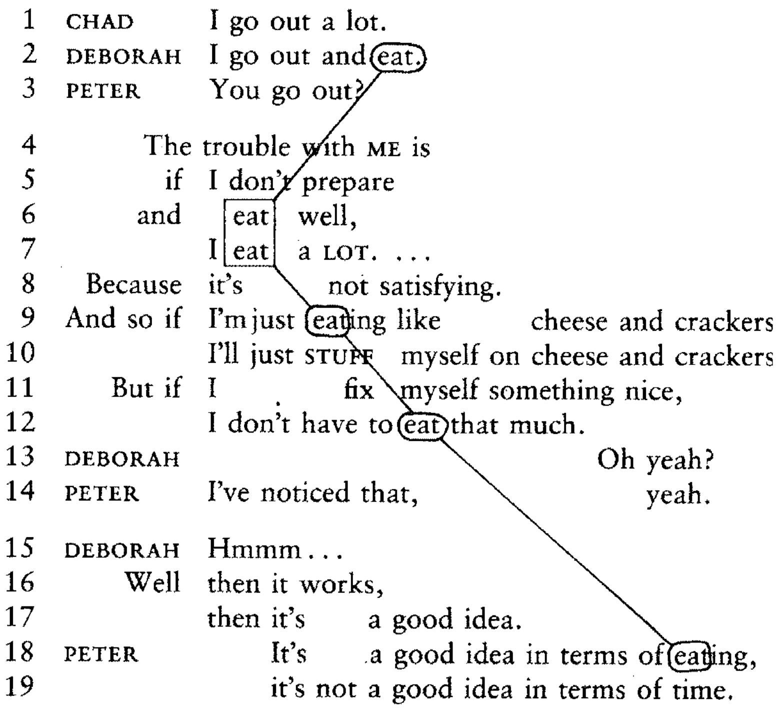
In her (1987) paper, Tannen explores lexical, syntactic, and prosodic repetition in conversation to show how repetition helps to lend a poetic quality to talk (cf. Tannen 1987: 574: "Repetition is a resource by which speakers create a discourse, a relationship, and a world"). Figure 2 exemplifies a specimen of lexical repetition in Tannen's data, and Figure 3 illustrates Tannen's notion of phonological repetition. Tannen suggests that repetition serves several functions simultaneously: production, comprehension, connection, interaction, and involvement. As for *production*,

> repetition enables a speaker to produce language in a more energy efficient, less energy-draining way. It facilitates the production of more language, more fluently ... the relative automaticity of repetition facilitates language production in conversation. (Tannen 1987: 581)

Levelt and Kelter (1982: 78) and especially Bock (1986: 379–380) have suggested similar accounts from a psycholinguistic perspective. Important along these lines is Tannen's notion of 'planning time': first, in a speaker-centered view, repetition "enables a speaker to produce fluent speech while formulating what to say next" (Tannen 1987: 582). Hence, repetition renders production less resource-demanding. Second, repetition also aids *comprehension*:

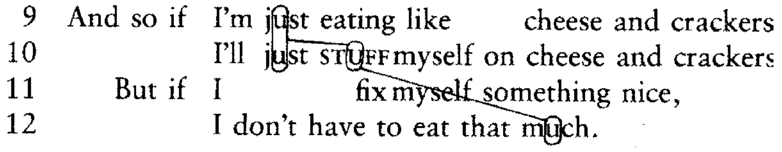
> The automatic nature of repetition and variation facilitates comprehension by providing semantically less dense discourse ... This redundancy in spoken discourse allows a hearer to receive information at roughly the rate the speaker is producing it. (Tannen 1987: 582)

Third, repetition achieves *connection* by serving a referential and tying function (cf. Halliday and Hasan 1976). "Repetition evidences a speaker's attitude, showing how it contributes to the meaning of the discourse ... repetition is a way of contributing to the rheme or comment" (Tannen 1987: 583).



1	CHAD	I go out a lot.
2	DEBORAH	I go out and eat.
3	PETER	You go out?
4	The	trouble with ME is
5	if	I don't prepare
6	and	eat well,
7		I eat a LOT. . . .
8	Because	it's not satisfying.
9	And so if	I'm just eating like cheese and crackers
10		I'll just STUFF myself on cheese and crackers
11	But if	I fix myself something nice,
12		I don't have to eat that much.
13	DEBORAH	Oh yeah?
14	PETER	I've noticed that, yeah.
15	DEBORAH	Hmmm . . .
16	Well	then it works,
17		then it's a good idea.
18	PETER	It's a good idea in terms of eating,
19		it's not a good idea in terms of time.

Figure 2. Exact lexical repetition in conversation (from Tannen 1989: 73)[5]



9	And so if	I'm just eating like cheese and crackers
10		I'll just STUFF myself on cheese and crackers
11	But if	I fix myself something nice,
12		I don't have to eat that much.

Figure 3. Phonological repetition in conversation (from Tannen 1989: 77)[6]

Fourth, on the level of *interaction*, "repetition not only ties parts of discourse to other parts, but ties participants to the discourse . . . linking speakers in a conversation" (Tannen 1987: 584). Repetition accomplishes the following interactional goals, among others: managing the floor, showing listenership, providing back-channel responses, humor and play, and ratifying another conversational party's contribution. By serving these and other functions in production, then, "repetition serves an overarching need for *interpersonal involvement*" (Tannen 1987: 584) by providing a resource (i) to create coherence, (ii) to keep the conversation going, and (iii) to connect to others, or at least to connect to what the other conversational parties have said by repeating what they have said. It is therefore not surprising, according to Tannen (1987: 601), that "some (and probably all) of conversation is also a system of pervasive parallelism," much like a poem.

Finally, mention should be made of how repetitions can be classified from a discourse-analytical perspective. Tannen (1987: 57) distinguishes the following types of repetition:

– *self-repetition* vs. *allo-repetition* (repetition of what others have said);

– *exact repetition* (the same words are repeated with the same rhythmic pattern) vs. *repetition with variation* (e.g. questions transformed into statements) vs. *free paraphrase*;

– *immediate repetition* vs. *delayed repetition*, where 'delayed' can refer to delay within a discourse or delay across days, weeks, months, and years" (Tannen 1989: 54).

In conclusion, workers in the field of discourse and conversation analysis have shown that repetition and persistence are tools by which certain interactional goals – for instance, interpersonal involvement – are achieved and by which the business of communicative, dialogic interaction is managed. Clearly, repetition affords pay-offs to both speakers (who can use repetition to aid memory and to free up planning time) and addressees (who are being involved). Conversationalists often employ repetition subconsciously and automatically, and sometimes consciously; ultimately, the mechanism of automaticity might be neurolinguistically conditioned (e.g. Tannen 1987). Tannen (1989: 95–97) also speculates that the pervasiveness of persistence in conversation may at least partly be due to the human drive to imitate, a drive which serves the purpose of learning (cf. Bock and Griffin 2000 with regard to 'implicit learning').

3. Corpus-linguistic approaches

Quantitative corpus analyses dealing with persistence – i.e. studies attempting to measure the extent of the phenomenon in corpus data – are not too numerous. Moreover, some of the studies that will be subject to review below (i.e. Poplack 1980, Weiner and Labov 1983, Poplack and Tagliamonte 1993, 1996) have stumbled across the phenomenon rather accidentally when parallelism in surface structure turned out to be a highly efficient predictor of the linguistic choices that speakers make.

In their study on "Statistical Dependence among Successive Occurrences of a Variable in Discourse," Sankoff and Laberge (1978) called attention to the fact that while it is linguistic standard practice to view successive occurrences of a variable as independent binomial trials (like two independent, unrelated throws of the dice), there may, in fact, exist interactions between such neighboring variables, depending on the syntagmatic proximity between them. To examine such effects, Sankoff and Laberge discuss three alternating variables in the Montreal French pronominal system: (i) *on*/*tu* – *vous*, (ii) *on*/*ils*, and (iii) *nous*/*on*. Their study is based on a database of 18,300 variables drawn from an unspecified corpus. Sankoff and Laberge conduct a comparatively simple analysis by straightforwardly counting variant switches between two successive variant sites and by distinguishing four types of syntagmatic proximity between two tokens (switching rate and type of syntagmatic proximity is then cross-tabulated): *embedding-constrained* tokens, where a single referent is the subject of two or more sentences, one of which is embedded in the other; *sequence-constrained* tokens, where a single referent is the subject in a sequence of two or more independent sentences which are conjoined or juxtaposed; *unconstrained* successive tokens, where two successive occurrences of the same variable are too distant to qualify for either of the above categories; and *hesitation* repetitions, where the subject variable is repeated. A *switch* of variant, as in (11) between *on* and *nous*, is more likely as the syntagmatic relationship between two tokens becomes weaker.

(11) ***On*** *a été avisé que* **nous** *étions dans une zone commerciale.*
'We were told that we were in a commercial zone.' (Sankoff and Laberge 1978: 121; translation mine)

For some sequence-constrained variables, Sankoff and Laberge (1978: 122–126) find that speakers switch about only one-third of the number of times they should if there were no dependence between two tokens. Thus (12) is a rather typical sequence-constrained succession of two identical tokens:

(12) *L'influence de la finance, c'est ça, **vous** payez tout en utilisant, **vous** en êtes pas privé.*
'That's the influence of finance: you pay while using it, you're not deprived of it.' (Sankoff and Laberge 1978: 120; translation mine)

According to Sankoff and Laberge (1978: 122), hesitation repetitions involve switching about as often as sequence-constrained variables do, while switching occurs even less often (thus, dependence between two successive tokens is even stronger) with regard to embedding-constrained variables. As for unconstrained successive tokens, Sankoff and Laberge argue that the switch rate is not significantly lower than one would expect anyway.

Although Sankoff and Laberge's (1978) method – counting switches while differentiating four types of proximity – is comparatively simple, they were among the first to conduct an empirical analysis of what impact previous variant choices can have on upcoming variant choices. Sankoff and Laberge presented evidence that this impact can indeed be sizable, depending on the syntagmatic proximity between tokens.

3.1. Poplack (1980)

Plural-marking (s) in Puerto Rican Spanish is variable. Thus, (13a) and (13b) are two equivalent ways of saying the same thing.

(13) a. *las casas bonitas*
b. *la*[ø] *casa*[ø] *bonita*__ (Poplack 1980: 61)

Poplack (1980) investigates various factors that she suspects to favor retention or deletion of the plural marker, as well as factors that help disambiguate noun phrases in case the plural marker is deleted. Using a corpus consisting of material of 18 speakers of Puerto Rican Spanish, Poplack conducted a multifactorial Variable Rule (Varbrul) analysis – the standard methodological tool in variationist sociolinguistics (see below [p. 53] for a discussion) – to determine the influence of five major factors on plural retention or deletion. The factors Poplack considered included:

Grammatical category (is the host to a potential plural marker an adjective, a noun, or a determiner?);

Nature of the following phonological element (pause, consonant, or vowel);

The nature of the following stress (weak or heavy);

Functional factors (is there morphological and/or non-morphological material available for disambiguation if the plural marker is deleted?);

Position (have plural (s) been deleted on tokens preceding the variable?).

For the purposes of the present study, let us focus on the position factor. Poplack (1980: 63) shows that "presence of a plural marker before the token favors marker retention of that token, whereas absence of a preceding marker favors deletion." The greatest effect is obtained when a marker immediately precedes the variable (the variable is then very unlikely to be realized as zero), while the scenario most favorable for marker deletion is when the markers on two preceding tokens have been deleted already: the deletion rate was over 90% when a variable was preceded by two deleted markers, and over 80% when it was preceded by only one (Poplack did not take into account tokens preceding the variable by more than two slot positions). In sum, the findings in Poplack (1980) suggest that there is a strong persistence effect operating in the retention or deletion of plural (s) in Puerto Rican Spanish.

3.2. Weiner and Labov (1983)

In 1983, Weiner and Labov conducted a study on "constraints on the agentless passive" in spoken English. Agentless passives, as in (14a), are different from agent passives, as in (14b). (14b), in turn, is another way of saying (14c).

(14) a. *The house was painted.*
b. *The house was painted by workers.*
c. *Workers painted the house.*

Agent passives as in (14b) are exceedingly rare in spoken data, according to Weiner and Labov. In contrast, "agentless passives" are comparatively common. These alternate with "generalized actives" that have 'semantically empty' pronouns – such as *they, somebody, someone, people* – as subjects. Semantically empty pronouns are [− definite, − specific] such that the referent is not known to the hearer, and the speaker does not have a particular referent in mind. (15) will illustrate the alternation between agentless passives (15a) and 'empty' actives (15b).

(15) a. *The liquor closet got broken into.*
b. *They broke into the liquor closet.* (Weiner and Labov 1983: 34)

Weiner and Labov assume rough semantic equivalence between agentless passives and generalized actives.

Using a corpus of 1489 agentless sentences – 528 (35%) of which were realized as agentless passives, and 961 (65%) as generalized actives – Weiner and Labov conduct a number of Variable Rule (Varbrul) analyses to determine the influence of several external and internal factors on the choice of generalized active or agentless passive in their corpus. The following external constraints were subject to analysis:

Careful vs. casual style. Weiner and Labov find that there is a significantly higher frequency of passives in careful style than in casual style, though they point out that in comparison to other variables, "the choice of active and passive is not an important stylistic factor in spontaneous speech" (Weiner and Labov 1983: 41).

Sex. This factor has no effect on the variable (Weiner and Labov 1983: 41).

Social class. Somewhat surprisingly, "working-class speakers use the passive significantly more than middle class" speakers (Weiner and Labov 1983: 41–42).

Age distribution. This factor has no statistically significant effect on the choice of passives over actives (Weiner and Labov 1983: 42–43).

With regard to the above external factors, which usually have massive influence on stable speaker variables, Weiner and Labov (1983) thus conclude that the passive does not seem to be a prominent sociolinguistic variable. Weiner and Labov then go on to examine the following internal constraints:

Given vs. new. Weiner and Labov define a given noun phrase as one "that has a coreferential noun phrase anywhere in the preceding five clauses" (Weiner and Labov 1983: 46). They find that if the logical object of a clause is given, it is realized as a subject of a passive construction more often than when it is new. The effect, according to Weiner and Labov, is statistically highly significant.

Parallelism in surface structure. This factor determines whether or not the logical object in a potential passive construction refers back to coreferential noun phrases in surface position. Weiner and Labov demonstrate that when coreferential noun phrases appear in subject position in preceding clauses, the logical object of an agentless clause is substantially more likely to be realized as a subject in a passive construction than when the noun phrase does not appear in subject position before. The effect is larger than the given vs. new effect and persists even when distance to the last coreferential noun phrase is taken into account.

Usage of a passive anywhere in the five preceding clauses, regardless of coreferentiality. According to Weiner and Labov (1983: 52), this is "an independent and powerful conditioning factor" – more powerful than given vs. new *or* parallelism in surface structure. Weiner and Labov take the strong showing of this factor as evidence that their assumption – that the choice of agentless passive is conditioned by syntactic, not semantic, considerations – is correct.

In summary, the authors present strong evidence that the single most powerful factor to influence the choice of actives vs. passives is repetition of previous structure. Weiner and Labov (1983: 56) conclude that "the distribution of information in discourse is not without influence, but it is a relatively minor factor compared to the more mechanical tendency to preserve parallel structure."

3.3. Estival (1985)

Building on Weiner and Labov's (1983) work on the choice between passives and actives, Estival conducted a study two years later "to isolate the effect of syntactic priming from the effect of other discourse factors" (Estival 1985: 7). What Weiner and Labov (1983) had called 'structural parallelism,' Estival (1985) renamed 'syntactic priming.'

Using a corpus of six interviews of unknown length and assuming syntactic priming to have a scope of five clauses (the criterion of Weiner and Labov 1983), Estival sets out to control the variation between actives and passives for various factors that would, as Estival argues, distort the effect of 'syntactic priming':

Logical Subject. Because Estival shows that the passive is unlikely to be used when the logical subject is one of the participants, she excludes all actives with first and second person logical subject from his data.

Repetitions. Speakers sometimes tend to use the same verb again, repeating the exact same verb from. Estival argues that such cases need to be excluded too since these would be instances of *lexical*, not *syntactic* priming. Estival finds that even if such purely lexical repetitions are excluded, there is still a significant effect of the presence or absence of a passive in the preceding context on the choice of a passive or active.

Logical object and grammatical subject. Estival shows that passives are more frequent in contexts where the subject of the passive or the object of the active is coreferential with a noun phrase in the preceding context (cf. Weiner and Labov 1983). Because this is an interaction between extraneous factors and what Estival considers the syntactic priming effect, she excludes all tokens with co-referential preceding passives from consideration. Even after this exclusion, however, there is still a significant effect of the presence of a passive in the preceding context on the choice of a passive or active.

Having thus taken into account the effect of discourse factors one-at-a-time, Estival conducts a Variable Rule (Varbrul) analysis taking into account these discourse factors simultaneously. Finding that the existence of a passive in preceding discourse has an effect on the choice between actives and passives (even if other factors favoring the choice of passives are excluded), Estival feels confident to "call the effect ... a syntactic priming effect, and to conclude that it is real" (Estival 1985: 21). This claim is almost certainly too strong. Estival has provided proof that the passive, once used, tends to persist in speech. To be sure, Estival has no evidence whatsoever that this is due to properties of the human speech production system, although it may be *speculated* that syntactic priming is involved here. But to prove this claim, Estival would have to present some independent psycholinguistic evidence which would probably have to be experimental in nature. Nonetheless, Estival's study is interesting for what it is worth: that passives, much like Weiner and Labov (1983) had shown, tend to be persistent in speech, even if other factors that favor application of the passive are accounted for.

3.4. Scherre and Naro (1991)

Scherre and Naro (1991) deal with the fact that in Brazilian Portuguese, much like in Puerto Rican Spanish, plural marking on verbs or predicate adjectives is optional and therefore variable, as illustrated in (16).

(16) a. *Os alunos não aceitaram isso*
+pl + pl +pl
'The students did not accept this'

b. *As pessoa*[ø] *nõ oide*[ø] *chegar*
+pl −pl −pl
'People cannot get there' (Scherre and Naro 1991: 23)

Using a corpus of 64 speakers and with approximately 4,800 relevant tokens (i.e. semantic plural verbs with or without morphological plural marking), Scherre and Naro set out to research structural parallelism in the variable. Somewhat poetically, Scherre and Naro suggest before they actually conduct their analysis that

> the principle that governs the real use of markers is something more like "birds of a feather flock together," that is, the more markers there are, the more likely another marker will be used; the fewer markers there are, the less likely another will be used ... Verbs that follow other verbs tend to mimic the marking of the previous utterance. (Scherre and Naro 1991: 24)

To add empirical evidence to this suggestion, Scherre and Naro classified all tokens in their database according to "whether the nearest preceding occurrence of a verb with the same plural subject was morphologically marked or not" (Scherre and Naro 1991: 24). Their dependent variable being whether or not the target verb or predicate adjective is morphologically plural-marked, Scherre and Naro first examine parallel effects at the discourse level (i.e., in the wider preceding context) and set up an independent variable with three factors:

1. The verb or predicate adjective is preceded by a marked verb or predicate adjective within the preceding 10 clauses;

2. The verb or predicate adjective is preceded by an unmarked verb or predicate adjective within the preceding 10 clauses;

3. The verb or predicate adjective is the first occurrence in the text, or the verb or predicate adjective is isolated, hence neither 1. nor 2. apply.

Scherre and Naro then perform various Variable Rule (Varbrul) analyses and demonstrate that the presence of a morphologically marked plural verb or predicate adjective in the preceding discourse context makes it a lot more likely that the target token will be morphologically plural marked as well. This effect holds regardless of whether (i) only tokens with an explicit plural subject are included in the analysis, (ii) only tokens with an explicit plural subject *or* explicit plural marking are included in the analysis, or (iii) all tokens are included in the analysis. Scherre and Naro also analyze parallel effects at the clausal level (i.e. in the immediately preceding syntagm). As for the marking of verbs, Scherre and Naro (1991: 28) postulate that "the marking of the last element in the noun phrase would spread to the verb." Indeed, there is evidence that there are significant parallel effects in this context, though these effects are slightly weaker at the clausal level than at the discourse level. As for predicate adjectives, the effects are quite similar, although marking of verbs, which are the nearest elements to the predicate adjective, exerts a stronger influence on the morphological marking of the predicate adjective than the marking of subjects, which are more distant.

In their conclusion, Scherre and Naro (1991: 30) make three points: First, the observed persistence effects sometimes contradict the functional principle of economy in language usage because "markers tend to occur precisely when they are not needed and tend not to occur when they would be useful" (Scherre and Naro 1991: 30). Second, much as Sankoff and Laberge (1978) have argued, successive occurrences of a variable should not be considered independent binomial tries. And third, Scherre and Naro argue that formal parallelism has been shown to be operative in so many phenomena and languages that "it should be considered a serious candidate for a universal of language use and processing" (Scherre and Naro 1991: 30).

3.5. Poplack and Tagliamonte (1993, 1996)

The subject of Poplack and Tagliamonte (1993, 1996) is past tense verb phrase marking. In Poplack and Tagliamonte (1996), the authors take a variationist approach to grammaticalization, investigating the past temporal reference sector of Nigerian Pidgin English. In Nigerian Pidgin English, there are six different past time reference markers, and it is the variation between them that is the subject of Poplack and Tagliamonte's study. Drawing on a corpus of informal conversations among 12 Nigerians that contains 4,759 verbal structures referring to the past, Poplack and Tagliamonte investigate the factors that have an impact on which form is selected to mark past time reference. These factors include the temporal relationship, temporal distance, lexical stativity, temporal disambiguation, negation, and the mark on the preceding verb. Poplack and Tagliamonte conduct six independent Variable Rule (Varbrul) analyses to pinpoint the contribution of the above six factors to the likelihood that each of the six candidate markers will be selected in a given past time reference context (Poplack and Tagliamonte 1996: 80). In a nutshell, they find that "the strongest predictor that each [of the candidate markers, BS] will be selected ... is after a verb on which it has already occurred" (Poplack and Tagliamonte 1996: 82). In a similar vein, Poplack and Tagliamonte (1993) examine the past temporal reference system in two corpora of "early" Black English (Samaná and the Ex-slave Recordings). Again, past time reference in these varieties is variable in that it may or may not be overtly marked on the verb. The factors that Poplack and Tagliamonte suspect influence the presence or absence of a marker on a verb are stativity/anteriority, discourse context, presence of temporal conjunctions, and, again, the mark on the preceding reference verb. After a series of Variable Rule (Varbrul) analyses on these factors, Poplack and Tagliamonte (1993) conclude that a "recurrent effect is contributed by the existence of a mark on a preceding reference verb ... a preceding mark increases the probability of a mark on the current verb, while a preceding zero leads to more zeros" (Poplack and Tagliamonte 1993: 197).

While Poplack and Tagliamonte (1993, 1996) do not have much to say about persistence in particular (after all, it was not one of their research questions), their research demonstrates that persistence can be observed in verb phrase marking.

3.6. Gries (2005)

Gries (2005) is a systematic, cutting edge corpus study of syntactic persistence. Published in the *Journal of Psycholinguistic Research*, the paper contributes to a methodological conversation in the psycholinguistic literature by arguing, *pace* Branigan et al. (1995), that syntactic priming as a psychological phenomenon can, in fact, be appropriately investigated on the basis of naturalistic corpus data, as opposed to experimental data. At the same time, Gries submits that previous psycholinguistic research has failed to take into account that some verbs may be more resistant or more responsive targets than other verbs thanks to idiosyncratic verbal associations with particular constructions. Other research questions addressed in the paper include the following: is priming long-lasting, or rather short-lived? Does the effect operate from comprehension to production as well as from production to production? How do morphosyntactic and lexical characteristics of the primes and targets impact effect size? Do corpus data show differences in effect size between alternations, and, given a specific alternation, does option A prime better than option B?

As case studies, Gries investigates the dative alternation, as in (6) (p. 17) above, and particle placement, as in (17) below, in the 1 million words British component of the International Corpus of English (ICE-GB):

(17) a. *John picked up the book.*
b. *John picked the book up*. (Gries 2005: 381)

As for the dative alternation, Gries extracted 3,003 prime-target pairs (i.e. textually subsequent constructions) from the corpus; the corresponding *N* for his investigation into particle placement was 1,797. This sizable database was then coded for the following independent variables:

Prime-target match. Does the target match the prime's dative construction (prepositional/ditransitive) or the prime's particle placement pattern?

Medium. Is the medium spoken or written (note that ICE-GB samples both registers)?

Textual distance between the prime and target, conceptualized as a variable with discrete categories.

Lemma and form matches. Does the target match the prime's verb lemma and/or grammatical form?

Same speaker. Is it the same speaker who produces both the prime and the target?

Verb type. Which specific verb is used in the prime and target?

Gries utilizes multivariate analysis methods to show that with regard to syntactic priming, “the general findings concerning particle placement are somewhat similar to those of the dative alternation” (Gries 2005: 382). In short, (i) there is a general priming effect; (ii) in both alternations, one of the two alternative constructions primes better than the other; (iii) verb form and verb lemma matches tend to enhance (albeit statistically insignificantly) the effect; (iv) in particle placement, the medium (spoken vs. written) has some effect on priming; (v) in the dative alternation, production-to-production priming is stronger than comprehension-to-production priming; and (vi), textual distance between the prime and the target does not significantly interact with the strength of the priming effect, although it turns out that priming is rather long-lasting and that the variable fares better if it is modeled logarithmically or quadratically rather than linearly. All of the above dovetails rather nicely with previous psycholinguistic research. Gries’ results with regard to the verb specificity of priming effects, however, strongly suggests that previous experimental research should have taken into account this factor. To illustrate this issue: the verb *give* is shown to strongly prefer the ditransitive, as in (18a), while the verb *sell* is strongly associated with the prepositional dative, as in (18b):

(18) a. *John gives Mary the book*
b. *John sells the book to Mary*

Gries’ point is that the verb *give* will resist prepositional dative priming, while the verb *sell* will be a comparatively ‘easy’ target for prepositional dative priming. This means that experiments that fail to control for such distinctive verbal construction preferences – and previous psycholinguistic studies have failed to do so – may obtain flawed effect sizes; ideally, experimenters will only want to include verbs that do not have a marked preference for either option. According to Gries (2005), therefore, corpus-based research is not merely an appropriate methodology to investigate syntactic priming – as a matter of fact, corpus study can fruitfully complement other approaches by adducing evidence that would otherwise be hard to come by (Gries 2005: 391).

4. A rival empirical phenomenon: *horror aequi*

To conclude this chapter, mention should be made of a principle that – at least at first glance – predicts the exact opposite of persistence of structure: that there is a "widespread (and presumably universal) tendency to avoid the use of formally (near-) identical and (near-) adjacent (non-coordinate) grammatical elements or structures" (Rohdenburg 2003: 236) when these are semantically unmotivated. Thus, according to Rohdenburg and Schlüter (2000), (19) is dispreferred:

(19) ?*She looked upon this solution* ***as as*** *good* ***as*** *that one* (Rohdenburg and Schlüter 2000: 466; emphases original)

The principle thought responsible for the unacceptability of (19) has been called the *horror aequi* principle (cf. Brugmann 1909, who is usually credited with having coined the term; Rohdenburg 1995, 1996; Rohdenburg and Schlüter 2000; Rohdenburg 2003; Mair 2003; Mondorf 2003; Vosberg 2003). *Horror aequi* is assumed to operate below the threshold of consciousness (Vosberg 2003: 321) and can manifest itself in two forms, strong and weak (cf. Mair 2003: fn. 2). In its weak form, it is not thought of as a hard grammatical constraint, but as one that can be weakened by a number of other factors (cf. Vosberg 2003: 321), such as the existence of intervening material (Rohdenburg 1995: 376), interaction with the so-called *complexity principle* (Rohdenburg 1995: 368; Rohdenburg 1996: 252; Rohdenburg 2003), or contexts of negation (Rohdenburg 1995: 378). Mondorf (2003: 279) speculates that ultimately, *horror aequi* is due to "the tendency to inhibit reactivation of neurons within a given time span in order to create refractory phrases and ...the tendency to create sufficiently distinct adjacent elements to facilitate recognition and processing."

Horror aequi is argued to be responsible, for instance, for the existence of the so-called Doubl-*ing* filter (see Ross 1972; Milsark 1988). In generative terms, there appears to exist a surface-structure constraint in English such that it is not acceptable "for a complement (as opposed to an object) marked with gerund participle inflection to be adjacent to its marked matrix clause verb when that verb is likewise in the gerund participle form" (Pullum and Zwicky 1999: 269). Hence, Pullum and Zwicky consider (20) ungrammatical:

(20) **Terry was starting reading aloud.* (Pullum and Zwicky 1999: 252)

In a similar vein, speakers of English avoid sequences of non-coordinated *to*-infinitives, in particular if they are not separated by intervening material (Gramley 1980; Rohdenburg and Schlüter 2000; Mair 2003; Vosberg 2003).

Further implications of *horror aequi* have been studied in a series of recent studies, many of them stemming from the Paderborn research project on determinants of grammatical variation in English. Fanego (1996a, 1996b), Rohdenburg (1995) and Rohdenburg (1998) have shown how the evolution of a number of complement structures has helped English to avoid identity phenomena. Rohdenburg and Schlüter (2000) present evidence how *horror aequi* can weaken or even neutralize conflicting pressures. Their examples include the replacement of the predicative marker *as* by *be* and the replacement of *to*-infinitives by *ing*-complements after a number of verbs. Surveying dependent interrogative clauses in the history of English, Rohdenburg (2003) demonstrates how the so-called complexity principle and *horror aequi* can interact such that both receding constructions tend to be preserved longer and incoming constructions become established faster in some functional niches. Mondorf (2003), in her study on factors determining the choice between analytic and synthetic comparison, shows empirically that adjectives ending in *-r* and *-re* are unlikely to take synthetic comparatives (for instance, *securer*) and that adjectives ending in *-st* are unlikely to take synthetic superlatives (for instance, *unjustest*). This is, according to Mondorf (2003), due to *horror aequi* and the pressure to avoid haplology.

5. Summary

In conclusion, language (especially spoken language) is inertial, repetitive, and characterized by a perseverance of linguistic patterns which can be lexical, phonological, morphological, or syntactic in nature. We have referred to this tendency as *persistence*.

Analytically, persistence may have a variety of underlying causes, a rough sketch of which is presented in Figure 4. Psycholinguists have sought to explain persistence by properties of the human language production system, which – under certain circumstances – is designed to be efficient at mechanically replicating already activated production or processing patterns from previous discourse. The system is comparatively less efficient when it has to activate new patterns from scratch. Thus, when there are alternatives, speakers are hard-wired to go for the already activated option. Discourse and con-

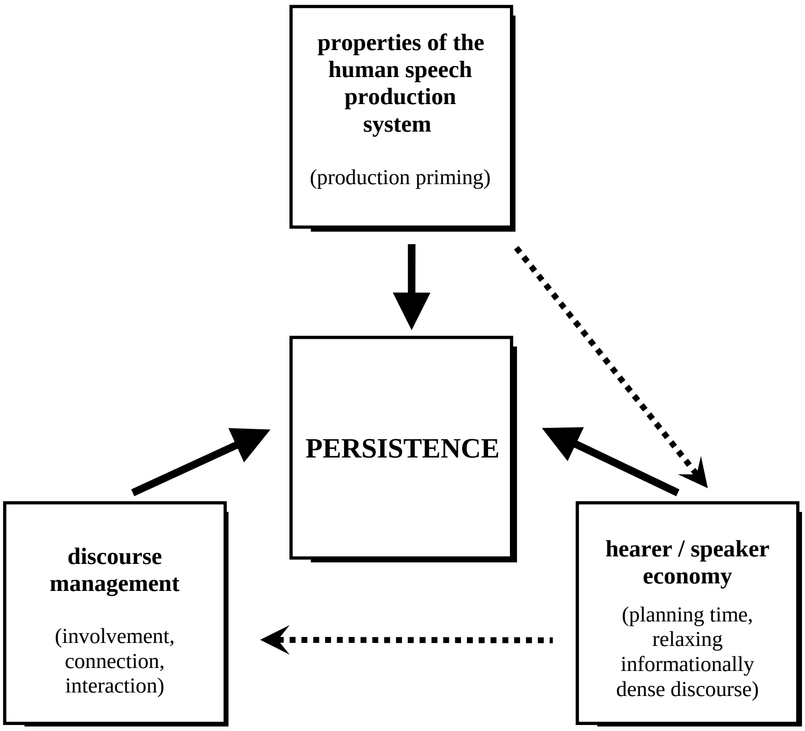

properties of the human speech production system
(production priming)
PERSISTENCE
discourse management
(involvement, connection, interaction)
hearer / speaker economy
(planning time, relaxing informationally dense discourse)

Figure 4. Factors contributing to the genesis of persistence phenomena

versation analysis have stressed the functional aspects of persistence, namely communicative goals and discourse management tasks (involvement, connection, interaction) which conversationalists can accomplish by being repetitive.

Both the discourse-analytic and the psycholinguistic literature agree that the redundancy provided by repetitiveness is speaker-economical in that it provides for planning time (e.g. Tannen 1987) and in that it makes speech more fluent (e.g. Levelt and Kelter 1982). By the same token, it is hearer-economical in that persistence reduces the processing load associated with informationally or computationally dense discourse (e.g. Tannen 1987, Branigan, Pickering, and Cleland 2000). Corpus linguists (rather than discourse analysts) have begun to contribute to the study of persistence by descriptively quantifying the effect persistence has on some variables in corpus data.

Thus, three root causes of persistence can be recognized: *properties of the human speech production and processing system*, accomplishment of *discourse management functions*, and maintaining *speaker-hearer economy*. These constituting factors are indicated by the bold arrows in Figure 4. At the same time, there are interactions between the factors themselves. For one thing, that our speech production and processing system is designed to mechanically replicate discourse-old patterns may be the very reason that persistence is also speaker-hearer economical: "Reusing recent materials may . . . be more economical than regenerating speech anew from a semantic base, and thus contribute to fluency" (Levelt and Kelter 1982: 78). In turn, the fact that persistence can be exploited for discourse management tasks – for instance, in order to create interpersonal involvement – is presumably also due to the fact that persistence is speaker-hearer economical. These interactions between the explanatory factors are indicated as dotted arrows in Figure 4.

We have also seen that there is a phenomenon, known as *horror aequi*, that is – at least at first glance – diametrically opposed to persistence, namely that identical linguistic patterns are sometimes dispreferred in adjacent position.

Chapter 3
Method and data

This chapter is designed to give an overview of the methodological and empirical foundations of the present study. I shall first spell out which grammatical alternations will be analyzed as dependent variables, and with regard to which factors (i.e. independent variables) these alternations are going to be investigated. I will then explain how the present study's database was extracted and coded, and which statistical methods were used in the analysis. Finally, I shall present the corpora that constitute the primary data sources of the present study.

1. Method

The main method that is utilized in the present study closely resembles the so-called Variable-Rules approach (cf. Sankoff and Labov 1979). The method provides a more or less theory-neutral heuristic tool of analysis which integrates probabilistic statements into the description of performance. It is applicable "wherever a choice can be perceived as having been made in the course of linguistic performance" (Sankoff 1998: 151; cf. Halliday 1991 for an overview of probabilistic grammar in corpus studies). In short, it will be the present study's job to quantify, probabilistically, how persistence impacts choice making in linguistic performance.

1.1. Dependent variables

In a variationist-probabilistic approach, the loci where persistence can be investigated are those identifiable occasions in the data where speakers demonstrably have the choice between using one variant or another. The notion of 'choice' implies that there is rough semantic equivalence (Labov 1969) between the patterns. This study, hence, will research persistence by investigating, as dependent variables, alternations between well-defined variant forms of variables that have been shown to be roughly equivalent semantically in previous scholarship. The following five alternations are among the most extensively researched in English and will serve as case studies for the analysis of persistence:

1. Comparison strategy choice: analytic vs. synthetic comparatives

(1) a. *Mary is **more ready** to do whatever she wants than Jim.*
b. *Mary is **readier** to do whatever she wants than Jim.*

2. Genitive choice: *s*-genitives vs. *of*-genitives

(2) a. *The **university's** budget is considerable.*
b. *The budget **of the university** is considerable.*

3. Future marker choice: BE GOING TO vs. WILL

(3) a. *Mary **is going to** talk to Jim.*
b. *Mary **will** talk to Jim.*

4. Particle placement: *V+NP+Part* vs. *V+Part+NP*

(4) a. *Mary **looked** the word **up**.*
b. *Mary **looked up** the word.*

5. Complementation strategy choice: *V+ger.* vs. *V+inf.*

(5) a. *Mary **began wondering** where Jim was.*
b. *Mary **began to wonder** where Jim was.*

There is near universal consensus that there is little, if any, semantic difference between any of the above (a) or (b) options in the vast majority of contexts. At any rate, they do not differ in truth conditions. Detailed discussions of the above alternations and their determinants as suggested in previous research will be provided in the respective empirical chapters.

1.2. Factors and independent variables

The present study assumes that for any given slot or variable (i.e. the variable under analysis, henceforth: CURRENT) in which either of two options can be employed, the likelihood for either option to be used is a function of several factors and factor groups, one of which is persistence.

1.2.1. 'Conventional' predictors

Crucially, the statistical models in this study will not only include persistence factors, but also variables meant to tap the major 'conventional' factors known to play a role in the respective alternations (these will be listed and discussed in the respective sections below). These may either probabilistically favor a given option, or disfavor it. For instance, in comparison strategy choice, the length of the adjective to take comparison will be considered (longer adjectives are known to prefer analytic comparison); in particle placement, the length of the direct object will be included as a factor (heavy objects tend to be postponed); and so on. The reason is that in order to be able to state anything of interest about the magnitude of persistence, it is necessary to relate its explanatory power to factors that have hitherto been claimed to influence the alternations under investigation in this study; otherwise, we would not know exactly how much consideration of persistence improves our ability to explain linguistic variation. Inclusion of such factors is also necessary since this procedure minimizes the likelihood that what appears to be persistence is actually a statistically spurious artifact of some other factor not included. Much like experimental researchers seek to control their research designs for various secondary factors, inclusion of baseline predictors in this study's corpus-based designs is meant to control for secondary intralinguistic factors, or, in short, for what I will refer to as 'baseline variation'; the factors and predictors that are responsible for this variation will correspondingly be referred to as 'baseline predictors.'

It is important for the reader to keep in mind, however, that this study will not for a moment claim to have explained any one of the alternations exhaustively. Rather, the point will be to sketch how persistence-related factors can constructively complement 'traditional' factors, and that a portion of what has been traditionally thought to be 'free' variation is actually not so free at all but governed by persistence.

1.2.2. Persistence-related predictors

The following intralinguistic factors or predictors will be standardly considered. All of them fall into the domain of α-persistence (β-persistence is too alternation-specific to be tapped by a generic variable).

WHICH VARIANT has been employed in the variable preceding CURRENT? (henceforth: PREVIOUS). Given two successive choice contexts in discourse, was the first one (if there was one – the only scenario where no such discourse-preceding variable can exist is if the given variable is the first one in the text under analysis) realized as in CURRENT or was the alternative option used? This is the most basic predictor of α-persistence. Consider (6), where there is a match between PREVIOUS (. . . *Matt'll* . . .) and CURRENT (. . . *we'll* . . .) with regard to the future marker chosen:

(6) *Matt'**ll** find this out, and, I mean, we'**ll** get involved in it* (CSAE 0906)

Hypothesis: Use of a given option in PREVIOUS increases the likelihood that the same option will be used in CURRENT.

TEXTUAL DISTANCE between CURRENT and PREVIOUS, i.e. between two choice contexts (henceforth: TEXTDIST). Previous corpus studies have found that the tendency towards surface parallelism weakens with increasing textual distance between two subsequent variables (for instance, Sankoff and Laberge 1978 and, with certain qualifications, Gries 2005). Experimental studies on priming have reported similar effects: in Branigan, Pickering, and Cleland (1999), for instance, priming weakened considerably when a fragment intervened between prime and target, and dissipated entirely when four fragments intervened (it is worth noting that in most experimental studies, intervening fillers are different constructions, whereas in corpus studies, material between two sites can be basically anything). TEXTDIST will be measured in the *natural logarithm* (henceforth: *ln*) of the number of interjacent words between PREVIOUS and CURRENT and is a proxy for recency of use of an alternating variable. The reason that this variable is going to be modeled logarithmically and not, say, in a linear fashion is that many psycholinguistic priming phenomena have been shown to decline this way; 'forgetting' functions are rarely linear (see, e.g. Cohen and Dehaene 1998 with regard to inappropriate repetitions due to brain damage; McKone 1995 with regard to decreasing exponential decay of repetition priming; Gries 2005 with regard to syntactic priming in corpus data). For illustration, consider again (6): textual distance between the two slots in this utterance is seven words (*. . . find this out, and, I mean, we . . .*), thus TEXTDIST would be *ln* 7 = 1.95.

Hypothesis: TEXTDIST interacts with PREVIOUS such that persistence effects are stronger if TEXTDIST is small.

LENGTH OF THE SENTENCE (in words) where the variable under analysis is embedded (henceforth: SENTENCELENGTH[7]). Sentence length will be taken to be a proxy for syntactic complexity of the environment where CURRENT is embedded (see Szmrecsanyi 2004 for a discussion of this method).

Hypothesis 1: If the two options employable in CURRENT differ in explicitness, higher syntactic complexity will make it more likely that the more explicit option will be used (cf. Rohdenburg 1996: 151).

Hypothesis 2: The longer SENTENCELENGTH, and hence, the higher syntactic complexity of the context where CURRENT is embedded, the more potent online processing constraints are (cf. Bortfeld et al. 2001: 141 on why disfluency increases as heavier demands are placed on the speech planning system) and hence, the more potent persistence is because its facilitative effect on online processing is exploited by speakers (cf. Tannen 1987, 1989; Branigan, Pickering, and Cleland 2000).

TYPE-TOKEN RATIO of the lexical environment where the variable under analysis is embedded (henceforth: TTR). TTR will be considered a proxy for lexical density. 'Lexical environment' refers to a textual context of 50 words before and 50 words after CURRENT.

Hypothesis 1: If the two options employable in CURRENT differ in explicitness, higher lexical complexity will make it more likely that the more explicit option will be used (cf. Rohdenburg 1996: 151).

Hypothesis 2: Similarly to SENTENCELENGTH, persistence is more powerful when TTR, and hence lexical density of the context where CURRENT is embedded, is high (cf. Tannen 1987 on why parallel patterns might be preferred in lexically dense contexts because of processing efficiency advantages).

TURN-TAKING. Was PREVIOUS in the same conversational turn as CURRENT (henceforth: SAMETURN), and was it produced by the same conversational interactant that produced CURRENT (henceforth: SAMESPEAKER)? These binary independents are about whether the effect size of persistence is sensitive to turn-taking (coded 1 if PREVIOUS and CURRENT are in the same turn [SAMETURN], or if PREVIOUS and CURRENT are produced by the same speaker [SAMESPEAKER], and 0 otherwise). (7) illustrates a case where two successive future marker choice contexts (*... it gonna be ...* and *... we'll do ...*), though successive in discourse, are neither located in the same conversational turn, nor are they produced by the same speaker (hence, both SAMETURN and SAMESPEAKER would be coded 0):

(7) JOE: *Or is it* ***gonna*** *be passthrough funds here at the bank? ...*
JIM: *Well, what we'****ll*** *do is ...* (CSAE 0906)

Hypothesis: Given previous research (for instance, Gries 2005), it is reasonable to expect that SAMETURN and SAMESPEAKER interact with PREVIOUS such that persistence within turns is stronger than persistence across turns, and that persistence is stronger when speakers repeat themselves than when they repeat what other parties have said.

In addition, several other persistence predictors tailored to the alternations studied will be introduced in the respective empirical chapters – for instance, predictors exploiting lexical effects (cf. Pickering and Branigan 1998) or predictors relating to β-persistence (e.g., in comparison strategy choice, does the token *more* trigger analytic comparison?).

1.2.3. Speaker characteristics

Whenever the demographically sampled – and sociologically annotated – section of the British National Corpus is analyzed (this will be the case for comparison strategy choice, future marker choice, and complementation strategy choice),[8] the following extralinguistic factors will be included in the analysis without any *a priori* hypotheses:

SPEAKER AGE (henceforth: AGE). Although priming effects appear to show comparatively little variation with age (Rastle and Burke 1996: 586), some experimental studies have found that elderly adults show greater priming effects than younger adults (for instance, Friederici, Schriefers, and Lindenberger 1998; Laver and Burke 1993).

SPEAKER SEX (henceforth: SEX). To the best of my knowledge, there is no major psycholinguistic evidence that priming effects are different between the sexes.

Appendix A illustrates, with the help of a concrete example, how the variables introduced in this section have been coded in practice.

1.3. Data extraction

In a first step, the relevant choice contexts were identified in the data. Where possible, this identification was performed automatically using Perl (*Practical Extraction and Report Language*, a programming language intended for text manipulation) scripts which parsed the data and identified and extracted the relevant sites automatically. When the patterns under analysis were too complex to be dealt with by software (such as, for instance, the alternation between the *of*-genitive and the *s*-genitive), or when nonexistent POS-tagging of the data source (as is the case for the CSAE and FRED) made automatic analysis impossible, extraction was performed manually, in which case coding protocols will be provided.

1.4. Coding

In a second step, the extracted alternation sites were coded for the independent or predictor variables (both persistence-related and persistence-unrelated ones). Again, wherever possible this was performed automatically by Perl scripts designed to analyze the linguistic environment of the choice contexts. For instance, all the factors listed in section 1.2 were coded automatically. A number of factors too complex to be coded by software – for instance, information status or variables that require POS-tagging for automatic analysis – were coded manually. More specific information will be given in the relevant empirical chapters below.

1.5. Statistical analysis

In a third step, the database was analyzed statistically to investigate the effect of persistence. Previous corpus-based studies approaching persistence quantitatively have used basically two types of statistical methods:

Determining switch rates. This is a comparatively straightforward method (cf. Sankoff and Laberge 1978 and Gries 2005). The basic idea is this: Given that two patterns A and B are alternating, it is determined whether switches from A to B and switches from B to A are less frequent than one would expect given the overall distribution of A and B patterns in the data. If a χ^2 test shows that there are significantly fewer switches than in a random distribution, this is taken to be evidence for persistence.

Multifactorial analyses. Poplack (1980), Weiner and Labov (1983), Estival (1985), Scherre and Naro (1991), and Poplack and Tagliamonte (1993, 1996) conducted multifactorial Variable Rule (Varbrul) analyses (see below [p. 53]) in which they included persistence as one factor among others. In this approach, if the Varbrul analysis shows that the realization of discourse-preceding variable sites has a significant effect on how the dependent variable under analysis is realized, persistence is (i) shown to be operative in the data and (ii) can be precisely quantified by means of probabilistic weights.

1.5.1. Switch rates

There are several limitations inherent in the switch rate method, the most important being that it is all but impossible to take into account other factors besides the sequential occurrence of A and B variants. For instance, textual distance between two occurrences of a variable – or any other of the factors listed in section 1.2 – can hardly be considered. Another issue is that persistence cannot be really quantified through this method, i.e. the relative power of the effect cannot be assessed straightforwardly. Yet, switch rate analyses are appealing because the result is relatively easy to visualize. Sankoff and Laberge (1978) and Gries (2005), for instance, visualize their findings with graphs such as the ones in Figure 5. These scatterplots display sequence-constrained switch rates as a function of variant proportion – more technically, they plot the relative frequency of switches from some variant B to some variant A, in relation to the total number of occurrences of variant A (in %, on the *y*-axis) against the share of A occurrences of the total number of A and B occurrences (in %, on the *x*-axis). Because what is really at issue here is the sequential configuration of the variable sites, scatterplots in the spirit of Figure 5 really display α-persistence.

How does it work? Assume a text with 20 variable slots where each slot may be realized by either linguistic variant A or linguistic variant B; for simplicity, let us also presume that in total, the text contains ten occurrences of variant A and ten occurrences of variant B. Let us further suppose three speakers – speaker 1, speaker 2, and speaker 3 – who, given the same text, choose to arrange variants A and B differently. After extracting all variables from each speaker's text and lining them up sequentially, the matrix in (8) emerges:

(8) Speaker 1: B B B B B B B B B B A A A A A A A A A A
Speaker 2: A B A B A B A B A B A B A B A B A B A B
Speaker 3: B B A A B B A A B B A A B B A A B B A A

It is intuitively evident that the speakers have rather different repetitiveness preferences. Speaker 1 is fairly persistent: her text exhibits only one switch, from B to A, in the middle of the text. Speaker 2, by stark contrast, appears to have a predilection for switching between linguistic variants, doing so basically at every opportunity. Speaker 3 seems to take the middle road. Figure 5, then, visualizes these differences: for all speakers, the distribution of A and B variants is 50:50, i.e. 50% (*x*-axis). However, speaker 1 switches from B to A only once while using ten A variants in all, hence her *y*-value is ${}^{1}/_{10} = 10\%$. Speaker 2 switches nine times and uses ten A variants, therefore her *y*-value is ${}^{9}/_{10} = 90\%$. Speaker 3 switches five times exhibiting a total of ten A variants, so speaker 3's *y*-value is ${}^{5}/_{10} = 50\%$. Note, now, that only speaker 3 matches

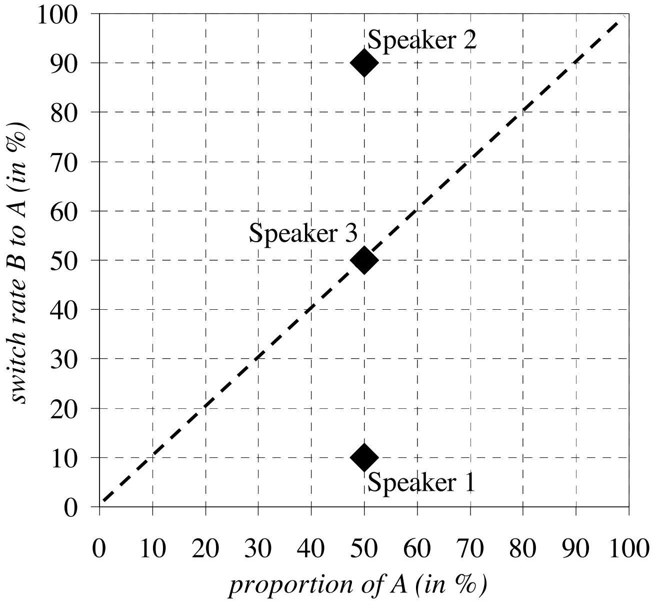


Figure 5. Sequence-constrained switches as a function of variant proportion (cf. Sankoff and Laberge 1978). Relative frequency of switches from B to A, in relation to the total number of A occurrences (in %), on *y*-axis; relative frequency of A occurrences (as percentage of all A and B occurrences) on *x*-axis. Dotted diagonal line represents null hypothesis that switch rate is proportional to variant proportions

the distribution of A and B variants to her switch rate – that is, speaker 3 is neither particularly persistent nor is she particularly non-persistent. In Figure 5, she is thus located right on the dotted diagonal line, which represents the null hypothesis that switch rates are proportional to the overall distribution of switched-to variants (A variants, in our case). Speaker 2's switch rate substantially exceeds the 50% threshold, which is why speaker 2 is highly non-repetitive. Speaker 3 switches considerably less often than one would expect given her overall distribution of variants – in other words, speaker 3 is highly persistent. The idea is that when aggregating data for many speakers or texts, dots in switch rate diagrams will cluster *below* the diagonal line when persistence effects are operative. The more the dots do cluster below the diagonal line, the more potent persistence is in the data. The present study will consistently make use of such switch rate scatterplots to provide readers with a first visual impression of the magnitude of α-persistence in the data. Note that since these diagrams are sensitive to intra-speaker persistence (i.e., self-repetition or production-to-production priming) only, they present a rather conservative estimate of α-persistence.

1.5.2. Logistic regression

However, for the main part, the name of the game in the present study will be multifactorial analysis – in particular binary logistic regression – rather than simple χ^2 tests for independence of switch rates. Logistic regression has the following advantages over traditional, univariate statistical analysis methods:

- logistic regression predicts an outcome (i.e. a linguistic choice) given several independent (or predictor) variables;
- it quantifies the influence of each predictor;
- it specifies the direction of the effect in which each predictor runs;
- it states how much of the empirically observable variation has been accounted for by all of the predictors considered;
- it specifies how well the model fares in predicting actual linguistic choices.

The Variable Rule (Varbrul) analyses familiar from the sociolinguistic literature (cf. Cedergren and Sankoff 1974) are basically an application of binary logistic regression. However, due to limitations of the Varbrul package, the present study will rely on the much more powerful logistic regression module of the SPSS package.[9]

In most simple terms, logistic regression models estimate which of two outcomes is more likely to occur given that one or more independent variables (which may be scalar, categorical, or both) influence the outcome. In this spirit, this study's analyses will seek to investigate how usage of linguistic option A in a given slot will influence the odds that linguistic option B will be used next time there is a choice. In other words, the basic question that will guide this study's logistic regression analyses is the following: How much does usage of a specific option discourage usage of the alternative option at the next opportunity?

A brief example may help illustrate the basic idea behind logistic regression: in New York City, class membership and gender (let's say these are the independent variables, both of which are dichotomous: middle-class vs. non-middle-class and male vs. female) influence the odds that a given speaker pronounces postvocalic *r* (this is the binary outcome, pronunciation of postvocalic *r* vs. non-pronunciation of postvocalic *r*). Now, a logistic regression model based on data from many speakers may estimate middle-class female speakers will pronounce postvocalic *r* with a probability of 70% ($p = 0.7$). As this probability exceeds the threshold (or cut value) of 50% (meaning that an outcome is more likely to occur than not to occur), the model thus actually predicts that if a speaker is middle-class and female, the outcome 'pronunciation of postvocalic *r*' will occur. In addition, the regression may indicate that being female increases the odds for postvocalic *r* pronunciation by 15%, while a middle class background increases the odds for postvocalic *r* pronunciation by 25%.[10] Logistic regression models thus actually classify cases – their goal is prediction of a single dichotomous dependent variable. In the present study, the job of logistic regression will be to predict a choice between two linguistic alternatives given (among many other variables) which option was chosen last time there was a choice.

Logistic regression rests on a number of assumptions, one of which is that there be no strong correlation *between* the independent variables. If this condition is not satisfied, the result is what is known as multicollinearity, which leads to unstable and unreliable estimates. Multicollinearity measures for all independent variables analyzed in the present study are reported in Appendix B – on the whole, multicollinearity is not a cause for concern in terms of the variables analyzed in the present study.

Key notions in logistic regression

Once a logistic regression model has been estimated, its validity must be compared to the original dataset from which it has been estimated to assess its quality. The present study will rely primarily on three criteria to assess logistic regression models:

1. *Predictive efficiency*. How well does the model work? How accurate is it in classifying the data? To answer these questions, the present study will report the percentage of correctly predicted cases vis-à-vis the baseline

prediction (*% correct (baseline)*). The percentage of correctly predicted cases indicates how many of the actual outcomes in the input data are correctly predicted by the model. A percentage of 80%, for instance, indicates that the model is successful at predicting outcomes 80% of the time. The baseline prediction indicates how a restricted, constant-only ('dumb') model would fare on the same dataset. By way of illustration, let us return to the example above and assume that according to the input data, people – regardless of class or gender – pronounce postvocalic *r* 55% of the time. Now, a restricted model would simply predict that all people are postvocalic *r* pronouncers, and would be correct 55% of the time. The measure thus indicates the improvement we achieve through an 'intelligent' model above and beyond a 'dumb' model.

2. *Variance explained.* Is the relationship between the independent variables in the model and the outcome strong enough for us to be interested in it? That is, is the model *substantially* significant? Unlike most multivariate linguistic research, this study will report *Nagelkerke's* R^2 values to answer these questions.[11] This measure approximates (there are certain issues with R^2 in logistic regression) what percentage of the variation in the dependent variable is accounted for by all included independent variables. R^2 can vary between 0 and 1, with an R^2 of 0 indicating that there is no relationship whatsoever between the independent and dependent variables, and an R^2 of 1 indicating that the model accounts for all of the variance in the dependent variable. Therefore, for instance, a R^2 of 0.8 indicates that roughly 80% of all the observed variation is accounted for by the model, regardless how good or how bad it fares at actually predicting outcomes. Usually, a model is considered *substantially significant* if $R^2 \geq 0.05$, that is, if the independents included explain at least 5% of the observable variance.

3. *Contribution of the independent variables.* If the overall model works well, how important is each of the independent variables? Which ones influence the dependent variable in a statistically significant way, and which ones are better or worse predictors of the dependent variable? To deal with these questions, this study will report *odds ratios* (also referred to as *exb*(b) values). An odds ratio is the number by which we would multiply the odds of an event occurring for each one-unit increase in the independent variable (for scalar independents), or for a categorical cod-

> ing of the independent (for categorical independents). Odds ratios have a lower boundary of 0, but no upper ceiling. An odds ratio smaller than 1 indicates that the odds of an outcome occurring decrease when the independent increases; an odds ratio greater than 1 indicates that the odds of an outcome occurring increase when the independent increases; and an odds ratio of 1 means that the independent has no influence whatsoever on the odds of an outcome occurring. If the independent in question is two-valued categorical, the odds ratio simply indicates the probability that an outcome will occur divided by the probability that it will not occur. Odds ratios, therefore, are somewhat similar to probabilistic weights usually reported in Varbrul analyses, with the difference being that odds ratios are about odds (i.e., about the probability of an event occurring divided by the probability of it not occurring) and probabilistic weights are about probabilities (i.e. the chance, in percent terms, of an event occurring).[12] Odds ratios will be tested for statistical significance to make sure their values are not accidental.[13]

I will also report results from omnibus tests for model coefficients to indicate the model's *overall* χ^2 (also known as *Hosmer and Lemeshow's G*), one of the usual significance tests for logistic regression models. This measure tests the predictive ability of all the independents included in the model and indicates whether a model is statistically significant overall.

Mention should be made that a given model may be good at one of the above criteria – predictive efficiency, variance explained, influence of individual independents – but bad at others (though normally these measures will be correlated). For instance, a model can have a good fit, but a low accuracy of prediction. For illustration, let us return to our initial example: the difference between predictive accuracy and predictive inaccuracy may be slight. A predicted probability of 0.49 for pronunciation of postvocalic *r* (cases near $p = 0.5$ are the difficult ones for logistic regression) leads us to classify a person as an postvocalic *r* non-pronouncer, and a predicted probability of 0.51 leads us to classify a person as an postvocalic *r* pronouncer. Thus, the difference between 0.49 and 0.51 makes a huge difference in terms of predictive accuracy, but it is really tiny in terms of variance explained, all other things being equal. Also, it is possible that a model is overall significant, but individual odds ratios are not found to be significant (or vice versa). This can happen because of several reasons. Ideally, one will consider significance of individual independent variables only when the model is overall significant.

Interaction terms in logistic regression

A powerful tool in logistic regression is provided by *interaction terms* or *interaction effects*. The most common variety of these are two-way interactions, where the effect of an independent variable (the 'focal' variable) on the outcome differs depending on the value of a third variable, a so-called 'moderator' variable. To take a hypothetical example: Assume that in logistic regression, it turns out that if a speaker from New York City has a middle-class background, this overall encourages postvocalic *r* pronunciation:

MIDDLE CLASS 2.0

The odds ratio associated with the variable MIDDLE CLASS is 2.0, hence when a speaker has a middle-class background the odds for pronunciation of postvocalic *r* generally double. This is the overall effect which social class has on the outcome 'postvocalic *r* pronunciation.' Assume now, though, that the effect social class has on postvocalic *r* pronunciation in New York City is not the same across all groups of speakers. For instance, female middle class speakers might differ from female non-middle class speakers more than male middle class speakers differ from male non-middle class speakers – in other words, the effect of social class on the likelihood for postvocalic *r* pronunciation differs according to gender. In regression terminology, gender then is said to *interact* with social class.

One will now want to increase the explanatory power of one's statistical model by considering this gender difference. In logistic regression, the most elegant way to do this is to include an interactional term – in this case, the interaction between MIDDLE CLASS and GENDER – in the model. For the hypothetical example at hand, the new, more precise estimate including the interaction between social class and gender might look as follows:

MIDDLE CLASS	1.5
MIDDLE CLASS ∗ GENDER(FEMALE)	2.5

What we called 'the effect of social class on postvocalic *r* pronunciation' before is now called, more precisely, 'the *main* effect of social class on postvocalic *r* pronunciation' (MIDDLE CLASS). Here, this main effect is associated with an odds ratio of 1.5, which applies conditioned that the interactional terms also included in the model are zero (i.e. conditioned that GENDER is

Table 1. Quick reference for important concepts in logistic regression

Basic design. In all cases, the dependent variable is a speaker's choice for one of two linguistic options.

% correctly predicted (predictive efficiency of model). The percentage of correctly predicted cases vis-à-vis the baseline prediction (*% correct (baseline)*) indicates how accurate the model is in predicting actual outcomes.

R^2 **(variance explained, substantial significance of model).** The R^2 value can range between 0 and 1 and indicates the proportion of variance in the dependent variable (i.e. in the outcomes) accounted for by all the independent variables included in the model. Bigger R^2 values mean that more variance is accounted for by the model and that the model is substantially more significant.

Model χ^2 **(statistical significance of model).** This measure tests whether or not all of the variables included in the model significantly contribute to explaining the variance in the dependent variable.

Odds ratio (exp(*b*); influence of independent on outcome). Odds ratios indicate how the presence or absence of a feature (for categorical independents) or an one-unit increase in an independent variable (for scalar independents) influences the odds for an outcome; it is the multiplicative factor by which the odds for a specific outcome are multiplied given an increase in, or the presence of, an independent. Because odds ratios can take values between 0 and ∞, three cases can be distinguished: (i) if $\exp(b) < 1$, an increase in the independent makes a specific outcome less likely; (ii) if $\exp(b) = 1$, the independent has no effect whatsoever on the outcome; (iii) if $\exp(b) > 1$, an increase in the independent makes a specific outcome more likely.

Interaction terms (differences between groups). Interaction terms are used to investigate how the influence of a particular independent (the 'focal' independent) depends on the value of a second independent (the 'moderator' independent). We distinguish between the *main effect* of the focal independent, which applies conditioned on the moderator being zero, and between the *interaction effect* between the focal and moderator. The exp(*b*) value associated with the interaction term is the multiplicative factor by which the main effect of the focal changes for a one-unit increase (for scalar independents) or for a categorical coding (for categorical independents) of the moderator.

zero or male). In other words, for male speakers, a middle class background increases the odds for postvocalic *r* pronunciation by a factor of 1.5, or 50%.

At the same time, the interactional term MIDDLE CLASS ∗ GENDER(FEMALE) indicates that the effect of social class on postvocalic *r* pronunciation is different for female speakers than for male speakers. More precisely, the main effect of MIDDLE CLASS is changed by a multiplicative factor of 2.5 if the speaker is female instead of male.

In a nutshell, what the above two terms and their odds ratios indicate is that in male speakers, a middle class background increases the odds for pronunciation of postvocalic *r* by a factor of 1.5; in female speakers, a middle class background increases the odds for pronunciation of postvocalic *r* by a factor of $1.5 \times 2.5 = 3.75$. This is another way of saying that social class has a stronger effect on postvocalic *r* pronunciation in female speakers than in male speakers. In statistical parlance, we say that gender interacts with social class such that social class has a different effect on the outcome depending on whether the speaker is female or male. In sum, interaction terms are an elegant device to capture differences between groups of cases.

Table 1 gives an overview of the most important terms and concepts in logistic regression.

2. Data

We have seen that persistence is a phenomenon which occurs because of the way language is produced and processed online and because of the way we manage discourse in talk. Therefore, spoken language (and not written language, though more readily available in corpus form) will constitute the present study's database. This database will include four major corpora of spoken English: *The British National Corpus* (BNC), the *Corpus of Spoken American English* (CSAE), the *Corpus of Spoken Professional American English* (CSPAE), and the *Freiburg Corpus of English Dialects* (FRED).

2.1. The British National Corpus (BNC)

The BNC, which was originally released in 1995, contains a spoken section of about 10 million words, which are part-of-speech (POS) tagged. It consists of spoken English of various kinds, produced by different speakers in various situations (for a detailed discussion of the corpus from an end-user's perspective, see Berglund 1999a). The spoken section of the BNC is subdivided into a *demographically sampled component* (henceforth: DS), consisting of "informal encounters recorded by a socially stratified sample of respondents, selected by age-group, sex, social class and geographic region" (Aston and Burnard 1998: 31), and into a *context-governed component* (henceforth: CG) of formal, pre-planned speech which has been categorized into four domains. For the remainder of this study, the DS and CG sections of the BNC will be treated as separate corpora, the first of which contains informal British English and the second formal British English. This is why differences or similarities between the DS and CG corpora are to a certain extent more reliable and valuable than between other corpora. This is because it is to be supposed that the transcription protocol that was applied is more uniform across these two corpora; as a matter of fact, they are subcorpora of the BNC corpus. The DS contains speaker information for the majority of speakers; in contrast, speaker information is available only for some speakers in the CG.

The BNC-DS and the BNC-CG will be analyzed with regard to comparison strategy choice, future marker choice, and the variation between infinitival and gerundial complementation.

2.2. The Corpus of Spoken American English (CSAE)

The Santa Barbara Corpus (hereafter: CSAE) was released in 2000 by a team of researchers led by Wallace Chafe and Jack DuBois. The version that will be used here is composed of the installments 1 and 2, spanning in all 41 texts/conversations. Designed primarily for conversation analysis purposes, this corpus is a comparatively small one (roughly 166,000 words of running text), but it is large enough for some of the purposes of this study. Moreover, it is currently the only publicly accessible corpus of American English conversation that the present author is aware of. Results obtained from the CSAE, then, may often prove to be statistically insignificant. I would like to stress very explicitly here, however, that this does not mean that they would not be significant if only the corpus were big enough, although we may not be able to prove that results can be generalized.

The CSAE will be investigated with regard to genitive choice, future marker choice, particle placement, and complementation strategy choice.

2.3. The Corpus of Spoken Professional American English (CSPAE)

The Corpus of Spoken Professional American English is a corpus of roughly 2 million words of American English, roughly half the size of each of the two BNC-based corpora. The corpus consists primarily of short interchanges by approximately 400 speakers that "are centered on professional activities broadly tied to academics and politics," as the publisher asserts. That means that the corpus is made up of official press conference transcripts released by the White House, as well as transcripts from faculty meetings and other committee meetings. As these transcripts are official or semi-official (and have probably not been transcribed by linguists), transcription is a somewhat problematic issue in the CSPAE: for instance, *gonna* does not occur in the corpus, as the form is apparently deemed to be too sub-standard for official releases. This study will make use of the POS-tagged version of the CSPAE.

The CSPAE will be analyzed with regard to comparison strategy choice, future marker choice, and the variation between infinitival and gerundial complementation.

2.4. The Freiburg English Dialect Corpus (FRED)

The aim of compiling FRED is to strengthen research on morphosyntactic variation in the British Isles (Anderwald and Kortmann 2002; Kortmann 2002, 2003, 2004). The corpus contains 370 texts and spans approximately 2.4 million words of running text, consisting of samples (mainly transcribed so-called "oral history" material) of dialectal speech from a variety of sources. The bulk of these samples was recorded between 1970 and 1990; in most cases, a fieldworker interviews an informant about life, work etc. in former days. The informants are typically elderly people with a working-class background. Speech styles are relatively formal due to the interview situation, though it is arguably less formal than the settings in the formal CG and CSPAE corpora used in this study. Age information is available for many speakers in the corpus; information on sex is available practically throughout. FRED will be investigated with regard to all five alternations subject to analysis in the present study.

Table 2. Corpora analyzed in the present study

	CG	CSPAE	DS	CSAE	FRED
comparison strategy choice	✓	✓	✓		✓
genitive choice				✓	✓
future marker choice	✓	✓	✓	✓	✓
particle placement				✓	✓
complementation strategy choice	✓	✓	✓	✓	✓

The data to be investigated in the present study hence spans (a) two major standard varieties (British English and American English) as well as several dialectal varieties spoken in the British Isles; and (b) three registers: informal, conversational spoken English; spoken English in interview situations; and more formal spoken English. Table 2 gives an overview of which corpora will be analyzed with regard to which alternation. Generally, an attempt was made to analyze as many data sources as possible for each alternation. However, particle placement and the genitive alternation are quite complex alternations where data extraction could not be performed automatically; this necessitated manual coding of manageable data sets. A decision was made to restrict these manageable data sets to FRED and the CSAE.

Chapter 4
Persistence in comparison strategy choice

This chapter will deal with persistence in one of the best-known and most extensively researched alternations in the grammar of English, namely the one between synthetic comparison in *-er*, as in (1a), and analytic comparison with *more*, as in (1b):[14]

(1) a. *If the new, **friendlier** systems do come onto the market, . . . people will just learn to use them.* (DS KRG 514)
b. *You talked a lot about computers being **more friendly** in the future than in the past.* (DS KRG 469)

This chapter has the same internal structure as the other four empirical chapters to follow: we shall first discuss the history of the alternation and which factors were claimed in previous scholarship to determine which form is chosen by speakers (sections 1 and 2). After outlining the specific methodology to deal with comparison strategy choice and the dataset that is going to be analyzed (section 2), I will present the findings, i.e. (i) to what extent baseline predictors explain the alternation (section 4.1), (ii) how much persistence-related predictors enhance our ability to account for the observable variation (section 4.2), and (iii) what role extralinguistic factors play (section 4.3). The chapter will be concluded by an interim summary (section 5).

1. Background and previous research

Historically, the analytic strategy with *more* is an innovation that is not attested prior to the thirteenth century (cf. Mitchell 1985: 84–85). By the beginning of the 16th century, though, it had become roughly as frequent as it is today (Pound 1901). The rivalry between the two types after Late Modern English is documented in Kytö and Romaine (1997); suffice it to say here that it is often speculated that one of the motivations for the genesis of the new strategy was that English was gradually shifting its syntax from synthetic to analytic, which is why the analytic comparison strategy was more consistent with the typology of English (cf. Kytö and Romaine 1997: 330). As for the situation in Present-Day English, Leech and Culpeper (1997) find that in their data of written English, disyllabic adjectives take analytic comparison in 60% of all cases. Leech and Culpeper also argue that there is an ongoing shift towards analytic comparison even in Present-Day English (similarly, Barber 1964: 52 and Potter 1969: 146–147), although at the same time analytic comparison is overall rather infrequent, especially in casual speech. Biber et al. (1999: 525) argue this is because especially in conversation, polysyllabic adjectives are rare.

The well-known rule of thumb governing the alternation in Present-Day English is that monosyllabic adjectives take synthetic comparison, adjectives with more than two syllables take analytic comparison, and disyllabic adjectives alternate in the comparison strategy they take (Lindquist 2000; Quirk et al. 1985; Bauer 1994; Sweet 1892). Leech and Culpeper (1997) conduct a study on some written corpora (BNC and LOB) and find that 99% of all monosyllabic adjectives, 42–51% of all disyllabic adjectives, and virtually no trisyllabic adjectives take synthetic comparison (Leech and Culpeper 1997: 355). Yet, there is substantial evidence that a range of monosyllabic adjectives can take analytic comparison (cf. Biber et al. 1999: 522; Quirk et al. 1985: 462; Leech and Culpeper 1997: 357; Mondorf 2002) and that even some trisyllabic adjectives can take synthetic comparison (cf. Biber et al. 1999: 522 and Mondorf 2003 contra Quirk et al. 1985: 462). Hence, there is variation between analytic and synthetic comparison in all disyllabic adjectives and in some monosyllabic and trisyllabic adjectives.[15] This variation will be investigated with regard to persistence in this chapter.

As far as the formal difference between the two comparison strategies is concerned, the alternation is for one thing a syntactic one (cf. Mondorf 2003; Hawkins 1999) in that the difference is the placement of the adjective and the existence of a filler-gap dependency in the synthetic variant. In addition, of course, there is also a difference in morphological marking (*-er*, ø), and a lexical (i.e. non-structural) difference in that analytic comparison comes with the additional token *more*. This makes the alternation susceptible to lexical priming. A word on terminology: While functional linguists would refer to the token *more* as rather 'functional', psycholinguists would classify *more* as rather 'lexical' (hence *lexical* priming).

Most recent empirical research on the issue has focussed on formal features of adjectives and the effect these have on the comparison strategy they take. Examples are Kuryłowicz (1964), Leech and Culpeper (1997), and Lindquist (2000); Mondorf (2000) falsifies an earlier claim that compound adjectives never take synthetic comparison; Mondorf (2002) investigates the influence of the presence of prepositional complements on strategy choice. Mondorf (2003), finally, is a study taking into account a vast number of factors influencing the alternation between the two comparison strategies in the BNC and some newspaper corpora.

2. Previously suggested factors

The following factors have been shown to influence the alternation between analytic and synthetic comparison when there is a choice:

Length. The basic determinant of the comparison strategy a given adjective may take is agreed to be its length in syllables.

Morphology. Disyllabic adjectives ending in *-y* have been shown to be solidly inflectional (Leech and Culpeper 1997: 359; cf. also Biber et al. 1999; Mondorf 2003; Lindquist 2000; Quirk et al. 1985; Bauer 1994; Sweet 1892). Also, prefixation with *un-* may cause a trisyllabic adjective to take synthetic comparison (e.g. *unhappier, untidier, unlikelier, unluckier, unrulier*; cf. Quirk et al. 1985: 462; Leech and Culpeper 1997: 358). Certain suffixes, finally, make synthetic comparison impossible, e.g. *-al* and *-ish* (cf. Mondorf 2003: 259).

Stress placement. Kuryłowicz (1964: 15) proposes a stress based rule according to which “the comparative in *er* is regular on adjectives stressed on the final syllable (e.g. *severer*), hence also with monosyllabic forms (*stronger*), but the periphrastic comparative in all other cases” (see Lindquist 2000; Quirk et al. 1985; Bauer 1994; Sweet 1892 for similar claims). Leech and Culpeper (1997) present empirical evidence, though, that there are many exceptions to Kuryłowicz’s rule.[16]

Syntactic function. Braun (1982: 116), Leech and Culpeper (1997: 366), and Mondorf (2003: 286–287) show that when the adjective occurs in predicative rather than attributive function, analytic comparison is favored. Thus *the gas lamps looked more friendly* (predicative) and *a friendlier place for wild life* (attributive) are typical.

Degree modifiers. Leech and Culpeper (1997: 367) and Lindquist (2000: 127) present evidence that there is a positive correlation between analytic comparison and a preceding degree modifier such as *much, even,* *far, a bit, a little, just that little bit, a whole lot, a thousand times, noticeably, slightly*, and *marginally*. Thus, *a much more ready acceptance*, but *a readier acceptance*.

Coordination and parallelism. Leech and Culpeper (1997: 367), Lindquist (2000: 129–130), and Mondorf (2003: 285) suggest that coordination of two adjective phrases favors choice of the same comparison strategy (e.g. *more interesting and more lively* rather than *more interesting and livelier*). This factor, of course, is basically the subject of the present study.

Cognitive complexity. Mondorf (2002) and Mondorf (2003: 252–253) argue that in cognitively demanding environments requiring increased processing efforts, "language users tend to make up for the additional effort by resorting to the analytic (*more*) rather than the synthetic (*-er*) comparative" due to Rohdenburg's complexity principle (Rohdenburg 1996: 151). This is because analytic comparison facilitates recognition of the relevant phrase and can thus help mitigate increased processing load (Mondorf 2003: 254). In terms of cognitive complexity, Mondorf (2002) suggests that the presence of a prepositional adjective complement (as in *it would be hard to find any couple* more proud *of their home than Michael and Kathleen*) makes analytic comparison more likely. Mondorf (2003: 254) shows that "the presence of a complement raises an adjective's proclivity towards the analytic comparative, claiming that in the case of infinitival complements, the effect of analytic comparison on parsing efficiency is threefold: (i) it unambiguously signals at the beginning of the phrase that a comparative follows; (ii) the analytic variant is more explicit and easier to parse due to its close form-function match; (iii) by simply choosing the analytic strategy, speakers can alert hearers to the fact that a cognitively complex adjective phrase may follow." Poutsma (1914) and Jespersen (1909) have argued along the same lines.

Frequency. It has been suggested that frequently used adjectives – which are typically also shorter – tend to take synthetic comparison, while less frequent adjectives tend to take analytic comparison (e.g. Sweet 1892: 327; Bolinger 1968: 120; Quirk et al. 1985: 463). Empirical evidence for this claim has been presented by Braun (1982) and Mondorf (2003).

Phonological factors. Mondorf (2003: 275–276) details three domains where phonological factors can interact with the choice between synthetic and analytic comparison: (i) the synthetic *-er* suffix can serve as a buffer between two adjacent stressed syllables (e.g. *a próuder cándidate*) and therefore help achieve stress clash avoidance; (ii) analytic comparison can help avoid haplology effects (as in *bitterer*; also cf. Sweet 1892: 327 and Jespersen 1909: 344); (iii) synthetic comparison is strongly dispreferred if the adjective ends in a /-pt, -lt, -ct/ consonant cluster.

Semantic and pragmatic factors. There are a number of semantic factors that also seem to influence the alternation: (i) adjectives denoting more concrete notions (e.g. *a fuller hotel*) tend to take synthetic comparison to a wider extent than do adjectives denoting more abstract notions (e.g. *the more bitter takeover battles of the past*) (cf. Mondorf 2003: 289–290); (ii) weakly gradable adjectives take analytic comparison more frequently than fully gradable adjectives (cf. Mondorf 2003: 289–290); (iii) synthetic comparison sometimes has the disadvantage of non-adjacency of adjective and complement, which violates iconic ordering (cf. Mondorf 2003: 291); (iv) Biber et al. 1999: 522; Curme (1931), Rohr (1929); Bolinger (1968) credit analytic comparison with some stylistic advantage in that the additional element allows the speaker to employ some extra stress or emphasis on the comparison. Contrarily, Jespersen (1909) argues that inflectional comparison is felt as more 'vigorous' and more 'emphatic.' Leech and Culpeper's data on superlatives empirically support Jespersen's claim (Leech and Culpeper 1997: 369–370).

3. Method, data and independent variables

3.1. Method and data

This chapter will investigate persistence in the comparison strategy taken in the following 112 adjectives, which have been shown to take both synthetic and analytic comparison in previous research (e.g. Bauer 1994: 55; Biber et al. 1999: 522; Leech and Culpeper 1997: 356–364; Mondorf 2003: 257 and 287; Quirk et al. 1985: 462):

able, acute, afraid, akin, ample, apt, aware, bitter, bizarre, blunt, bold, brittle, cheap, cheeky, clear, clever, common, compact, complete, correct, costly, cosy, crazy, cruel, curt, dead, deadly, dense, empty, exact, extreme, feeble, fierce, fit, fond, free, friendly, full, gentle, guilty, handsome, handy, humble, hungry, intense, just, keen, kindly, likely, little, lively, lonely, lovely, lowly, lucky, mature, mellow, narrow, nimble, noble, obscure, odd, pale, pleasant, polite, poor, precise, profane, profound, prone, proud, queer, quiet, rare, ready, real, remote, rich, right, risky, robust, rude, secure, severe, sexy, shallow, sick, silly, simple, sincere, slender, slow, sober, solid, sound, stable, stupid, subtle, sure, tender, trendy, tricky, true, ugly, unhappy, unwise, used, wealthy, wicked, worthy, wrong, yellow

Extraction of the relevant forms was performed automatically. A Perl script identified the above adjectives in the dataset and extracted them if they were either preceded by the token *more* or if they carried an *-er* suffix. This method yielded an accuracy rate of approximately 94%. The following false positives were then weeded out of the database manually: (i) constructions where *more* was not a modifier of the adjective but a determiner of the noun phrase (e.g. *to play more friendly matches*); (ii) constructions with the pattern *more* Adj *than* Adj (e.g. *more dead* than *alive*); (iii) tokens that were not actually adjectives at all (for instance, the noun *fitter* is frequent in FRED).

Due to the extremely low number of relevant tokens in the CSAE (16), the corpus had to be excluded from analysis in this chapter. Analysis of the remaining corpora (CSPAE, FRED, DS, CG) yielded 1,794 relevant tokens in all, a breakdown of which is displayed in Table 3. Overall, the most frequent alternating adjective is *cheap* with 533 occurrences (29.7%) over all corpora; next frequent are *likely* (212 occurrences, 11.8 %) and *clear* (93 occurrences; 5.2%). What can clearly be seen is that analytic comparison is a good deal

Table 3. Comparison strategy choice: distributional variation across corpora

corpus	*N*	*N* analytic	*N* synthetic
CG	884	390 (44.1%)	494 (55.9%)
DS	521	104 (20.0%)	417 (80.0%)
CSPAE	219	107 (48.9%)	112 (51.1%)
FRED	170	37 (21.6%)	133 (77.8%)
total	**1,794**	**638 (35.6%)**	**1,156 (64.4%)**

more common in the more formal corpora (CG, CSPAE) than in FRED and the informal DS corpus. In the formal corpora, the distribution of analytic and synthetic comparison is roughly 50:50, in the other two corpora it is more like 20:80. Among other factors that may contribute to this skewing, this is most probably correlated to the fact that in the two formal corpora, adjectives are on average longer than in the informal corpora: mean length (in syllables) of the inflected form is 2.38 in the CG and 2.44 in the CSPAE, while it is only 2.18 in the DS and 2.23 in FRED.

3.2. Independent variables

In addition to most variables discussed in chapter 3, the following factors specific to comparison will be included in this chapter's analysis (Table 4 gives an overview).[17]

3.2.1. Previously suggested and persistence-unrelated predictors

LENGTH of the synthetically inflected form in syllables (henceforth: LENGTH). For instance, *cheaper* has a length of two syllables.
Hypothesis: The longer the adjective, the greater the odds for analytic comparison.

MORPHOLOGY (henceforth: MORPH). Does the adjective which takes comparison (i) begin in *un-* (as *unhappy*) or (ii) end in *-y* (as *lucky*; coded 0 if such affixes were not present and 1 if one of the affixes was present)?
Hypothesis: Presence of such affixes makes synthetic comparison more likely.

STRESS PLACEMENT (henceforth: STRESS). If the adjective which takes comparison is polysyllabic, is it stressed on the final syllable (as *complete*; coded 1 for final stress and 0 otherwise)?
Hypothesis: Final stress increases the odds for synthetic comparison.

FREQUENCY (henceforth: FREQUENCY). What is the text frequency of the base form of the adjective under analysis in the spoken section of the BNC? *Poor*, for instance, has a text frequency of 1,031 words per million in the spoken section of the BNC.

Hypothesis: The higher the text frequency of a given adjective, the greater the odds for synthetic comparison.

SYNTACTIC FUNCTION (henceforth: SYNFUN). Does the adjective occur in attributive function, as in (2a), or in another function, as in (2b) (coded 0 for attributive function and 1 for other functions)?

(2) a. *and that was the* ***poorer*** *people anyway* (CG D8Y)
b. *it's certainly* ***cheaper*** *than three hundred and fifty* (CG F7C)

Because some readers might feel that coding this feature is tricky, Cohen's κ, which measures the proportion of the best possible improvement over chance, was used to evaluate intercoder reliability. A second coder, a native speaker and trained linguist, re-coded a random subset of the CG sample (N = 50, ca. 10% of the entire CG sample); comparison of the two samples yielded a simple agreement rate of ca. 96% and an 'excellent' (cf. Orwin 1994) κ value of approximately 0.90. See Appendix C for the – relatively simple – coding scheme.
Hypothesis: Predicative usage increases the odds for analytic comparison.

DEGREE MODIFIERS (henceforth: DEGREE). Is the adjective preceded by one of the following degree modifiers, as in (3): *much, even, far, bit, little, lot, times, noticeably, slightly, marginally* (coded 0 for degree modifiers not present and 1 for degree modifiers present)?

(3) *he is much* ***cheaper*** *than in well, in Leverington* (DS KB7)

Hypothesis: A preceding degree modifier makes analytic comparison more likely.

PRESENCE OF VERBAL COMPLEMENTS (henceforth: COMPLEMENT). Is the adjective followed by prepositional or infinitival complements (coded 0 for verbal complements not present and 1 for verbal complements present)? (4) illustrates:

(4) *and we came to the conclusion it was* ***cheaper to print*** (CG D8Y)

Hypothesis: Verbal complements increase the odds for analytic comparison.

3.2.2. Additional, persistence-related predictors

PRESENCE OF TRIGGERS in the preceding context. Referring to sites not necessarily alternating, this independent pertains to β-persistence. The variable is also sensitive to whether parallelism in coordinated adjective phrases (cf. Leech and Culpeper 1997; Lindquist 2000; Mondorf 2003) obtains. Two scenarios will be distinguished:

– *items triggering analytic comparison* (henceforth: ATRIGGER). Does the token *more* occur in a context of (a) more than 75 words (or not at all), (b) 75 words, (c) 25 words, (d) 5 words prior to CURRENT? Note that if CURRENT takes analytic comparison, the token *more* – which then necessarily accompanies CURRENT – does not count in the tally. (5) is an example of *more* occurring in a context of five words prior to CURRENT:

(5) *it was a, it was developed* ***more****, it was* ***more compact*** *you know at the ending, yeah* (DS KPV)

– *items triggering synthetic comparison* (henceforth: STRIGGER). Does a synthetic comparative (not necessarily one that could also be analytic) occur in a context of (a) more than 75 words (or not at all), (b) 75 words, (c) 25 words, (d) 5 words prior to CURRENT? (6) is an example of a synthetic comparative (*louder*) occurring in a context of five words prior to CURRENT (*quieter*):

(6) *can you have it* ***louder*** *and quieter* (DS KB8)

Hypothesis: Recent usage of *more* will trigger analytic comparison, recent usage of a token ending in *-er* will trigger synthetic comparison.

Two of the general variables discussed in chapter 3 – SAMETURN, SAMESPEAKER – cannot be considered in this chapter's analysis. The reason is the comparatively low text frequency of alternating adjectives: with textual

Table 4. Comparison strategy choice: independent variables considered

variable	type	coding method
	a. previously suggested and persistence-unrelated independents	
SENTENCELENGTH*	scalar	software
TTR*	scalar	software
LENGTH	scalar	manual
MORPH	two-way categorical	manual
STRESS	two-way categorical	manual
FREQUENCY	scalar	software
SYNFUN	two-way categorical	manual
DEGREE	two-way categorical	manual
COMPLEMENT	two-way categorical	manual
	b. persistence-related independents	
PREVIOUS*	two-way categorical	software
TEXTDIST*	scalar	software
ATRIGGER	four-way categorical	software
STRIGGER	four-way categorical	software
	c. speaker characteristics	
AGE*	scalar	software
SEX*	two-way categorical	software

* independent variable discussed in chapter 3, section 1.

distance (TEXTDIST) between two relevant comparison slots averaging between 2,408 words (CG) and 15,299 words (FRED), SAMETURN and SAMESPEAKER – variables meant to help investigate interaction between persistence and turn-taking – have to be omitted.

4. Results

4.1. Baseline variation

In a first step, a logistic regression model was estimated including those variables claimed in previous research to be influencing the alternation: LENGTH, MORPH, STRESS, FREQUENCY, SYNFUN, DEGREE, and COMPLEMENT. In

Table 5. Comparison strategy choice: odds ratios associated with baseline predictors

	CG	CSPAE	DS	FRED
SENTENCELENGTH	0.99	1.00	1.00	**0.98** *
TTR	**1.04** *	0.99	1.00	0.94
LENGTH	**0.11** ***	**0.09** ***	**0.07** ***	**0.14** ***
MORPH(1)	**0.27** ***	0.00	**0.13** ***	0.94
STRESS(1)	**0.12** ***	0.00	**0.08** *	0.13
FREQUENCY	**0.99** ***	1.00	**0.99** ***	1.00
SYNFUN(1)	**0.37** ***	**0.28** *	1.45	0.43
DEGREE(1)	0.85	2.44	0.60	**15.78** *
COMPLEMENT(1)	**0.31** ***	0.65	0.84	0.62
model intercept	∞ ***	∞	∞ ***	∞ ***
N	884	219	521	170
model χ^2	549.36 ***	191.53 ***	221.15 ***	50.86 ***
R^2	0.620	0.777	0.547	0.398
% correct (baseline)	85.1 (55.9)	89.5 (51.1)	90.4 (80.0)	81.2 (78.2)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for synthetic comparison in *-er*.

addition, SENTENCELENGTH and TTR, two variables not discussed in previous research though unrelated *per se* to persistence, were included. The model, displayed in Table 5, will serve as the benchmark against which a model admitting persistence and speaker variables will be compared. Note that here and in the following, the value in brackets following categorical independents indicates which category of the independent has been tested. Therefore, MORPHOLOGY(1) tests the presence (as opposed to the absence) of affixes on an adjective.

Let us first assess the overall quality of the model. As the model χ^2 values indicate, the model is statistically significant in all corpora. This means that on the whole, the independent variables entered have a significant effect on the outcome. As for variance explained, the R^2 values are fair to decent, ranging from 0.398 for FRED to 0.777 for the CSPAE. This means that the model accounts for between 40% (FRED) and 77% (CSPAE) of the variation between synthetic and analytic comparison. Overall, it can be seen that more variation is accounted for in the formal corpora (CG: 0.620, CSPAE: 0.777) than in the less formal corpora (DS: 0.547, FRED: 0.398). As for predictive efficiency, the model correctly predicts between 81.2% (FRED) and 90.4% (DS) of all actual outcomes.

As for the influence of individual independent variables on the outcome 'synthetic comparison,' LENGTH is most powerful and statistically highly significant, in all corpora under analysis. The variable's odds ratio averages 0.1, which means that as the length of the inflected adjective increases by one syllable, the odds that the adjective will take synthetic comparison (rather than analytic comparison) decrease by 90%. This is another way of saying that the longer adjectives are, the more likely they are to take analytic comparison, which is to be expected.

MORPH, STRESS, and FREQUENCY run in the same direction (which is not the theoretically expected one[18]) throughout but achieve statistical significance in the CG and DS only. The presence of a *-y* suffix or an *un-* prefix (MORPH) substantially decreases the odds for synthetic comparison (by a factor of 0.3 and 0.1 in the CG and DS, respectively). This means that the presence of such affixes actually increases the odds for analytic comparison, this contradicts Leech and Culpeper (1997), Biber et al. (1999), Mondorf (2003), Lindquist (2000), Quirk et al. (1985), Bauer (1994), and Sweet (1892). Likewise, if an adjective is stressed on the final syllable (STRESS), this actually decreases the odds for synthetic comparison by a factor between 0.12 and 0.08, which is contrary to what Kuryłowicz (1964), Lindquist (2000), Quirk et al. (1985), Bauer (1994), and Sweet (1892) have suggested. Also, each one-per-million-word increase in text frequency (FREQUENCY) decreases the odds for synthetic comparison by 1%, an effect that is rather potent given that some of the adjectives differ in hundreds of per million words frequency points. Contrary to what Sweet (1892), Bolinger (1968), Quirk et al. (1985), Braun (1982), and Mondorf (2003) have claimed, this suggests that the more frequent an adjective is, the more likely it is to take analytic comparison.

In the corpora where SYNFUN has a statistically significant effect on the outcome – the CG and CSPAE – the variable is associated with an exp(b) value of about 0.3. Thus, as soon as an adjective occurs in non-attributive function (for instance, predicatively, as in *John is more friendly*), this decreases the odds that the adjective will take synthetic comparison by about 60–70%, a finding which is consonant with Braun (1982), Leech and Culpeper (1997), and Mondorf (2003). Verbal complements (COMPLEMENT; as in *it's cheaper to print*) have a significant effect only in the CG (although the effect of the variable runs in the same direction in the other corpora): the presence of a verbal complement decreases the odds for synthetic comparison (in the CG, by about 70%). Given Mondorf (2003), Poutsma (1914), and Jespersen (1909), this is as expected. Results concerning degree modifiers (DEGREE) are not conclusive. In FRED – the only corpus where the variable reaches statistical significance – the presence of degree modifiers (as in *John is a much friendlier person than Mary*) hugely increase the odds for synthetic comparison, which would contradict Leech and Culpeper (1997) and Lindquist (2000). In the CG and DS, however, the variable has the opposite effect but does not achieve statistical significance. SENTENCELENGTH and TTR, finally, seem to have an only moderate influence on the outcome and reach statistical significance only rarely. In FRED, a one-word increase in sentence length decreases the odds for synthetic comparison by 2% and therefore increases the odds for analytic comparison. Assuming that SENTENCELENGTH is a proxy for syntactic complexity and that the more explicit analytic option is preferred in more complex environments (cf. Rohdenburg 1996), this effect is as hypothesized. In the CG, a one-point increase in TTR increases the odds for synthetic comparison by 4%. In other words, the synthetic, lexically more economic option seems to be preferred in otherwise lexically dense contexts.

In all, most of the hypotheses from previous research are borne out by this study's analysis, and some are not. A possible explanation for this mixed picture is that most previous research on comparison has investigated written English; the database examined here, though, contains spoken English. Circumstantial evidence for this may be that overall, the model explains more of the variation in the formal corpora (i.e. material that is comparatively more similar to written material, which was in the focus of most previous research) than in the less formal corpora.

4.2. Persistence-induced variation

Now, what role does persistence play in comparison strategy choice? The importance of α-persistence can be gauged from the scatterplots in Figure 6. These plots lump together speakers from all the corpora under analysis in this chapter and compare their switching rate to their usage of the switched-to strategy (note that the null hypothesis is that switch rates are proportional to the usage of the switched-to strategy, which is indicated by the dotted line). As is evident, we can visually reject the null hypothesis: speakers cluster

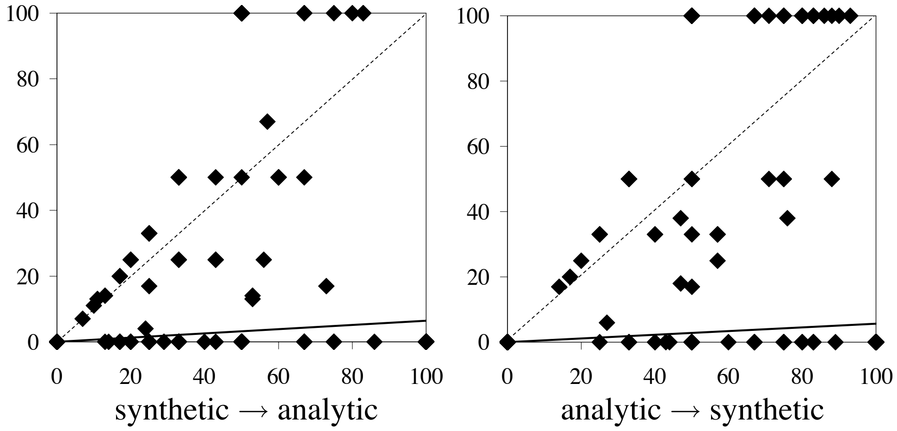


Figure 6. Switches between comparison strategies as a function of strategy proportion (relative frequency of switches, in %, on y-axis; relative frequency of the switched-to strategy, in %, on x-axis). Each dot represents one speaker. Dotted diagonal line represents null hypothesis that switch rate is proportional to variant proportions. Heavy line indicates linear trend (synthetic → analytic: $y = 0.06x$, analytic → synthetic: $y = 0.06x$)

below the diagonal line.[19] This means that overall, switches from analytic → synthetic comparison and switches from synthetic → analytic comparison are rarer than the null hypothesis visualized by the dotted line would predict. This is strong evidence for α-persistence. The heavy lines are regression lines providing information about actual overall switch rates across speakers. Observe that overall, speakers switch between the comparison strategies less than one tenth of the time they would if there were no α-persistence. Table 6 provides more detailed information about switch rates across corpora. These rates are indeed remarkably homogeneous, ranging from 4% to 12%, and differences between corpora and between switch directions thus appear substantially insignificant.

Next, the variables pertaining to the domain of persistence – PREVIOUS, TEXTDIST, ATRIGGER, STRIGGER – were entered into logistic regression (Table 7). This step yields a substantial increase of both predictive efficiency and variance explained; the change in model χ^2 is statistically significant throughout.[20] In the CG, CSPAE, and DS, predictive efficiency is up some

Table 6. Linear regression estimates of switch rates in comparison strategy choice across corpora (y is the relative frequency of A → B switches, in %; x is the relative frequency of B forms, in %; the expected linear relationship, uninfluenced by persistence, is $y = x$)

corpus	analytic → synthetic	synthetic → analytic
CG	$y = 0.06x$	$y = 0.06x$
DS	$y = 0.06x$	$y = 0.09x$
CSPAE	$y = 0.12x$	$y = 0.07x$
FRED	$y = 0.04x$	$y = 0.06x$

1 to 2 percent points; in FRED, the model now predicts comparison strategy choice correctly in 98% of all cases, which is up 16 percent points compared to our earlier model. Also, variation accounted for (R^2) is up some 3–7 percent points in the CG, CSPAE, and CG, while in FRED, it is up over 50 percent points. In FRED, the model accounts now for 93% of the observed variation. However, the rather low number of observations in FRED ($N = 169$) and the fact that few individual persistence predictors are selected as significant in the corpus suggest that these differences should not be exaggerated.

4.2.1. α-persistence

PREVIOUS, the main α-persistence predictor, reaches statistical significance in the CG and DS. The variable has a huge effect: if PREVIOUS – the last occurrence of an alternating adjective taking comparison – is analytic, this decreases the odds for synthetic comparison in CURRENT by some 97–99%. This holds conditioned on TEXTDIST being zero. The interaction term PREVIOUS $*$ TEXTDIST, then, takes into account the effect of an increasing textual distance between PREVIOUS and CURRENT: for every one-unit increase in the *ln* of intervening words between PREVIOUS and CURRENT, PREVIOUS' odds ratio increases by a multiplicative factor of between 1.35 (CG) and 1.93 (DS). In plain English, this means that in the CG, the main effect of PREVIOUS is smaller (0.03) but declines more slowly (1.35), while in the DS, PREVIOUS's main effect is bigger (0.01) but declines faster (1.93).

It will be helpful to present this relationship graphically. Figure 7 takes a closer look at the interplay between PREVIOUS and TEXTDIST exemplarily in the CG and DS, for which a sufficient number of observations is available.

Table 7. Comparison strategy choice: odds ratios associated with persistence-related predictors in logistic regression (baseline predictors are included, but not displayed)

	CG	CSPAE	DS	FRED
PREVIOUS(ANA)	**0.03** ***	7.85	**0.01** **	6.33
PREVIOUS(ANA) $*$ TEXTDIST	**1.35** *	0.67	**1.93** *	0.68
STRIGGER	–	–	–	– *
STRIGGER(75)	1.95	5.31	4.95	0.00
STRIGGER(25)	2.31	3.38	2.08	0.00
STRIGGER(5)	1.17	∞	1.57	0.00
ATRIGGER	– *	–	– ***	–
ATRIGGER(75)	2.18	∞	2.52	0.00
ATRIGGER(25)	1.50	0.92	**0.16** ***	0.00
ATRIGGER(5)	**0.22** **	2.2	**0.02** ***	0.00
model intercept	51.28 *	∞	∞ ***	∞
N	533	218	422	169
model χ^2	381.78 ***	213.31 ***	213.27 ***	157.52 ***
R^2	0.688	0.832	0.647	0.932
% correct (baseline)	86.1 (57.8)	91.3 (51.4)	92.4 (81.8)	97.6 (78.1)

* significant at p<.05, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for synthetic comparison in *-er*.

The figure plots textual distance between PREVIOUS and CURRENT against the percentage of persistent PREVIOUS / CURRENT pairs (thus, the intuitive interpretation of the y-axis is that it indicates the strength of α-persistence).[21] The graphs confirm visually that as TEXTDIST increases, the proportion of matches between PREVIOUS and CURRENT decreases. What is interesting is the nature of this decrease. The graphs offer a logarithmic, or decreasing exponential, decline function (heavy line) and a linear decline function (dotted line). In both corpora, the logarithmic estimate appears to fit the data better, both visually and statistically, as the following curve fits[22] demonstrate:

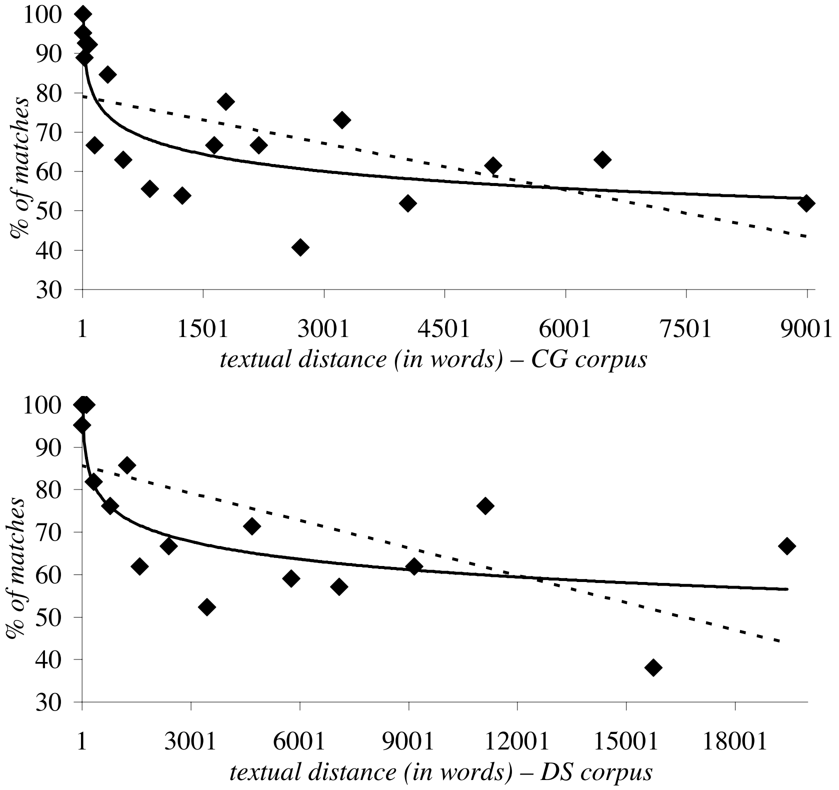


Figure 7. Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same comparison strategy is used) as function of textual distance between CURRENT and PREVIOUS. Heavy line represents logarithmic estimate of the relationship, dotted line represents linear estimate of the relationship

	CG	DS
adjusted R^2 linear	0.30 **	0.39 ***
adjusted R^2 logarithmic	0.67 ***	0.73 ***
df	17	17

Thus, the forgetting function that describes the decline of α-persistence is indeed best thought of as logarithmic.

4.2.2. β-persistence

The two variables hypothesized to be sensitive to β-persistence are STRIGGER and ATRIGGER. STRIGGER is not selected as significant in logistic regression, but ATRIGGER is: the presence of the token *more* in CURRENT's immediately preceding context does have a statistically significant effect (independent of the threshold levels discussed below) on the comparison strategy employed in CURRENT in the two BNC-based corpora, the DS and CG. Here, the pattern conforms with expectations: the more recently the token *more* has been used, the smaller the odds are that CURRENT will take synthetic comparison.[23] Take, for instance, the DS: if the token *more* has been used no less recently than 5 to 25 words prior to CURRENT, the odds for synthetic comparison diminish by 84% (compared to when there is no *more* at all in a horizon of 75 words). If the token *more* has been used no less recently than 1 to 5 words prior to CURRENT, the odds for synthetic comparison shrink by 98% (again, compared to when there is no *more* at all in a horizon of 75 words). Similarly, in the CG, the presence of the token *more* 5 to 1 words prior to CURRENT has a statistically significant effect in that the odds for synthetic comparison shrink by 78%.

There is no immediately obvious way how persistence in comparison strategy choice differs along variety lines. However, a curious pattern emerges with regard to REGISTER: if we take the DS and FRED to represent more informal speech and the CG and CSPAE to represent more formal, carefully planned speech, it appears from Table 7 that PREVIOUS is somewhat more potent in the formal corpora and ATRIGGER and STRIGGER are more influential in the informal corpora. Thus, in analytic vs. synthetic comparison, α-persistence appears to be more powerful in formal registers, and β-persistence appears to be more powerful in informal registers.

4.3. Inter-speaker variation

I will now report estimates of a logistic regression where in addition to baseline and persistence-related independents, the speaker variables SEX and AGE and some interactions between them and other variables are included (Table 8). This regression is on the DS database only. Given comparatively low case numbers overall and less availability of speaker information in the other cor-

Table 8. Comparison strategy choice: odds ratios associated with speaker predictors in logistic regression on the DS (baseline predictors and persistence-related predictors are included, but not displayed)

AGE	**1.12** *
SEX(F)	0.35
AGE ∗ PREVIOUS(ANA)	2.42
AGE ∗ STRIGGER(1)	**0.91** $^+$
AGE ∗ ATRIGGER(1)	0.94
SEX(F) * PREVIOUS(ANA)	0.00
SEX(F) * STRIGGER(1)	∞ ***
SEX(F) * ATRIGGER(1)	0.41
AGE ∗ PREVIOUS(ANA) ∗ TEXTDIST	**0.89** $^+$
model intercept	0.61
N	191
model χ^2	133.36 ***
R^2	0.826
% correct (baseline)	95.3 (82.2)

$^+$ marginally significant at $p < .15$, * significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for synthetic comparison in *-er*.

pora, analysis of the other datasets would have been too restricted to arrive at statistically reliable results. Also note that for simplicity, the three distance thresholds in ATRIGGER and STRIGGER have been conflated into a single dichotomy, namely whether or not a trigger had been used no less recently than 75 words prior to CURRENT (coded 0 for such a trigger not present, and 1 for present).

Adding the speaker variables (and their interactions with persistence-related variables) improves the model significantly.[24] Predictive efficiency is enhanced by two percent points and is now an excellent 95.3%. Variance explained increases from $R^2 = 0.628$ to $R^2 = 0.826$, so that the model reported in Table 8 accounts for roughly 80% of the variation between synthetic and analytic comparison. As for the speaker variables, the main effect of SEX does not come close to reaching statistical significance (p = 0.68). AGE, in contrast, is significant. The variable's odds ratio of 1.12 indicates that for every one year increase in a speaker's age, the odds for synthetic comparison rise by 12% (conditioned on ATRIGGER and STRIGGER being zero). Obviously, this is an apparent-time phenomenon: the older speakers are, the more likely they are to use synthetic comparison. This would seem to support claims that there is an ongoing shift towards analytic comparison (Leech and Culpeper 1997; Barber 1964; Potter 1969).

As for the interactional terms included, three of these are statistically significant or marginally significant and therefore merit more detailed discussion. To start with, the interaction term AGE ∗ STRIGGER shows how the impact of STRIGGER on the odds for synthetic comparison in CURRENT depends on how old the speaker is. The odds ratio of the interaction is 0.91, hence for every 1-year increase in AGE, the odds ratio associated with STRIGGER changes by a multiplicative factor of 0.91. This means that the older speakers are, the less powerful is the influence of a synthetic trigger (STRIGGER) on the comparison strategy chosen for CURRENT.

Second, the interaction term SEX ∗ STRIGGER checks how the impact of the presence or absence of STRIGGER on the odds for synthetic comparatives differs as a function of the sex of the speaker. Its value indicates that the odds ratio associated with STRIGGER increases infinitely if the speaker is female, compared to a male speaker. This means that somewhat surprisingly, female speakers almost categorically employ synthetic comparison when some STRIGGER was present in the immediately preceding discourse. Another way of putting this is that STRIGGER is more influential in female speakers than in male speakers.

Lastly, the marginally significant exp(b) value of 0.89 associated with the three-way interaction AGE ∗ PREVIOUS ∗ TEXTDIST indicates that for every one-year increase in a speaker's age, the weakening effect TEXTDIST has on persistence between PREVIOUS and CURRENT loses 11% of its power. What does this mean? The forgetting function that describes the decline of persistence with increasing textual distance between CURRENT and PREVIOUS is more level in old speakers than in young speakers. Figure 8 seeks to visualize this (admittedly complicated) relationship in the DS. In principle, it plots the percentage of PREVIOUS / CURRENT pairs against textual distance between PREVIOUS / CURRENT, much like Figure 7 (p. 79) does; the difference is that Figure 8 distinguishes between speakers that are older than 37 years (which is the mean age in the DS database for the present comparison study) and speakers younger than 38 years, presenting separate logarithmic estimates of the relationship between α-persistence and textual distance (or recency of use).

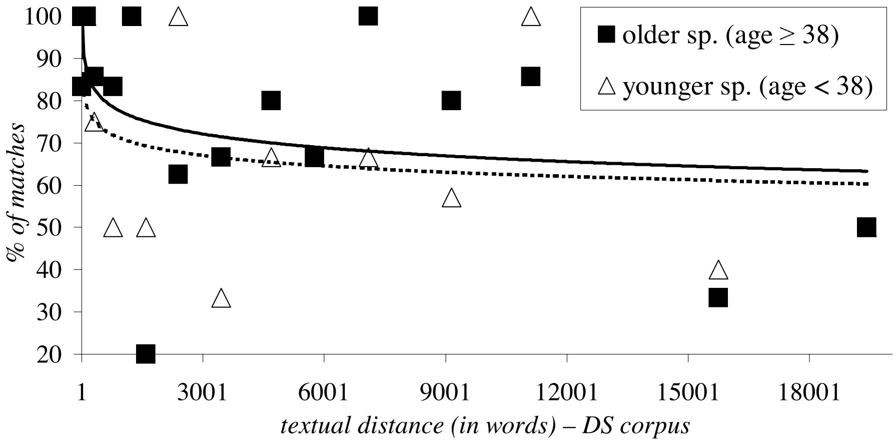


Figure 8. Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same comparison strategy is used) as function of textual distance (in words) between CURRENT and PREVIOUS in the DS. Heavy line represents logarithmic estimate of the relationship in older speakers, dotted line represents logarithmic estimate of the relationship in younger speakers

As is evident from the graph, persistence declines faster in younger speakers than in older speakers. Take a textual distance of approximately 1,500 words between PREVIOUS and CURRENT: while in the production of older speakers, there is still a 75% likelihood that PREVIOUS and CURRENT match at this textual distance, the corresponding figure is only 70% in younger speakers.

5. Summary

The analysis in this chapter has shown that consideration of persistence-related factors enhances the researcher's ability to make predictions about whether speakers will employ synthetic or analytic comparison when they have the choice. More specifically, I believe that I have delivered evidence for the following claims:

Clearly, length of the adjective is the key conventional predictor of whether or not the adjective will take synthetic comparison. Also conforming with the literature is the importance of the syntactic function the adjective serves and whether or not it takes complements. At the same time, according to the data three of the baseline predictors – text frequency, the morphology of the adjective, and stress placement – do not have the effect suggested in previous scholarship. The analysis also tested two predictors not hitherto discussed – TTR as a proxy for lexical density and sentence length as a proxy for syntactic complexity – and found that their effect can be significant: the longer sentences are – and thus, by inference, the more complex they are syntactically – the smaller the odds for (less explicit) synthetic comparison; the higher type-token-ratios of the surrounding material are, the greater the odds for (lexically more concise) synthetic comparison. Overall, the independent variables proposed in previous research fit better to and correctly predict more of the variation in more formal data than in more informal data.

What about persistence? Overall, speakers switch between analytic and synthetic comparison only about one tenth of the time they should if switching were governed by chance. This overall switch rate seems to be unaffected by the corpus examined, and neither analytic nor synthetic comparison appears to be substantially more persistent than the alternative strategy. In logistic regression, we obtained significant evidence for α-persistence: if PREVIOUS is analytic rather than synthetic, the odds that the speaker will choose to employ synthetic comparison in CURRENT decrease substantially. I also showed that there is a forgetting function such that speakers forget about previous comparison choices as time passes by. As expected, persistence declines logarithmically rather than linearly. As for β-persistence, I presented evidence that triggers can manipulate the odds for the comparison strategy chosen for CURRENT, and that they do so increasingly as textual distance between the trigger item and CURRENT decreases. While we were unable to obtain evidence that the presence of a generic adjective taking synthetic comparison in the discourse increases the odds for synthetic comparison in CURRENT, the presence of the token *more* in the discourse immediately preceding CURRENT can clearly increase the odds for analytic comparison in CURRENT significantly. A possible reason why the token *more* triggers analytic comparison, but the affix *-er* does not trigger synthetic comparison, is that for one thing, *more* is a lexical item of its own; second, it has more phonological substance than *-er*, which is phonetically often just realized as [ə]. In psycholinguistic terms, then, it should not be surprising that *more* is a better prime than *-er*.

The relative importance of α-persistence and β-persistence turned out to be different between formal and informal registers. In formal data, α-persistence seems to be more important, whereas β-persistence appears to have more of an effect on informal discourse. Also, on aggregate, persistence in comparison strategy choice is stronger in the informal data than in the formal data.

Inclusion of the speaker variables AGE and SEX additionally improved the quality of this chapter's modeling of speakers' choices. First, the effect a synthetic trigger has on CURRENT decreases with increasing age, thus the main effect of β-persistence seems to become weaker the older speakers are. Second, the effect of a synthetic trigger on the comparison strategy chosen for CURRENT is greater in female speakers than in male speakers (in other words, β-persistence appears to be more potent in female speakers than in male speakers). There was no significant evidence for any interaction between the effect of the analytic-lexical trigger *more* and the speaker variables, nor between α-persistence and the speaker variables. However, there is apparent-time evidence that analytic comparison is indeed spreading. Finally, I showed that α-persistence declines faster in younger speakers than in older speakers, where the phenomenon appears to be more long-lived.

Chapter 5
Persistence in genitive choice

This chapter will investigate the alternation between the inflected genitive (henceforth: the *s*-genitive), as in (1a), and its periphrasis with *of* (henceforth: the *of*-genitive), as in (1b):

(1) a. ***anthropology's history*** *is indeed implicated in the scientific construction …* (CSAE 1034)
b. *it forces us to rethink …* ***the history of American anthropology*** (CSAE 1034)

Besides complementation strategy choice and comparison strategy choice, genitive choice is probably one of the most extensively researched areas of syntactic variation in the grammar of English.

1. Background and previous research

In historical terms, the *of*-genitive is the incoming form, which appeared during the 9th century. During the wholesale reorganization of the Old English case system, the *of*-genitive – which had by then established itself as an alternative to the *s*-genitive – was subject to a dramatic surge in usage, to an extent that the *s*-genitive was even close to extinction (Jucker 1993: 121). Intriguingly, though, the *s*-genitive recovered during the Modern English period and is even argued to be spreading right now (for instance, Potter 1969: 105–106; Dahl 1971: 141; Raab-Fischer 1995; Rosenbach 2003: 394–395).

In modern English, the two genitives – but particularly the *s*-genitive – are argued to encode a "a grab-bag" (Givón 1993: 264) of semantic and pragmatic relations. For the *s*-genitive alone, Quirk et al. (1985: 321–322) list eight different meanings (possessive, subjective, objective, the genitive of origin, descriptive, the genitive of measure, the genitive of attribute, the partitive genitive). It is fairly uncontroversial that objective relationships – e.g. *the imprisonment of the man* ($\sim$ someone imprisoned the man) – are more often than not encoded by the *of*-genitive, while subjective relationships – e.g. *the plane's arrival* ($\sim$ the plane arrived) – can be encoded by both genitives (Altenberg 1982; Biber et al. 1999: 303–302; Quirk et al. 1985: 1279–1281).

Some researchers claim that possessive relations have a privileged status in the semantics of the *s*-genitive (cf. Taylor 1989), though others stress that the two genitives convey generally the same meaning (cf. Altenberg 1982: 11; Chomsky 1970; Jespersen 1909: 312). It has also been proposed that the two constructions are semantically actually empty (for instance, Hudson 1984; Kempson 1977), or that they have an exclusively syntactic function (Chomsky 1986: 192). More recently, Rosenbach (2003) has suggested that iconic/natural principles are the reason why the *s*-genitive is favored with animate/topical possessors and with what she calls 'prototypical possessive relations.' In the very same monograph, though, Stefanowitsch (2003: 413) argues that the two genitives are distinct semantic-role constructions, with the *s*-genitive encoding a possessor-possessee relation and the *of*-genitive a part-whole relation unless, crucially, "the head noun itself specifies a different relation." Given this definitorial mess, it is not surprising that one occasionally encounters defeatist statements such as "any attempt to sum up 'the meaning' of the *s*-genitive is doomed" (Strang 1968: 109).

What is clear, however, is that of the many contexts in which either *s*-genitives or *of*-genitives can be observed, not all are choice contexts where the two genitives are semantically interchangeable, or where they both could even be used. Instead, there is a range of contexts where one of the two genitives has become obligatory. For instance, a well-known knock-out context for the *s*-genitive are partitive constructions of the quantitative and qualitative types: *a glass of water* vs. **a water's glass* (cf. Quirk et al. 1985: 1277–1278). Nonetheless, Quirk et al. (1985: 321) argue that "in many instances there is a similarity of function and meaning" between the *s*-genitive and *of*-genitive, and that often, "the two forms are equivalent in meaning and are both perfectly acceptable." Similarly, Jucker (1993: 121) observes that "a fairly large area of overlap . . . exists between the two constructions." It is these choice contexts that we shall seek to investigate with regard to persistence.

Formally and structurally, the two genitives clearly differ in their syntactic structure. This is why, psycholinguistically, the genitive alternation is susceptible to syntactic priming. Observe, however, that the *of*-genitive comes with the extra lexical/functional token *of*, which is why lexical priming, too, could potentially be involved in genitive choice. Finally, the two genitives differ in complexity and explicitness: "The *s*-genitive is characteristically more compact and less explicit in meaning. The nature of the connection to the head noun is left unspecified with the *s*-genitive, whereas postmodifiers usually contain more signals of syntactic/semantic relationships" (Biber et al. 1999:

300). On the other hand, the postmodification characteristic of *of*-genitive "produces a less dense and more transparent means of expression" (Biber et al. 1999: 302).

2. Previously suggested factors

When there is truly a choice between the two genitives, the literature lists the following factors as influencing this choice:

Phonology. A final sibilant in the possessor encourages use of the *of*-genitive. It follows that morphologically, a regular plural ending also encourages use of the *of*-genitive (Altenberg 1982).

Stylistics. The more informal the setting, the greater the preference for the *s*-genitive, and vice versa (Altenberg 1982: 284). Jucker (1993) showed that the *s*-genitive is more frequent in down-market and mid-market newspapers than in up-market papers, and more frequent in sports sections than in news sections. The tendency for *s*-genitive to be more frequent in less formal contexts may have to do with expressivity: according to Dahl (1971: 172), when the *s*-genitive is used with inanimate objects, this is usually done because of "the tendency to brevity and greater expressive force" (similarly, Biber et al. 1999: 302).

Regional differences. The *s*-genitive is more frequent in American English than in British English (for instance, Rosenbach 2003: 395–396; Hundt 1998).

Lexical factors. The lexical class of the dependent noun has generally been considered the most important factor determining genitive choice. The more human and animate a possessor, or the more it conveys the idea of animate things and human activity, the more likely it is to take the *s*-genitive (cf. Altenberg 1982: 117–148; Biber et al. 1999: 302–303; Dahl 1971: 140; Jucker 1993: 126–128; Quirk et al. 1985: 1277; Taylor 1989: 668–669). Rosenbach (2005) presents experimental as well as corpus evidence that although animate possessors tend to be shorter than inanimate ones, animacy and weight (see below) are, in fact, independent factors. Further, there are also some idiomatic expressions where the *s*-genitive is preferred, and it can often be found in expressions of time and measure (e.g. *the final quarter's profits*) and in some spatial adverbial expressions (Dahl 1971). More generally, Osselton (1988) argues that it is the general topic with which the speaker or writer is engaged which determines what nouns can take the *s*-genitive. Thus, “in a book on phonetics, *sound* will get its genitive [i.e. the *s*-genitive, BS] . . . and in a book on economics you can expect to find *a fund’s success, the pound’s strength, inflation’s consequences*,” and so on (Osselton 1988: 143).

Syntactic factors. Heavy restrictive postmodification of the possessum favors the *s*-genitive. The reason is that usage of the *of*-genitive in such contexts is likely to be understood as non-restrictive. The reverse holds when the possessor is postmodified, in which case the *of*-genitive is somewhat preferred (Quirk et al. 1985: 1281–1282; Altenberg 1982: 76–110). Therefore, (2b) is preferred to (2a), and (3b) is definitely preferred to (3a):

(2) a. *the arrival of a friend which had been expected for several weeks*
b. *a friend’s arrival which had been expected for several weeks*

(3) a. **a friend’s arrival who had been studying for a year at a German university*
b. *the arrival of a friend who had been studying for a year at a German university* (Quirk et al. 1985: 1281–1282)

At the same time, the principle of end weight (Behaghel 1909/1910) is operating in genitive choice: more complex, ‘heavier’ constituents tend to be placed towards the end, thus if the possessor is heavy, there is a general preference for the *of*-genitive; if the possessum is heavy, there is a general preference for the *s*-genitive (Biber et al. 1999: 304; Altenberg 1982: 76–79; Quirk et al. 1985: 1282; Rosenbach 2005). Hawkins (1994) – among many others – has claimed that end weight facilitates parsing while Wasow (1997) has suggested a speaker-centered account according to which the principle of end weight facilitates sentence planning. Hawkins (1994) has also argued that animacy (much like information status) is actually epiphenomenal in that animate entities are usually shorter, which is why they tend to be coded with the *s*-genitive; Rosenbach (2005) presents evidence against this view.

Discourse flow. Givenness, or end-focus, also plays a role in the choice of the genitive. Thus, if the possessor is discourse-new, the *of*-genitive is preferred; if the possessum is discourse-new, the *s*-genitive is preferred (Biber et al. 1999: 305; Quirk et al. 1985: 1282).

Structural parallelism. According to Altenberg (1982), the genitive construction just used tends to be repeated, when possible, at the next opportunity. Altenberg (1982: 290) concludes that "parallelism is an important secondary factor . . . , a factor which extends and supports a choice determined by other primary factors in the context." Needless to say, it is precisely this factor that the present chapter seeks to investigate in detail.

No multivariate analysis of the factors determining genitive choice has, to my knowledge, been conducted so far.

3. Method, data and independent variables

3.1. Method and data

As has been pointed out above, this study's analysis will be one of choice contexts only. Therefore, all those cases where there is no (potential) variability between the *of*-genitive and the *s*-genitive will be omitted. No variability means that usage of the other genitive would make the expression ungrammatical or very odd. This judgement cannot be made by software, which is why the analysis in this chapter relies to a large extent on manual coding of manageable datasets. Thus, the entire CSAE as well as a subset of FRED[25] was parsed manually to identify the genitive choice contexts in the data.

More specifically, according to the coding protocol, the following *of*-genitive contexts were *not* coded as variable: (i) cases where the *s*-genitive has a partitive or appositive meaning and where use of the *s*-genitive would result in odd final prominence (e.g. *the part of a problem* vs. **the problem's part*; (ii) partitive constructions of the quantitative and qualitative types; (iii) when the possessum of the *of*-genitive is indefinite (e.g. *a friend of the president*); (iv) in the cases of certain possessum nouns (such as *idea, issue*, etc.) which are not used with an *s*-genitive and build compound nouns instead (e.g. *the issue of the budget deficit was discussed this morning*). Correspondingly, the following *s*-genitives were *not* coded as variable: (i) local genitives

Table 9. Genitive choice: distributional variation across corpora

corpus	*N*	*N s*-genitive	*N of*-genitive
CSAE	332	160 (48.2%)	172 (51.8%)
FRED	1,818	1,084 (59.6%)	734 (40.4%)
total	**2,150**	**1,244 (57.8%)**	**906 (42.1%)**

(e.g. *we are meeting at Paul's tonight*); (ii) when the relation existing between the possessor and possessum simply cannot be paraphrased by an *of*-genitive (e.g.*the world's best universities*). The so-called independent genitive – whether analytical or synthetic – was coded whenever possible (e.g. *her memory is like that of an elephant* vs. *her memory is like an elephant's*).

A Perl script then extracted the manually identified variables and coded them for the standard variables, such as PREVIOUS, and most of the variables specific to genitive choice. In a final step, the genitive variables in the database were then post-coded manually for lexical class (see below, section 3.2). This procedure yielded a database of 2,150 genitive choice contexts, which Table 9 breaks down according to corpus and genitive type. As can be seen, optional *of*-genitives and optional *s*-genitives are virtually equally frequent in the CSAE. In FRED (at least in the subset analyzed), there appears to be a slight preference for the *s*-genitive, though we will see later (section 4.1) that there are significant differences between dialect areas with regard to such preferences.

3.2. Independent variables

In addition to the standard variables discussed in chapter 2 (PREVIOUS, TEXTDIST, SENTENCELENGTH, TTR, SAMETURN, SAMESPEAKER), the following independents will be included in multivariate analysis:

3.2.1. Previously suggested and persistence-unrelated predictors

LEXICAL CLASS of the possessor (henceforth: LEXCLASS). For this variable, all possessor noun phrases in the database were coded following the classification of possessor NPs suggested in Altenberg (1982: 120–

123), which recalls the noun classes in Silverstein's (1976) animacy hierarchy: inanimate > animal > human common NP > proper NP.

	example	coding
animate human individual proper NPs	*Tom, God*	5
animate human individual common NPs	*friend, woman*	4
animate human collective NPs	*nation, parish*	3
animate animal NPs	*dog, cow*	2
inanimate abstract NPs	*morning,yesterday*	1
inanimate concrete NPs	*world, sea, study*	0

Hypothesis: Altenberg (1982), among many others, has claimed that the higher the lexical class of the possessor noun phrase is positioned in the above table, the greater the likelihood for the *s*-genitive. I expect to obtain the same relationship in my analysis.

LENGTH (in words) of the possessor phrase and the possessum[26] phrase (henceforth: POSSESSORLENGTH and POSSESSUMLENGTH). Consider (4): the possessor phrase – *a bloody scoundrel* – commands three words, the possessum phrase – *last refuge* – commands two words.

(4) *Politics is* ***the last refuge of a bloody scoundrel****, ain't it? I say that's* ***the last refuge of a bloody scoundrel*** *– politics.* (FRED SFK035)

Hypothesis: In accordance with the principle of end weight (cf. Behaghel 1909/1910), we expect that (i) the longer the possessor phrase, the greater the likelihood that the *of*-genitive will be used (because it places the possessor phrase second); and (ii) the longer the possessum phrase, the greater the likelihood that the *s*-genitive will be used (because it places the possessum phrase second).

PHONOLOGICAL SHAPE of the possessor (henceforth: FINALSIB). Does the possessor phrase end in the grapheme <s>, as the possessor phrase in (5) does (coded 1 if the possessor phrase ends in <s>, and 0 if it does not)?
Hypothesis: If the possessor phrase ends in the grapheme <s>, we expect, given the literature, the *s*-genitive to be dispreferred.

(5) *I'm going to be saving a lot of money working here so if I'm making decent money I'll be able to uh . . . get something on my own. . . . With* ***the help of my parents*** *of course.* (CSAE 0404)

INFORMATION STATUS of the possessor and the possessum (henceforth: POSSESSORGIV AND POSSESSUMGIV). Has the lemma of the head of the possessor phrase or the lemma of the head of the possessum phrase been mentioned in a discourse context of 100 words (this is equivalent to 4 to 7 sentences) prior to CURRENT? To illustrate: the lemma of the head of the possessor phrase in *the farmers' bright red barns* would be *farmer*, the lemma of the head of the possessum phrase would be *barn* (coded 1 if the possessor or possessum is discourse-old, and 0 if the possessor or possessum is discourse-new).
Hypothesis: If the possessor has *not* been mentioned (if, therefore, it is discourse-new), we expect the *of*-genitive to be more likely; if the possessum is discourse-new, we expect the *s*-genitive to be more likely.

FRED DIALECT AREA (henceforth: FRED-AREA). This variable is obviously relevant for FRED only and is meant to tap into how genitive choice differs across the dialect regions (Hebrides, Midlands, North, Southwest, Wales, and Southeast) sampled in the FRED corpus subset under analysis.

3.2.2. *Additional, persistence-related predictors*

TEXTUAL DISTANCE to the last occurrence of the token *of* (henceforth: TEXTDIST-OF). This is a β-persistence variable. It is conceivable that generic, non-genitive occurrences of the token *of* (as in *stories* of *her travels*) make it more likely, for instance via lexical priming, that a speaker will tend to go for an *of*-genitive instead of an *s*-genitive next time he or she has a choice. TEXTDIST-OF measures the textual distance, in the *ln* of interjacent words, between CURRENT and the last generic occurrence of *of.* By way of illustration, in (6) a generic occurrence of *of* is followed by an optional *of*-genitive five words later (thus, TEXTDIST-OF would be *ln* 5 = 1.61).

(6) *So, you got this Oscar there, swimming around in the tank, . . . with like, . . . a goldfish sticking out* ***of*** *his mouth, you know, the* ***the head of a goldfish****, so you could see the little goldfish's eyes . . .* (CSAE 0403)

Hypothesis: As a working hypothesis (and to some extent contrary to Bock 1989, who did not find that closed-class items have an effect on the strength of priming), we expect that as TEXTDIST-OF decreases – and thus, as an *of*-genitive trigger gets closer to CURRENT – the odds for an *of*-genitive in CURRENT increase.

IDENTITY of the possessor or the possessum in PREVIOUS and CURRENT (henceforth: POSSESSORID and POSSESSUMID). This variable determines whether two neighboring genitive constructions dominate (i) exactly the same possessor phrase (POSSESSORID) or (ii) exactly the same possessum phrase (POSSESSUMID) (coded 1 if the possessor/possessum phrases are identical, and 0 if they are not). Example (7) illustrates a case where two successive genitive constructions dominate the same possessor phrase and the same possessum phrase.

(7) *Politics is* ***the last refuge of a bloody scoundrel****, ain't it? I say that's* ***the last refuge of a bloody scoundrel*** *– politics.* (FRED SFK035)

Hypothesis: If the possessor/possessum phrases are identical, we expect – in accordance with Cleland and Pickering (2003), who showed that production priming is stronger when the head words in the prime and the target match – that persistence between PREVIOUS and CURRENT is stronger than it would be otherwise.

Heavy restrictive postmodification of either the possessor or the possessum is quite rare in the data – there is only one such case in the CSAE (*she gave me the name of this other who's Amherst Architect*), and only a handful of occurrences in the FRED dataset. For reliable statistical analysis of the variable, the number of observations is too low, which is why the variable had to be dropped from analysis. Table 10 summarizes the independents considered in this chapter.

Table 10. Genitive choice: independent variables considered

variable	type	coding method
	a. previously suggested and persistence-unrelated independents	
SENTENCELENGTH*	scalar	software
TTR*	scalar	software
LEXCLASS	six-way ordinal	manual
POSSESSORLENGTH	scalar	software
POSSESSUMLENGTH	scalar	software
FINALSIB	two-way categorical	software
POSSESSORGIV	two-way categorical	software
POSSESSUMGIV	two-way categorical	software
FRED-AREA	five-way categorical	software
	b. persistence-related independents	
PREVIOUS*	two-way categorical	software
TEXTDIST*	scalar	software
TEXTDIST-OF	scalar	software
SAMETURN*	two-way categorical	software
SAMESPEAKER*	two-way categorical	software
POSSESSORID	two-way categorical	software
POSSESSUMID	two-way categorical	software

* independent variable discussed in chapter 3, section 1.

4. Results

4.1. Baseline variation

Let us first establish how the baseline predictors play out in genitive variation (Table 11). Let us begin by noting which predictors are *not* selected as significant in logistic regression: the two information status variables POSSESSUMGIV and POSSESSORGIV. As operationalized through these two variables, information status does not seem to influence genitive choice (this statement will be somewhat qualified in section 4.2 below).

The odds ratios associated with SENTENCELENGTH and TTR dovetail nicely with the working hypothesis. Exp(*b*) values associated with SENTENCELENGTH are greater than 1 in both corpora (though significantly so only in

Table 11. Genitive choice: odds ratios associated with baseline predictors in logistic regression

	CSAE	FRED
SENTENCELENGTH	**1.02** *	1.01
TTR	0.99	**0.96** **
LEXCLASS	1.17	**0.36** ***
POSSESSORLENGTH	**6.58** ***	**1.78** *
POSSESSORGIV(1)	1.21	1.07
POSSESSUMLENGTH	1.01	**1.33** *
POSSESSUMGIV(1)	0.44	0.71
FINALSIB(1)	**3.48** *	**2.72** ***
POSSESSORLENGTH * LEXCLASS	**0.68** ***	1.05
POSSESSUMGIV(1) * LEXCLASS	1.26	**0.83** *
FRED-AREA	n.a.	– ***
FRED-AREA (Hebrides)	n.a.	**2.03** *
FRED-AREA (Midlands)	n.a.	**2.58** ***
FRED-AREA (North)	n.a.	1.07
FRED-AREA (Southwest)	n.a.	1.57
FRED-AREA (Wales)	n.a.	**2.74** ***
model intercept	0.15	12.07 *
N	332	1,818
model χ^2	102.72 ***	1,208.28 ***
R^2	0.355	0.656
% correct (baseline)	70.8 (51.8)	86.6 (59.6)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for the *of*-genitive.

the CSAE) while those associated with TTR are smaller than 1 in both corpora (though significantly so only in FRED). Recall now that we have seen that the *of*-genitive is the syntactically and lexically more explicit option. The more explicit option, then, is preferred when sentence length increases (thus, when syntactic complexity of the surrounding material is higher), while the *of*-genitive is dispreferred when type-token ratios decrease (thus, when lexical density of the surrounding lexical material is lower). This relationship is as expected.

Lexical class (LEXCLASS) is highly significant in FRED and has the following impact on genitive choice: for each one-point rise of the possessor in Altenberg's sixfold animacy hierarchy (i.e. when animacy of the possessor increases), the odds for the *of*-genitive diminish considerably (exp(*b*) = 0.36), exactly as hypothesized. For some reason, lexical class does not appear to be that relevant for genitive choice in the CSAE.

The length of the possessor phrase (POSSESSORLENGTH) also influences genitive choice in the expected fashion. For each additional word the possessor phrase commands, the odds for the *of*-genitive increase about sixfold in the CSAE (exp(*b*) = 6.58) and about twofold in FRED (exp(*b*) = 1.78). In accordance with the principle of end weight, long possessors indeed tend to be placed towards the end by means of the *of*-genitive. By much the same token, we had expected increased length of the possessum phrase to be negatively associated with the *of*-genitive. The opposite is true, as the odds ratios slightly greater than 1 (exp(*b*) = 1.01/1.33) demonstrate. It appears that there is a preference to code both long possessum phrases and long possessor phrases (the latter to a much greater extent) by the *of*-genitive.

The phonological shape of the possessor phrase is also a significant predictor. If the possessor (phrase) ends in a sibilant (more precisely, in terms of this study's research design, if it ends in the grapheme <s>), the odds for the *of*-genitive increase about threefold (exp(*b*) = 3.48/2.72). This finding, too, is fully consonant with the literature on genitive choice.

Two moderately interesting interaction terms turned out to be significant in logistic regression. First, the interaction POSSESSORLENGTH ∗ LEXCLASS is associated with a significant odds ratio of 0.68 in the CSAE. This means that increasing length of the possessor phrase and increased animacy of the possessor phrase work against each other. Second, the interaction POSSESSUMGIV(1) ∗ LEXCLASS is significant in FRED. One interpretation of its odds ratio of 0.83 is that when the possessum is discourse-old, animacy of the possessor has more influence on genitive choice than when the possessor is discourse-new. Finally, in FRED, dialect areas account for a good deal to genitive variation (FRED-AREA). Taking the Southeast as the statistical 'baseline' area (this is meant as an atheoretical, purely statistical procedure, and implicates no substantive claim of any sort), it emerges that compared to the Southeast, dialect speakers from the Hebrides, the Midlands, and Wales have a significant preference for the *of*-genitive.

When the relative importance of the individual predictors is compared, two things deserve attention:

- In the CSAE, length of the possessor phrase is the single most important determinant of genitive choice, followed by sentence length (i.e. syntactic complexity of the surrounding material) and whether or not the possessor phrase ends in a sibilant.
- In FRED, animacy of the possessor phrase is by far the most important determinant of genitive choice. Second is whether the possessor phrase ends in a final sibilant, followed by the two weight variables.[27]

The factors so far discussed help predict between 71% (CSAE) and 87% (FRED) of speakers' linguistic choices accurately, and account for between 36% (CSAE) and a very decent 66% (FRED) of the observable variance in genitive choice. For some reason the varieties of English sampled in FRED conforms much better with claims in the literature on genitive variation than the very conversational American English sampled in the CSAE.

4.2. Persistence-induced variation

Let us now determine how the picture changes when persistence-related predictors are factored in. Evidence on this point is shown in Figure 9, which plots switch rates between the *of*-genitive and the *s*-genitive for speakers in the dataset under analysis. All but one speaker switch less often from the *s*-genitive to the *of*-genitive than we would expect if switch rates were proportional to overall variant proportions (indicated by the dotted lines), and *all* speakers switch from the *of*-genitive to the *s*-genitive less often than expected.

The heavy lines indicate regression estimates of actual switch rates. By being virtually horizontal, these statistically confirm that speakers practically do not switch between the two genitives. More detailed information on switch rates, broken up according to corpus and switch direction, is provided in Table 12. In terms of switch rates, no substantial differences exist between FRED and the CSAE or between switch directions: the linear regression estimates indicate that switch rates are in the $0.01x$ to $0.05x$ range. Given that the 'natural' switch rate would be $1x$, both genitives are extremely sticky.

For a more fine-grained analysis, Table 13 regresses the persistence-related predictors against CURRENT. This considerably enhances both predictive efficiency and explanatory power, compared to the baseline model; the increase in the model χ^2 is statistically significant.[28] In FRED, the model now ex-

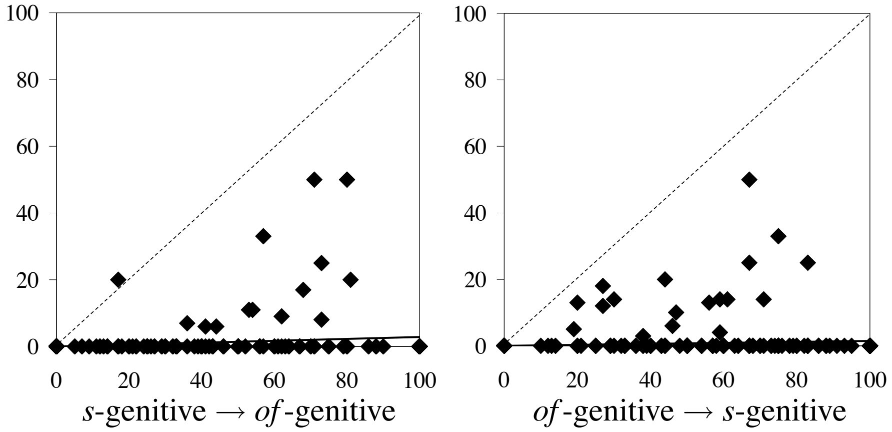


Figure 9. Switches in genitive choice as a function of overall proportion of genitives (relative frequency of switches, in %, on *y*-axis; relative frequency of the switched-to genitive type, in %, on *x*-axis) in both FRED and the CSAE. Each dot represents one speaker. Dotted diagonal line represents null hypothesis that switch rate is proportional to variant proportions. Heavy line indicates linear trend (*s*-genitive → *of*-genitive: $y = 0.03x$, *of*-genitive → *s*-genitive: $y = 0.02x$)

Table 12. Linear regression estimates of switch rates in genitive choice across corpora (y is the relative frequency of A → B switches, in %; x is the relative frequency of B forms, in %; the expected linear relationship, uninfluenced by persistence, is $y = x$)

corpus	*s*-genitive → *of*-genitive	*of*-genitive → *s*-genitive
CSAE	$y = 0.05x$	$y = 0.01x$
FRED	$y = 0.02x$	$y = 0.02x$

plains a satisfactory 71% of the observable variance (up from approximately 66%) and predicts speakers' genitive choices accurately in 88% of all cases (up from 87%). In the CSAE, the statistical improvement is better than in FRED: variance explained rises from approximately 36% to 53%, and predictive efficiency is improved by 4 percent points to approximately 75%.

Table 13. Genitive choice: odds ratios associated with persistence-related predictors in logistic regression (baseline predictors are included, but not displayed)

	CSAE	FRED
PREVIOUS(S-G.)	**0.02** ***	**0.00** ***
PREVIOUS(S-G.) ∗ TEXTDIST	1.19	1.05
PREVIOUS(S-G.) ∗ SAMETURN(1)	**4.65** *	0.73
PREVIOUS(S-G.) ∗ SAMESPEAKER(1)	0.38	1.15
PREVIOUS(S-G.) ∗ TTR	1.02	**1.09** **
PREVIOUS(S-G.) ∗ POSSESSORLENGTH	1.23	**1.64** *
PREVIOUS(S-G.) ∗ POSSESSUMLENGTH	1.52	**1.67** *
POSSESSORID	**0.00** ***	**0.00** ***
POSSESSUMID	∞ ***	∞ ***
TEXTDIST-OF	0.80	**0.87** *
model intercept	2.58	905.90 ***
N	295	1,654
model χ^2	150.68 ***	1238.05
R^2	0.534	0.711
% correct (baseline)	75.3 (53.6)	87.9 (59.2)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for the *of*-genitive.

4.2.1. α-persistence

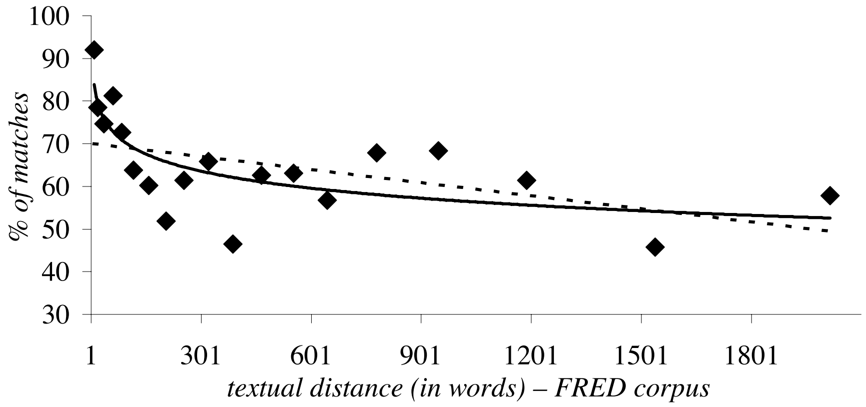
The main effect of PREVIOUS – the α-persistence variable in this study – is highly significant and extraordinarily sizable in both FRED and the CSAE. This was to be expected, given that we had already seen in our discussion of switch rates (Figure 9) that speakers are highly disinclined to switch between the two genitives. In the CSAE, if an *s*-genitive was employed in the last variable site in discourse (and conditioned on all other interactional factors discussed below being zero), the odds for an *of*-genitive in CURRENT are reduced by a considerable 98% (exp(b) = 0.02). The corresponding main effect of PREVIOUS on CURRENT in FRED is such that statistically, an *s*-genitive is actually *never* followed by an *of*-genitive (exp(b) = 0.00) when interactional factors are controlled for.

I now discuss how the main effect of PREVIOUS changes when persistence interacts with other variables. The interaction between α-persistence and the two turn-taking variables (PREVIOUS $*$ SAMETURN and PREVIOUS $*$ SAMESPEAKER, respectively) did not turn out to be statistically significant in logistic regression. In a similar vein – probably due to comparatively low *N*s – the interaction between PREVIOUS and TEXTDIST (the interaction between persistence and recency of use of PREVIOUS) missed statistical significance. Notice, however, that the (insignificant) odds ratios associated with that interaction are greater than 1 (1.19 and 1.05, respectively), which shows the theoretically expected effect.

Yet, Figure 10 demonstrates that there actually is a relationship between the strength of α-persistence and textual distance between PREVIOUS and CURRENT. When plotting the strength of persistence on the *y*-axis against textual distance between two genitive choice contexts on the *x*-axis, it becomes apparent that at least in FRED (the data for the CSAE are more erratic, presumably due to low *N*s), the percentage of matched pairs (pairs where the same genitive is employed in both CURRENT and PREVIOUS) clearly does not bob around randomly. Instead, the more recently a genitive choice was made, the more likely speakers are to go for the same genitive at the next opportunity. A statistical analysis of the curve fits[29] shows that the forgetting function that describes this relationship is best described as logarithmic, at least in FRED (in the CSAE, both fits are unacceptably bad).

	FRED	CSAE
adjusted R^2 linear	0.20 *	-0.04
adjusted R^2 logarithmic	0.54 ***	0.00
df	17	17

POSSESSORID and POSSESSUMID appear to influence genitive choice in a rather unanticipated way. Recall that we had initially assumed that if two neighboring genitive variables involve the same possessor phrase (POSSESSORID) or the same possessum phrase (POSSESSUMID), or both, α-persistence between PREVIOUS and CURRENT would be even stronger than otherwise (cf. Cleland and Pickering 2003). However, we actually obtained no such interaction effect in logistic regression – the interaction terms PREVIOUS(S-G.) $*$ POSSESSORID/POSSESSUMID were so insignificant that they had to be dropped from analysis altogether. However, as it turns out, POSSESSORID and POSSESSUMID are very highly significant predictors of genitive choice



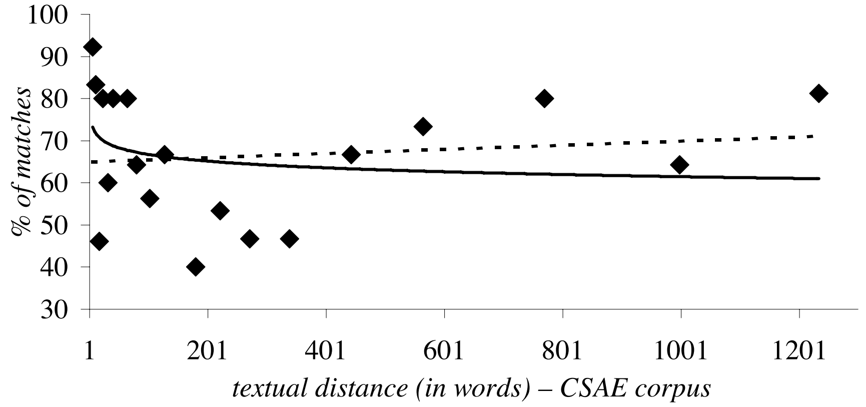


Figure 10. Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same genitive type is used) as function of textual distance between CURRENT and PREVIOUS. Heavy line represents logarithmic estimate of the relationship, dotted line represents linear estimate of the relationship

on their own (note that because no interaction with PREVIOUS is involved here, this is not really related to persistence). What does this mean? The only way to make sense of this is to interpret these variables as tapping information status (as might be remembered, we could not obtain evidence for any relevance of information status to genitive choice from considering the variables POSSESSORGIV and POSSESSUMGIV): the odds ratio associated with POSSESSORID is exactly 0, therefore if two neighboring genitive constructions command exactly the same possessor phrase, the second genitive, according to the data, will be an *s*-genitive. As the possessor phrase is given, the *s*-genitive establishes old-before-new order. The inverse holds for POSSESSUMID: if two neighboring genitives command exactly the same possessum phrase, the odds for an *of*-genitive in the second slot increase manifold (exp(b) = ∞), according to this study's analysis. This means that when the possessum is given, the second of two successive genitives will virtually always be an *of*-genitive. From a discourse flow perspective, the *of*-genitive is optimal under such circumstances since it places the possessum before the possessor.

The interaction between PREVIOUS and TTR returns exp(b) values greater than 1 in both FRED (1.09; significant) and the CSAE (1.02; insignificant). Because the main effect of PREVIOUS is smaller than 1, this means that α-persistence actually weakens in lexically complex contexts. Given Tannen (1987), we would have expected the inverse relationship.

4.2.2. β-persistence

The model also included one β-persistence predictor: TEXTDIST-OF. The variable measures the *ln* of textual distance between the slot for which a genitive choice has to be made, CURRENT, and the last generic occurrence of the token *of* (for instance, as in *stories* of *her travels*). In logistic regression, the predictor has the expected effect on genitive choice and returns remarkably similar odds ratios in both FRED (where it is statistically significant) and the CSAE (where with p = 0.09, it misses statistical significance narrowly). For every one-unit increase in the *ln* of textual distance between CURRENT and the last generic occurrence of the token *of*, the odds for the *of*-genitive decrease by between 20% (CSAE; exp(b) = 0.80) and 13% (FRED; exp(b) = 0.87). This is equivalent to saying that as recency of use of the token *of* increases, the odds for the *of* genitive at the next opportunity rise, too. Figure 11 visualizes this relationship by plotting the relative frequency (in percent) of the *of*-genitive against textual distance to the last generic occurrence of the token *of*. In both corpora, this relationship is overall such that the relative frequency of the *of*-genitive is higher when textual distance is smaller.[30]

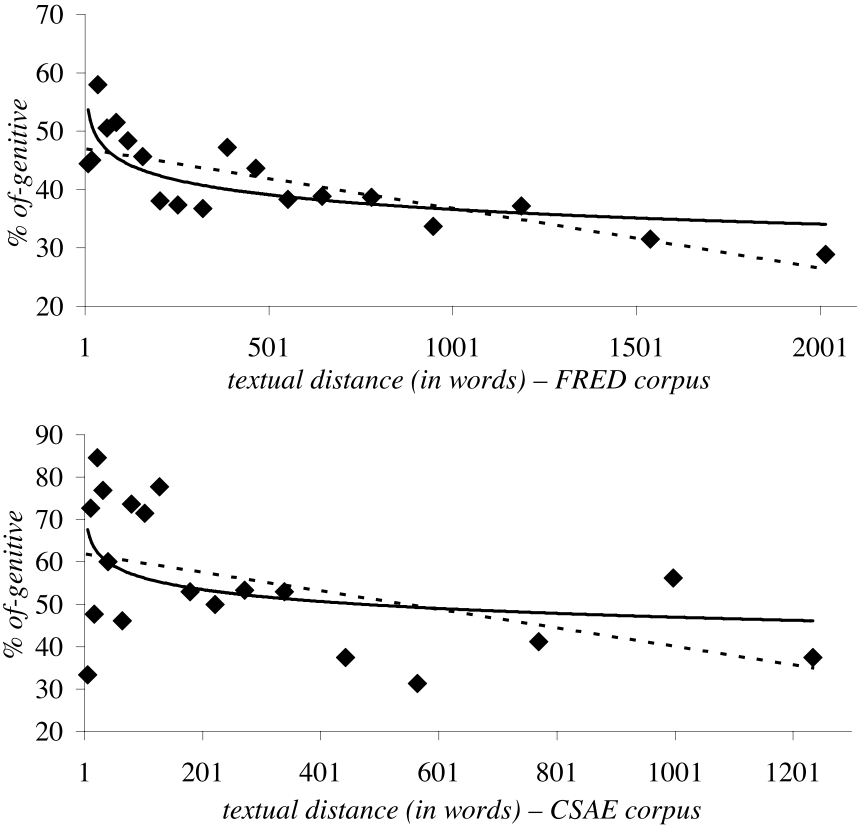


Figure 11. Share of the *of*-genitive (on *y*-axis) as a function of textual distance to the last *of* trigger (on *x*-axis). Heavy line represents logarithmic estimate of the relationship, dotted line represents linear estimate of the relationship

5. Summary

In conclusion, the analysis of variation in genitive choice in FRED and the CSAE seems to suggest the following:

Almost all of the predictors traditionally discussed in the literature on genitive choice turned out to have the expected effect in logistic regression. First, the more animate the possessor, the more likely the *s*-genitive. Second, the longer the possessor, the more likely the *of*-genitive. That both of these predictors turned out to be significant individually in FRED strongly supports claims that animacy and weight are independent factors (cf. Rosenbach 2005). Along these lines, we also found that, interestingly, increasing length of the possessum is positively correlated with the *of*-genitive (according to the principle of end weight, it shouldn't be). This essentially means that long and heavy genitives (i.e. genitives with long possessor and/or possessum phrases) in general have a preference for the *of*-genitive. Presumably, this is because the *of*-genitive is the more explicit option (cf. Biber et al. 1999: 300) and thus might help ease the processing load implicated by heavy genitives. Third, if the possessor ends in the grapheme <s> (which almost always means that it ends in a final sibilant), the *of*-genitive is preferred as well, a finding once again which, given the literature, should surprise no one. I moreover tested how SENTENCELENGTH (as a proxy for syntactic complexity) and TTR (as a proxy for lexical complexity) impact genitive choice. The results dovetailed nicely with this study's working hypothesis: the more explicit *of*-genitive is preferred when syntactic complexity is high (cf. Rohdenburg 1996); it is dispreferred when lexical complexity is low.

Information status is an issue of its own. It is received wisdom that the *of*-genitive is preferred with discourse-old possessums and the *s*-genitive with discourse-old possessors (in both scenarios, given-before-new order is established). Accordingly, an attempt was made to operationalize information status through two variables (POSSESSORGIV and POSSESSUMGIV), which checked whether the lemma of the head of the possessor/possessum phrase were mentioned in the previous discourse – in other words, if they were given. The two variables were not even remotely significant in regression. On the other hand, the model included two variables (POSSESSORID and POSSESSUMID) that were supposed to interact with persistence. They did no such thing. However, it turns out that these two variables unexpectedly tapped information status: being sensitive to whether the *whole* possessor or possessum phrase was used previously, this study's regression estimates suggested that for a given slot in which the same possessor/possessum phrase is repeated, that genitive type will be chosen which establishes old-before-new order. Ergo, the findings suggest that the *whole* possessor or possessum phrase (as operationalized through POSSESSORID and POSSESSUMID) is relevant with regard to information status, but a single head noun (as operationalized though POSSESSORGIV and POSSESSUMGIV) is not.

We saw that persistence is an important determinant of genitive choice. I demonstrated that if one only relied on baseline predictors of genitive choice, one would miss out between 4% (FRED) and even 20% (CSAE) of the variation, variation which persistence is responsible for – in other words, one would erroneously assume that this variation is free. According to this study's analysis, it is highly patterned.

Four observations with regard to persistence in genitive choice strike me as especially important. First, switch rates between the two genitives are exceedingly low, thus genitives are 'sticky' variables. By the same token, in logistic regression the main effect of a previous choice on an upcoming choice (α-persistence) is huge: in FRED, usage of an *s*-genitive reduces the odds that an *of*-genitive will be used next time by 98%; the corresponding figure for the CSAE is 100%. This effect, however, weakens as textual distance between two genitive sites increases. The forgetting function that describes this relationship is logarithmic. Furthermore, α-persistence is weaker in informationally dense environments (i.e. when TTR is high). This contradicts the working hypothesis that parallel patterns are preferred in lexically dense contexts because of processing efficiency advantages (cf. Tannen 1987). There is also evidence for β-persistence: the token *of*, used in non-genitive contexts (e.g. *stories* of *her travels*) can trigger an *of*-genitive in choice contexts. This effect becomes stronger as recency of use of a non-genitive *of* increases. Let me also add that in genitive choice, α-persistence appears to be vastly more powerful than β-persistence.

Chapter 6
Persistence in future marker choice

This chapter will investigate persistence effects in a comparatively neglected grammatical alternation in the grammar of English, that is, in the choice speakers have between the future marker families BE GOING TO and WILL (/SHALL). Each of these highly grammaticalized options to overtly express futurity in English has semi-institutionalized variant forms in transcribed corpora: *be going to*, as in (1a), and *gonna*, as in (1b); *will*, as in (1c), cliticized *'ll*, as in (1d), *won't*, as in (1e), and *shall/shan't*, as in (1f).

(1) a. *Matt **is going to** go to London tomorrow.*
b. *Matt **is gonna** go to London tomorrow.*
c. *Matt **will** go to London tomorrow.*
d. *Matt'**ll** go to London tomorrow.*
e. *Matt **won't** go to London* tomorrow.
f. *Matt **shall/shan't** go to London tomorrow.*

1. Background and previous research

Almost needless to say, the WILL/SHALL variants are the older forms. While Danchev and Kytö (1994) have shown that BE GOING TO must have developed into a future marker prior to the middle of the 17th century, Mair (2004) reports that "a marked rise in frequency did not occur until the end of the 19th century, but continues unabated in the present."

Previous research on the alternation between BE GOING TO and WILL/SHALL has primarily dealt with the following issues: (i) alleged semantic and/or pragmatic differences between BE GOING TO and WILL/SHALL, (ii) stylistic, regional, or sociolinguistic variation, or (iii) text frequencies, both synchronically and diachronically. In summary, WILL/SHALL is agreed to be the unmarked or simplex future which is employed to make a "plain statement about the future" (Close 1988: 51), with a possible overtone of obligation or volition (Kytö 1990: 277 and Wekker 1976: 40). BE GOING TO, in contrast, is generally assumed to suggest "prior intention, imminence, or inevitability" (Nicolle 1997: 355), "dynamic current orientation" (Haegeman 1983:

157), "future culmination of present intention or cause" (Haegeman 1989: 293; similarly, Nicolle 1997: 373), immediate or proximal futurity, inceptive present, and intentionality (Binnick 1971), or that there are "indications in the present that something will happen" (Wekker 1976: 124). Whatever the differences in meaning between the two options may be, it has proven notoriously hard to pin them down: after all, the actual choice for one or the other construction "has a scarcely perceptible effect on meaning" (Quirk et al. 1985: 218), which is why "it is difficult to discover any simple sentences in which either *will* yields a clearly definable sense which *going to* does not" (Hall and Hall 1970: 138). Similarly, Danchev et al. (1965: 384) argue for overall synonymy, and Palmer (1974: 163) asserts that "in most cases, there is no demonstrable difference between *will* and *be going to*." Haegeman (1989) has argued that whatever the difference is between BE GOING TO and WILL, it must be pragmatic rather than truth-conditionally semantic. In sum, the assumption of rough semantic equivalence between BE GOING TO and WILL/SHALL can certainly be justified.[31]

The alternation between BE GOING TO and WILL/SHALL is for one thing a syntactic one. The two markers differ in the complexity of the auxiliary node and in the complementation of the auxiliary (bare infinitive vs. *to*-infinitive). Additionally, the two constructions differ in the lexical and grammatical material they are composed of: BE GOING TO necessarily involves the primary verb *to be*, the verb *to go* (or a reduction thereof) and the infinitive marker *to* (or a phonological reduction thereof). WILL/SHALL is clearly more economic syntactically and lexically. Viewed from a production priming perspective, this is an alternation where syntactic priming is most indistinguishable from lexical priming.

2. Previously suggested factors

A couple of factors influencing the alternation between BE GOING TO and WILL/SHALL have been identified in previous research.

Register. It has been shown that as the informality of the setting increases, the contracted and/or cliticized variant forms such as *won't* or *'ll* gain in frequency (see, among others, Close 1988). At the same time, BE GOING TO is in general more widespread in informal settings (Berglund 1999b, 2000a, 2000b; Close 1988; Mair 1997a; Wekker 1976).[32]

Variety of English. All other things being equal, BE GOING TO is demonstrably more frequent in American English than in British English (see, among others, Biber et al. 1999; Hundt 1997; Mair 1997b; Tottie 2002).

Type of syntactic environment. It is well known that in standard English, WILL/SHALL is bad in some temporal and conditional subclauses, so that BE GOING TO is sometimes the only option there (cf. Binnick 1971; Comrie 1982, 1985; Danchev et al. 1965; Declerck 1991; Hall and Hall 1970; Wekker 1976). Szmrecsanyi (2003) found that in syntactically dependent and syntactically complex environments in general, there is a preference for BE GOING TO. Berglund (1999b), Berglund (2000b), and Szmrecsanyi (2003), finally, report that BE GOING TO is preferred over WILL in contexts of negation.

To my knowledge, no multifactorial analysis of these factors has been conducted so far.

3. Method, data and independent variables

3.1. Method and data

The following variant forms of BE GOING TO and WILL will be considered in this chapter's analysis:

- *be going to* + inf.
- *be gonna* + inf.
- *will* + inf.
- *'ll* + inf.
- *won't* + inf

Due to exceedingly low frequencies and its marginal status in present-day spoken English (cf. Kjellmer 1998; Tottie 2002; Trudgill 1984) *shall* (and *shan't*) will not be considered in this chapter. Also, past tense forms with BE GOING TO (as in *I've forgotten what* I was gonna *say,* DS KB0) will be

Table 14. Future marker choice: distributional variation across corpora

corpus	*N*	*N* BE GOING TO	*N* WILL
CG	39,774	10,497 (26.4%)	29,277 (73.6%)
DS	39,640	11,223 (28.3%)	28,417 (71.7%)
CSPAE	18,377	5,332 (29.0%)	13,045 (71.0%)
CSAE	1,354	570 (42.1%)	784 (57.9%)
FRED	4,861	856 (17.6%)	4,005 (82.4%)
total	**104,006**	**28,478 (27.4%)**	**75,528 (72.6%)**

excluded from analysis since these are *a priori* not possible with WILL. The condition of interchangeability underlying this study's method would not be satisfied had they been included (see Berglund 1999a for a similar coding decision). As far as the analysis of the data is concerned, no difference will be made between the individual variant forms. That is, *be going to* and *gonna* on the one hand and *will*, *'ll* and *won't* on the other hand will be considered true variant forms, with no distributional idiosyncracies of their own. This is certainly an abstraction, but one that is justifiable.

Extraction of the above forms from the texts, and classification into one of the two paradigms, was conducted automatically using Perl scripts. For the POS tagged corpora, this method yielded an accuracy rate of 98% (the error rate is mainly due to incorrect tagging of the corpora). For the corpora without POS tagging, the accuracy rate was approximately 91%, since the script could not identify and omit spatial forms of *be going to* (e.g. *I am going to school*). These spatial occurrences were discarded from the database manually.[33]

Analysis of the corpora yielded a database of 104,006 future marker instances. Table 14 gives a breakdown. As for the CG, DS, CSPAE, and CSAE, the shares conform with what has been reported in previous research on these corpora (cf. Berglund 1997, 1999a, 1999b, 2000a, 2000b; Szmrecsanyi 2003): BE GOING TO is more frequent in American English than in British English, and more frequent in more informal speech than in more formal speech. As for FRED, BE GOING TO is strikingly infrequent in this corpus. Since the collection consists of material produced by comparatively old speakers, this may be linked to the fact that BE GOING TO has been spreading in apparent time during the past century (cf. Krug 2000; Mair 2004).

3.2. Independent variables

In addition to the standard variables discussed in chapter 3 (TTR, SENTENCE-LENGTH, PREVIOUS, TEXTDIST, SAMETURN, SAMESPEAKER, AGE, SEX), the following independents were included in the analysis.

3.2.1. Previously suggested and persistence-unrelated predictors

CONTEXTS OF NEGATION (henceforth: NEGATION). Is the future marker in CURRENT negated by *not*, as in (2a), or by a *not*-contracted auxiliary, as in (2b), or is the future marker variant *won't* used, as in (2c) (coded 0 for affirmative contexts and 1 for negated contexts)?

(2) a. *those ministers from the South* ***will not*** *be conducting morning worship tomorrow* (DS KBK)
b. *cos the walls* ***ain't gonna*** *be done* (DS KB6)
c. *cos you* ***won't*** *be that late with Marge in bed* (DS KBF)

Hypothesis: BE GOING TO is favored in contexts of negation.

FRED DIALECT AREA (henceforth: FRED-AREA). This variable is relevant for FRED only and is sensitive to how future marker choice differs across the dialect regions sampled in FRED.

3.2.2. Additional, persistence-related predictors

ALLITERATIONS to BE GOING TO and WILL (henceforth: G-ALLIT and W-ALLIT), respectively. In a context of 50 words before and 50 words after CURRENT, how many tokens are there that start in <g> (G-ALLIT) or <w> (W-ALLIT), respectively? To illustrate: in (3), there are three words that start in <w> (excluding, of course, the future marker *will*), hence (3) would be coded '3' with regard to W-ALLIT.

(3) ***W****ell, through the Fed,* ***w****hat I think* ***w****hat* ***will*** *happen …* (CSAE 0906)

For one thing, discourse analysts have argued that speakers often try to use an option (if they have one) that is sound coordinated with things in its neighborhood (cf., for instance, Sacks 1971; Tannen 1989). Second, there is psycholinguistic evidence that the human speech production system has a tendency for phoneme perseveration (cf., for instance, Dell 1986; Cohen and Dehaene 1998), which is probably why alliteration 'sounds good,' and can be exploited as a stylistic device. G-ALLIT and W-ALLIT, then, clearly fall under the scope of β-persistence. Not only the context *before* CURRENT, but also the context *after* CURRENT will be considered because speakers, in all likelihood, also anticipate upcoming speech while choosing what option to employ in CURRENT (Dell 1986: 285).

Hypothesis: As G-ALLIT increases, the odds for BE GOING TO increase; as W-ALLIT increases, so do the odds for WILL.

HORROR AEQUI contexts (henceforth: G-HORRORAEQUI and W-HORRORAEQUI). Here, I define a *horror aequi* context as a context where identical morpho-grammatical marking or identical phonological material occur in a context of no more than 5 words prior to CURRENT. Thus, G-HORRORAEQUI indicates whether an *-ing* form occurs in such a horizon, as in (4a) where the ending in *riding* possibly encourages the non-use of a BE GOING TO marker. W-HORRORAEQUI is about whether a word ending in *-ll* occurs in such a context as in (4b) where the ending in *well* possibly encourages the non-use of a WILL marker (coded 0 for such features not occurring, and 1 for such features occurring in a context of 5 words prior to CURRENT):

(4) a. *and that they're rid**ing** a lot, they**'ll** just, let the college kids do em* (CSAE 0408)

b. *we**ll**, we're **gonna** have to find somewhere, to get, something* (CSAE 0408)

Hypothesis: Horror aequi-contexts discourage usage of the future marking options affected.

PRESENCE OF THE VERB *to go* in the preceding context (henceforth: G-TRIGGER). Do the tokens *go, goes, went, going*, or *gone* occur in a context of (a) 75 words, or (b) 25 words, or (c) 5 words prior to CUR-

Table 15. Future marker choice: independent variables considered

variable	type	coding method
	a. previously suggested and persistence-unrelated independents	
SENTENCELENGTH*	scalar	software
TTR*	scalar	software
NEGATION	two-way categorical	software
	b. persistence-related independents	
PREVIOUS*	two-way categorical	software
TEXTDIST*	scalar	software
G-ALLIT	scalar	software
W-ALLIT	scalar	software
G-HORRORAEQUI	two-way categorical	software
W-HORRORAEQUI	two-way categorical	software
GO-TRIGGER	four-way categorical	software
SAMETURN*	two-way categorical	software
SAMESPEAKER*	two-way categorical	software
	c. speaker characteristics	
AGE*	scalar	software
SEX*	two-way categorical	software

* independent variable discussed in chapter 3, section 1.

RENT?[34] (5) is an example of *go* occurring in a context of 25 words prior to CURRENT:

(5) *you* ***go*** *look, and every horse's hoof is shaped different. . . . Every horse is* ***gonna*** *have a little different shape.* (CSAE 0408)

Hypothesis: The presence of the verb *to go* may trigger a BE GOING TO based future marker in a nearby choice context through lexical priming or similar mechanisms, a triggering effect which would qualify as β-persistence in the present study.

Table 15 summarizes the independents considered in this chapter.

4. Results

4.1. Baseline variation

In a first step, a logistic regression model was estimated on the basis of independents that have been discussed in previous research, and that are not related to persistence: SENTENCELENGTH, TTR, and NEGATION. The model is displayed in Table 16. Although model χ^2 is often significant, explanatory power is generally poor: R^2 values range between 0.001 (CSAE) and 0.04 (CSPAE), i.e. the models account for less than 5% of the observable variance between GOING TO and WILL. Predictive efficiency is even more disappointing: nowhere does consideration of SENTENCELENGTH, TTR, and NEGATION even slightly enhance our ability to predict future marker choice over and above the baseline prediction.

As for the individual independents, there appears to be a tendency (statistically significant in the CSPAE only) for increased sentence length to negatively affect the odds for usage of a WILL-based marker. In the CSPAE, as sentence length increases by one word, the odds for WILL decrease by 1%. This relationship is as expected given previous research. The picture with regard to TTR is mixed: the variable is significant in the CG, CSPAE, and DS. In the former two corpora (both of which are formal) as TTR increases by one unit, the odds for WILL increase by 3–4%. Indicating that the lexically more compact option is preferred in lexically more dense contexts, this relationship would be as hypothesized. In the DS, however, a one-point increase in TTR actually decreases the odds for WILL; thus in the DS, WILL tends to be dispreferred in lexically more dense contexts. Finally, the impact of contexts of negation (as in *John will not marry Mary*) on future marker choice (NEGATION) is statistically significant in all corpora save the CSAE. In the CG, CSPAE, and DS, a context of negation hugely increases the odds for WILL. In FRED, by contrast, contexts of negation actually decrease the odds for WILL by 32%. These differences are due to the different prominence of *won't* across corpora. If *won't* is removed from the database, the following odds ratios for NEGATION emerge:

CG	0.43
CSPAE	7.59
DS	0.08
CSAE	0.19
FRED	0.09

Table 16. Future marker choice: odds ratios associated with baseline predictors in logistic regression

	CG	CSPAE	DS	CSAE	FRED
SENTENCELENGTH	1.00	**0.99** ***	1.00	0.99	1.00
TTR	**1.03** ***	**1.04** ***	**0.99** ***	1.01	0.99
NEGATION(1)	**2.09** ***	**16.75** ***	**2.05** ***	1.06	**0.68** ***
FRED-AREA	n.a.	n.a.	n.a.	n.a.	**–** ***
model intercept	0.40 ***	0.21 ***	3.27 ***	0.93	6.85 ***
N	39,775	18,377	39,640	1,354	4,861
model χ^2	493.27 ***	539.66 ***	380.08 ***	1.17	53.34 ***
R^2	0.018	0.041	0.014	0.001	0.018
% correct (baseline)	73.6 (73.6)	71.0 (71.0)	71.7 (71.7)	57.9 (57.9)	82.4 (82.4)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for WILL future marking.

This suggests that with intrinsically negated *won't* excluded, a context of negation increases the odds for BE GOING TO considerably, a finding which is consonant with previous research. An exception is the CSPAE, where even when *won't* is excluded, contexts of negation increase the odds for WILL. This anomaly is discussed in some detail in Szmrecsanyi (2003: 303–305); suffice it to say here that speakers in the CSPAE have a marked preference for the collocation *will not*.

In summary, these findings indicate that independents discussed in previous research, while demonstrably influencing the choice, do a rather inadequate job of explaining the observable variance between BE GOING TO and WILL. Variation in future marker choice – unlike, say, variation in comparison strategy choice – still is an alternation that is not very well understood.

A word is due on FRED: there is significant variation between dialect areas with regard to future marker choice. Taking the Southeast as statistical baseline area (as before, in an entirely arbitrary fashion), this variation manifests as follows in logistic regression:[35]

Hebrides	1.71
North	1.62
Wales	1.98

Thus, compared to the Southeast, WILL is significantly more frequent in the Hebrides, in the North, and in Wales. Other regional differences are not significant.

4.2. Persistence-induced variation

How does persistence affect future marker choice? I begin to show the importance of α-persistence in the data by presenting the scatterplots in Figure 12. These graphs compare each speaker's switching rate (BE GOING TO → WILL or vice versa) to his or her overall usage proportion of the two marker families. Had there been no persistence, dots would have clustered close to the diagonal, dotted line – but clearly, this is not the case. The vast majority of speakers are heavily clustered *below* the diagonal line. This means that in most speakers' production, there are considerably fewer switches from BE GOING TO to WILL than pure chance would predict: once they have made a marker choice, speakers actually tend to stick to that marker, a behavior which I have termed α-persistence.

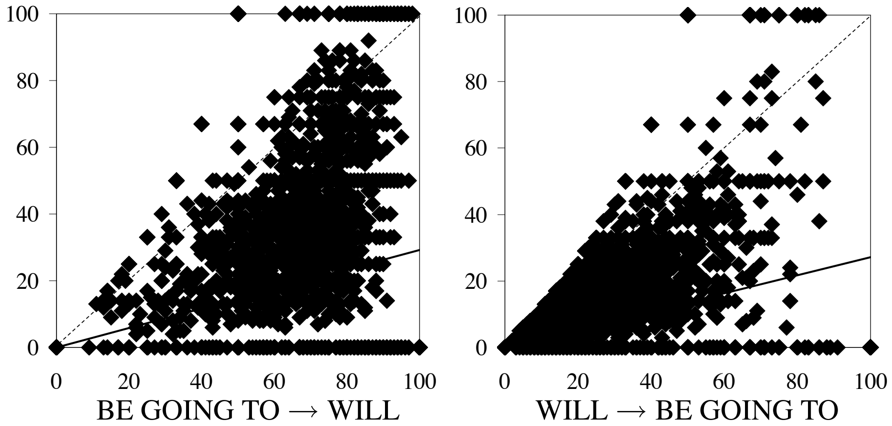


Figure 12. Switches in future marker choice as a function of overall proportion of markers (relative frequency of switches, in %, on y-axis; relative frequency of the switched-to marker family, in %, on x-axis) in the entire database under analysis. Each dot represents one speaker. Dotted diagonal line represents null hypothesis that switch rate is proportional to variant proportions. Heavy line indicates linear trend (BE GOING TO → WILL: $y = 0.29x$, WILL → BE GOING TO: $y = 0.27x$)

Table 17 provides the equations of the regression lines for BE GOING TO → WILL-switches (Figure 12) and for WILL → BE GOING TO-switches (not displayed in Figure 12). The regression estimates in this table confirm that speakers, on average, switch only about 30% of the time they would if there were no persistence. There are differences between corpora with regard to the strength of persistence exhibited: Overall, switch rates are lower (i.e. persistence is more powerful) in the two corpora of British English than in the two corpora of American English. Switch rates are highest in FRED. Differences in switch rates between the two future marker forms are somewhat erratic across corpora.[36]

Next, I estimated logistic regression models including persistence-related independents. Inclusion of persistence-related independents is a huge leap forward in terms of variance explained and predictive efficiency (albeit from, as we have seen, an admittedly low level). Variance explained (R^2) is now a moderately satisfactory 45% in the CSPAE; in the two BNC-based corpora, it is roughly half of that. Variance explained is still outright bad in the CSAE

Table 17. Linear regression estimates of switch rates in future marker choice across corpora (y is the relative frequency of A → B switches, in %; x is the relative frequency of B forms, in %; the expected linear relationship, uninfluenced by persistence, is $y = x$)

corpus	BE GOING TO → WILL	WILL → BE GOING TO
CG	$y = 0.01x$	$y = 0.25x$
DS	$y = 0.26x$	$y = 0.23x$
CSPAE	$y = 0.35x$	$y = 0.35x$
CSAE	$y = 0.28x$	$y = 0.39x$
FRED	$y = 0.40x$	$y = 0.55x$

and especially in FRED, where it is only ca. 6%. In all, though, inclusion of persistence-related independents yields a statistically significant model χ^2 increase.[37]

4.2.1. α-persistence

As for α-persistence, consider PREVIOUS, which has a statistically significant main effect on future marker choice throughout. The odds ratio associated with the variable is generally smaller than 0.05, meaning that when PREVIOUS is a BE GOING TO marker, the odds for WILL in CURRENT decrease by over 95%. In other words, it is unlikely that a BE GOING TO marker is followed by a WILL marker in the following slot.

This claim, however, only holds conditioned on the interactional factors discussed below being zero. If they are not, the following picture emerges: the interaction term PREVIOUS $*$ TEXTDIST indicates how the effect of PREVIOUS actually depends, among other factors, on the textual distance between PREVIOUS and CURRENT. Where this interactional term is significant, the odds ratio associated with PREVIOUS change by a multiplicative factor of between 1.20 and 1.36 for every one-unit increase in the *ln* of the textual distance between PREVIOUS and CURRENT. This means that α-persistence weakens as the textual distance between two successive future marker sites in discourse increases. This is, after all, the hypothesized relationship between TEXTDIST and α-persistence. Figure 13 takes a closer look at the nature of this relationship by plotting the percentage of PREVIOUS/CURRENT matches as a function of non-logged textual distance (note that the intuitive

Table 18. Future marker choice: odds ratios associated with persistence-related predictors in logistic regression (baseline predictors are included, but not displayed)

	CG	CSPAE	DS	CSAE	FRED
PREVIOUS(G)	**0.01** ***	**0.002** ***	**0.002** ***	**0.05** ***	**0.04** **
PREVIOUS(G) ∗ TEXTDIST	**1.43** ***	1.06	**1.30** ***	1.00	**1.20** ***
PREVIOUS(G) ∗ SENTENCELENGTH	1.00	1.00	1.00	1.00	1.00
PREVIOUS(G) ∗ TTR	**1.07** ***	**1.08** ***	**1.07** ***	**1.04** *	1.03
PREVIOUS(G) ∗ SAMETURN(1)	0.91	**0.54** ***	**0.67** ***	**0.56** *	0.83
PREVIOUS(G) ∗ SAMESPEAKER(1)	**0.83** ***	1.02	**0.79** ***	**0.64** *	1.08
G-ALLIT	**1.47** ***	**2.08** ***	**1.27** ***	**0.94** *	0.96
W-ALLIT	**1.00** ***	**1.06** ***	**1.02** **	**1.05** *	0.98
G-HORRORAEQUI	**0.89** *	**0.72** ***	1.06	1.27	1.18
W-HORRORAEQUI	**2.22** ***	**1.31** ***	**2.10** ***	**1.47** *	**1.80*** **
GO-TRIGGER	– ***	– ***	– ***	–	– *
GO-TRIGGER(5)	**0.66** ***	0.65	1.08	1.62	**2.53** ***
GO-TRIGGER(25)	**0.54** ***	**0.29** ***	**0.66** ***	1.08	0.98
GO-TRIGGER(75)	**0.65** ***	**0.60** ***	**0.80** ***	1.17	0.85
model intercept	8.33	0.31 ***	0.00	3.32	17.57
N	38,944	18,376	39,450	1,294	4.515
model χ^2	8,487.39 ***	7,028.41 ***	5,682.91 ***	125.38 ***	177.52
R^2	0.286	0.454	0.193	0.124	0.064
% correct (baseline)	78.1 (73.6)	80.5 (71.0)	74.2 (71.7)	65.6 (57.5)	82.5 (82.5)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for WILL future marking.

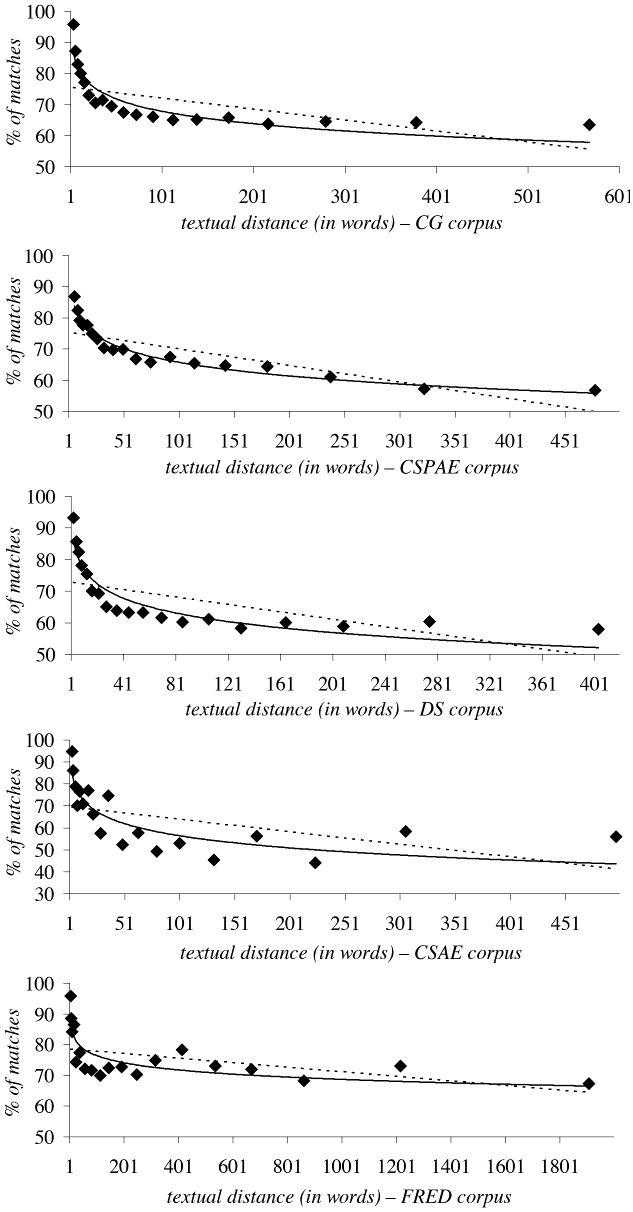


Figure 13. Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same future marker is used) as function of textual distance between CURRENT and PREVIOUS. Heavy line represents logarithmic estimate of the relationship, dotted line represents linear estimate of the relationship

interpretation of the y-axis is that it indicates the strength of α-persistence). As can be seen, this relationship is nicely logarithmic (or decreasing exponential). While this impression is already quite clear visually, it is confirmed statistically when comparing curve fits:[38]

	CG	CSPAE	DS	CSAE	FRED
adj. R^2 linear	0.30 **	0.64 ***	0.34 **	0.22 *	0.19 *
adj. R^2 logarithmic	0.83 ***	0.96 ***	0.86 ***	0.71 ***	0.63 ***
df	17	17	17	17	17

So, in short, the strength of α-persistence depends on recency of use of PREVIOUS, and the forgetting function that describes this relationship is logarithmic.

The term PREVIOUS $*$ TTR indicates how the effect of PREVIOUS depends on the lexical complexity of the environment where CURRENT is embedded. The value of PREVIOUS $*$ TTR is greater than 1 everywhere and often statistically significant. This means that as lexical complexity of CURRENT's environment increases, the odds ratio associated with PREVIOUS increases. Hence (because exp(b) of the main effect of PREVIOUS is smaller than 1), α-persistence is actually weakened in lexically more complex environments.

The effect of PREVIOUS $*$ SAMETURN is associated with an exp(b) value of between 0.64 (CSPAE) and 0.91 (CG). The impact of PREVIOUS on CURRENT, therefore, increases if PREVIOUS was in the same turn as CURRENT. Similarly, the value of PREVIOUS $*$ SAMESPEAKER indicates that when PREVIOUS was produced by the same speaker who produced CURRENT, the odds ratio of PREVIOUS is multiplied by a factor between 0.64 (CSAE) and 0.83 (CG). As expected, therefore, α-persistence is stronger (i) when PREVIOUS and CURRENT are located in the same conversational turn, and (ii) when PREVIOUS and CURRENT are produced by the same speaker.

The interaction term PREVIOUS $*$ SENTENCELENGTH is not selected as significant anywhere, therefore no interaction between persistence and syntactic complexity can be observed.

4.2.2. β-persistence

G-ALLIT and W-ALLIT – variables measuring the number of words in CURRENT's environment that begin in <g> and <w>, respectively – turn out to behave almost utterly contrary to expectations. We had expected that contexts with lots of words starting in <g> (thus contexts where G-ALLIT is high) would favor BE GOING TO markers, and contexts with lots of words starting in <w> (and thus, contexts where W-ALLIT is high) would favor WILL markers. It turns out that only in the CSAE is this the case. In this corpus, an increasing number of words starting in <g> decreases the odds for usage of WILL (by 6% per word that starts in <g>), and an increasing number of words starting in <w> increases the odds for usage of WILL (by 5% per word that starts in <w>). In the other corpora where the variable is significant, <g>-alliterations not only seem to *encourage* usage of WILL, but they also do so more strongly than do <w>-alliterations.

The effects of G-HORRORAEQUI and W-HORRORAEQUI are also unexpected, but their behavior has an explanation. We hypothesized that tokens that have the same endings as the full forms of future marker paradigms in CURRENT's immediately preceding context would make speakers avoid identity effects by resorting to the alternative option. As can be seen, the opposite is true. The presence of a token ending in *-ing* (G-HORRORAEQUI) in CURRENT's immediately preceding context actually decreases the odds for a WILL-based marker in CURRENT by between 12% and 28%. In a similar vein, the presence of a token ending in *-ll* in CURRENT's immediately preceding context actually increases the odds for a WILL-based marker in CURRENT. This means that rather than avoiding identity effects, speakers actually seek them.

Finally, let us look at the effect of the presence of the verb *to go* (or one of its inflected forms) in CURRENT's preceding context. Overall, the predictor is significant throughout except in the CSAE. Individual threshold levels (i.e. whether *go* was used in a context of 5, 25, or 75 words prior to CURRENT) are selected as significant less often, but where they are, their effect is as hypothesized:[39] if the verb *go* has just been used, the odds for WILL decrease. By implication, this means that the verb *go* triggers BE GOING TO (β-persistence). Surprisingly, no substantially interesting differences can be observed between the individual threshold levels of GO-TRIGGER.

4.3. Inter-speaker variation

The database available is large enough to conduct an analysis of how the speaker variables AGE and SEX influence persistence in the DS and CSAE.

Table 19. Future marker choice: odds ratios associated with speaker predictors in logistic regression (baseline predictors and persistence-related predictors are included, but not displayed)

	DS	CSAE
AGE	**1.01** ***	**1.03*****
SEX(F)	1.04	1.13
AGE ∗ PREVIOUS(G)	1.01	1.05
AGE ∗ GO-TRIGGER(1)	1.01	1.01
SEX(F) ∗ PREVIOUS(G)	0.91	1.70
SEX(F) ∗ GO-TRIGGER(1)	1.00	0.95
AGE ∗ TEXTDIST ∗ PREVIOUS(G)	0.99	**0.99** *
model intercept	0.00	2.58
N	17.394	861
model χ^2	2,461.01 ***	141.50 ***
R^2	0.190	0.203
% correct (baseline)	74.0 (71.9)	67.5 (55.7)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for WILL future marking.

In this spirit, logistic regressions were run on these two corpora.[40] As can be seen from Table 19, inclusion of the speaker variables and terms results in moderate increases in variance explained in the CSAE (from approximately 12% to 20%), while R^2 remains virtually unchanged in the DS. In the CSAE, predictive efficiency is enhanced slightly; in the DS, it is even worsened somewhat. Overwhelmingly, odds ratios for individual variables are statistically insignificant. What follows is thus a discussion of tendencies, not necessarily of statistical facts. Some of these tendencies, however, appear to be worth mentioning.

The main effect of AGE is that increasing age significantly favors usage of WILL based markers in both the DS and the CSAE. This is obviously an apparent time effect supporting claims that BE GOING TO is spreading. The exp(b) values associated with SEX may hint that female speakers have a preference for WILL-based markers compared to male speakers. The interaction term AGE ∗ PREVIOUS has values of greater than 1 in both corpora, a fact which may justify a very tentative claim that the main effect of PREVIOUS in-

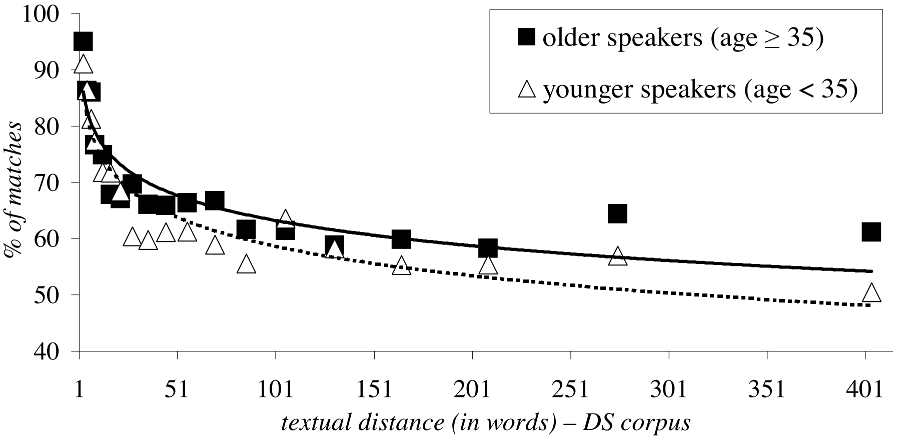


Figure 14. Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same future marker is used) as function of textual distance between CURRENT and PREVIOUS in the DS. Heavy line represents logarithmic estimate of the relationship in older speakers, dotted line represents logarithmic estimate of the relationship in younger speakers

creases as speakers' age increase – in other words, persistence weakens with increasing age. The term SEX ∗ PREVIOUS has contradicting values in the two corpora and is thus not interpretable. SEX ∗ GO-TRIGGER is smaller than 1 in both corpora, suggesting that β-persistence may be slightly weaker in women than in men.

The interactional term AGE ∗ TEXTDIST ∗ PREVIOUS is associated with exp(b) values slightly smaller than 1 in both corpora and is statistically significant in the CSAE. Three-way interactions are notoriously hard to interpret, but what this one actually suggests is that as TEXTDIST increases, the main effect of PREVIOUS decreases more slowly when subjects are older than when speakers are younger. What does this mean?

α-persistence, according to these estimates, declines more slowly in older speakers than in younger speakers. Figure 14 visualizes this finding by plotting – much like Figure 13 (p. 122) – the percentage of matching PREVIOUS / CURRENT pairs in the DS (for which a sufficiently large number of observations for such an analysis is available) against textual distance between them. At the same time, Figure 14 provides separate regressions for younger speakers (i.e. speakers who are younger than 35 years, which is the mean age in the DS database on future marker choice) and for older speakers. Observe, now, that the forgetting curve that describes the decline of persistence in older speakers is slightly more level than the corresponding curve for younger speakers. Take, for instance, a textual distance of approximately 400 words between PREVIOUS and CURRENT: while in the production of older speakers, there is still a 53% likelihood that PREVIOUS and CURRENT match (i.e. that they are persistent), the corresponding likelihood for younger speakers is only 49%.

5. Summary

This chapter has reported the following observations with regard to the alternation between BE GOING TO and WILL:

Intralinguistic factors that have previously been cited to (partially) explain the variation between BE GOING TO and WILL – such as whether or not a future marker is used in negated context, or the syntactic complexity of the surrounding material – have a low explanatory yield. According to the regression estimates, these factors explain 5% of the observable variation at most. In a nutshell, sentence length, and thus increased syntactic complexity, can favor application of a BE GOING TO marker, and contexts of negation in general favor usage of BE GOING TO when *won't* is controlled for. Increased lexical complexity can have differing effects, depending on the data source.

How is persistence relevant to future marker choice? Switch rates between BE GOING TO and WILL are, on average, only one third of what would be the 'natural' switch rate (cf. Sankoff and Laberge 1978: 122 for a similar observed switch rate with regard to switches from *on* to *tu–vous*). This means that speakers have a very marked tendency to avoid switching between future markers – in plain words, there is a good deal of persistence in the expression of futurity.

In logistic regression, consideration of persistence-related variables enhances the quality of our modeling of speakers choices, but substantially so only in the three major corpora, the CG, CSPAE, and DS. α-persistence factors generally account for approximately 7–11% (though only for 4% in FRED) of the observable variance. What marker was used in the preceding future marker slot has, on the whole, a sizable impact on the marker that will be chosen for an upcoming future marker slot. The exact magnitude of this impact depends, though, on several factors:

1. Recency of use (operationalized here as textual distance between two successive marker slots) turned out to be a significant factor. A previous marker choice influences an upcoming choice to a greater extent if the previous choice was recent. As hypothesized, α-persistence often declines logarithmically.

2. My findings also indicated that increased lexical complexity of the environment where a future marker is going to be used decreases the chance that PREVIOUS will be used again. This finding is unexpected, and contradicts Tannen (1987), who claimed that parallel patterns might be preferred in lexically dense contexts due to an advantage in processing efficiency.

3. The effect of a previous choice on an upcoming choice is stronger when (i) the previous future marker occurrence was in the same turn as the upcoming choice, and when (ii) the previous future marker occurrence was produced by the same speaker who is faced with the upcoming choice. Therefore, persistence across turns is weaker than persistence within turns, and intra-speaker persistence (‘self-repetition’) is stronger than inter-speaker persistence (‘allo-repetition’), which once again dovetails nicely with previous studies (for instance, Gries 2005).

β-persistence is more powerful in the two formal corpora (CG and CSPAE) than in the informal corpora. My analysis sought to measure β-persistence through consideration of an amalgam of factors:

To start with, Sacks (1971) has argued that if speakers have a choice, they go for an option that is sound coordinated with material in its neighborhood; psycholinguists have argued that the human speech production system is geared towards phoneme perseveration (for instance, Dell 1986). Operationalizing these notions as the number of words in a future marker slot’s context that start in <g> or <w>, it emerged that this is only partially true: In a given context, as the number of words beginning in <w> increases, so do the odds for choice of a WILL-based marker. However, in general, an increased number of tokens beginning in <g> does not increase the odds for BE GOING TO, as it should if speakers also tried to sound coordinate BE GOING

TO markers with their environment. Sound coordination, therefore, is a rather weak predictor of future marker choice.

Secondly, the presence or absence of a form of the verb *to go* in a future marker slot's preceding context has the expected impact on future marker choice: if *go* has been used recently, BE GOING TO is preferred over WILL. This is a clear instance of β-persistence. However, it appears to make no big difference if *go* has been used in a context of 5, 25, or 75 words before CURRENT. Future research will have to use broader threshold levels, or operationalize the variable in a scalar fashion.

The analysis also considered a variable that was supposed to be sensitive to *horror aequi* effects, namely, the presence or absence of tokens in a future marker slot's immediately preceding context that have endings identical to the full forms of the two future marker families: *-ing* and *-ll*. The hypothesis was that the presence of such identical material would make speakers resort to the future marker option that would avoid an adjacent identity effect. As a matter of fact, We could detect no such *horror aequi* effects: presence of material ending in *-ll* actually *encourages* usage of WILL, and presence of material ending in *-ing* encourages usage of BE GOING TO, according to the data. What was thought to be *horror aequi* thus turned out to be β-persistence. With hindsight, the variable name G/W-HORRORAEQUI is actually a misnomer.

Finally, the data seem to suggest that α-persistence appears to decline more slowly in old speakers than in young speakers; this is another way of saying that α-persistence is more long-lasting in older speakers. Moreover, the analysis suggested – much as some previous studies did (for instance, Mair 2004) – that BE GOING TO is spreading in apparent time.

Chapter 7
Persistence in particle placement

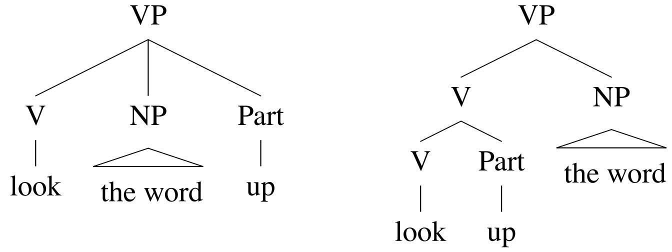
This chapter will research persistence in the word order alternation that can be observed for transitive, separable phrasal verbs – “type II transitive phrasal verbs” in Quirk et al.’s (1985:1153) diction. Consider (1a), where the verb and its particle are separated (henceforth: *V+NP+Part*), and (1b), where they are adjacent (henceforth: *V+Part+NP*):

(1) a. *When you* ***take*** *the absolute value* ***off*** (CSAE 0713)
b. *What I have to do, is* ***take off*** *the distributor wire* (CSAE 0593)

1. Background and previous research

While the above two word order patterns are certainly semantically equivalent, they are different formally, and maybe pragmatically and discourse-functionally. What exactly these differences are has been the subject of a sizable body of research on particle placement[41] in English. Fairly basic is Bruce Fraser’s and Dwight Bolinger’s research on phrasal verbs (Fraser 1965, 1974; Bolinger 1971). For recent perspectives on particle placement, both functional and generativist, see the monograph by Dehé et al. (2002). Also interesting, particularly from a variationist and cognitive perspective, is Stefan Th. Gries’ work on particle placement (Gries 1999, 2002, 2003a). For the purposes of the present study, suffice it to point out that variation in particle placement is a syntactic and positional alternation. Consider (2):[42]

(2)



There is *per se* no lexical/functional material to be repeated (unlike with, for instance, analytic and synthetic comparatives, two subsequent phrasal verbs do not necessarily have any lexical/functional material in common). Therefore, if production priming is involved in how the alternation plays out in discourse, it is primarily syntactic priming: Konopka and Bock (in press) show experimentally that the syntax of phrasal verbs can be primed, and that the strength of the priming effect does not depend on whether the phrasal verb is idiomatic or literal.

2. Previously suggested factors

A vast number of factors contributing to particle placement have been suggested in the literature, mostly based on introspective evidence. The following review will focus on factors that have been successfully backed up by at least some empirical evidence.

Stress quality of the direct object. It has been argued that the *V+Part+NP* pattern is preferred when the direct object is strongly stressed (for instance, Van Dongen 1919; Kruisinga and Erades 1953).

Length and/or complexity of the direct object. The longer (in terms of phonetic material) and the more complex (in terms of the presence of embedded clauses) the direct object is, the greater the preference for *V+Part+NP* ordering (cf. Van Dongen 1919: 351; Kennedy 1920: 30; Hawkins 1994; Biber et al. 1999: 932–933; Quirk et al. 1985: 1154). Consider (3):

(3) a. *Mary **looked up** the word which she had heard the other day while talking to her neighbor.*

b. ?*Mary **looked** the word which she had heard the other day while talking to her neighbor **up**.*

Of course, this phenomenon is not restricted to particle placement, but related to the more general notion of syntactic weight and end weight. The literature on these phenomena is extensive, yet it essentially boils down to Behaghel's (1909/1910) principle of end weight ("Gesetz der wachsenden Glieder"; cf. the discussion of end weight in chapter 5, section 2).

Presence of a directional prepositional phrase after the verb phrase. The *V+NP+Part*-pattern is preferred if the construction is followed by a directional prepositional phrase (Gries 1999; Biber et al. 1999; Gries 2002, 2003a). Consider (4):

(4) a. *Mary **put** the cup **back** into the cupboard.*
b. ?*Mary **put back** the cup into the cupboard.*

Idiomaticity of the construction. According to Gries (1999), Biber et al. (1999: 933), and Quirk et al. (1985: 1155), phrasal verb constructions with an idiomatic meaning, as in (5a), prefer the *V+Part+NP* pattern, whereas phrasal verbs where the particle has literal – that is, spatial – meaning, as in (5b), prefer the *V+NP+Part* pattern.

(5) a. *I carry out my duties.*
b. *I carry my garbage out.*

News value (topicality, givenness) of the direct object. If the direct object is discourse-old, the *V+NP+Part* pattern is preferred; if it is discourse-new, there is a preference for the *V+Part+NP* pattern (cf. Kruisinga and Erades 1953, Bolinger 1971, Chen 1986). As Gries (1999: 111–112) points out, this factor accounts for a number of further, more subtle distributional differences between the two patterns, but these will be omitted here.[43]

3. Method, data and independent variables

3.1. Method and data

This chapter will investigate particle placement with regard to 263 transitive phrasal verbs, which are listed in Appendix D. The alternation in particle verbs is a relatively complex one which cannot be handled wholly by software: while the *V+NP+Part* pattern is easy to identify automatically, the *V+Part+NP* pattern is a major problem for two related reasons: first, automatic identification of the object noun phrase would required syntax-tagged data; second, because the object noun phrase cannot be identified, software cannot distinguish between intransitive phrasal verb usages (e.g. *I took over*)

Table 20. Particle placement: distributional variation across corpora

corpus	*N*	*N* (*V+Part+NP*)	*N* (*V+NP+Part*)
CSAE	187	123 (65.8%)	64 (34.2%)
FRED	1,168	281 (24.1%)	887 (75.9%)
total	**1,355**	**404 (29.8%)**	**951 (70.2%)**

from transitive usages (e.g. *I took over the chair*) in data that is not syntax-tagged. For this reason, the analysis in this chapter relies to a large extent on manual coding of manageable datasets. The entire CSAE as well as a subset of FRED[44] was parsed manually to identify the above phrasal verbs in the data.

Because the verb-particle order is virtually categorically *V+NP+Part* if the direct object of transitive phrasal verbs is a pronoun (e.g. *he looked it up*; see, for instance, Kennedy 1920; Quirk et al. 1985: 1154; Biber et al. 1999: 934), cases where the direct object was a pronoun were excluded from the tally. Also, idiomatic conventions do not always allow an alternative positioning of the particle (e.g. *I was crying my eyes out* vs. ?*I was crying out my eyes*), which is why such cases were also excluded from the tally.

A Perl script was then used to extract the variables and to code them for the standard variables (e.g. SENTENCELENGTH, TTR, PREVIOUS, TEXTDIST, SAMETURN, SAMESPEAKER). In a second step, the database was then coded manually for the variables specific to particle placement (see section 3.2). This yielded a database of 1,355 alternating phrasal verbs (Table 20 gives an overview). Unfortunately, this number of observations is too low to reliably consider the speaker variables AGE and SEX in this chapter.

The proportions in Table 20 are interesting: the received wisdom is that "conversation opts for a high frequency of mid-position because clauses are generally short and because the connection between the verb and the adverbial particle can be marked by intonation" (Biber et al. 1999: 934). The numbers show that this is true for FRED, but not for the very conversational CSAE: in the latter, *V+Part+NP* is more frequent than the alternative pattern. Conceivably, Biber et al. (1999) included cases where the direct object was a pronoun (these are not included in the tally). Because personal pronouns are known to be particularly frequent in conversation, this could explain the differential.

3.2. Independent variables

In addition to the standard variables discussed in chapter 3 (TTR, SENTENCELENGTH, PREVIOUS, TEXTDIST, SAMETURN, SAMESPEAKER), the following predictors will be included in the multivariate analysis. These have been shown by previous research to discriminate sufficiently between the two placement patterns (for instance, Gries 2003a: 165).

3.2.1. Previously suggested and persistence-unrelated predictors

DEFINITENESS OF THE DIRECT OBJECT (henceforth: DEFINITEDO). Does the phrasal verb construction under analysis contain a direct object that is determined by a definite determiner – i.e., *the, this, that, these, them, those*, as in (6) (coded 1 for a definite determiner not present and 0 otherwise)?

(6) *send* ***the*** *girl in first* (FRED SAL26)

This variable is one way to check on the news value of the direct object. A test of intercoder reliability of this coding, which was computed by having a second scorer (a trained linguist) code a random subset of ca. 10% of the CSAE database ($N = 102$), yielded a simple agreement rate of approximately 96% and an 'excellent' (cf. Orwin 1994) Cohen's κ value of 0.91. See Appendix C for the feature's coding scheme.
Hypothesis: If the determiner of the direct object is definite, there is a preference for the *V+Part+NP* pattern (cf. Gries 2003a: Table 2).

NEWS VALUE OF THE DIRECT OBJECT (henceforth: NEWSVALUEDO). This variable is another, more direct way to assess the news value of the direct object. It is coded 0 if the referent of the direct object is not mentioned in the preceding five sentences, and it is coded 1 if the referent is mentioned in the preceding five sentences, as in (7):

(7) *Well, make these little* ***passes*** *or something that say, one free lunch with the teacher. Or one free lunch to sit with whoever you want at lunchtime. I get to sit with my friends at lunch. Or, what, think up whatever little privileges like that? Um if there's anything in the classroom that they really like to do, have that*

> *be a privilege, that nobody can do, unless they have this **pass**. And then **give those passes out** for good behavior at the end of the day.* (CSAE 0523)

Hypothesis: If the direct object is mentioned in the preceding discourse, the *V+NP+Part* pattern is more likely.

LENGTH OF THE DIRECT OBJECT in syllables (henceforth: SYLLABLES-DO). For instance, the direct object in (8) contains 3 syllables:

(8) *And of course they filled **the bucket** up* (FRED SFK011)

Hypothesis: The longer the direct object, the greater the preference for the *V+Part+NP* pattern due to the principle of end weight (cf. Behaghel 1909/1910).

COMPLEXITY of the direct object (henceforth: COMPLEXITYDO). Does the direct object of the phrasal verb contain embedded clauses, as in (9) (coded 0 for embedded clauses not present and 1 for embedded clauses present)?

(9) *pick out the ones **that you are going to use for seed*** (FRED HEB021)

Hypothesis: The presence of embedded clauses in the direct object will make the *V+Part+NP* pattern more likely.

PRESENCE OF DIRECTIONAL PREPOSITIONAL PHRASES after the phrasal verb construction (henceforth: DIRECTIONALPP). Is the phrasal verb phrase followed by a directional prepositional phrase, as in (10) (coded 1 if one is following, and 0 otherwise)?

(10) *We were sending cattle off **to the mainland*** (FRED LAN012)

Hypothesis: The *V+NP+Part*-pattern will be more likely if there is a directional prepositional phrase.

LITERALNESS of the phrasal verb (henceforth: LITERALNESS). Does the phrasal verb have a rather literal/spatial meaning, as in (11a), or a rather idiomatic meaning, as in (11b) (coded 0 if the construction has a rather idiomatic meaning and 1 if it has a rather literal meaning)?

(11) a. *I'd have to* ***get*** *a step ladder* ***out*** (CSAE 0514)
b. *let us kind of* ***figure out*** *a classic um diet uh quota for you* (CSAE 1047)

All verb occurrences were coded individually, taking into account their respective context. Because coding for this feature reliably can admittedly be problematic, a test of intercoder reliability was, once again, performed. After initially poor Cohen's κ values in the 0.5 range, recoding by a trained linguist of a random sample of ca. 10% ($N = 102$) of the FRED database yielded, after a good deal of training, a simple agreement rate of approximately 87% and a moderately satisfactory Cohen's κ value of 0.74 (see Appendix (2) for the feature's final coding scheme).
Hypothesis: Constructions with more literal or spatial meanings will prefer the *V+NP+Part* pattern.

DISTINCTIVE COLLOSTRUCTION STRENGTH of the phrasal verb (henceforth: DISTINCTIVENESS). Biber et al. (1999: 933) point out that "there is considerable variability among individual phrasal verbs in their preference for . . . particle placement." In order to account for this variability, the analysis in this chapter will, for every individual phrasal verb, incorporate results from Gries and Stefanowitsch's (2004) so-called 'distinctive collexeme analysis.' Gries and Stefanowitsch (2004) extracted 700 verbs from the ICE-GB corpus and determined the collostructional strengths associated with them (i.e. basically whether and to what extent each of these verbs prefers the *V+Part+NP* or *V+NP+Part* pattern) by means of a statistical analysis.[45] My analysis will operationalize Gries and Stefanowitsch's findings through the scalar variable DISTINCTIVENESS, which can take values between 0 and 100. Low values close to 0 indicate that the verb under analysis has a preference for the *V+NP+Part* pattern, and high values close to 100 indicate that the verb under analysis has a preference for the *V+Part+NP* pattern. To illustrate: *find out* has a fairly high distinctiveness score (99.99) and is therefore strongly associated with the *V+NP+Part* pattern (as in *the examiner'd find these little faults out*

[FRED SAL030]). The inverse is true for the verb *send back*, which has a comparatively low distinctiveness score (1.49).

FRED DIALECT AREA (henceforth: FRED-AREA). This variable is obviously relevant for FRED only and is sensitive to how particle placement differs across the dialect regions (Hebrides, Midlands, North, and Southeast) sampled in the FRED corpus subset under analysis.

3.2.2. Additional, persistence-related predictors

TEXTUAL DISTANCE to the last generic non-separated pattern or to the last generic separated pattern (henceforth: TEXTDIST-SEP and TEXTDIST-NONSEP, respectively). The idea behind these two β-persistence predictors is that there exist phrasal or prepositional constructions where the word order of object and particle is not optional (for instance, when the object is a pronoun, or when a phrasal verb is used intransitively, or when there is a prepositional verb), and that hence do not qualify as dependent variables in the sense of this study. Nonetheless, these non-optional patterns may influence optional orderings and help trigger one or the other pattern. TEXTDIST-NONSEP and TEXTDIST-SEP measure the textual distance (as with TEXTDIST, in the *ln* of interjacent words) between CURRENT and the last such trigger site in the discourse. By way of illustration, consider (12) which contains a potential *V+Part+NP* trigger, that is, the prepositional verb *to look at* where the preposition and the verb are not separated. This prepositional verb is being used 47 words prior to where a phrasal verb with optional word order, *to send in*, is being used (hence, the value of TEXTDIST-NONSEP would be *ln* 47).

(12) *I guess they just* ***look at the quality*** *of the facility. Not necessarily the wiring plumbing and heating, but, is it clean, and is it safe, and that sort of thing. They're supposed to be there the fifth, and he said, that the lady told him, that it usually takes a week, for them to* ***send a report in*** (CSAE 0906)

In (13), a phrasal verb with a non-optional pattern (*put it on*, where the particle and the verb are separated by the pronoun) is being used three words prior to CURRENT (hence, the value of TEXTDIST-SEP would be *ln* 3):

(13) LINDSEY: *With the cast, to* ***put it on****.*
MARCIA: *To actually* ***put the cast on*** (CSAE 0533)

Hypothesis: As the value of TEXTDIST-NONSEP decreases – and thus, as the distance to a *V+Part+NP* trigger decreases, the odds for the *V+Part+NP* pattern increase. The corresponding relationship is expected to hold for the relationship between TEXTDIST-SEP and the odds for the *V+NP+Part* pattern. In sum, we expect generic non-separated patterns to trigger *V+Part+NP* particle placement, and generic separated patterns to trigger *V+NP+Part* particle placement.

SAME VERB LEMMA in both PREVIOUS and CURRENT (henceforth: VLEMMAID). This variable involves whether two successive transitive phrasal verb constructions do in fact involve the same phrasal verb (though not necessarily the same verb form; coded 1 if the lemma is the same, and 0 if it is not). (14) illustrates the case where two successive transitive phrasal verb slots are in fact filled by the same phrasal verb lemma:

(14) PETE: *you wanna* ***try on*** *the men's clothes?*
JAMIE: *the one suggested that, so you wouldn't be so bored. . . . so they* ***tried on*** *the men's clothes, and they had a very small selection of men's clothes* (CSAE 0513)

Hypothesis: Pickering and Branigan (1998) and Gries (2005) showed that production priming is stronger when the priming verb lemma and the target verb lemma are the same. This is why we conjecture that if the verb lemma matches between two successive variables, α-persistence is even stronger than it would be otherwise.

Table 21 gives an overview of the variables considered in this chapter.

Table 21. Particle placement: independent variables considered

variable	type	coding method
a. previously suggested and persistence-unrelated independents		
SENTENCELENGTH*	scalar	software
TTR*	scalar	software
DEFINITEDO	two-way categorical	manual
SYLLABLESDO	scalar	manual
LITERALNESS	two-way categorical	manual
COMPLEXITYDO	two-way categorical	manual
DIRECTIONALPP	two-way categorical	manual
NEWSVALUEDO	two-way categorical	manual
DISTINCTIVENESS	scalar	software
FRED-AREA	four-way categorical	software
b. persistence-related independents		
PREVIOUS*	two-way categorical	software
TEXTDIST*	scalar	software
SAMETURN*	two-way categorical	software
SAMESPEAKER*	two-way categorical	software
VLEMMAID	two-way categorical	software
TEXTDIST-NONSEP	scalar	software
TEXTDIST-SEP	scalar	software

* independent variable discussed in chapter 3, section 1.

4. Results

4.1. Baseline variation

We begin by examining how good a job predictors hitherto discussed in the literature do in predicting particle placement. Table 22 gives the corresponding logistic regression estimates. In this study's data, the baseline predictors explain between 31–37% of the variance, hence variance explained is moderate, and so is predictive efficiency.

To start with, dialect areas in FRED play a role in predicting particle placement. The values for FRED-AREA in Table 22 take the Southeast as baseline area (as before, in an entirely arbitrary fashion) and measure deviations in

Table 22. Particle placement: odds ratios associated with baseline predictors in logistic regression

	CSAE	FRED
SENTENCELENGTH	1.00	**1.02** ***
TTR	**1.06** *	**0.94** ***
DEFINITEDO	1.61	**1.38** *
SYLLABLESDO	**0.67** ***	**0.68** ***
LITERALNESS	1.80	**2.24** ***
DIRECTIONALPP	**10.05** ***	**3.42** *
COMPLEXITYDO	0.00	**0.05** **
NEWSVALUEDO	**3.05** **	**1.62** ***
DISTINCTIVENESS	**0.99** *	**0.99** ***
FRED-AREA	n.a.	– ***
FRED-AREA (Hebrides)	n.a.	**0.67** ***
FRED-AREA (Midlands)	n.a.	0.76
FRED-AREA (North)	n.a.	1.24
model intercept	0.06	424.3 ***
N	187	1,167
model χ^2	48.14 ***	339.14 ***
R^2	0.314	0.377
% correct (baseline)	73.8 (65.8)	82.2 (75.9)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for the *V+NP+Part* pattern.

the other dialect areas. Compared to the Southeast, then, there is a significant dispreference for the *V+NP+Part* pattern in the Hebrides (0.67). A similar effect, though insignificant, is shown in the Midlands. In the North of England, finally, there is an (insignificant) tendency for the *V+NP+Part* pattern to be preferred when compared to the Southeast.

The most potent among the independents is SYLLABLESDO, that is, the length of the direct object in syllables. The odds ratio associated with the variable is 0.7, hence for every one-syllable increase in length of the direct object, the odds for the *V+NP+Part* pattern decrease by 30%. The related variable, COMPLEXITYDO, has an even more sizable effect, but on a lower significance level (it is even insignificant in the CSAE) due to a too small number of observations: When the direct object contains an embedded clause, the odds for *V+NP+Part* decrease by approximately 95%.

LITERALNESS and NEWSVALUEDO are also potent predictors. If the phrasal construction employed in CURRENT has a rather literal, spatial meaning (as in *John brings the garbage out*), the odds for the *V+NP+Part* pattern increase by up to 124% (significantly so in FRED only). NEWSVALUEDO – whether or not the direct object is discourse-old – also has the hypothesized effect: if the direct object is discourse-old, the odds for the *V+NP+Part* pattern multiply by a factor of 1.62 in FRED and even 3.05 in the CSAE. DEFDO – whether or not the object noun phrase contains a definite determiner, as in *John picks up the letter* – is related to NEWSVALUEDO, but is a less powerful predictor of particle placement (and a significant one in FRED only). If the object noun phrase is indefinite, the odds for *V+NP+Part* are increased by 40% in FRED. Also favoring the *V+NP+Part* is the presence of a directional prepositional phrase after the phrasal verb construction (as in *Mary put the cup back into the cupboard*): this factor boosts the odds for the *V+NP+Part* pattern more than threefold in FRED, and more than tenfold in the CSAE – in both cases, significantly so.

In FRED, increased sentence length (SENTENCELENGTH) significantly favors the *V+NP+Part* pattern. I will venture a tentative explanation for this finding in the conclusion of this chapter. TTR turns out to be a significant predictor in both corpora, though it has opposite effects: in FRED, increased lexical density favors usage of the *V+Part+NP* pattern, while in the CSAE it favors usage of the *V+NP+Part* pattern. Therefore, this variable is not interpretable. Anyway, there is no logical way in which lexical density could interact with a word order alternation such as particle placement.

Gries and Stefanowitsch's collostruction strength scale (DISTINCTIVENESS), finally, also turns out to be a significant predictor of particle placement. For each 1-unit increase in the scalar variable DISTINCTIVENESS, the odds for the *V+NP+Part* pattern decrease by 1%. This is the expected relationship. On the whole, the predictor DISTINCTIVENESS accounts for 5% of the observable variance in particle placement in this study's data. Along these lines, it is important to note that Gries and Stefanowitsch's 'distinctive collexeme' scores were derived from the ICE-GB, a corpus of spoken and written Standard British English. Given that these scores were applied to a corpus of conversational American English and to a corpus of English dialects, the share of variance accounted for by the variable is actually considerable.

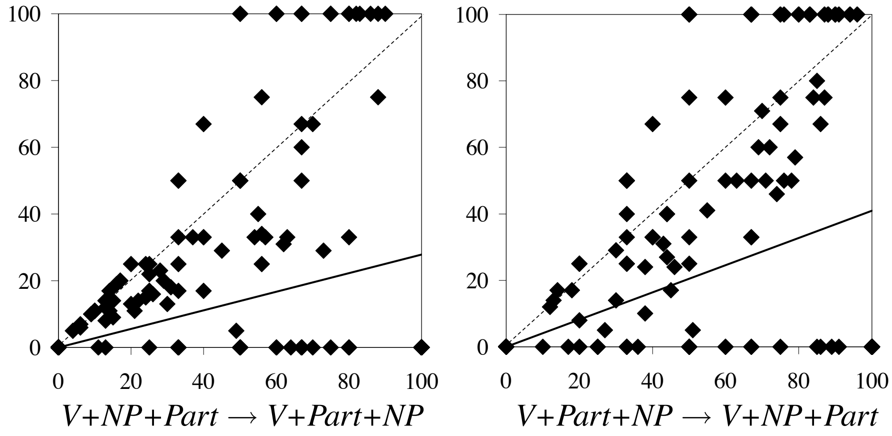


Figure 15. Switches in particle placement as a function of overall proportion of placement strategy (relative frequency of switches, in %, on y-axis; relative frequency of the switched-to placement strategy, in %, on x-axis) in both FRED and the CSAE. Each dot represents one speaker. Dotted diagonal line represents null hypothesis that switch rate is proportional to variant proportions. Heavy line indicates linear trend (*V+NP+Part* → *V+Part+NP*: $y = 0.28x$, *V+Part+NP* → *V+NP+Part*: $y = 0.41x$)

In sum, there are no surprises with regard to the predictors discussed above: they are well behaved findings consistent with previous research. The relative effect sizes obtained are roughly comparable to those reported by Gries (2003a, 2003b), as is the predictive efficiency of the model.

4.2. Persistence-induced variation

This section will discuss how persistence affects particle placement choice. Consider first Figure 15, which sums up switching rates in the corpora analyzed in this chapter. Again, it is evident that speakers switch far less between the two particle placement strategies than pure chance would predict: with coefficients of the trend lines ranging between 0.27 and 0.40, speakers switch only about one third of the time they should if the null hypothesis (i.e. that there was no persistence) held. Saying that switch rates are lower than predicted by the null hypothesis is another way of saying there is more per-

Table 23. Linear regression estimates of switch rates in particle placement across corpora (y is the relative frequency of A → B switches, in %; x is the relative frequency of B forms, in %; the expected linear relationship, uninfluenced by persistence, is $y = x$)

corpus	*V+NP+Part* → *V+Part+NP*	*V+Part+NP* → *V+NP+Part*
CSAE	$y = 0.16x$	$y = 0.41x$
FRED	$y = 0.50x$	$y = 0.41x$

sistence (more α-persistence, to be precise) in speakers' production than is expected given the null hypothesis.

At the same time, Table 23 suggests that there are differences between the CSAE and FRED with regard to switch rates: switch rates from *V+Part+NP* → *V+NP+Part* are exactly the same in both corpora (0.41), but *V+NP+Part* → *V+Part+NP* switch rates are different. In the CSAE, there are considerably fewer *V+NP+Part* → *V+Part+NP* switches than in FRED. This means that in the CSAE, the pattern *V+NP+Part* pattern is more 'sticky' than in FRED.

To further examine the magnitude of persistence in the data, the variables pertaining to the domain of persistence (PREVIOUS, TEXTDIST, TEXTDIST-NOSEP, TEXTDIST-SEP, SAMETURN, SAMESPEAKER) were entered into logistic regression (Table 24). In both FRED and the CSAE, this step improved the model significantly.[46] As a result, consideration of persistence-related independents improves variance explained (R^2 is now approximately 0.44 in both corpora) and predictive efficiency in both corpora.

4.2.1. α-persistence

The main effect associated with PREVIOUS, the primary α-persistence variable, has somewhat different effect sizes in the CSAE and in FRED. In the CSAE, the odds ratio of 0.78 indicates that if a phrasal verb with variable object-particle patterning takes the *V+Part+NP* pattern, the odds that the other pattern (*V+NP+Part*) will be used next time decrease by 22%. This effect is not statistically significant, however. The corresponding decrease in FRED is 99% and statistically significant. These percentages hold conditioned that the interactional factors in the regression model are zero. If they are not, the impact of PREVIOUS on CURRENT is modulated in the following ways:

Table 24. Particle placement: odds ratios associated with persistence-related predictors in logistic regression (baseline predictors are included, but not displayed)

	CSAE	FRED
PREVIOUS(*V*+*Part*+*NP*)	0.78	**0.01** ***
PREV.(*V*+*Part*+*NP*) ∗ TEXTDIST	1.00	1.00
PREV.(*V*+*Part*+*NP*) ∗ SENTENCELENGTH	0.96	**1.02** *
PREV.(*V*+*Part*+*NP*) ∗ TTR	1.01	**1.01** *
PREV.(*V*+*Part*+*NP*) ∗ SAMETURN(1)	0.59	1.02
PREV.(*V*+*Part*+*NP*) ∗ SAMESPEAKER(1)	1.67	0.65
PREV.(*V*+*Part*+*NP*) ∗ VLEMMAID	**0.07** **	0.41
TEXTDIST-NONSEP	**1.01** *	1.01
TEXTDIST-SEP	1.00	1.00
PREV.(*V*+*Part*+*NP*) ∗ LITERALNESS(1)	2.02	**2.63** *
model intercept	0.83	∞ ***
N	152	1,048
model χ^2	57.97 ***	367.38 ***
R^2	0.434	0.445
% correct (baseline)	77.0 (63.8)	85.2 (76.3)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for the *V*+*NP*+*Part* pattern.

In FRED, there is a significant interaction (1.02) between PREVIOUS and SENTENCELENGTH such that as sentence length increases – and hence, as syntactic complexity of the environment surrounding CURRENT increases – the impact of PREVIOUS on CURRENT decreases. This finding is not expected in that it implies that α-persistence is actually *weakened* in syntactically complex environments, which is contrary to this study's working hypothesis.

Second, and again in FRED, TTR significantly interacts with PREVIOUS (1.01) such that as TTR increases – and hence, as lexical density of the environment surrounding CURRENT increases – the impact of PREVIOUS on CURRENT decreases. Again, this finding contradicts this study's working hypothesis that persistence is stronger in lexically complex environments because it is supposed to relax informationally dense contexts.

Third, we obtain a somewhat surprising interaction (significant in FRED only, but leaning in the same direction in the CSAE) between PREVIOUS and LITERALNESS: if the phrasal verb employed in CURRENT has a rather literal meaning, the odds ratio associated with PREVIOUS increases by a multiplicative factor of between 2.0 (CSAE) and 2.6 (FRED). In plain words, α-persistence is apparently weaker if the phrasal verb has a more literal meaning. In this case, the semantics of the target slot seems to override α-persistence.

Lastly, the CSAE exhibits a significant interaction such that the strength of α-persistence is also dependent on whether the same verb lemma is employed in both PREVIOUS and CURRENT. If it is not, the odds ratio associated with the main effect of PREVIOUS is 0.78, as we have seen before. If the verb lemma is identical, however, the odds ratio associated with PREVIOUS is $0.78 \times 0.07 = 0.05$. This means that α-persistence in the CSAE is much stronger if PREVIOUS and CURRENT share the same verb lemma. For some reason, α-persistence is not modulated this way in FRED.

No significant interaction between the two turn-by-turn variables (SAMETURN and SAMESPEAKER) and α-persistence could be obtained in logistic regression. Likewise, there was no significant interaction effect between PREVIOUS and TEXTDIST – that is, between the strength of α-persistence and textual distance between two subsequent variables. This is very likely to be a consequence of the overall comparatively low number of observations on which the regression estimates are based. Notwithstanding lacking significance in logistic regression, Figure 16 illustrates that in fact there is a relationship between PREVIOUS and TEXTDIST by plotting the strength of persistence, on the *y*-axis, against textual distance between two successive choice contexts on the *x*-axis. In FRED, the proportion of identical particle placement choices in PREVIOUS and CURRENT decreases as (non-logged) textual distance between PREVIOUS and CURRENT increases, as expected. In FRED at least, this relationship is best described as logarithmic, as the following curve fits[47] demonstrate:

	FRED	CSAE
adjusted R^2 linear	0.04	-0.02
adjusted R^2 logarithmic	0.33 **	-0.03
df	17	17

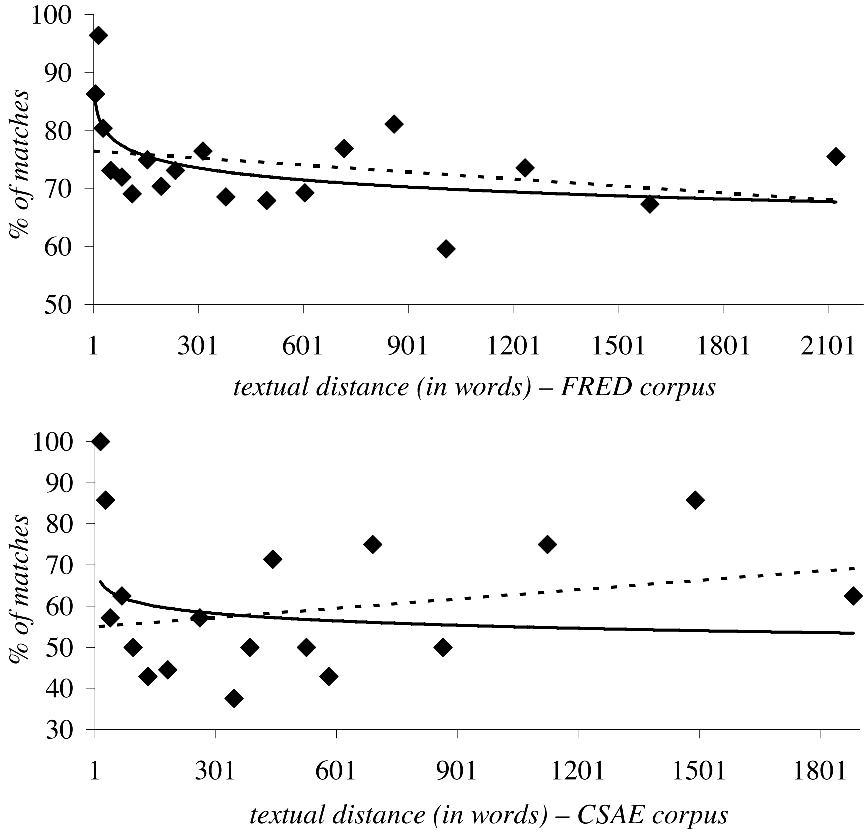


Figure 16. Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same particle placement strategy is used) as function of textual distance (in words) between CURRENT and PREVIOUS. Heavy line represents logarithmic estimate of the relationship, dotted line represents linear estimate of the relationship

In the CSAE, the fit of both curves is rather unacceptably bad. This is very likely due to low *N*s, which make the elimination of statistical outliers harder. Even so, it is clear that α-persistence in particle placement declines with increasing textual distance between two successive variable slots, as it should given our working hypothesis.

4.2.2. β-persistence

In logistic regression, I included two variables that were hypothesized to be sensitive to β-persistence, TEXTDIST-NONSEP, and TEXTDIST-SEP.

TEXTDIST-SEP does not nearly reach significance in either of the two corpora; moreover, the odds ratio associated with the variable (1.00) suggests that it would not matter even if it was significant. In contrast, TEXTDIST-NONSEP – the textual distance between CURRENT and the last generic non-separated pattern, such as a prepositional verb (e.g. *I look at the house*) – is significant in CSAE, and has the same effect in both corpora (1.01): as the *ln* of textual distance between CURRENT and the last generic non-separated pattern increases by one unit, the odds for *V+NP+Part* pattern, i.e. the separated pattern in CURRENT, increase by 1%. In a nutshell, this means the closer a generic non-separated pattern is to the variable site, the greater the odds for the *V+Part+NP* pattern, i.e. the non-separated pattern. This is β-persistence: a non-variable pattern can trigger a certain pattern in a slot when there is a choice.

5. Summary

The factors argued in previous research to influence particle placement do a decent job in explaining the alternation – they account for about one third of the observable variance. As for predictive efficiency, Gries (2003b: 130) notes that his discriminant analyses of spoken data correctly predicted 79.5% of the outcomes. My models including traditional predictors predict between 74% and 82% of the outcomes correctly (cf. Table 22). Thus, it seems that this study's findings are roughly consonant with the previous multivariate literature on particle placement.[48]

Given this literature, there were also no surprises with regard to the effects of individual predictors. In sum, the length of the direct object, in syllables, is the most potent predictor of particle placement. Next is whether or not the phrasal construction has a rather literal meaning – a semantic factor – and whether or not the referent of the object noun phrase is discourse-old, which is a discourse-functional variable. The presence of a directional prepositional phrase after the direct object and the presence of a definite determiner in the direct object phrase also have the expected effect on particle placement.

Increased sentence length – hence, by inference, increased syntactic complexity – appears to slightly favor usage of the *V+NP+Part* pattern, at least in FRED. This is rather puzzling: Hawkins (1994) has demonstrated that if the direct object of a transitive phrasal verb is longer than one word, *V+Part+NP* is always easier to parse than *V+NP+Part*. To ease the processing load in syntactically complex environments, one would actually expect language users to favor the *V+Part+NP* pattern in syntactically complex environments. Note now, however, that the *V+NP+Part* pattern has one advantage over its alternative: it more clearly delimits the direct object phrase from the syntagmatic environment because the 'moved' particle serves as a verb bracket. In this respect, the *V+NP+Part* pattern might help to ease the processing load in syntactically complex environments. The interplay between parsing ease of the phrasal verb and its direct object, syntactic complexity of the syntagm where the phrasal verb is embedded (conceptualized here through the variable SENTENCELENGTH), and the length of the direct object is most likely a very complex one. In the configuration of the research design used to analyze FRED, at any rate, increased complexity of the embedding syntagm seems to favor the *V+NP+Part* pattern.

We obtained contradictory effects for increased lexical density. Finally, consideration of idiosyncratic collostructional preferences of individual phrasal verbs for one or the other particle placement pattern – Gries and Stefanowitsch's 'distinctive collexeme' score – yields an increase in explained variance of approximately 5 percent points, which is considerable. As for FRED specifically, it emerged that compared to the Southeast, the *V+NP+Part* pattern is dispreferred in the Hebrides and in the Midlands, but slightly preferred in the North of England.

Next, we saw that speakers switch between the two particle placement strategies a good deal less than we would expect if usage of the two patterns was unaffected by persistence; the lower the switch rates, the greater α-persistence. Switch rates in the direction *V+Part+NP* → *V+NP+Part* were the same in both FRED and the CSAE (0.41). Switch rates in the other direction (*V+NP+Part* → *V+Part+NP*) differed between FRED and the CSAE; the average switch rate in this direction was 0.28. This suggests that the *V+NP+Part* pattern may be more 'sticky' than the *V+Part+NP* pattern.

Logistic regression also provided evidence for both α and β-persistence. Compared to a model including persistence-related variables, a model not including such variables misses some 4–14% of the observable variance. A good deal of this extra explanatory power is due to α-persistence: as a main effect, usage of the *V+Part+NP* pattern decreases the odds that the rival pattern will be used at the next possible opportunity by between 22% and 90%, depending on the corpus. However, a number of factors interact and interfere with this main effect, thus modulating α-persistence. To begin with, α-persistence is comparatively stronger in syntactically more complex environments (of which increased sentence length was taken as a proxy). This is unexpected, given that we assumed persistence to ease production as well as comprehension, and that this facilitative effect should be especially important in syntactically complex environments. Relatedly, as lexical density increases, α-persistence weakens. Again, this is surprising since we assumed repetitive, persistent production to relax informationally dense discourse. Again, with regard to particle placement, the opposite seems to be true, according to the data. Moreover, persistence in particle placement is weaker when the phrasal verb has a more literal, spatial meaning. It seems that this is a case where semantics can override sequential dependencies. We also saw that in the CSAE, α-persistence between two subsequent phrasal verb constructions is considerably stronger when the same verb lemma is employed in both constructions (cf. Pickering and Branigan 1998 and Gries 2005). A possible psycholinguistic account for this finding is that if the same lemma is used, there is also repetition or lexical priming between two subsequent tokens, which may amplify syntactic priming in particle placement. Last but not least, the interaction effect between recency of use of a pattern and α-persistence turned out to be insignificant in logistic regression; yet, plotting the percentage of persistent pairs against textual distance between the pair (Figure 16) indicated that there is such a relationship, at least in FRED: α-persistence between two phrasal verb constructions is stronger when previous usage was recent. The decline function that describes this relationship is best described as logarithmic, rather than linear.

The importance of β-persistence varies between FRED and the CSAE: while virtually non-existent in FRED, it accounts for 4–5% of the observable variance in particle placement in the CSAE. β-persistence manifests as follows: given a phrasal verb construction, the more recent a generic non-separated pattern – for instance, a transitive prepositional verb (*I look at the house*) – was used, the more likely it is that the non-separated *V+Part+NP* pattern will be used in the variable slot. Thus, generic non-separated patterns – although not variable themselves – trigger the *V+Part+NP* pattern when there is a particle placement choice. This is true for production in the CSAE, but for some reason not in FRED.

In sum, the analysis presented here demonstrated that more variation in particle placement can be accounted for and that speakers' choices can be predicted substantially more accurately if persistence-related factors are considered.

Chapter 8
Persistence in complementation strategy choice

This chapter will investigate persistence in the choice of nonfinite verbal constructions, that is, in the binary variation after a number of head verbs between infinitival complementation (henceforth: *V+inf.*) as in (1a), and gerundial complementation (henceforth: *V+ger.*) as in (1b):

(1) a. *By giving 4 or 5 of those, then the intelligent adult will* ***start to think*** *about this*. (CSPAE Comm597)
b. *We have* ***to start thinking*** *creatively* ... (CSPAE Facmt97)

1. Background and previous research

There appears to be more or less of a consensus that the complementation patterns themselves have semantic content. In this spirit, Quirk et al. (1985) state that

> where both constructions ... are admitted, there is usually felt to be a difference of aspect or mood which influences the choice. As a rule, the infinitive gives a mere 'potentiality' for action, as in *She hoped to learn French*, while the participle gives a sense of the actual 'performance' of the action itself, as in *She enjoyed learning French*. (Quirk et al. 1985: 1191)

Yet, it has proven to be notoriously hard to pin down these differences (cf. Quirk 1974: 66–67: "There ought to be a big award for anyone who can describe exactly what makes him say 'I started to work' on one occasion and 'I started working' on another"). There is a voluminous literature on semantic or pragmatic differences between the two complementation types, a survey of which would go beyond the brief of this book – and be that as it may, Mair (2003: 329) submits that "this particular fragment of English grammar is in a state of flux diachronically, with *-ing*-forms gradually encroaching on the infinitive," which "bedevils any attempt at an 'exact' synchronic description." Suffice it, then, to point out here that according to the literature (Fanego 1997; Quirk et al. 1985: 1192–1193; Řeřicha 1987: 30), four classes of verbs – of which the first two classes will be investigated here – can potentially take both complementation types:

- emotive verbs (*dread, hate, like, loathe, love, prefer*);
- aspectual verbs of beginning, continuing, and ending (*start, begin, continue, cease*, etc.);
- retrospective verbs (*forget, remember, regret*);
- some verbs of effort (*try*, *intend*, *attempt*).

Of these four classes, verbs of beginning, continuing, and ending seem to have been the focus of interest of most researchers (for instance, Duffley 1999; Mair 2002, 2003; Řeřicha 1987). After such head verbs, it is generally agreed that despite structural and semantic constraints, there is a considerable range of variation in which the two complementation patterns compete. In contrast, the classes of retrospective verbs and verbs of effort are somewhat problematic from a variationist perspective since these verbs are strongly semantically conditioned; this is why the latter two classes of verbs, except for *intend* (where the choice between infinitival and gerundial complementation is sufficiently optional), will not be considered in this chapter.

There are two ways in which persistence can influence the alternation between *V+inf.* and *V+ger.* complementation. For one thing, the two complementation patterns are of course different syntactically, so psycholinguistically any given complementation site can be a target for syntactic priming. On the other hand, the two complementation patterns differ in the lexical and morphological material with which they are coded. Infinitival complementation involves the infinitive marker *to*, which gerundial complementation does not. Gerundial complementation is thus more compact lexically, although it affixes the complement with the morpheme {ing}. Therefore, both lexical priming and morphological priming are potentially relevant to the alternation.

2. Previously suggested factors

With the exception of *horror aequi* contexts, infinitival complementation is generally always possible, while there are some constraints on the syntactic environments where gerundial complementation can occur (cf. Mair 2003: 333). These constraints are the following:

Horror aequi contexts. *V+ger.* tends to be avoided when the head verb itself is used in the progressive or as a participle (e.g. ?*I was starting wondering*) (cf. Mair 2003: 333). In a similar vein, *V+inf.* tends to be avoided if the head verb is itself in the infinitive (e.g. ?*to start to wonder*; cf. Vosberg 2003: 322; Fanego 1997). *Horror aequi* comes pretty close to being a knock-out factor, but since it is predominantly referred to as a stylistic constraint in the literature, it will be considered a non-categorical constraint in this chapter.

Adverbials between the head verb and complement. *V+ger.* is avoided when an adverbial is placed between the head verb and its verbal complement (cf. Dixon 1991: 178). Thus, according to Mair (2003: 333), *they began in the following years to sell the product* is preferred to *they began in the following years selling the product*. Note that this factor is going to be omitted in this chapter's analysis for lack of relevance. Examination of a sample of 100 *start* + ger. and 100 *start* + inf. forms in the CSPAE revealed that there was not a single intervening adverbial in this sample. Presumably, the factor is more relevant in written language than in spoken language.

Hypothetical meaning. Biber et al. (1999: 757–758) produce evidence that "75% of the occurrences of *like* + *to*-clause ... are preceded by *would* ... In contrast, *-ing*-clauses rarely occur with a hypothetical meaning." Thus, *V+inf.* is preferred in hypothetical contexts.

Stative complements. According to Řeřicha (1987: 130), *V+ger.* is avoided after the verbs *begin* and *start* when the following verb is stative. Presumably, this constraint also applies to all other verbs where the complementation pattern is variable. This is because the constraint has probably less to do with the head verb, but with the fact that the *-ing* form is related to the progressive construction (cf. Freed 1979: 72–73; Palmer 1974: 171), which is also impossible with stative verbs.

3. Method, data and independent variables

3.1. Method and data

Loci of variation in the sense of the present study are verbs whose complementation behavior is maximally unconditioned by semantic factors, *viz.*

emotive verbs and aspectual verbs. This means that this chapter will analyze the complementation patterns of the following 11 verbs: *begin, cease, continue, dread, hate, intend, like, loathe, love, prefer,* and *start*.

3.1.1. Data extraction

Identifying *V+inf.* and *V+ger.* patterns requires manual disambiguation, unless the data source is POS tagged.[49] Because not all of the corpora subject to analysis in this chapter are POS tagged, I used two different methods to extract the relevant alternation sites from the data:

– *POS tagged corpora (DS, CG, CSPAE)*: for these corpora, extraction was performed automatically. A Perl script identified all instances of the above head verbs that were followed by either a gerund or an infinitive phrase and extracted them. This method yielded an accuracy rate of 96–97%; errors were largely due to incorrect POS tagging in the data source(s).

– *Non-tagged corpora (FRED, CSAE)*: for these corpora, the data were first parsed manually to identify the relevant alternation sites. The datasets analyzed were the CSAE in its entirety and a manageable subset of FRED.[50] The alternation sites (i.e. all occurrences of the above head verbs that were either followed by a gerund or an infinitive phrase) were manually tagged, so that a Perl script could then retrieve and extract them.

3.1.2. Data coding

Another Perl script then coded the extracted sites for the standard independent variables (see chapter 3, section 1.2) as well as for those independents specific to complementation choice (see section 3.2 below). This method yielded a total of 9,520 tokens, a breakdown of which is presented in Table 25. Two things about this table strike me as noteworthy: first, the proportion of *V+inf.* to *V+ger.* is roughly 50:50 in the less formal corpora (DS, CSAE, FRED), while it is approximately 80:20 in the more formal corpora. Clearly, *V+ger.* is more common in more informal registers. Second, there is really no clear way in which American English differs from British English with regard to complementation strategy choice: it is true that in the CSAE (informal American English), *V+ger.* has the highest text frequency (52.0%) in the dataset, *Table 25*. Complementation strategy choice: distributional variation across corpora

corpus	*N*	*N* infinitival	*N* gerundial
CG	4,551	3,554 (78.1%)	997 (21.9%)
DS	2,027	1,011 (49.9%)	1,016 (50.1%)
CSPAE	2,135	1,837 (86.0%)	298 (14.0%)
CSAE	102	49 (48.0%)	53 (52.0)%
FRED	705	325 (46.1%)	380 (53.9%)
total	**9,520**	**6,776 (71.2%)**	**2,744 (28.8%)**

yet in the CSPAE (formal American English), *V+ger.* has actually the lowest text frequency (14.0%). Judging from these numbers, it appears that stylistic stratification is stronger in American English than in British English.

3.2. Independent variables

The following predictors, which are tailored to the alternation between *V+ger.* and *V+inf.*, will be considered in addition to the standard variables discussed in chapter 3:

3.2.1. Previously suggested and persistence-unrelated predictors

STATIVE COMPLEMENTS (henceforth: STATIVE-COMPL). Is the complement one of the following stative verbs[51], as in (2)?

> *abhor, adore, astonish, be, believe, concern, contain, cost, deserve, desire, detest, dislike, doubt, equal, feel, fit, forgive, guess, hate, imagine, impress, include, intend, involve, know, lack, like, love, matter, mean, need, owe, own, perceive, please, possess, prefer, presuppose, realize, recall, recognize, regard, remember, require, require, resemble, satisfy, seem, smell, sound, suppose, taste, understand, want, wish*

(2) *So, I would* ***like to know*** *who the enemy is*. (CSPAE Comm897)

(coded 1 if the complement is stative, and 0 otherwise)

Hypothesis: Following Řeřicha (1987: 130), we expect that *V+ger.* is avoided when the complement is stative.

HYPOTHETICAL MEANING (henceforth: HYPOTHETICAL). Is the VP used in a hypothetical context, i.e. is the head verb preceded by *would, would not, wouldn't*, or *'d not*, as in (3)?

(3) *. . . he voiced an opinion he* ***would not like to be put to sleep*** *if at all possible.* (CSPAE Wh97a)

(coded 1 if the context is hypothetical, and 0 otherwise)
Hypothesis: According to Biber et al. (1999: 757–758), *V+ger.* is unlikely in hypothetical contexts.

HORROR AEQUI (henceforth: TO-HORRORAEQUI and ING-HORRORAEQUI). Is the head verb itself an infinitive, as in (4), or is it an *-ing* form, as in (5)?

(4) *The President has indicated that he was going* ***to start using*** *a cane Monday.* (CSPAE Wh97a)

(5) *The states* ***are just starting to test*** *that idea.* (CSPAE Wh97a)

(coded 1 if the head verb is an infinitive/*-ing* form, and 0 otherwise)
Hypothesis: If the head verb itself is an *-ing* form, we expect a *horror aequi* effect such that *V+ger.* is then avoided (cf. Mair 2003: 333). By the same token, we expect that *V+inf.* is avoided when the head verb is itself used in the infinitive.

TYPE OF THE HEAD VERB (henceforth: VERB). It is likely that the 11 head verbs under analysis in this chapter differ in their collostructional preferences. As leaving this verbal variation unaccounted would cause unnecessary statistical noise, the 11-way categorical variable VERB will control for this variation.

MORPHOLOGICAL FORM of the head verb (henceforth: MORPHOLOGY). Morphologically, is the head verb used

1. in its base form (e.g. *we start wondering*; note that the base form is not necessarily an infinitive),

2. in the 3rd person singular (*he starts wondering*),
3. in its past or past participle form (*he started wondering*)?[52]

Though not discussed in the literature (except for *horror aequi* effects), it may be that the morphological form of the head verb has an impact on which complementation type is chosen.

FRED DIALECT AREA (henceforth: FRED-AREA). This variable is relevant for FRED only and is sensitive to how complementation choice differs across the dialect regions (Hebrides, Midlands, North, and Southeast) sampled in the FRED subset under analysis.

3.2.2. *Additional, persistence-related predictors*

SAME VERB LEMMA in PREVIOUS and CURRENT (henceforth: VLEMMA-ID). This predictor concerns whether two successive complementation slots involve the same head verb lemma (though not necessarily the same head verb form – coded 1 if the lemma is the same, and 0 if it is not).
Hypothesis: Pickering and Branigan (1998) and Gries (2005) showed that production priming is stronger when the priming verb lemma and the target verb lemma are the same. Thus, when the head verb lemma is the same in two successive complementation slots, α-persistence between these sites is even stronger than it would be otherwise.

SAME VERB MORPHOLOGY in PREVIOUS and CURRENT (henceforth: VMORPHID). Complementing VLEMMAID, this predictor shows whether two successive variable complementation slots do in fact have the same morphological verb form (though not necessarily the same head verb lemma, see below – coded 1 if the verb morphology is the same, and 0 if it is not). For example, in (6) two complementation sites (*start seeing* and *start having*) occur in proximity. They involve the same verb lemma (*start*) as well as the same head verb morphology (the base from of the verb *start*). In addition, of course, they also involve the same complementation pattern, *V+ger.*

(6) *I think it may be 10 years from now when people* ***start seeing*** *the long-term effects from it, and you* ***start having*** *problems with it.* (CSPAE Comm 8a97)

Hypothesis: In analogy to VLEMMAID, it is to be expected that α-persistence between two slots is stronger than it would be otherwise if the two head verbs have the same morphological form (cf., for instance, Gries 2005).

TEXTUAL DISTANCE to last occurrence of the token *to* or to the last infinitive VP (henceforth: TEXTDIST-INF). This is a β-persistence variable. For the POS tagged corpora (DS, CG, CSPAE), this variable measures the *ln* of the textual distance between CURRENT and the last generic infinitive phrase (as in *he stopped the car to smoke a cigarette*); for FRED and the CSAE, it measures the *ln* of the textual distance between CURRENT and the last generic occurrence of the token *to* (as in *he went to the store*).
Hypothesis: Infinitive phrases or *to* tokens can trigger *V+inf.* Thus, if textual distance between CURRENT and an infinitival trigger is small, the odds for *V+inf.* in CURRENT should increase.

TEXTUAL DISTANCE to last word ending in *-ing* or to the last gerund (henceforth: TEXTDIST-ING). This is the gerundial counterpart to TEXT-DIST-INF. For the POS tagged corpora, it measures the *ln* of the textual distance between CURRENT and the last generic gerund form (as in *working usually tired him*); for FRED and the CSAE, it measures the *ln* of the textual distance between CURRENT and the last token ending in *-ing* (as in *he was fighting many battles*).
Hypothesis: The odds for *V+ger.* in CURRENT increase when a gerundial trigger has been used recently – i.e. when TEXTDIST-ING is small.

NUMBER OF WORDS STARTING IN <t> in the discourse preceding CURRENT (henceforth: T-ALLIT). In a context of 50 words before CURRENT, how many tokens are there that – like *to* – start in <t>?[53] T-ALLIT is yet another β-persistence predictor. There is both discourse analytic evidence (cf. Sacks 1971; Tannen 1989 on 'sound coordination') and psycholinguistic evidence (cf. Dell 1986; Cohen and Dehaene 1998 on 'phoneme perseveration') that speakers prefer alliterating options.

Table 26. Complementation strategy choice: independent variables considered

variable	type	coding method
a. previously suggested and persistence-unrelated independents		
SENTENCELENGTH*	scalar	software
TTR*	scalar	software
STATIVE-COMPL	two-way categorical	software
HYPOTHETICAL	two-way categorical	software
TO-HORRORAEQUI	two-way categorical	software
ING-HORRORAEQUI	two-way categorical	software
VERB	11-way categorical	software
MORPHOLOGY	three-way categorical	software
FRED-AREA	five-way categorical	software
b. persistence-related independents		
PREVIOUS*	two-way categorical	software
TEXTDIST*	scalar	software
SAMETURN*	two-way categorical	software
SAMESPEAKER*	two-way categorical	software
VLEMMAID	two-way categorical	software
VMORPHID	two-way categorical	software
TEXTDIST-INF	scalar	software
TEXTDIST-ING	scalar	software
T-ALLIT	scalar	software
c. speaker characteristics		
AGE*	scalar	software
SEX*	two-way categorical	software

* independent variable discussed in chapter 3, section 1.

> *Hypothesis:* Infinitival complementation categorically involves the token *to*; speakers are more likely to use the infinitival option when T-ALLIT is high – in other words, when there are many words in CURRENT's context that start in <t>, as does infinitival *to*.

Table 26 summarizes the independent variables considered in this chapter.

4. Results

4.1. Baseline variation

As usual, I begin by reviewing how well persistence-unrelated factors – especially those discussed in previous research – actually account for the observable variation in the present study's database. I estimated logistic regression models for every individual corpus, the parameters of which are displayed in Table 27.

As for the overall quality of these models, variance explained (R^2) encompasses a surprisingly narrow 54–56% range in the three major corpora (the CG, CSPAE, and the DS), which is a decent level of explanatory power. In the CSAE, the model explains a much better 77% of the observable variance in complement choice, while in FRED, the figure is 32% only. The picture with regard to predictive efficiency is similar: in the CG, CSPAE, the DS, and the CSAE, the model gets between roughly 80–90% of speakers' actual choices right, while the figure is only 72% in FRED. In sum, the model works best for the CSAE, decently for the three major corpora, and not so well for FRED.

With regard to the individual predictors, SENTENCELENGTH (and, by implication, syntactic complexity) is significant in the CSAE and FRED only and has opposed effects in the two corpora. Therefore, the effect of the predictor is unclear. By contrast, increased TTR values (thus, increased lexical density) favors the *V+ger.* pattern across all corpora under analysis, albeit significantly so only in the CSPAE and DS. Therefore, speakers appear to resort to the option that is more economic lexically (*V+ger.*) in contexts that are lexically dense anyway.

Given the literature, there was good reason to believe that *V+ger.* would be avoided with stative complement verbs. This expectation is indeed borne out in the CSPAE and CSAE: in the former, a stative complement verb significantly reduces the odds for *V+ger.* by 90% (exp(b) = 0.10); the corresponding figure for the CSAE is even a categorical – and highly significant – 100% (exp(b) = 0.00). However, the effect of the predictor is just the reverse in the other corpora. In the CG, a stative complement significantly *increases* the odds for *V+ger.* by 65%; in the DS by a highly significant 143%; and in FRED by 60% (though insignificantly). This means that we are dealing with a very clear pattern of regional stratification here: in the two American

Table 27. Complementation strategy choice: odds ratios associated with baseline predictors in logistic regression

	CG	CSPAE	DS	CSAE	FRED
SENTENCELENGTH	1.00	0.99	1.00	**1.08** *	**0.99** *
TTR	1.01	**1.04** *	**1.04** ***	1.02	1.01
STATIVE-COMPL	**1.65** *	**0.17** *	**2.43** ***	**0.00** ***	1.60
HYPOTHETICAL	**0.05** *	**0.10** ***	**0.03** ***	0.07	**0.39** ***
TO-HORRORAEQUI	**2.70** *	**3.05** ***	1.54	∞ ***	0.67
ING-HORRORAEQUI	**0.05** *	**0.03** ***	**0.02** ***	**0.00** **	**0.08** ***
VERB	**–** ***	**–** ***	**–** ***	**–** **	**–** ***
MORPHOLOGY	**–** ***	**–** ***	**–** ***	–	**–** ***
FRED-AREA	n.a.	n.a.	n.a.	n.a.	**–** ***
model intercept	0.05 ***	0.02 ***	0.03 ***	0.01	0.15
N	4,551	2,136	2,022	101	709
model χ^2	2,065.38 ***	761.47 ***	1,101,25 ***	86.09 ***	194.64 ***
R^2	0.561	0.541	0.560	0.765	0.321
% correct (baseline)	86.4 (78.1)	90.7 (86.0)	79.5 (50.0)	88.1 (51.5)	72.2 (53.7)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for *V+ger*.

NOTE: Estimates for individual category levels of VERB, MORPHOLOGY, and FRED-AREA are not displayed for reasons of space.

English corpora, the predictor has the expected effect, while in the three corpora sampling English spoken in the British Isles, it has not.

Note, now, that Řeřicha's (1987) original hypothesis was that gerundial complementation is avoided after the verbs *begin* and *start* only, not after verbs whose complementation pattern is variable in general. Some readers may object that Řeřicha (1987) was right, and that the reason for the mixed picture is that I did not restrict my analysis of the effect of stative complements to *begin* and *start* only. To deal with this, I conducted another logistic regression run on the CG and DS database that included *begin* and *start* observations only. This resulted in the following odds ratios[54] for the effect of stative complements:

	CG	DS
STATIVE-COMPL	0.28 ***	0.22 ***

Thus, much as originally claimed by Řeřicha, stative complements after *begin* and *start* do indeed reduce the odds for *V+ger.*, also in the British English corpora. In the American English corpora, stative complements have this effect even with other head verbs; in the British English corpora, they do not.

The predictor HYPOTHETICAL has the same effect across the board. If the head verb is used in a hypothetical context (as in *I would start to wonder*), the odds for *V+ger.* are reduced substantially, as expected (though in the CSAE, the predictor is not significant). The effect size is such that a hypothetical context reduces the odds for *V+ger.* by between 51% (FRED) and 97% (DS). Again, there is a moderate pattern of stylistic and regional stratification: the predictor appears to have (i) a stronger effect in the two informal standard corpora (exp(*b*) = 0.03 / 0.07) than in the two formal standard corpora (exp(*b*) = 0.05 / 0.10), and (ii) it seems to be slightly more influential in the two Standard British English corpora (exp(*b*) = 0.05/0.03) than in the two standard American English corpora (exp(*b*) = 0.10 / 0.07).

The two *horror aequi* predictors also dovetail nicely with our expectations. TO-HORRORAEQUI (i.e. whether the head verb itself is an infinitive, as in *John had to start wondering*) turns out to be significant in the CG, CSPAE, and CSAE. In these corpora, if the head verb itself is an infinitive, the odds for *V+ger.* increase manifold. This is evidence for *horror aequi*: speakers avoid two infinitives in adjacency when they can. Much the same goes for ING-HORRORAEQUI – when the head verb itself is an *-ing*-form, as in *John was starting to wonder*. This is the only predictor that is selected as significant throughout. It is also the predictor that is associated with the biggest effect sizes throughout. More specifically, if the head verb itself is an *-ing* form, the odds for *V+ger.* decrease by between 92% (FRED) and a categorical 100% (CSAE). Thus, *horror aequi* strongly discourages speakers from using two adjacent *-ing* forms if they can avoid it.

The variable VERB is sensitive to how the eleven head verbs lumped together in the analysis differ in their preferences for *V+ger.* or *V+inf.* The following tabular displays odds ratios for every individual head verb except *begin*, which is, entirely arbitrarily, taken as the statistical baseline verb:[55]

	CG	CSPAE	DS	CSAE	FRED
cease	3.66	(n.s.)	∞	–	(n.s.)
continue	(n.s.)	0.23	(n.s.)	(n.s.)	(n.s.)
dread	5.79	–	(n.s.)	–	–
hate	9.74	(n.s.)	6.82	(n.s.)	(n.s.)
help	(n.s.)	0.00	3.75	0.00	(n.s.)
like	2.17	(n.s.)	3.99	(n.s.)	(n.s.)
loathe	0.00	0.00	–	–	–
love	8.09	(n.s.)	8.38	(n.s.)	(n.s.)
prefer	(n.s.)	(n.s.)	(n.s.)	(n.s.)	(n.s.)
start	21.74	10.24	31.29	98.12	16.49

This means that compared to *begin*,

- *cease* has a marked preference for *V+ger.*;
- *continue* has a preference for *V+inf.* in the CSPAE;
- *dread* and *hate* have a preference for *V+ger.* in the CG;
- *help* has a strong preference for *V+inf.* in the two corpora of American English, but a preference for *V+ger.* in the DS;
- *like* prefers the *V+ger.* pattern, at least in the two British English corpora;
- *loathe* is only observed with *V+inf.* (moreover, *loathe* is only attested in the two formal corpora);
- *love*, in the two British English corpora, has a strong preference for *V+ger.*;
- *prefer* does not differ significantly from *begin*;

– *start* has an exceedingly strong association with *V+ger.* throughout (i.e. *I began to wonder* and *I started wondering* are typical). While this finding is consonant with Biber, Conrad, and Reppen's (1998, 99) claim that "relative to *start*, *begin* has a greater preference for the *to*-clause pattern," the strength of the skewing is somewhat surprising, given that *begin* and *start* are regularly lumped together analytically.

I also suggested that the morphological shape of the head verb might influence complementation preferences. Below I give odds ratios associated with the 3rd person singular and the past/past participle form (recall that the impact of the infinitive form and the *-ing* form are covered by the variables TO-HORRORAEQUI and ING-HORRORAEQUI already). The base form of the head verb (as in *we start wondering*) serves as the statistical baseline form.[56]

	CG	CSPAE	DS	CSAE	FRED
3rd person sg.	0.42	0.29	0.44	(n.s.)	0.21
past/past participle	0.71	2.15	0.36	(n.s.)	0.39

For one thing, compared to the base form, the 3rd person singular form seems to consistently discourage use of the *V+ger.* pattern; thus, *Jim starts to wonder* is somehow more typical than *Jim starts wondering*. Second, the past/past participle form also discourages the use of the *V+ger.* pattern, with one exception: in the CSPAE, it increases the odds for *V+ger.* Therefore, in the CSPAE, *I started wondering* is more typical than *I started to wonder*; in most of the other corpora, the relationship is the other way round.

Finally, in FRED there is variation between dialect areas with regard to complementation preferences. Taking the Southeast as the statistical baseline area (as before, in an entirely arbitrary fashion), this variation manifests as follows in logistic regression:[57]

Hebrides	1.76
Midlands	0.37
North	2.72
Southwest	0.53
Wales	1.52 (n.s.)

Therefore, compared to the Southeast, *V+ger.* is significantly more frequent in the Hebrides, and a lot more frequent in the North of England. In the Midlands and in the Southwest, by contrast, *V+ger.* is less likely to occur than

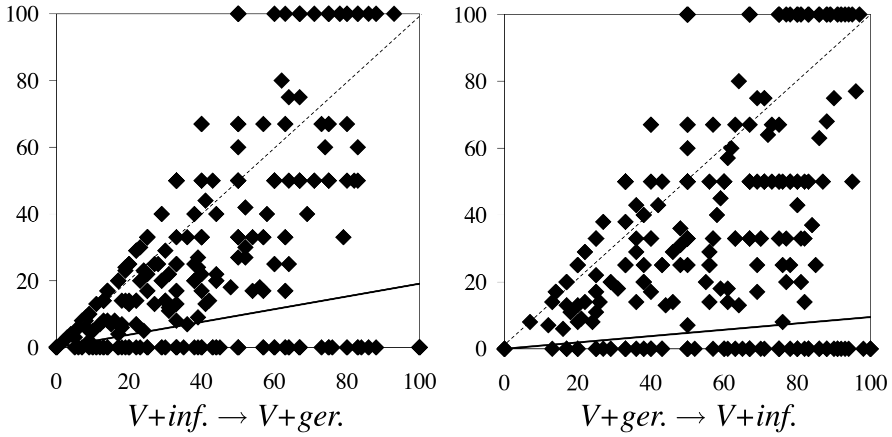


Figure 17. Switches in complementation strategy choice as a function of overall proportion of complementation strategy (relative frequency of switches, in %, on y-axis; relative frequency of the switched-to complementation strategy, in %, on x-axis) in the dataset under analysis. Each dot represents one speaker. Dotted diagonal line represents null hypothesis that switch rate is proportional to variant proportions. Heavy line indicates linear trend (*V+inf.* → *V+ger.*: $y = 0.19x$, *V+ger.* → *V+inf.*: $y = 0.10x$)

in the Southeast. There is no significant difference between Wales and the Southeast.

4.2. Persistence-induced variation

How readily do speakers switch between infinitival and gerundial complementation? Consider Figure 17, which plots, in the dataset under analysis in this chapter, each individual speaker's switching rate (*V+inf.* → *V+ger.* and vice versa) against his or her overall usage proportion of the two complementation strategies. Were there no persistence, dots should cluster close to the diagonal, dotted line, which indicates a hypothetical, persistence-unaffected 'natural' switch rate. Actually, however, most speakers switch less than that: in both switching directions, the dots clearly cluster below the diagonal line, but especially so for *V+ger.* → *V+inf.*, where the dots nicely scatter all over the lower right half of the graph. Observe, therefore, that switch rates

Table 28. Linear regression estimates of switch rates in complementation strategy choice across corpora (y is the relative frequency of A → B switches, in %; x is the relative frequency of B forms, in %; the expected linear relationship, uninfluenced by persistence, is $y = x$)

corpus	*V+inf.* → *V+ger.*	*V+ger.* → *V+inf.*
CG	$y = 0.18x$	$y = 0.08x$
DS	$y = 0.14x$	$y = 0.11x$
CSPAE	$y = 0.35x$	$y = 0.09x$
CSAE	$y = 0.18x$	$y = 0.09x$
FRED	$y = 0.33x$	$y = 0.40x$

from *V+ger.* → *V+inf.* are lower than those from *V+inf.* → *V+ger.*, both intuitively and statistically (the regression line on *V+inf.* → *V+ger.* is twice as steep as the one on *V+ger.* → *V+inf.*). This observation also plays out in Table 28, which displays regression estimates of switch rates on a per-corpus basis. Switch rates are overall highest (thus, persistence is weakest) in FRED. Switch rates are a good deal lower in the other corpora. Also, in all corpora except FRED, *V+ger.* is more 'sticky' than *V+inf.* It is moreover worth mention that while switch rates from *V+ger.* → *V+inf.* are remarkably homogeneous in the four non-dialect corpora, switch rates from *V+inf.* → *V+ger.* appear to be higher in the formal corpora (CG: $y = 0.18 \times x$; CSPAE: $y = 0.35 \times x$) than in the informal corpora (DS: $y = 0.14 \times x$; CSAE: $y = 0.18 \times x$). If this is statistically real, it would mean that infinitival complementation – while less sticky than gerundial complementation – is less persistent in formal, more carefully planned speech than in conversation.

Table 29 shows how persistence-related predictors (i.e. PREVIOUS, TEXTDIST, VLEMMAID, VMORPHID, SAMETURN, SAMESPEAKER) behave in logistic regression. First of all, compared to the model displayed in Table 27, inclusion of the above predictors significantly enhances our understanding of how speakers select complementation strategies; the increase in model χ^2 is statistically significant throughout.[58] The increases in explanatory power (R^2) and predictive efficiency are as follows:

	CG	CSPAE	DS	CSAE	FRED
explanatory power (R^2)	+ 4.5%	+ 3.8%	+ 5.5%	+ 23.5%	+ 10.6%
predictive efficiency	+ 1.7%	+ 1.0%	+ 2.3%	+ 11.9%	+ 2.6%

Table 29. Complementation strategy choice: odds ratios associated with persistence-related predictors in logistic regression (baseline predictors are included, but not displayed)

	CG	CSPAE	DS	CSAE	FRED
PREVIOUS(*V+inf.*)	0.13	**0.001** *	**0.01** ***	**0.00** ***	3.00
PREVIOUS(*V+inf.*) ∗ TEXTDIST	**1.31** ***	0.94	**1.47** ***	∞	**1.32** *
PREVIOUS(*V+inf.*) ∗ VLEMMAID(1)	0.77	**0.58** *	0.97	0.00	0.45
PREVIOUS(*V+inf.*) ∗ VMORPHID(1)	0.99	1.04	1.13	**0.00** *	**2.9** *
PREV.(*V+inf.*) ∗ VMORPHID(1) ∗ VLEMMAID(1)	**0.51** *	0.48	0.67	0.00	0.37
PREVIOUS(*V+inf.*) ∗ SAMETURN(1)	1.23	1.29	**0.17** *	∞	0.61
PREVIOUS(*V+inf.*) ∗ SAMESPEAKER(1)	0.89	0.60	0.72	0.00	0.61
PREVIOUS(*V+inf.*) ∗ SENTENCELENGTH	1.00	0.99	**1.01** **	0.03	0.99
PREVIOUS(*V+inf.*) ∗ TTR	1.00	**1.12** *	1.02	17.07	0.96
TEXTDIST-ING	1.03	**0.85** *	**0.90** *	∞	1.06
TEXTDIST-INF	1.05	0.94	0.94	0.00	1.15
T-ALLIT	**0.86** ***	**0.88** ***	**0.88** ***	0.00	**0.89** ***
model intercept	0.59	28.69	0.57	0.00	0.06
N	3,928	2.134	1,876	71	555
model χ^2	1,943.98 ***	825.56 ***	1,160.35 ***	98.07 ***	213.76 ***
R^2	0.606	0.579	0.615	1.000	0.427
% correct (baseline)	88.1 (78.8)	91.7 (86.0)	81.8 (50.2)	100.0 (53.5)	74.8 (53.5)

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for *V+ger.*

So, the contribution of persistence to observed variance in complementation choice is approximately 4% in the two formal corpora, the CG and CSPAE. It is slightly bigger in the informal DS (5.5%), considerable in FRED (11%), and huge in the CSAE (24%). Overall, variance explained is now approximately 43% in FRED, and ca. 60% in the CG, DS, and CSPAE. The model for the CSAE accounts for all of the observable variation between infinitival and gerundial complementation in the corpus ($R^2 = 100\%$), which may sound more impressive than it actually is: the regression on the CSAE is based on 71 observations only, and although the model still is highly significant, one should proceed with caution when interpreting results based on such small case numbers. In all, it seems safe to say that persistence is clearly less influential in more formal registers than in more informal ones, as was to be expected.

4.2.1. α-persistence

What role does α-persistence play in complementation strategy choice? In order to answer this question, consider the impact of PREVIOUS, as well as the several interaction terms including PREVIOUS. To begin with, PREVIOUS is significant in three of the five corpora under analysis. The main effect of the predictor boils down to this: given two successive head verbs whose complementation type is optional, if *V+inf.* was used for the first head verb, the odds that *V+ger.* will be used for the second head verb decrease by

- 99.9% in the CSPAE;
- 99% in the DS;
- a categorical 100% in the CSAE (i.e. if the interactional factors are controlled for, *V+ger.* is never followed by *V+inf.* in the CSAE).

compared to if *V+inf.* had been used for the first head verb.

These percentages are controlled for the interactional factors in Table 29; let us now sort out the impact of each of these interactional factors on the strength of persistence between two successive variable sites. First, the interaction term with PREVIOUS and TEXTDIST is selected as significant in three of the five corpora and interacts with PREVIOUS in the hypothesized way: for each one-unit increase in the *ln* of textual distance between PREVIOUS and

CURRENT, the main effect of PREVIOUS is changed by a multiplicative factor of 1.31 (CG), 1.47 (DS), or 1.32 (FRED). In plain English, this means that α-persistence is stronger when textual distance between two complementation sites is small. Figure 18 visualizes this relationship by plotting, for every individual corpus except the CSAE (for which the number of observations is too low), the strength of persistence (on the *y*-axis) against the non-logged textual distance between the members of the pairs. Clearly, the likelihood that the complementation strategy in PREVIOUS and CURRENT matches is greater when PREVIOUS and CURRENT are textually close. And throughout, a logarithmic estimate of the relationship fits the data better[59] than a linear estimate:

	CG	CSPAE	DS	FRED
adjusted R^2 linear	0.47 ***	0.16	0.26 *	0.03
adjusted R^2 logarithmic	0.84 ***	0.62 ***	0.74 ***	0.38 ***
df	17	17	17	17

Thus, there is a forgetting function such that speakers 'forget' about previous choices as the discourse proceeds. This forgetting function appears to be logarithmic.

VLEMMAID determines whether two successive head verbs with optional complementation patterns involve the same head verb lemma. Although the interaction between PREVIOUS and VLEMMAID is significant only in the CSPAE, the fact that the exp(*b*) value associated with the interaction is smaller than 1 throughout leaves us good reason to believe that there is sufficient substance to the interaction between PREVIOUS and VLEMMAID to be interesting. In a nutshell, α-persistence is even stronger when the lemmas of two successive head verbs match (as in *I think it may be 10 years from now when people* start seeing *the long-term effects from it, and you* start having *problems with it* [CSPAE Comm8a97]) than it would be otherwise. In the CG, for instance, if there is a verb lemma match between PREVIOUS and CURRENT, α-persistence is 42% stronger than if the verb lemmas do not match. By parallel logic, VMORPHID indicates whether the morphological makeup of two successive head verbs with optional complementation patterns match (i.e. whether, for instance, both head verbs are affixed with 3rd person sg. -*s*). In analogy to VLEMMAID, I had expected that if the morphology is identical, persistence would be stronger than otherwise. VMORPHID is significant in FRED and the CSAE only. In the CSAE, the variable has the expected effect, though recall that due to low case numbers, results from the CSAE are

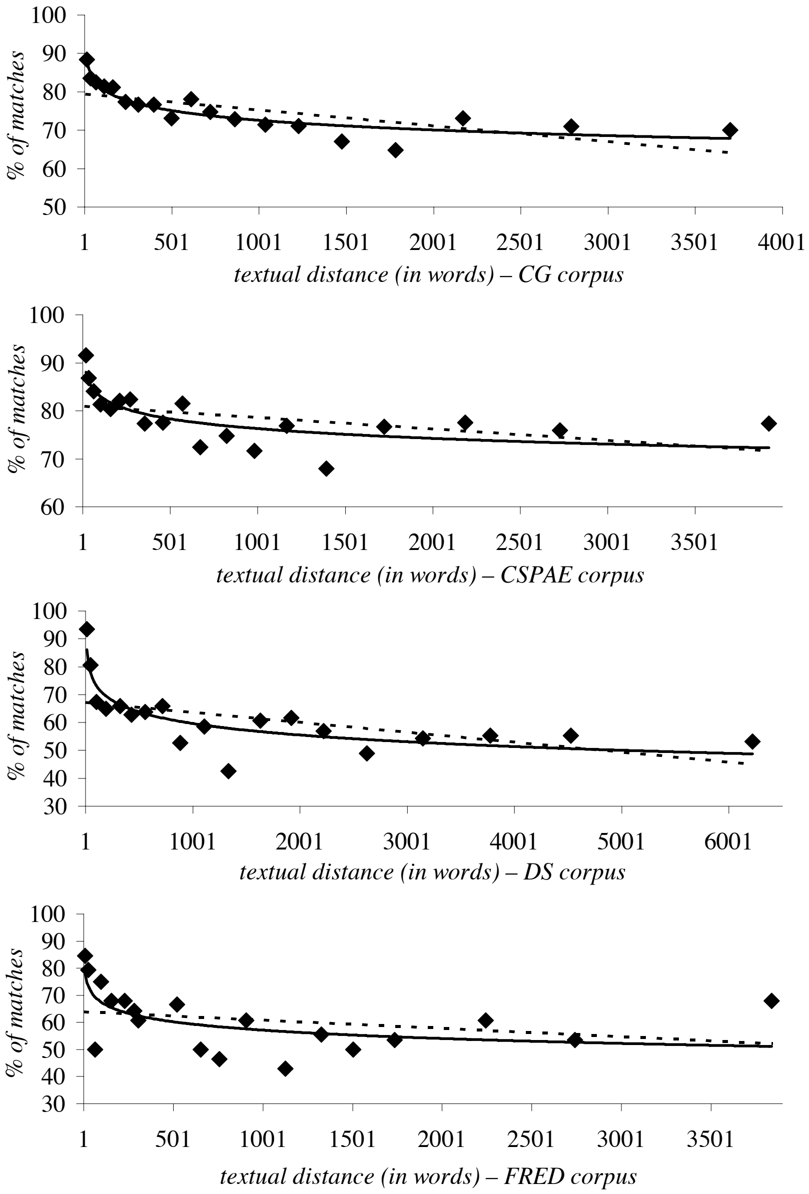


Figure 18. Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same complementation strategy is used) as function of textual distance between CURRENT and PREVIOUS. Heavy line represents logarithmic estimate of the relationship, dotted line represents linear estimate of the relationship

not overly reliable. In FRED, the effect of VMORPHID is at odds with this study's hypothesis: matching morphology here weakens persistence.

What, we may now ask, happens when both the verb lemma *and* the verb morphology of PREVIOUS and CURRENT correspond? This condition is captured in the three-way interaction PREVIOUS $*$ VMORPHID $*$ VLEMMAID. This interaction turns out to be significant in the CG only, yet it is associated with similar exp(*b*) values throughout. For illustration, consider the facts in the CG:

- as a main effect, if *V+inf.* was used in PREVIOUS, the odds for *V+ger.* in CURRENT decrease by 87%;
- if *V+inf.* was used in PREVIOUS and if PREVIOUS's and CURRENT's verb lemma match, the odds for *V+ger.* in CURRENT decrease by 90%;
- if *V+inf.* was used in PREVIOUS and if PREVIOUS's and CURRENT's verb morphology match, the odds for *V+ger.* in CURRENT decrease by 88%;
- if *V+inf.* was used in PREVIOUS and if PREVIOUS's and CURRENT's verb lemma *and* verb morphology match, the odds for *V+ger.* in CURRENT decrease by 93%.

With the situation being similar statistically in the other corpora, the conclusion is that α-persistence is more powerful than otherwise if PREVIOUS's and CURRENT's verb lemma and/or verb morphology are identical.

Persistence in complementation strategy choice appears to only weakly interact with turn-taking. SAMETURN is significant only in the DS: given two successive head verbs with optional complementation, if *V+inf.* is used for the first head verb *and* if the second head verb is in the same conversational turn as the first one, α-persistence is 83% stronger compared to when there was a trade of turns in between. As for speaker change (SAMESPEAKER), no statistically significant findings are obtained. Note though that with exp(*b*) values consistently being smaller than 1, there is some reason to assume that persistence is stronger when PREVIOUS and CURRENT are produced by the same speaker compared to when they are not.

Lexical complexity (TTR) and syntactic complexity (SENTENCELENGTH) also turn out to be only moderately associated with persistence strength. In the DS, increasing syntactic complexity significantly weakens persistence (exp(*b*) = 1.01), and so does increased lexical complexity in the CSPAE (exp(*b*) = 1.12).

4.2.2. β-persistence

We sought to conceptualize β-persistence through three predictors, TEXTDIST-ING, TEXTDIST-INF, and T-ALLIT. TEXTDIST-INF is never significant, hence the hypothesized ability of generic infinitives or of generic occurrences of the token *to* to trigger infinitival complementation cannot be confirmed by this study's analysis.

The idea behind TEXTDIST-ING was that a generic gerund or *-ing* form (as in *I was out* teaching *this afternoon* [DS KBW]) might trigger *V+ger.* after a nearby head verb whose complementation is optional. This hypothesis is indeed borne out by this study's analysis of the CSPAE and DS, where TEXTDIST-ING is associated with statistically significant exp(*b*) values of 0.85 and 0.90, respectively. Hence, for every one-unit increase in the *ln* of textual distance between CURRENT and the last generic gerund or *-ing* form, the odds for *V+ger.* in CURRENT decrease by 10–15%. This is tantamount to saying that the more recent a generic *-ing* form was used, the greater the odds for *V+ger.* Figure 19 visualizes this relationship in the two corpora where the interaction was selected as significant by plotting the share of *V+ger.* against textual distance to the last generic *-ing* form. As can be seen, from the second measuring point onwards, the share of *V+ger.* decreases quite steadily as textual distance to the last *-ing* trigger increases; this is as expected. Why is it, though, that in both graphs, *V+ger.* is so rare at the first measuring point? This is, of course, *horror aequi*: if an *ing*-form has just been used, the likelihood that it will be used again in CURRENT is much lower than otherwise. Thus, to be precise, *-ing* forms trigger *V+ger.* in nearby slots unless they are immediately adjacent to such a slot.

V+inf. necessarily involves the token *to*, which starts in <t>. T-ALLIT measures the number of other tokens in PREVIOUS's preceding phonetic environment which also start in <t>. It turns out that in all corpora except the CSAE, there is a statistically significant tendency for increasing values of T-ALLIT to be negatively correlated with *V+ger.* In a remarkably uniform fashion, T-ALLIT is associated with an exp(*b*) value of between 0.86 and 0.88. This means that for every additional item in PREVIOUS's context that starts in <t>, the odds for *V+ger.* decrease by between 11–14%. Another way of saying this is that when the *V+inf.* option – by virtue of containing the token *to* – is better sound coordinated with its environment, it is preferred over *V+ger.*

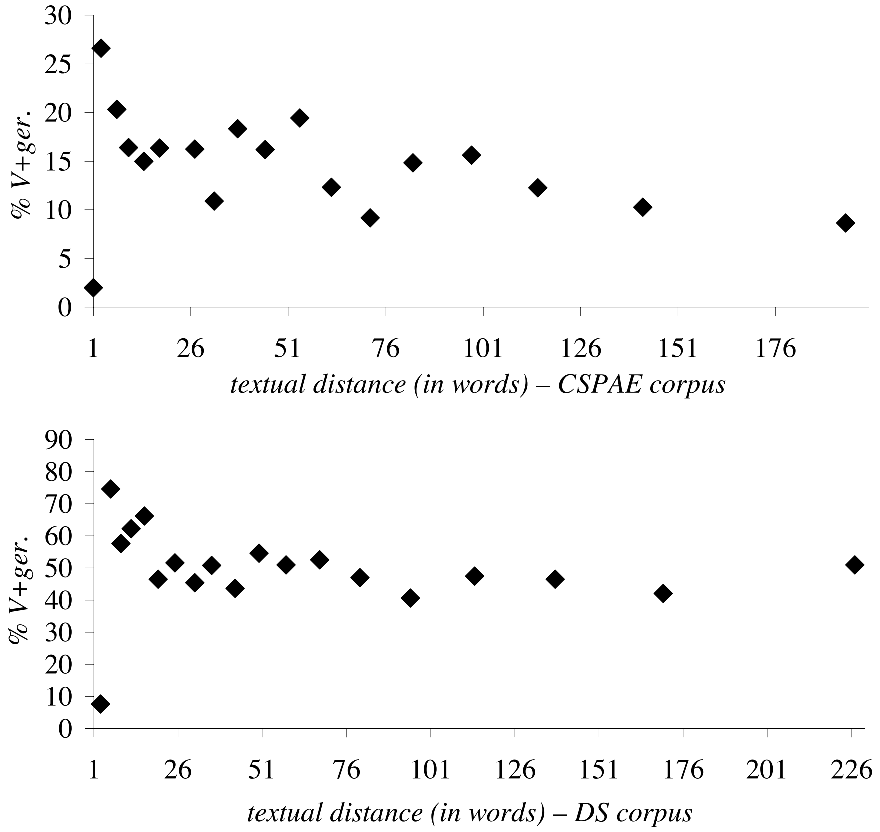


Figure 19. Share of gerundial complementation (on y-axis) as a function of textual distance to the last *-ing* trigger (on x-axis)

4.3. Inter-speaker variation

How are AGE and SEX relevant to persistence in complementation strategy choice? Table 30 presents a model where these two predictors were additionally regressed against CURRENT in the DS database. This step improves model χ^2 significantly (step χ^2 = 12.59, df = 6, $p = 0.05$). Overall, this model accounts for 67% of the observable variation between infinitival and gerundial complementation, which is a gain of 5.5% vis-à-vis a DS model not considering speaker variables (cf. Table 29).

Table 30. Complementation strategy choice: odds ratios associated with speaker predictors in logistic regression in the DS (baseline predictors and persistence-related predictors are included, but not displayed)

AGE	1.01
AGE $*$ PREVIOUS(*V+inf.*)	**1.07** $^+$
AGE $*$ PREVIOUS(*V+inf.*) $*$ TEXTDIST	**0.99** *
SEX(MALE)	1.23
SEX(MALE) $*$ PREVIOUS(*V+inf.*)	0.08
SEX(MALE) $*$ PREVIOUS(*V+inf.*) $*$ TEXTDIST	1.53
model intercept	0.12
N	902
model χ^2	558.72***
R^2	0.669
% correct (baseline)	84.4 (50.6)

$^+$ marginally significant at $p < .10$, * significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$. Predicted odds are for *V+ger.*

The main effect of SEX is not significant, nor is any interaction term including SEX. The main effect of AGE is not significant either, although the variable misses the cut-off level for statistical significance narrowly. However, two interaction terms with AGE were significant or marginally significant:

– AGE $*$ PREVIOUS: the term is associated with a marginally significant exp(b) value of 1.07. This means that for every one-year increase in speakers' age, the effect PREVIOUS has on CURRENT decreases by 7%. In other words, the main effect of α-persistence is weaker in older speakers than it is in younger speakers.

– AGE $*$ PREVIOUS $*$ TEXTDIST: This three-way interaction indicates that for every one-year increase in speakers' age, the multiplicative factor describing how persistence declines with increasing textual distance between PREVIOUS and CURRENT decreases by 1 percent point. This is another way of saying that the effect of textual distance, or recency of use, on persistence is different for older speakers than it is for younger speakers.

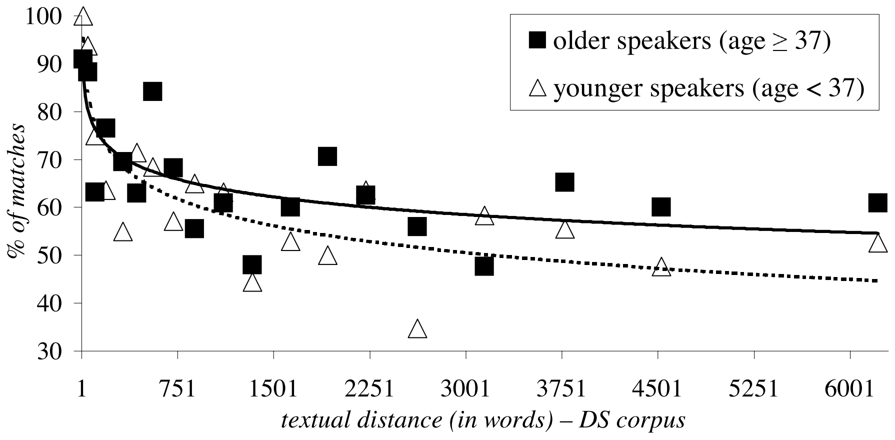


Figure 20. Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same complementation strategy is used) as function of textual distance between CURRENT and PREVIOUS in the DS. Heavy line represents logarithmic estimate of the relationship in older speakers, dotted line represents logarithmic estimate of the relationship in younger speakers

This means that while α-persistence is *per se* weaker in older speakers, this study's analysis, once again, finds that it declines more slowly in old speakers than in young speakers. Figure 20 is an attempt to shed light on this relationship: the graph plots separate forgetting functions for older speakers (i.e. for speakers that are older than 36 years, which is the mean age in the DS dataset on *V+ger.* vs. *V+inf.*), and for younger speakers (i.e. speakers that are younger than 37 years). What can be seen is that initially, younger speakers start off with a 95% match between PREVIOUS and CURRENT when PREVIOUS and CURRENT are textually very close; the corresponding match for older speakers is only ca. 90%. However, the initially higher level of persistence in younger speakers declines faster, while the curve is somewhat more level for older speakers.

No statistically or substantially significant interaction between AGE and any of the β-persistence predictors (TEXTDIST-ING, TEXTDIST-INF, T-ALLIT) could be obtained.

5. Summary

The analysis in this chapter has investigated the alternation between infinitival and gerundial complementation with regard to persistence on the basis of eleven head verbs whose complementation is roughly optional. My core findings are the following.

A model not including persistence factors accounts for 55% of the observable variance between *V+ger.* and *V+inf.* in the three major corpora. In FRED, variance explained is considerably lower, in the CSAE, considerably higher. Such a non-persistence model can be summarized as follows:

1. *V+ger.*, which is arguably the more compact option, is preferred in lexically complex/dense contexts.

2. In the corpora of American English, stative complements generally reduce the odds for *V+ger.* In the corpora of British English, stative complements have this effect after the head verbs *begin* and *start* only.

3. Throughout, *V+ger.* is significantly less likely than *V+inf.* in hypothetical contexts.

4. *Horror aequi* is an important determinant of complementation choice: if the head verb is an infinitive, *V+inf.* is dispreferred, and – even more clearly – if the head verb is an *-ing* form itself, *V+ger.* is dispreferred.

5. The eleven head verbs analyzed differ in their individual preferences for either *V+ger.* or *V+inf.* Moreover, their morphological shape also seems to manipulate the likelihood that either *V+ger.* or *V+inf.* will be used.

Persistence appears to be a major determinant of the alternation between *V+ger.* and *V+inf.* For one thing, I demonstrated that speakers are significantly disinclined to switch between *V+ger.* and *V+inf.* when they can avoid it (Figure 17 presented evidence on this point). At the same time, it appears that *V+ger.* is 'stickier' than *V+inf*. Also, *V+ger.* is stickier in formal, more carefully planned speech than in conversational speech.

In a similar vein, the logistic regression estimates showed that both α-persistence and β-persistence are an integral part of a statistical model seeking to predict complementation choice accurately. A model heeding persistence explains up to 24% more of the observable variance than a model that does not, though I found that persistence is responsible for more variance in the informal corpora than in the more formal corpora.

On the whole, the main effect of α-persistence is such that when a speaker used *V+inf.* last time there was a choice, the odds that he or she will use the other complementation type next time are 99% lower than the odds that he or she will use *V+inf.* again. This overall α-persistence effect is modulated, however, by a number of secondary factors. First, α-persistence weakens with increasing distance between two successive head verb slots; the function that describes this decline appears to be logarithmic. Second, as expected, α-persistence is even stronger when two successive head verbs' lemmas match (cf. Pickering and Branigan 1998 and Gries 2005). We obtained somewhat mixed results with regard to verb morphology identity (which fails to replicate Gries 2005 but is consonant with Pickering and Branigan 1998 insofar as Pickering and Branigan did not obtain evidence that matching morphosyntax enhanced syntactic priming). Third, there is a tendency for α-persistence between two successive head verbs to be weaker when a trade of turns has occurred in between them, or when speaker change has been effected in the meantime (cf. Gries 2005). This is also in accordance with this study's hypotheses, though both of these turn-taking factors are rather weak determinants of α-persistence in complementation strategy choice. Finally, there is some evidence that both increased lexical complexity and syntactic complexity work against α-persistence. This last finding is at odds with this study's initial hypothesis that higher lexical and syntactic complexity would strengthen persistence for functional reasons.

What about β-persistence? The analysis in this chapter was unable to confirm the suspicion that infinitives in general can trigger *V+inf.* complementation. However, we saw that generic gerunds or *-ing* forms can trigger *V+ger.*: the more recently an *-ing*-form was used, the greater – on aggregate – the odds are for *V+ger.* in CURRENT. This statement, however, must be qualified in one important way: if an *-ing*-form has been used extremely recently, *V+ger.* complementation is markedly less likely than otherwise. This is the well-known *horror aequi* effect already discussed above. Furthermore, I showed that the odds for *V+inf.* increase in contexts where a lot of tokens start in alveolar stops, presumably because the infinitive marker *to* is better sound coordinated in and alliterates with such linguistic environments. This finding confirms that both sound coordination (for instance, Sacks 1971) and phoneme perseveration (for instance, Dell 1986) are relevant to complementation strategy choice.

Finally, I was also able to show that speaker age is a determinant of the strength of persistence. On the one hand, I submitted that as speakers' age increases, the level of persistence their speech exhibits decreases, i.e. persistence is, on the whole, weaker in old speakers than in young speakers. On the other hand, however, I showed that persistence declines at a faster rate in younger speakers (despite the fact that persistence starts off from a higher level in younger speakers) than it does in older speakers. In short: speech produced by older speakers is less persistent and less inertial on the whole, but it is persistent over a longer time span.

Chapter 9
Discussion of findings

> The overall picture produced by an analysis that pays attention to all the relevant factors is, admittedly, complex and intricate . . . but it is, I believe, the only kind of analysis that can achieve descriptive adequacy and explanatory power. It is language itself that is immensely complex. (Wierzbicka 1998: 151)

This chapter is an attempt at generalization. By adopting a bird's eye view on the findings reported in the five preceding empirical chapters, it will pick out and discuss salient patterns relating to persistence. Issues that will be addressed include the following: How, and how much, does persistence contribute to linguistic variation (section 1)? What are the parameters of α-persistence (section 2)? How does β-persistence show in the data (section 3)? Do speaker characteristics interact with persistence (section 4), and – in a similar vein – are register and regional differences relevant (section 5)? Methodologically, this chapter's focus on findings that have turned out to be statistically significant in the empirical chapters of this study.

1. The contribution of persistence to linguistic variation

As pointed out in the Introduction, this study's main motivation for investigating persistence was to examine, in the spirit of Labov (1969), how much persistence helps us explain – that is, understand – linguistic variation. Figure 21 is a somewhat impressionistic attempt to provide a partial answer to this question in one single graph, which plots average share of variance explained (for an explanation of this term, see Table 1 [p. 58]) by core persistence predictors across the alternations analyzed in the present study.[60] The bigger this share, the more powerful is persistence (more precisely, α-persistence) in the respective alternation.[61]

With respect to the percentages in Figure 21, the five alternations investigated fall into three groups. The share of variance accounted for by

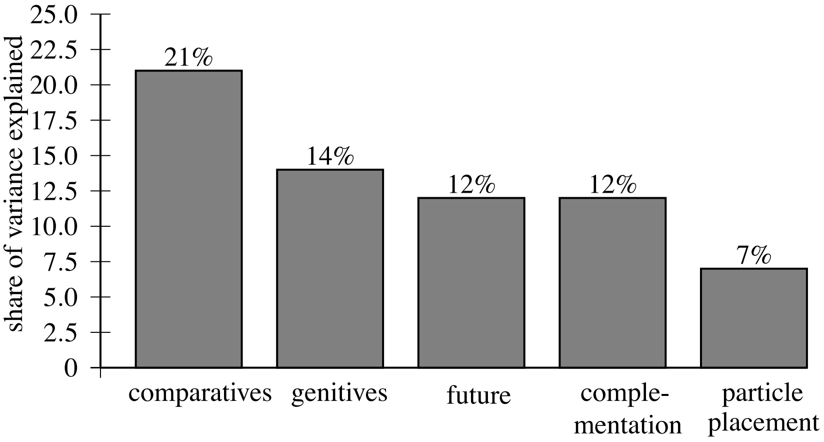


Figure 21. How much variance is caused by persistence? Nagelkerke R^2 (variance explained) in persistence-only models predicting CURRENT on the basis of PREVIOUS, TEXTDIST, and TEXTDIST ∗ PREVIOUS across alternations. Figures are averages across all corpora studied for each individual alternation

PREVIOUS, TEXTDIST, and TEXTDIST ∗ PREVIOUS is most substantial (approximately 21%) in comparison strategy choice; it is smallest in particle placement (7%). Genitive choice, future marker choice, and complementation strategy choice cover quantitatively the middle ground (12–14%). I would like to submit that this ordering is not entirely accidental. Comparison strategy choice is an alternation where a good deal of lexical and morphological material – primarily *-er* and *more* – is subject to repetition. By contrast, there is no substance to be entrenched in particle placement, which is a purely positional alternation. This reading of Figure 21 dovetails nicely with the fact that even in particle placement and complementation strategy choice, persistence effects are strengthened when there is a verb lemma match (cf. below, section 2.3), and hence, when substance is involved. Given this line of reasoning, it is indeed a bit surprising that future marker choice is, according to Figure 21, only moderately affected by persistence: this is, after all, the alternation where most lexical and morphological material is subject to repetition. In psycholinguistic terms, it is the alternation where syntactic priming is most indistinguishable from lexical priming. Still, there seems to be a relationship such that as the extent of linguistic substance which is subject to persistence increases, the effect size of persistence observable in corpus data increases as well.

At this point, it will be instructive to put the present study's findings into perspective quantitatively with psycholinguistic investigations of priming. Recall that Bock (1986) – which is the seminal study on syntactic priming (see chapter 2, section 1.4.2 for a review) – examined to what extent prepositional datives, as in (1a), or double-object datives, as in (1b), are subject to syntactic priming.

(1) a. *A rock star sold some cocaine to an undercover agent*
b. *A rock star sold an undercover agent some cocaine* (Bock 1986: 361)

Bock (1986: Table 1) found, among other things, that prepositional primes or double objects primes increased the probability of corresponding targets by between 22–23%. These probabilities can be translated into an odds ratio of approximately 0.31, such that if the prime was a prepositional dative, the odds that subjects would produce a double-object dative decreased by 69%. Apart from the fact that it is problematic to compare experimental findings to corpus findings, this odds ratio is not directly comparable to the odds ratios obtained in the course of the present study for methodological reasons. What is crucial is that Bock's design did not take into account factors such as syntactic or lexical complexity, so it would be distorting to compare estimates including these factors to the effect obtained by Bock (1986). This is why the logistic regression models presented earlier were, once again, reestimated under inclusion of (i) all of the baseline predictors, which, much like Bock's set up, control for persistence-unrelated intralinguistic factors and (ii) PREVIOUS (the 'prime' – the 'target', in psycholinguistic parlance, would be CURRENT) as the only persistence-related predictor.

The resulting recalculated odds ratios associated with PREVIOUS are displayed in Table 31. As can be seen, except for comparison strategy choice, the odds ratios obtained through corpus study are, on average, somewhat larger than Bock's odds ratio of 0.31. This means that on aggregate, Bock obtained a more sizable (though really not a dramatically more sizable) effect than did the present study. It is worth noticing that the two alternations in the present study which are most similar (in terms of, e.g. the amount of morpho-lexical material repeated) to the alternation between double-object and prepositional

Table 31. Odds ratios associated with PREVIOUS in logistic regression. Underlying models were estimated using baseline predictors and PREVIOUS as the only persistence-related predictor

	CG	CSPAE	DS	CSAE	FRED	**avg.**
comparison	0.20***	0.25*	0.35*	–	0.36*	**0.29**
genitive	–	–	–	0.24***	0.51***	$\boxed{\mathbf{0.38}}$
future	0.30***	0.32***	0.40***	0.33***	0.43***	**0.36**
part. placement	–	–	–	(n.s.)	0.49***	$\boxed{\mathbf{0.49}}$
complementation	0.38***	0.47***	0.46***	(n.s.)	0.46***	**0.44**

* significant at $p < .05$, ** significant at $p < .01$, *** significant at $p < .005$.

datives are probably particle placement and genitive choice. The average odds ratios of 0.49 and 0.38, respectively, associated with these alternations clearly exceed Bock's 0.31 value. What is the significance of this discrepancy? That the present study obtained a smaller effect than Bock (1986) should, as a matter of fact, not be too surprising: corpus study cannot control for various intralinguistic and extralinguistic factors the way a carefully set-up experiment can (as might be recalled, this is one of the reasons the present study never pretended to research priming). Given this inherent disadvantage of corpus study in comparison to experimental research designs, the fact that Bock's (1986) priming effect is not dramatically bigger than this study's persistence effect is, in fact, reassuring.

2. The parameters of α-persistence

Let us now pull together the core findings of this study with regard to α-persistence, i.e. the tendency for two successive choice contexts to influence each other. For one thing, the discussion of switch rates demonstrated that in every alternation studied, speakers switch markedly less between two options than pure chance would predict. We obtained the highest switch rate for switches from WILL to BE GOING TO in FRED, where speakers switch approximately 55% of the time they would if their choices were governed by chance. The overall lowest switch rate obtained for switches from the *of*-genitive to the *s*-genitive in the CSAE, where speakers only switch approximately 1% of the time they would if their behavior were governed by chance alone. On aggregate, switch rates were highest in particle placement, and lowest in genitive choice.

Second, in logistic regression, α-persistence accounted for between 1.8% (CSPAE, comparison strategy choice) and 20% (CSAE, genitive choice) of the observable variation. The key predictor utilized to tap the main effect of α-persistence (i.e. the effect of α-persistence when secondary factors, such as textual distance, are controlled for) was PREVIOUS. In all, 18 logistic regression runs[62] including this predictor were conducted, and in 10 of these the predictor was significant. Where significant, the exp(*b*) value associated with PREVIOUS ranged between 0 (FRED, genitive choice) and 0.05 (CSAE, future marker choice). This means that given two successive choice contexts, if option A was used in the first one, the odds that speakers would switch to option B in the second one were between 95–100% lower than the odds that speakers would stick to option A. The effect of PREVIOUS was generally a bit weaker in FRED than in the other corpora.

However this may be, this study's analyses have recurrently demonstrated that the main effect of α-persistence, as conceptualized through PREVIOUS, interacts with several secondary factors. These will be discussed below.

2.1. Textual distance

We had noted in chapter 2 that the persistence of syntactic priming is a rather controversial topic in the current psycholinguistic literature. Can the present study shed some light on this issue? Textual distance between two successive choice context slots turned out to be an important – perhaps the most important – determinant of α-persistence. In 8 out of 18 logistic regression runs, the predictor TEXTDIST interacted significantly with PREVIOUS such that as textual distance between two slots increased, the magnitude of α-persistence decreased. In other words, α-persistence turned out to be stronger when PREVIOUS and CURRENT were textually close than when they were textually distant.

Given the psycholinguistic literature (for instance, Cohen and Dehaene 1998; McKone 1995; Gries 2005), we had assumed *a priori* that persistence would decline comparatively faster immediately after a choice has been made than when the choice was made a longer time ago. Thus, in logistic regression, TEXTDIST was modeled logarithmically, i.e. such that very small textual distances between two slots would have more weight than larger distances.

The subsequent analyses provided independent evidence that this assumption was justified: even in those datasets where the interaction between PREVIOUS and TEXTDIST was not selected as significant in logistic regression (due to low *N*s, for instance), a logarithmic forgetting function described the decline of persistence best. This is strong evidence for the important role priming plays in persistence: whatever the contribution of discourse-functional factors to persistence, this kind of logarithmic decline cannot be explained by functional factors. It must be due to the design of the human speech processing and production system.

The logarithmic nature of the decline of α-persistence begs the question how long-lived α-persistence actually is. There are different ways to answer this question on the basis of naturalistic data. The one that will be used here relies on the forgetting functions that have been presented throughout this study. For instance, the way α-persistence in future marker choice declines in the DS corpus can be described mathematically by equation 9.1, which is the formal statement of the logarithmic estimate presented in Figure 13 (p. 122):

$$p = -6.82 \times ln(x) + 93.11 \tag{9.1}$$

where p is the probability (in percent) that the same future marker is employed in two successive slots, and x is the textual distance (in words) between these two slots. p can be straightforwardly interpreted to be indicative of the strength of α-persistence: the larger p, the stronger α-persistence.

Observe now that there exists a 'natural' probability P (in percent) that the options (in this case, future markers) employed in two successive slots match. This probability follows from the relative frequency of two binary options in a given dataset and would obtain if there were no persistence effects. P can be approximatively stated as 9.2:

$$P = \left[\left(\frac{a}{N}\right)^2 + \left(\frac{b}{N}\right)^2 \right] \times 100 \tag{9.2}$$

where N is the total number of relevant slots in the dataset, a is the number of slots where option A has been employed, and b is the number of slots in which option B has been employed. 9.2 can be derived as follows: the probability that option A is chosen in one single trial is a/N; the probability that option A is chosen in two successive trials is $a/N \times a/N$, hence $(a/N)^2$; the corresponding probability for option B is $(b/N)^2$. The accumulated probability, then, that either option A *or* option B is chosen in two successive trials is $({}^{a}\!/_{N})^2 + ({}^{b}\!/_{N})^2$. By multiplying this probability by 100, we obtain the corresponding probability in percent terms. To illustrate: according to Table 14, the DS dataset on future marker choice analyzed in chapter 6 is based on *N* = 39,640 relevant slots, in 11,223 of which a BE GOING TO marker has been employed (*a*), and in 28,417 of which a WILL marker has been employed (*b*). Therefore, equation 9.3 yields *P* in the DS dataset on future marker choice:

$$P = \left[\left(\frac{11,223}{39,640} \right)^2 + \left(\frac{28,417}{39,640} \right)^2 \right] \times 100 \approx 59.4 \qquad (9.3)$$

Hence, the 'natural' probability that one will get the same future marker in two successive slots in the DS is 59.4%. I now define that persistence is neutralized at a given textual threshold z when the corresponding forgetting function returns the 'natural' match probability *P*. Thus, taking the DS dataset on future marker choice as an example again,

$$p = -6.82 \times ln(z) + 93.11 \overset{!}{=} P \overset{!}{=} 59.4 \qquad (9.4)$$

Solving for z leaves us with $z \approx 140$. Hence, 140 words between PREVIOUS and CURRENT is the textual threshold in this dataset after which persistence is completely neutralized because the probability that there is a future marker match ceases to be greater than it would be if there were no persistence effect. Given an average speech rate of 120 words per minute (Biber et al. 1999: 27), this is equivalent to a bit more than 1 minute of talk. This train of thought is visualized in Figure 22, which plots the forgetting function in future marker choice in the DS (heavy line) and the 'natural' probability of a match in two successive slots (dotted horizontal line). The threshold where α-persistence is neutralized is where the two lines intersect ($z = 140$).

Analogous forgetting functions, 'natural' match probabilities, and threshold levels for all those datasets which have been analyzed in the course of the present study and for which *N*s are sufficiently large[63] are displayed in Table 32. There are some noteworthy differences between corpora and alternations with respect to the longevity of persistence. Persistence is most short-lived in the DS dataset on future marker choice (Table 32c), where it disappears entirely after a bit more than 1 minute. It is most long-lived in the CG dataset on comparison strategy choice (Table 32e),[64] where persistence does not entirely dissipate until 110 minutes after a choice has been made. The other alternations cover the middle ground. It is quite remarkable that

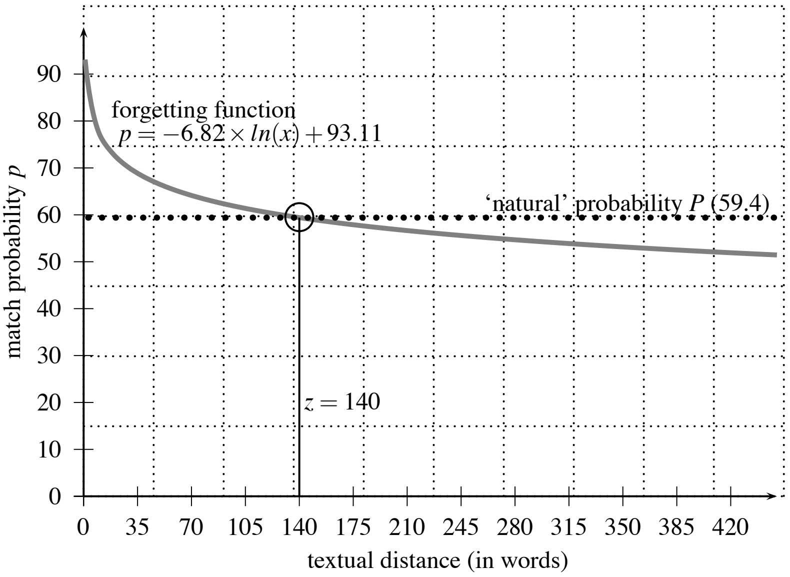


Figure 22. The longevity of α-persistence in future marker choice in the DS: forgetting function, 'natural' match probability, and textual threshold z

persistence in future marker choice is the most short-lived, for it means that in future marker choice, persistence is most potent initially (cf. also section 1), but apparently evaporates quickly. It is also noteworthy that Table 32 exhibits a slight tendency (comparison strategy choice is an exception here) for persistence to decline more quickly in datasets containing informal registers than in datasets sampling formal speech.

The thresholds in Table 32 notwithstanding, it should be kept in mind that due to the logarithmic nature of the forgetting functions, most of the effect of α-persistence declines in a comparatively brief interval just after a choice has been made. How brief, we may now ask, is 'brief'? One possible answer to this question is to define that most α-persistence has declined when the forgetting function becomes more level than it is steep, i.e. when its mathematical derivative becomes smaller than -1 (this means that a tangent on the forgetting function would have an angle to the abscissa of less than 45°). Conveniently, the textual thresholds are equivalent to the coefficients of the $ln(x)$ terms in the forgetting functions in Table 32: thus, in compari-

Table 32. The longevity of α-persistence: forgetting functions across corpora

corpus	forgetting function	'natural' match	threshold	
		probability P	*words*	*minutes*
	a. comparison strategy choice			
CG	$-6.28 \times ln(x) + 110.3$	50.6	13,000	110
DS	$-6.02 \times ln(x) + 116.0$	68.0	3,000	24
	b. genitive choice			
FRED	$-5.78 \times ln(x) + 96.5$	51.8	2,300	19
CSAE	$-2.29 \times ln(x) + 77.3$	50.0	∞	∞
	c. future marker choice			
CG	$-5.79 \times ln(x) + 94.57$	61.1	300	3
CSPAE	$-6.38 \times ln(x) + 95.17$	58.8	300	2 ½
DS	$-6.82 \times ln(x) + 93.11$	59.4	150	1
CSAE	$-8.03 \times ln(x) + 93.45$	51.2	200	1 ½
FRED	$-3.35 \times ln(x) + 91.85$	71.0	500	4 ½
	d. particle placement			
FRED	$-2.99 \times ln(x) + 90.62$	63.4	9,000	75
CSAE	$-2.60 \times ln(x) + 73.03$	55.0	1,000	9
	e. complementation strategy choice			
CG	$-3.64 \times ln(x) + 97.70$	65.8	6,400	53
CSPAE	$-2.93 \times ln(x) + 96.57$	75.9	1,200	10
DS	$-5.97 \times ln(x) + 100.89$	50.0	5,000	42
FRED	$-4.46 \times ln(x) + 87.96$	50.3	4,600	29

NOTE: the time estimates assume a speech rate of 120 words per minute (cf. Biber et al. 1999: 27).

son choice, most α-persistence will have declined after about six words after PREVIOUS, in particle placement after about three words, etc. In all, it can be seen that no matter which grammatical alternation is examined, most of the α-persistence effect declines within an interval of 10 words after PREVIOUS. This is equivalent to ca. 5 seconds of speech.

Some readers might wonder whether persistence isn't either incredibly long-lived, given some of the thresholds in Table 32, or, alternately, incredibly short-lived, given the second criterion that we have just established. As a matter of fact, from a psycholinguistic perspective, these thresholds are not entirely implausible (discourse analysts have not explicitly addressed the issue, but they appear to view repetitiveness as a rather local phenomenon). The discussion in chapter 2, section 1.4.2 (p. 17) has demonstrated that the time course of (syntactic) priming is a somewhat confused issue in the psycholinguistic literature – findings range from extremely short durations (e.g. Branigan, Pickering, and Cleland 1999) to a longevity of over *a week* in aphasic patients (Saffran and Martin 1997). As a matter of fact, it has been suggested that experimental research may actually be ill-equipped to settle this issue (cf. Bock and Griffin 2000: 179: "with existing [i.e., experimental/laboratory, BS] data . . . it is impossible to assess the normal time course of priming under carefully controlled conditions"). Given these claims, the corpus findings presented here suggest the following: persistence declines significantly after a relatively short period after a choice has been made (10 words, or 5 seconds of talk), but on the whole it is fairly long-lived before the effect dissipates entirely.

2.2. Turn-taking

Logistic regression showed that both turn-taking and speaker change interact with α-persistence. More specifically, according to the data, both of these discourse mechanisms appear to *weaken* persistence.

Turn trading (SAMETURN) was significant determinant of persistence in 5 out of 14 regression runs.[65] With the exception of the CSAE dataset on genitive choice, α-persistence was consistently stronger when two successive choice context slots were located in the same conversational turn. More precisely, given two successive slots and given (i) that option A was used in the first slot and (ii) that both slots were located in the same turn, the odds that a speaker would switch to option B were between 83% (DS, complementation strategy choice) and 33% (DS, future marker choice) lower than when both slots were located in different conversational turns. Interestingly, in the CSAE dataset on genitive choice, a trade of turns seemed to actually strengthen persistence; on a rather speculative note, this finding might be due to the strongly conversational nature of the data sampled in the CSAE, which possibly renders dialogue in the corpus especially amenable to inter-speaker involvement and allo-repetition.

Speaker change necessarily implies trading turns, but two successive slots may be produced by the same speaker with a trade of turns having, or having not, occurred in the meantime. On these grounds, this study has distinguished analytically between SAMETURN and SAMESPEAKER. Speaker change, too, clearly weakens persistence. SAMESPEAKER was selected as significant in 4 out of 14 regression runs. In summary, given two successive slots and given (i) that option A was used in the first slot and (ii) that both slots are produced by the same speaker, the odds that the speaker would switch to option B were between 83% (DS, complementation strategy choice) and 17% (CG, future marker choice) lower than when the two slots, though successive in discourse, were produced by different speakers.

In most general terms, then, this is robust evidence that both cross-speaker persistence and same-speaker persistence can be empirically observed. My results more specifically indicate that persistence is weaker across turns than within turns, and that comprehension-to-production priming is weaker than production-to-production priming, which means – in short – that speakers prefer repeating themselves over repeating what others have said. This finding is consonant with previous research (for instance, Gries 2005). It is also worth pointing out that speaker change and turn-taking cannot be epiphenomenal in that they are really other ways of measuring textual distance between two slots (for instance, it might be argued that if two subsequent variables are located in different turns, they will be textually less close than two variables in the same turn): consider that more often than not, logistic regression selected textual distance *in addition* to SAMETURN and SAMESPEAKER as statistically significant. Thus, statistically speaking, SAMETURN and SAMESPEAKER have an effect on the strength of persistence over and above the effect of textual distance. Therefore, the empirical significance of SAMETURN and SAMESPEAKER is evidence that discourse factors do interfere with persistence.

2.3. Matching lemmas and matching morphology

Pickering and Branigan (1998) and Gries (2005) have claimed that syntactic priming is stronger if the priming verb lemma and the target verb lemma is identical. Thus, a corresponding variable (VLEMMAID) was included in the logistic regression estimates on particle placement and complementation strategy choice. VLEMMAID checked whether PREVIOUS and CURRENT involve the same verb lemma. In 2 out of 7 relevant regression runs, VLEMMAID was significant. In the CSAE database on particle placement, the odds for a particle placement strategy switch between two successive phrasal verb slots were 93% lower if the two slots had been filled by the same phrasal verb. The corresponding figure for the CSPAE database on complementation strategy choice is still a considerable 42%. Thus, the present study of persistence replicates the findings of both Pickering and Branigan (1998) and Gries (2005): matching verb lemmas provide an extra stimulus to reuse the linguistic strategy that had been used before.

In analogy to matching verb lemmas, we also hypothesized that matching verb morphologies would strengthen α-persistence (cf. Gries 2005) and tested this study's databases on complementation strategy choice to that effect (predictor VMORPHID). The results were mixed: in the CSAE database on complementation strategy choice, matching morphology indeed significantly strengthened persistence; in the FRED database though, it significantly weakened persistence. Finally, it is worth mentioning that there was a significant three-way interaction in the CG dataset on complementation strategy choice: if both the head verb lemma *and* morphology match between two successive complementation slots, α-persistence was substantially stronger than it would have been otherwise.

2.4. Syntactic and lexical complexity

One of this study's initial working hypotheses was that increased syntactic complexity (operationalized as sentence length, in words [SENTENCELENGTH]) and increased lexical complexity (operationalized as the type-token ratio of the lexical context where CURRENT is embedded [TTR]) would intensify persistence. The idea was that in cognitively complex environments, speakers would economize by functionally exploiting the pay-offs that persistence affords according to previous scholarship: persistence provides for planning time (e.g. Tannen 1987), increases fluency (e.g. Levelt and Kelter 1982), and can, by virtue of the redundancy that it is associated with, reduce processing load (e.g. Tannen 1987; Branigan, Pickering, and Cleland 2000).

Indeed, SENTENCELENGTH was selected as a significant moderator variable of PREVIOUS in 2, TTR in 6 out of 18 regression runs. However, whenever these terms turned out to be significant, their effect was precisely contrary to this study's hypothesis: according to the data, both increased lexical and syntactic complexity have a *weakening* effect on persistence. For every one-word increase in length of the sentence where a choice context slot is embedded, the effect of α-persistence between that slot and the last choice context slot weakens by between 1% (DS, complementation strategy choice) and 2% (FRED, particle placement). Correspondingly, for every one unit-increase in type-token ratio, the effect size of α-persistence is reduced by between 1% (FRED, particle placement) and 12% (CSPAE, complementation strategy choice).[66] How come increased lexical or syntactic complexity has this unanticipated effect on persistence? With hindsight, I submit that increased lexical and syntactic complexity might be indicative not only of increased lexical or syntactic complexity, but of better monitored and better planned speech as well. It is a well-known fact that speakers parse their inner speech and inspect their speech programs prior to articulation (Postma 2000: 105; Levelt 1983: 96). Crucially, monitoring depends "on the level of formality required by the context of discourse" (Levelt 1989: 461), which means that "contextual factors determine which aspects of speech will be given most scrutiny by the speaker" (Levelt 1989: 463). My claim, then, is that speakers try to deal with the increased computational load that comes with a syntactically/lexically complex environment by increasingly planning and monitoring their production. Apparently, this side-effect of increased complexity is more dominant than the potential other effect – that is, trying to deal with the increased cognitive load by exploiting the pay-offs of persistence. If this argument is correct, it is actually to be expected that better monitoring and planning have a weakening effect on persistence, a phenomenon which is at least partly a matter of the subconscious and whose surface manifestation (repetitiveness) is frowned upon by most prescriptivist traditions. Once again, it should be clarified that increased sentence length cannot be a mere epiphenomenon of increased textual distance between two variables – note that the predictor TEXTDIST always controlled for textual distance in regression, over and above the effect of increased sentence length.

3. β-persistence

β-persistence is the tendency of speakers to use a given linguistic option in a choice context when they have recently produced or were recently exposed to some not necessarily variable linguistic pattern that shares one or more lexical, morphological, phonological, or structural characteristics with a variable option. To recapitulate, the difference between α-persistence and β-persistence is that if both the 'prime' and the 'target', in psycholinguistic parlance, are variants of the same linguistic variable, we are dealing with α-persistence. If they are not, the relationship between prime and target by definition falls under the scope of β-persistence (cf. chapter 1, section 1.1). The distinction between α-persistence and β-persistence is methodologically indispensable for operating the variationist machinery that has been utilized throughout this study.

This study has sought to tap β-persistence through consideration of two major factor groups: triggers, which may be lexical or morphosyntactic in nature, and sound coordination variables, i.e. variables checking whether a specific option would be better integrated phonologically into its environment.

Two research questions guided this study's investigation of β-persistence in comparison strategy choice: first, can the presence of items ending in *-er* trigger synthetic comparison (which would also affix an adjective with *-er*) in a choice context nearby? And second, can the presence of the lexical item *more* trigger analytic comparison in a close-by choice context? While we found no proof that *-er* can trigger synthetic comparison, the analysis uncovered significant evidence that the token *more* can trigger analytic comparison: for instance, in the DS database, if *more* was used up to 25 words prior to CURRENT, the odds for analytic comparison increase by 84%; if *more* was used up to 5 words prior to CURRENT, the odds for analytic comparison increase by even 98%.

In genitive choice, it turned out that the more recently the token *of* was used, the greater the odds for the *of*-genitive. In FRED, for every one-unit decrease in the *ln* of textual distance between a choice context and an occurrence of *of*, the odds for the *of*-genitive increase by 13%.

As for future marker choice, I tested what effect a nearby occurrence of the verb *go* has on choice contexts. On the whole, the verb *go* helps trigger BE GOING TO in a close-by future marker slot. In the CG, for example, if a form of the verb *go* was used up to 25 words prior to CURRENT, the odds for WILL decrease by 46%.

We tested for two purely syntactic or positional triggers in this study's analysis of particle placement: textual distance to the last generic pattern where a particle or preposition either preceded a direct object (e.g. *I look* *at the waiter*), or did not (e.g. *I brought it out*). The hypothesis was that these patterns would trigger the corresponding separated or non-separated particle placement in transitive phrasal verbs. In summary, a generic non-separated pattern can indeed trigger *V+Part+NP* in a nearby choice context (significantly so in the CSAE), but there was no evidence that a generic separated pattern can trigger *V+NP+Part*.

As for complementation strategy choice, this study investigated whether a generic *-ing* form makes *V+ger.* more likely, and, correspondingly, whether a generic infinitive increases the odds for *V+inf.* While (somewhat surprisingly) it did not appear that infinitives can trigger *V+inf.*, I showed that *-ing* forms seem to be able to trigger *V+ger.* In the CSPAE, for instance, for every one-unit decrease in the *ln* of textual distance between a generic *-ing* form and a choice context, the odds for *V+ger.* increase by 15%; the corresponding figure in the DS is 10%.

What impact does sound coordination have on speakers' linguistic choices? For one thing, I tested whether in future marker choice, an increasing number of words that start in <w> in a future marker slot's immediate environment favor usage of WILL in that slot. The answer is yes: for every additional context word that starts in <w>, the odds for WILL increase by 2% (DS), 5% (CSAE), or even 6% (CSPAE). There was no corresponding evidence that an increased number of words starting in <g> can trigger BE GOING TO. Second, in complementation strategy choice, I checked whether an increased number of words starting in <t> in a slots's contextual environment would make the *V+inf.* option more likely (this would be the case by virtue of the infinitive marker *to*, which would then be better sound coordinated with its context). In four of the five corpora studied, this hypothesis was borne out: for every additional word in a slot's context starting in <t>, the odds for *V+inf.* increase by between 14% (CG) and 11% (FRED). My findings support discourse-analytic claims that speakers prefer options which are sound coordinated with their neighborhood (cf. Sacks 1971; Tannen 1989) and psycholinguistic assertions that the human speech production system has a tendency for phoneme perseveration (cf. Dell 1986; Cohen and Dehaene 1998).

In sum, there is a broad variety of ways (possibly an infinite variety) how *β*-persistence can interfere with linguistic choices that speakers make, and I have no intention of conveying the impression that I have researched and discussed these ways exhaustively. Of the two factor groups, triggers and sound coordination, triggers however appear to be empirically more important. Still, some readers might wonder whether β-persistence does not go against some previous psycholinguistic findings, particularly against Bock and Loebell (1990). Bock and Loebell (1990: Experiment 3) found that the infinitive phrase in *Susan brought a book to study* did not appreciably prime the prepositional phrase in *Susan brought a book to Stella* when compared to a double-object control such as *Susan brought the student a book*. It is true that certain kinds of β-persistence are indeed not predicted, given this particular experiment. Yet, it is important to bear in mind that the present study is a study of naturalistic corpus data, not a carefully controlled experiment. In summary, the crux of the matter appears to be the following: β-persistence is a rather robust effect that is observable in corpus data, and its existence calls for further – possibly experimental – research to elucidate its exact nature.

4. Inter-speaker variation

Except for particle placement and genitive choice (alternations for which the available database was too small for reliable statistical analysis), we also investigated how persistence effects differ among different groups of speakers. The predictors that were analyzed were AGE and SEX. In logistic regression, inclusion of these predictors often increased predictive efficiency and explanatory power.

4.1. Age

My core findings with respect to how AGE interacts with persistence can be summarized as follows:

- In comparison strategy choice, increasing age weakens the influence of triggers (β-persistence) on any given choice context by 9% for every one-year increase in age.
- In future marker choice and especially complementation strategy choice, there is a tendency for α-persistence to weaken with increasing age. For instance, in complementation strategy choice, we observed that for every one-year increase in the speaker's age, the effect PREVIOUS has on CURRENT decreases by 6%.

- In comparison strategy choice, future marker choice, and complementation strategy choice, the forgetting function that describes the decline of α-persistence is more level and less elastic in older speakers than in younger speakers.

- More often than not, FRED exhibited the lowest level of persistence (for an overview, consider, for instance, Table 31). It is very likely that this has less to do with the data sampled (dialect speech), but rather with the much higher mean age of speakers in the corpus.

On aggregate, age seems to have a weakening effect on persistence. There is a twist to this statement, however, which boils down to this: in younger speakers, α-persistence tends to be stronger when textual distance between two slots is small, but it declines faster than in older speakers. In turn, the speech produced by older speakers is *per se* less persistent, but it is persistent at this lower level over a longer time span.

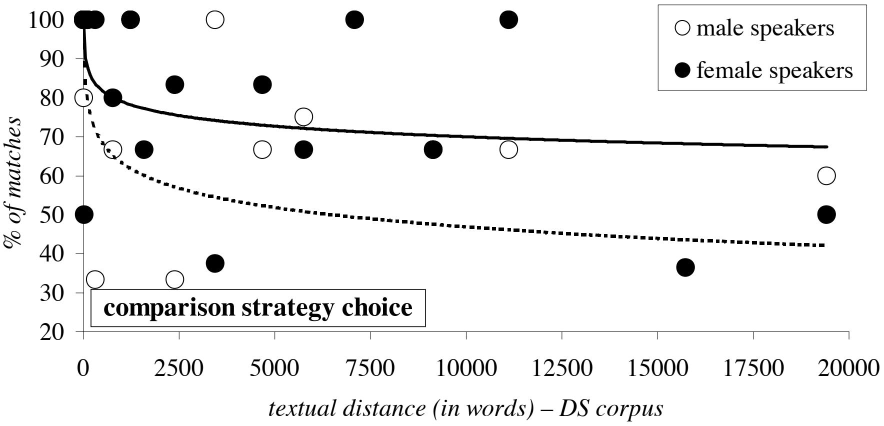
Now, do these findings make sense given what is known about the effects of age on speech production, especially priming? The psycholinguistic literature offers somewhat conflicting views: according to Rastle and Burke (1996: 586), repetition priming shows little variation with age, but Laver and Burke (1993) found that elderly adults produce overall bigger semantic priming effects than younger adults. Likewise, Friederici, Schriefers, and Lindenberger (1998) – a study of syntactic priming, type I – observed larger priming effects for elderly adults than for young adults. The empirical majority view seems to be that if anything, priming effects are *stronger* in older adults. It is fair to say that my results are at odds with this view.

Recall, however, that according to the data, older speakers show reduced persistence effects initially, but that in the long run, persistence declines more slowly than in younger speakers. That persistence starts off from a lower level in older speakers might have to do with memory limitations, which have been implicated in numerous studies of elderly adults' speech processing (for instance, Zurif et al. 1995; see Kemper 1992: 222–225 for a review). The reason for the slower, more inertial decline of persistence in older speakers, in turn, may be due to an amalgam of factors. For one thing, elderly adults have an overall lower propensity for implicit learning (cf. Bock and Griffin 2000), which may imply that their speech is less sensitive to contextual influences. On the other hand, it is known that elderly adults have a reduced ability to inhibit irrelevant information (Hasher and Zacks 1988) – and arguably, previous linguistic choices become increasingly irrelevant as time passes by. Younger speakers may be better able to delete data about previous choices from their memory after a certain time span. Finally, and maybe most generally, old age slows down reaction times and processing rates (for instance, Hertzog 1991). Thus, the decline of persistence, and the adaptation of speech to new contextual environments, may simply be slowed down in older speakers. The present study cannot settle this issue conclusively; future research may want to further investigate the question of how age impacts persistence or priming.

In sum, the clear empirical impact of age on persistence is further proof that persistence is to a high degree due to properties of the human speech production system, which is subject to change over a lifetime. At any rate, no interaction between age and persistence would follow from functional accounts of repetitiveness.

4.2. Sex

The variable SEX did not make a strong showing in this study's analyses. Among the only noteworthy effects is that in comparison strategy choice, there is a tendency for β-persistence to be more influential in female than in male speakers, though this relationship is exactly the reverse in future marker choice. Could it be that the difference SEX may make is too subtle to be detected in logistic regression (for instance, because of too low *N*s)? To look into that question, I conducted some further analyses. Figure 23 plots different forgetting functions for male and female speakers in the DS for comparison strategy choice and complementation strategy choice (in future marker choice, there was no difference whatsoever between male and female forgetting functions). According to Figure 23, the forgetting functions that describe the decline of persistence in female speakers are more level than the corresponding functions for male speakers. Hence, on a very speculative note (without having much hard evidence except for Figure 23), it might be the case that persistence declines differently in female speakers than in male speakers. As far as I know, in the literature on priming no such claim is on record, though more generally it is known that psycholinguistically, there are some differences between the sexes (for instance, men are slightly more disfluent than women (cf. Bortfeld et al. 2001). Be that as it may, SEX was certainly one of the weakest predictors included in this study's investigation.



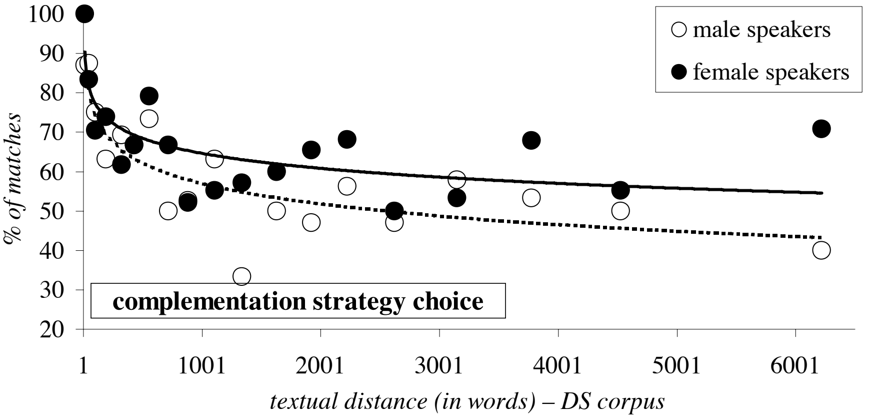


Figure 23. Percentage of persistent pairs (i.e. PREVIOUS / CURRENT pairs where the same option is used) as function of textual distance between CURRENT and PREVIOUS in the DS. Heavy line represents logarithmic estimate of the relationship in female speakers, dotted line represents logarithmic estimate of the relationship in male speakers

5. Register and regional variation

The selection of data analyzed in the present study (five corpora sampling different spoken registers and varieties) was at least partly motivated by the intention to determine whether persistence effects differ (i) depending on the formality of the speech situation and (ii) depending on the specific variety studied (American English, Standard British English, or English dialects).

The most comprehensive result in this respect is that register and variety seem to have less influence on persistence than one might think. Especially with regard to differences between varieties, hardly any substantially or statistically significant variation was evident. Most noteworthy, perhaps, was that for every alternation under analysis, switch rates were slightly higher in the American English corpora than in the British English corpora (recall that higher switch rates mean a lower level of persistence); switch rates tended to be highest in FRED, and also in FRED, the impact of textual distance on persistence tended to be lowest. With respect to FRED it should be reiterated, however, that the reason that this corpus often behaved differently than the other corpora probably has less to do with the kind of English sampled in this corpus (dialect speech) than the fact that speakers in FRED are, on average, much older (and, perhaps, male to a higher extent) than in the other corpora. As we have seen in section 4, age is a speaker characteristic that does have an impact on how persistence plays out – not only in FRED, but in general. The observation, then, that persistence has a quite uniform effect across varieties may be seen as evidence that persistence is an universal tendency which is not restricted to any particular variety.

What about register? It was generally conspicuous that in the formal corpora (CG and CSPAE), the persistence-related predictors were significant less often significant than in the informal corpora (DS and CSAE). While, for instance, it was to be expected that the turn-taking variables SAMETURN and SAMESPEAKER would be less relevant in more formal settings due this register's less interactional, dialogic nature, it is not self-obvious that, for instance, textual distance between two slots should be less relevant to persistence in formal settings than in informal settings. On the whole, it seems that in more formal settings, persistence is sensitive to fewer determining factors than in less formal settings. A notable exception to this tendency is that lexical complexity – as operationalized through the predictor TTR – turned out to be significant more often in the formal corpora than in the informal corpora, which is why lexical complexity seems to be a more powerful determinant in more planned speech. The following register differences also strike me as interesting:

- α-persistence accounts for less variance in the more formal corpora than in the less formal corpora (5.4% vs. 7.6% on average);

- β-persistence accounts for more variance in the more formal corpora than in the less formal corpora (10.5% vs. 7.3% on average);
- there appear to be no substantial differences as to the overall extent of variance caused by persistence between more formal and more informal corpora;
- there is a slight tendency for persistence to be more long-lived in more formal speech.

In other words, while the absolute proportion of variance for which persistence is responsible is approximately the same in more formal and in more informal settings, α-persistence is comparatively more important in informal settings, and β-persistence in more formal settings.

Let me suggest a theory to account for this difference. Once again, I would like to venture a tentative explanation in terms of self-monitoring and speech planning (cf. Levelt 1989): more formal speech (speech in settings such as press conferences, committee meetings, lectures, etc.) is in all likelihood better planned and monitored than more casual speech. For one thing, this increased planning and monitoring may include intentional stylistic efforts to avoid repetitiveness. Second, it stands to reason that more conscious speech planning should weaken an (at least partially) implicit phenomenon such as persistence. These two factors might explain why α-persistence is weaker in more formal registers. But why does β-persistence appear to be stronger in more formal settings? My claim is that it may be easier to avoid (if such an effort is at all possible) repetitiveness between a manageable number of choice contexts (α-persistence) than within a potentially infinite number of ties between a choice context and its linguistic context (β-persistence). It may be that speakers just "cannot help" being β-persistent, and they might be even more β-persistent than otherwise when they try to avoid being α-persistent. What is more, some manifestations of β-persistence – for example, alliterating speech – are stylistically even welcome and not stigmatized by prescriptivist traditions.

If we take the kind of speech sampled in the informal DS and CSAE as the casual end of a continuum spanning from informal to formal data, the kind of formal spoken language contained in the CG and CSPAE is far from being as formal as it could get. Given this, it will be worthwhile to digress briefly in order to determine – rudimentarily, at least – to what extent persistence manifests itself in written data, i.e. data that tends toward the formal end of

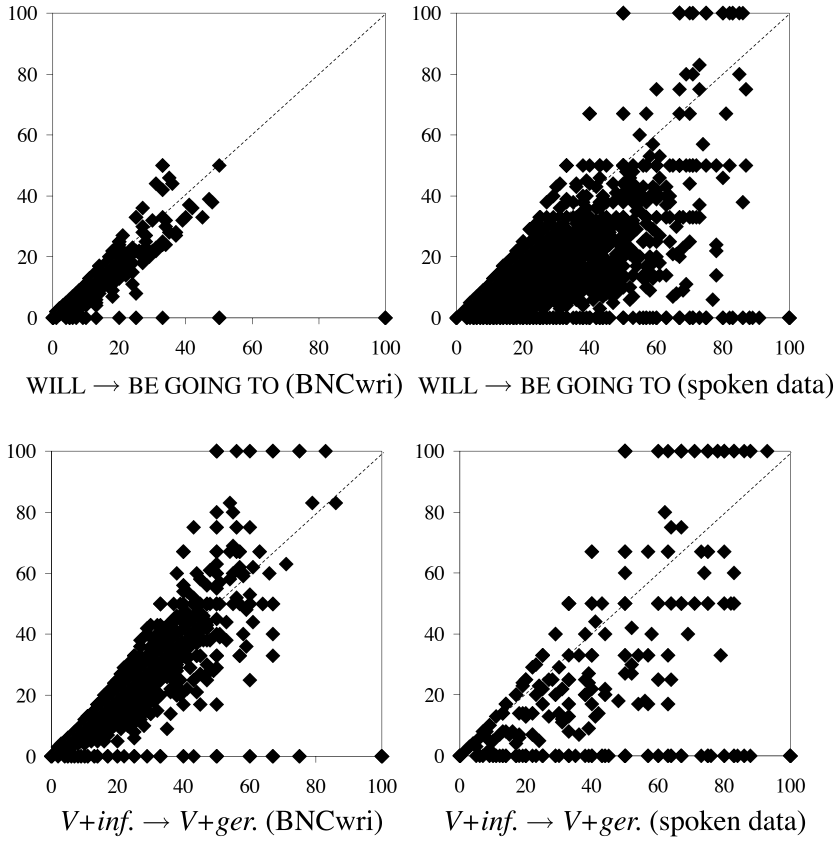


Figure 24. Switches in future marker choice (above) and complementation strategy choice (below) as a function of overall proportion of a given strategy (relative frequency of switches, in %, on *y*-axis; relative frequency of the switched-to complementation strategy, in %, on *x*-axis) in a random sample of the BNCwri (left) and in spoken data (right). Each dot represents one speaker or author. Dotted diagonal line represents null hypothesis that switch rate is proportional to variant proportions

the aforementioned continuum. To this purpose, two of the alternations investigated in the present study, future marker choice and complementation strategy choice, were also studied in a random sample of the written section of the BNC (BNCwri). To provide a rough impression of the relevance of persistence in written data, Figure 24 compares selected switch rates in this written data source (left graphs) to the corresponding switch rates that were presented already in the scatterplots in figures 12 (p. 119) and 17 (p. 167). Recall, now, that the more the dots (which, in the case of the BNCwri, represent individual authors and which in spoken data represent individual speakers) are clustered below the diagonal line, the lower is the authors' or speakers' propensity to switch between alternative options to say the same thing. This is another way of saying that the more the dots are clustered below the diagonal line, the stronger α-persistence is. What can clearly be seen is that in the written data source, the dots do not really cluster below the diagonal line, as they do in the spoken data sources – rather, they actually cluster close to the diagonal line. In other words, authors, unlike speakers, are not particularly disinclined to switch between two options of saying the same thing (but, true enough, they do not have a marked inclination for switching either). What we see in the two scatterplots on written data instead is actually a distribution which one should expect to obtain given a pure chance distribution. Thus, I would like to submit that persistence is what makes a basic difference between spoken and written language, a good deal of which is presumably due to the nature of online production and processing.

6. Summary

By way of summary, Figure 25 decomposes the different factor groups that impact on speakers' linguistic choices among alternatives. These choices are, for one thing, influenced by 'conventional' (or baseline) intralinguistic factors, such as parsing factors (for instance, end weight in particle placement) or information status (for instance, given vs. new ordering in genitive choice or particle placement). But, on the other hand, any given linguistic choice is also impacted by persistence, which may – to reiterate – manifest itself in two varieties: α-persistence, which is the effect previous choices have on upcoming choices, and β-persistence, which is the effect of (not necessarily optional) characteristics of the contextual environment on a given choice slot. Further, we have seen that persistence itself is moderated by an amalgam of factors such as textual distance between two slots, turn-taking, matching lemmas and matching morphology of two choice contexts, and syntactic and lexical complexity of the surrounding linguistic material. These secondary factors do not exert a direct influence on what option speakers choose – instead, their leverage interacts with persistence only. A third factor group that the present study

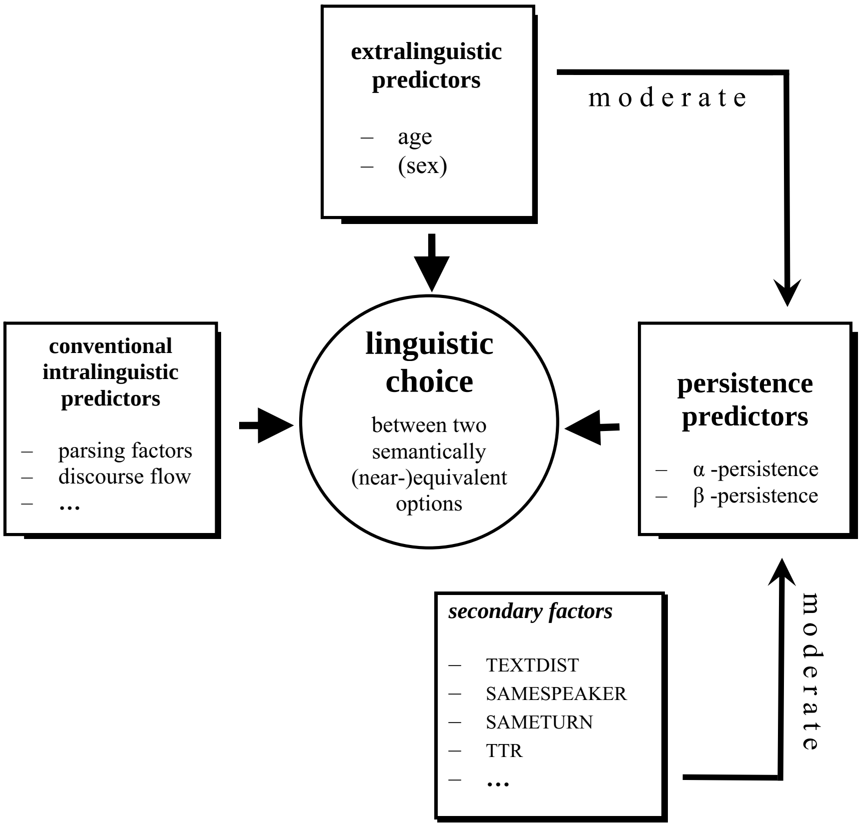


Figure 25. Predicting linguistic choices: 'conventional' intralinguistic predictors, extralinguistic predictors, and persistence predictors

considered were extralinguistic variables (more specifically, sex and age). On the one hand, these factors impact linguistic variation as primary factors – for example, think of the apparent time variation which we observed in comparison strategy choice and future marker choice. Intriguingly, however, age in particular simultaneously moderates persistence as a secondary factor (recall that we have obtained different forgetting functions for older and younger speakers). This makes persistence an unique factor since, to the best of my knowledge, it has hitherto not been observed that, for instance, information status factors have differential effects in older and in younger speakers. Finally, while there is a clear (but not overpowering) tendency that persistence effects differ across spoken registers, there was hardly any evidence that there are differences in persistence along variety lines. It seems to be the case, however, that persistence is a characteristic of spoken language in particular.

Chapter 10
Conclusion

Analyzing five classical alternations in the grammar of English on the basis of a sizable and diverse database, I believe that I have demonstrated empirically that speakers are, indeed, creatures of habit who tend to reuse linguistic patterns from previous discourse. At a minimum, the present study would seem to have shown that repetitiveness is sufficiently patterned to serve as an explanatory factor in empirical approaches to linguistic variation, and that (naturalistic) corpus data can match (experimental) psycholinguistic data. The specific variationist approach I have suggested indicated that consideration of the factor can enhance the explanatory power of linguistic model building sizably.

More specifically, I hope that my analysis has provided substance to the following claims. First, successive variable sites in discourse are not statistically independent of each other. For one thing, we saw that switch rates between two alternative linguistic options are considerably lower than chance switch rates. Also, logistic regression clearly showed that given option A was used in the first of two successive variable sites in discourse, the odds that option B would be used in the second site are reduced substantially. I termed this instantiation of persistence α-persistence. At the same time, the magnitude of α-persistence is itself a function of several determinants such as, among other things, textual distance between two successive variable sites (persistence declines logarithmically with increasing textual distance), whether two successive variable sites involve the same verb lemma (if they do, persistence is stronger), whether two successive variable sites are in the same conversational turn, and whether they are produced by the same conversational party (in both cases, persistence is more powerful if the answer is yes).

A second way in which persistence impacts speakers' choices is the following: given a variable site where speakers have a choice between two or more options, that choice is not only influenced by other variable sites; rather, it is also affected by non-variable linguistic patterns that share structural or lexical characteristics with one of the options. This instantiation of persistence is what I have called β-persistence. We have seen, for instance, that a non-comparative occurrence of the token *more* (as in *I would like more soup*) can help trigger an analytic comparative in a variable slot nearby. Finally, the analysis has suggested that extralinguistic factors such as speaker age (and possibly speaker sex) can interact in a somewhat complex fashion with persistence.

Needless to say, the present study has had to leave several potentially interesting questions unanswered. For instance, I wish to stress that the present study certainly did not offer an exhaustive account of all the (possibly infinite) ways in which β-persistence may manifest itself. Future research should empirically test further factors for their explanatory power. By virtue of its corpus-based method, the present study also had to leave open the question of precisely to what extent persistence is due to hard wiring (i.e. properties of the human speech production and processing system), and how much of the effect is due to soft wiring (i.e. social technology to manage discourse, especially conversation, and functional factors such as speaker-hearer economy). This is an issue that, I believe, only experimental research can attempt to settle satisfactorily, though this study's results suggest that both realms are involved in the phenomenon. For instance, there is no way that the logarithmic shape of the forgetting functions could follow from discourse-analytic accounts of repetitiveness; instead, this kind of logarithmic decline strongly hints that decaying energy and activation levels in the speech production system are at issue here. By the same token, that age seems to impact the magnitude of persistence likewise indicates that the phenomenon is to some degree due to the human speech processing system, which, as human 'hardware', is subject to aging. On the other hand, the observation that turn-taking, for instance, moderates the magnitude of persistence implicates that persistence may simultaneously be a functional phenomenon which is involved in how speakers manage the business of conversation. A related issue (which, truth be told, the present study has carefully avoided so far) is whether persistence is actually an intralinguistic or extralinguistic factor. In my view, persistence is extralinguistic inasmuch as is due to psycholinguistic mechanisms relating to the human speech production and processing system. At the same time, it is intralinguistic to the extent that it serves functions in discourse. Thus, for the same reason that the present study cannot disentangle the psycholinguistic and discourse-functional root causes of the phenomenon, it cannot at this point give a more satisfying answer than that persistence is both extralinguistic and intralinguistic.

I should now also add a word about *horror aequi*, the principle stating that speakers tend to "avoid the use of formally (near-) identical and (near-) adjacent (non-coordinate) grammatical elements or structures" (Rohdenburg

2003: 236). As I have set forth in chapter 2, persistence and *horror aequi* may at first glance appear to be two diametrically opposed and empirically incompatible tendencies. Note though that this study's subsequent analyses have demonstrated that they actually are not, for two reasons. First, while testing for a handful of potential *horror aequi* effects, the only relatively strong one corroborated empirically in the present study was the dispreference of *V+ger.* + *V+ger.* and *V+inf.* + *V+inf.* in complementation strategy choice (which is the textbook example of *horror aequi*). Thus, although it is fair to say that the present study did not look as hard for *horror aequi* effects as it did for persistence effects, the latter seem to be the empirically more relevant ones. Second, *horror aequi* has an exceedingly limited textual scope, both according to the literature and according to my analysis – in complementation strategy choice, for instance, the *horror aequi* effect only stretches over a couple of interjacent tokens at maximum. By stark contrast, we saw that persistence effects may not dissipate before tens of minutes of talk. In a very real sense, then, *horror aequi* and persistence play in different leagues, both empirically and – in all likelihood – psycholinguistically.

How are the present study's results relevant to linguistic practice? As has been set forth in the Introduction already, the strong empirical showing of persistence plays methodological havoc with a standard assumption underlying most empirical linguistic research: namely, that an occurrence of a linguistic pattern can and should be considered the result of a new throw of the dice, and that it can be investigated in isolation and out of the wider discourse context. This is, first, a problem for qualitative linguistic inquiry where, often, a data fragment is investigated asking, 'why did the speaker use this specific option, instead of the alternative one, here?' The present study leaves us good reason to believe that the answer might often be as simple as 'because the speaker had just used that option – or some trigger – before.' What does this mean? To illustrate, consider the data fragment in (1):

(1) *And that then* ***starts to provide*** *some feedback for our contractor.* (CSPAE Comm597)

Why did the speaker use infinitival complementation in (1)? How does the (semantic) context in (1) license this kind of complementation? A qualitative analysis would probably point out that *start to provide feedback* "implies only an entry into the initial phase of an activity" and not "the initial phase of a repeated activity" (Řeřicha 1987: 130) – hence infinitival complementation. However, the action referred to in (1) does not seem to give a mere potentiality for action but rather a sense of the actual performance of *providing feedback*, so following Quirk et al. (1985: 1191) the analyst might conclude that there is actually no good reason to use infinitival complementation. In other words, a qualitative analysis would be challenged to offer a conclusive account for (1). The present study, by contrast, would seem to have suggested that, first of all, more context[67] needs to be considered, as in (2):

(2) *Then, I think the thing would be is to* ***start to move*** *to take a look at how we would* ***start to put*** *some specifics, especially those ones that we want to sample on a regular basis. And that then* ***starts to provide*** *some feedback for our contractor.* (CSPAE Comm597)

Crucially, (2) elucidates that gerundial complementation had just been used immediately before (*start to move* and *start to put*), with the analyses in this study having demonstrated that under such circumstances, the odds for infinitival complementation in the choice context under analysis increase by more than 90% (cf. Table 29; in addition, note that there is an infinitive trigger [*to sample*] in immediate adjacency). This kind of argument should, at a minimum, complement a qualitative analysis: given two or more conflicting factors, as seems to be the case in (1), persistence may tip the balance in favor of either factor, which seems to be exactly what is happening in (2).

Secondly (and maybe more importantly), persistence also poses a problem to quantitative linguistic studies (i.e. studies that seek to identify text frequencies of some morpheme, lexeme, or construction) in that text frequencies may be misleading unless, among other things, textual distances between the individual hits are factored in. An example will illustrate: Szmrecsanyi (2003: Table 2) (cf. Berglund 1999b for similar figures) claimed that in the DS, the distribution of BE GOING TO and WILL/SHALL is roughly 28:72. This figure, of course, does not take persistence into account. If the researcher chooses to exclude, for instance, all cases where two successive future marker slots are located in the same turn (because, as we have seen, persistence is powerful in such contexts), the distribution changes to roughly 30:70, a difference which is highly significant.[68] Further, if the researcher opts to exclude all cases where textual distance between two future marker hits is less than 150 words (recall that according to Table 32, this is the textual threshold after which persistence dissipates entirely), the ratio changes to 32:68, which is, again, a significant difference.[69] For a similar skewing, take the distribution of BE GOING TO and WILL in the CSAE: according to Table 14 (p. 112), this distribution is roughly 42:58; however, this figure does not control for persistence either. If the analyst controls for persistence by excluding all occurrences that are closer than 200 words to a previous occurrence (again, it should be kept in mind that according to Table 32, this is the textual threshold after which persistence dissipates entirely in the CSAE), the distribution changes to roughly 55:45 – meaning that all of a sudden, BE GOING TO turns out to be the actually dominant future marking strategy.[70] In other words, accounting for persistence in the domain of future marker reference yields distributions where BE GOING TO is really more frequent than has hitherto been thought. While it is possible to construct even more extreme examples, the above case studies go to show that text frequencies may be distorted (or at least influenced) by persistence, with persistence – much like restarts, for instance – ordinarily not being a factor that researchers interested in text frequencies would like to have in their statistics. It may therefore be a worthwhile aim for future research to develop a comprehensive algorithm to control text frequencies for the distorting effect of persistence.

On a more general, theoretical level, persistence has the potential to be of theoretical interest to linguists engaged in very diverse research programs. First of all, behaviourists – had they not disappeared from the linguistic scene long ago – would find the stimulus-response pattern of persistence, repetitiveness, and prime-target pairs absolutely intriguing. Certainly however, the present study seems to have demonstrated that persistence, as an explanatory factor, is immediately relevant to all those who seek to account for the choices speakers make in the spirit of variationism or probabilistic grammar. Along somewhat different lines, persistence may alternatively be thought of as a type of short-term entrenchment. ‘Entrenchment’ (originally a Cognitive Grammar term) is a mechanism due to which the effect of discourse frequency on mental representations is such that these representations are strengthened through their activation in use (cf. Langacker 1987: 59–60; Hudson 1997: 82–83). It is true that entrenchment is understood to be a mechanism operating over longer intervals of time, possibly a speaker’s lifespan – in contrast, persistence is a phenomenon that dissipates after minute intervals (cf. chapter 9, section 2.1). Yet, persistence, too, is due to linguistic patterns, or representations thereof, being activated through use. I suggest, then, that the cognitive mechanism of entrenchment is one long-term manifestation of persistence, and that persistence is among one of the short-term mechanisms which relate to entrenchment in the long run; in a nutshell, it may make sense to refer to persistence as "micro-entrenchment," and to entrenchment as "macro-persistence."

Cognitive grammar aside, persistence is obviously interesting to mainstream functionalists since issues such as online processing constraints, economy, and discourse management are, as we have seen, involved in motivating surface structure. But also for less mainstream, more extreme functionalists who view grammar as an emergent system of meaningful repetition and as a "vast collection of hand-me-downs that reaches back in time to the beginnings of time" (Hopper 1998: 150), persistence should be a worthwhile phenomenon to consider.

Maybe surprisingly, the existence of the phenomenon can even be seen as underlining the validity of the generative enterprise, for two reasons. First, persistence or parallelism in surface structure can potentially yield linguistic outcomes that are dysfunctional: Scherre and Naro (1991: 30), for instance, have noted that due to speakers' inclination to maintain surface parallelism, morphological "markers tend to occur precisely when they are not needed and tend not to occur when they would be useful". Thus, persistence and functional factors can very well work against each other – for instance, in contexts where functional factors would license some option A, but due to persistence it is option B that is actually used. *Ex negativo*, this can be interpreted as evidence that grammar cannot be motivated functionally alone, hence the need for formal analysis. Second (and relatedly), the fact that speech generation is sometimes heavily inertial and mechanical (insofar as the human speech processing system is skewed toward repetition) can be construed as evidence (albeit somewhat indirect) for the Autonomy of Syntax Hypothesis. The point is that if speakers cannot help being persistent and repetitive (a claim that the present study certainly has not contradicted empirically), the cognitive module which is responsible for syntax must be, to some extent at least, self-contained.

Last but not least, persistence could also have implications for historical linguistics: the multiplicative and self-enforcing effect of persistence, coupled with logarithmic forgetting functions, might very well be involved in the s-curve patterns so often observable in language change. This is an intriguing issue which, needless to say, would be worth exploring in future research.

Appendix A
Sample coding

To illustrate the coding of the factors discussed in chapter 3, section 1, consider the following chunk of conversation (Corpus of Spoken American English, text 0906):

```
   1    JIM:  <<POUND +five +five +five,
   2          ... +five +five .. +five +five POUND>>.
   3          ... Is what the phone number will be.
   4          (H) .. But,
   5          .. this is .. things that,
   6          (H) I'm I'm .. anxious for Matt to get here,
   7          because I'm getting,
   8          (H) .. I'm getting tuckered out,
   9          [trying to] .. get all these nitty-gritty things,
  10    JIM:  It's not tough to get the phone number.
  11          (H) But see they need the phone number in order
  12          to order letterhead,
  13    JOE:  [Mhm].
  14    JIM:  [(H)] in order to or- .. % have business cards.
  15          (H) .. LCL in [2..2] in Chicago needs that,
  16    JIM:  .. to get that process [3going3].
  17    JOE:                         [3Hm3].
  18    JIM:  [4Uh,
  19    JOE:  [4(TSK) (H) Uh is there LCL4] accounts gonna
  20          be maintained here,
  21    JIM:  cause cause actually4] --
  22    JOE:  I mean,
  23          or .. is there gonna be a separate,
  24          (H) They're gonna have an account in Chicago,
  25          for the funds to pass through?
→ 26          Or is it gonna be passthrough funds
  27          here at the bank? Or,
  28          (H) is that %= --
  29    JIM:  ... Well,
  30          .. w- .. what we'll do is,
  31          ... those'll probably
  32          wire transfer [out].
  33    JOE:                [Through Bolt]mans or something,
  34    JIM:  Well,
  35          .. through the Fed,
  36          what --
  37          (H) I think what will happen,
  38          (H) but we --
  39          .. Matt'll find this out,
  40          and, I mean, we'll get involved in it
```

In this passage, there is a good deal of variation between the future marker paradigms BE GOING TO (*gonna*) and WILL (*will* and *'ll*). There are in all 10 variables, in 4 of which a BE GOING TO marker and in 6 of which a WILL marker is used. Let us suppose, now, that the independent variable to be analyzed, CURRENT, is the variable in line 26, where JOE uses the BE GOING TO variant *gonna* (*Or is it* gonna *be passthrough funds here at the bank?*), instead of some WILL marker:

- the value of PREVIOUS, then, is BE GOING TO as well, since the last future marker slot – the discourse-preceding variable – in line 24 (*... they're* gonna *have ...*) is also realized by the BE GOING TO variant *gonna*;
- TEXTDIST is ln 15 (≈ 2.7), since the textual distance between the last future marker slot (*gonna* in line 24) and the variable under analysis is 15 words;
- the sentence in which CURRENT is embedded (*Or is it gonna be passthrough funds here at the bank?*) contains 9 words (excluding the future marker *gonna* itself and its auxiliary *is*), thus SENTENCELENGTH is assigned a value of 9;
- TTR ranges somewhere between 1 and 100, depending on the wider lexical context;
- the value of SAMETURN is 1 ('yes') since the discourse-preceding future marker variable slot, in line 24, is located in the same turn as the variable under analysis;
- the value of SAMESPEAKER is 1 ('yes') as well since the discourse-preceding future marker variable slot, in line 24, is produced by the same speaker – JOE – who produces CURRENT;
- according to the speaker table accompanying the CSAE, speaker JOE is male and 45 years old, thus SEX is 'male', and AGE is '41.'

All of the above coding would have been conducted automatically using Perl scripts.

Appendix B
Multicollinearity statistics

The following tables report Variance Inflation Factors (VIFs) for all variables that were entered into logistic regression. Variance Inflation Factors are familiar from multiple regression, but are applicable to logistic regression too; they measure the strength of inter-relationships among explanatory variables in a multivariate model. Increasing Variance Inflation Factors indicate increasing regression coefficients, which may result in more unstable estimates. Variance Inflation Factors exceeding a value of 10 are commonly considered to indicate multicollinearity, but values above 2.5 may already be a cause for concern.

VIFs in comparison strategy choice

	DS	CG	CSPAE	FRED
SENTENCELENGTH	1.086	1.172	1.084	1.133
TTR	1.081	1.287	1.145	1.131
LENGTH	2.269	3.070	4.728	2.211
MORPH	2.401	4.265	3.702	2.010
STRESS	1.207	1.560	3.088	1.428
FREQUENCY	1.042	1.180	1.399	1.151
SYNFUN	1.123	1.619	1.251	1.168
DEGREE	1.064	1.496	1.084	1.143
COMPLEMENT	1.331	2.080	1.211	1.437
PREVIOUS	1.115	1.575	1.058	1.110
TEXTDIST	1.931	1.547	1.169	1.401
ATRIGGER	1.924	1.372	1.131	1.411
STRIGGER	1.180	1.690	1.078	1.250
AGE	1.059	1.163	–	1.173
SEX	1.028	1.448	–	1.598

VIFs in genitive choice

	FRED	CSAE
SENTENCELENGTH	1.048	1.137
TTR	1.104	1.147
LEXCLASS	1.249	1.174
POSSESSORLENGTH	1.080	1.157

POSSESSUMLENGTH	1.016	1.054
FINALSIB	1.026	1.038
POSSESSORGIV	1.096	1.095
POSSESSUMGIV	1.081	1.210
FREDAREA	1.025	–
PREVIOUS	1.138	1.228
TEXTDIST	1.366	2.005
TEXTDIST-OF	1.068	1.179
SAMETURN	1.428	2.171
SAMESPEAKER	1.274	1.323
POSSESSORID	1.177	1.114
POSSESSUMID	1.170	1.246

VIFs in future marker choice

	CG	CSPAE	DS	CSAE	FRED
SENTENCELENGTH	1.012	1.036	1.025	1.109	1.029
TTR	1.042	1.038	1.016	1.136	1.048
NEGATION	1.008	1.004	1.016	1.019	1.005
PREVIOUS	1.045	1.174	1.032	1.123	1.025
TEXTDIST	1.577	1.670	1.524	1.486	1.619
G-ALLIT	1.177	1.082	1.088	1.125	1.074
W-ALLIT	1.010	1.030	1.022	1.058	1.033
G-HORRORAEQUI	1.012	1.024	1.023	1.029	1.009

VIFs in particle placement

	FRED	CSAE
SENTENCELENGTH	1.077	1.121
TTR	1.087	1.136
DEFINITEDO	1.022	1.145
SYLLABLESDO	1.044	1.186
LITERALNESS	1.107	1.153
COMPLEXITYDO	1.047	1.109

DIRECTIONALPP	1.036	1.115
NEWSVALUEDO	1.087	1.093
DISTINCTIVENESS	1.073	1.194
FRED-AREA	1.212	–
PREVIOUS	1.107	1.115
TEXTDIST	1.492	1.814
SAMETURN	1.589	1.435
SAMESPEAKER	1.143	1.247
VLEMMAID	1.126	1.509
TEXTDIST-NONSEP	1.030	1.128
TEXTDIST-SEP	1.129	1.145

VIFs in complementation strategy choice

	CG	CSPAE	DS	CSAE	FRED
SENTENCELENGTH	1.087	1.049	1.045	1.500	1.060
TTR	1.146	1.032	1.054	1.312	1.106
STATIVE-COMPL	1.055	1.011	1.044	1.249	1.064
HYPOTHETICAL	1.636	1.417	1.225	1.378	1.247
TO-HORRORAEQUI	1.672	2.443	1.533	1.351	1.858
ING-HORRORAEQUI	1.879	2.332	1.540	1.223	1.172
VERB	1.239	1.116	1.225	1.798	1.421
MORPHOLOGY	1.240	1.268	1.151	1.548	1.436
FRED-AREA	–	–	–	–	1.118
PREVIOUS	1.136	1.030	1.087	1.339	1.069
TEXTDIST	1.742	1.887	1.524	2.114	1.377
SAMETURN	1.745	2.041	1.419	1.413	1.417
SAMESPEAKER	1.377	1.534	1.155	1.384	1.202
VLEMMAID	1.475	1.227	1.149	1.674	1.395
VMORPHID	1.656	1.306	1.199	1.472	1.125
TEXTDIST-TO	1.703	2.430	1.496	1.697	1.864
TEXTDIST-ING	1.721	2.112	1.525	1.179	1.058
T-ALLIT	1.181	1.099	1.082	1.681	1.096
AGE	1.053	–	1.048	–	–
SEX	1.105	–	1.034	–	–

Appendix C
Coding schemes in particle placement

The author and the second coder followed the following coding schemes:

Syntactic function (attributive vs. predicative) of adjectives

Code ‘0’ for attributive function (e.g. *the* green *house is there*, *I like* red *cars*). Code ‘1’ for predicative function (e.g. *the house is* green, *the car seems* nice, *Jim became* angry).

Definiteness of the direct object of transitive phrasal verbs

Code definite direct objects of phrasal verbs as ‘0’ (e.g. *Jim looked up* the word). Direct objects tend to be preceded by a definite article, or by some kind of genitive or possessive pronoun. Code indefinite direct objects of phrasal verbs as ‘1’ (e.g. *Jim looked up* a word). Indefinite objects tend to be preceded by an indefinite article, or by no article at all.

Literal vs. idiomatic phrasal verbs

If the phrasal verb is literal, code ‘1’. Literal phrasal verbs are verbs where the meaning of the whole verb is more or less the semantic sum of the verb and the particle. Often, literal phrasal verbs are phrasal verbs where some spatial movement is involved (for instance, *to bring in* is the semantic sum of *to bring* and *in*; also, some spatial movement is involved). If the phrasal verb is idiomatic, code ‘0’. A phrasal verb is idiomatic if the meaning is more than the semantic sum of verb and particle (if one needs to have learned its idiomatic meaning, therefore). Most often, idiomatic phrasal verbs are not spatial (for instance, to *figure out* means something else than the semantic sum of *to figure* and *out*, and also, there is no spatial movement involved).

Appendix D
Phrasal verbs analyzed

NOTE: absolute frequencies in the dataset in brackets.

back up (1)
beat up (1)
bellow out (1)
blow out (3)
blow up (5)
boil up (2)
break apart (1)
break down (3)
break off (1)
break up (3)
breed off (1)
brick up (1)
bring back (18)
bring down (6)
bring in (17)
bring out (5)
bring over (5)
bring up (28)
brush off (1)
build up (2)
buy up (2)
call out (2)
call up (5)
calm down (1)
carry away (1)
carry back (1)
carry down (2)
carry in (3)
carry on (1)
carry out (4)
carry up (3)
chalk on (1)
chew up (2)
chop off (1)
chop up (5)
chuck in (3)
chuck out (4)
chuck over (1)
chuck up (1)
chuckle up (1)
clean down (1)
clean off (5)
clean out (18)
clean up (4)
clear away (1)
clear out (1)
clear up (5)
close down (7)
close up (1)
cock up (1)
coil up (1)
coke up (1)
count out (3)
cry out (1)
cut down (2)
cut off (10)
cut out (8)
cut up (8)
dig in (1)
dig out (1)
dish out (2)
do up (1)
drag off (1)
drain off (1)
draw out (1)
draw up (2)
dress up (3)
drive down (2)
drive in (2)
drop in (4)
drop off (4)
eat up (1)
edge up (1)
fetch in (1)
figure out (3)
fill in (9)
fill out (6)
fill up (25)
find out (2)
finish up (2)
fire down (1)
fix up (2)
flip over (5)
get back (8)
get down (2)
get in (21)
get off (4)
get on (1)
get out (10)
get up (3)
give away (2)
give back (2)
give out (4)
give up (9)
grind up (1)
grow up (1)
hand back (1)
hand off (1)
hand over (1)
hang out (1)
hang up (3)
haul up (1)
have on (2)
heave in (1)
heave out (1)
help out (1)
hold back (1)
hold out (4)
hold up (7)
hook up (1)
keep away (1)
keep down (1)
keep in (1)
keep out (2)
keep up (16)
kick out (5)
kick up (2)
kill off (1)
knock out (1)
knock down (1)
lace up (1)
lay on (1)
lay up (1)
leave out (2)

let down (4)
let in (1)
let off (2)
let out (11)
lift out (1)
lift up (17)
light up (2)
load up (1)
lock up (1)
look up (2)
lower down (5)
lug down (2)
make out (5)
make up (17)
mix up (1)
mop up (1)
move up (1)
open up (1)
partition off (1)
pass away (2)
pass down (1)
patch up (1)
pay back (2)
pay in (1)
pay off (3)
pay up (1)
phone up (3)
pick out (22)
pick up (83)
play up (2)
polish up (1)
pour in (1)
pour out (1)
pull away (2)
pull back (1)
pull down (11)
pull in (9)
pull off (6)
pull out (24)
pull together (1)
pull up (29)
pump up (2)
push away (2)
push up (3)
put away (3)
put back (2)
put down (11)
put in (16)
put off (5)
put on (75)
put out (17)
put up (36)
raise up (1)
read out (3)
reckon out (1)
reckon up (1)
regurgitate up (1)
rent off (1)
rig up (2)
ring up (2)
roll over (1)
roll up (1)
root up (1)
run down (3)
save up (5)
scrape off (1)
screw up (1)
sell off (1)
sell out (1)
send away (5)
send back (1)
send down (5)
send in (5)
send off (2)
send up (3)
serve out (1)
serve up (1)
set down (1)
set up (13)
shoot out (4)
shout out (1)
shove down (1)
shove in (1)
shut off (1)
slip off (1)
smuggle in (1)
sort out (3)
spit out (1)
stack up (1)
start off (1)
start up (5)
stem up (1)
stick in (1)
stick up (2)
sweep out (1)
sweep up (1)
swill round (1)
take away (15)
take back (14)
take down (18)
take in (35)
take off (57)
take on (8)
take out (66)
take over (23)
take up (28)
throw away (1)
throw down (4)
throw in (1)
throw on (1)
throw out (6)
throw up (2)
tie in (2)
tie up (7)
tighten up (1)
tip up (3)
total up (1)
trim up (1)
try on (4)
try out (4)
turn down (2)
turn in (3)
turn off (2)
turn on (13)
turn out (5)
turn up (5)
wake up (1)
wash out (5)
wash up (1)
wear out (2)
wear up (1)
weigh out (1)
weigh up (1)
wind up (5)
wipe out (2)
work back (1)
work out (2)
write down (10)
write out (2)

Notes

1. Consider, along these lines, the recent monograph on "Determinants of Grammatical Variation in English" by Rohdenburg and Mondorf (2003). This is a collection of state-of-the-art variationist research papers dealing with numerous alternation phenomena in English and focussing on "major extra-semantic and largely neglected factors determining grammatical variation" (Rohdenburg and Mondorf 2003: 1). Yet while *horror aequi* receives ample attention in the volume, persistence – as a phenomenon that from a psycholinguistic perspective is comparatively better documented – is not referred to even once as an explanatory factor.
2. See Tannen (1987: 577–579) for a comprehensive overview. To cite just some studies: Keenan (1977: 125) notes that "one of the most commonplace observations in the psycholinguistic literature is that many young children often repeat utterances addressed to them." Merritt (1982) shows that children in primary school use repetition and reformulation to get their instructors' attention. Conversely, Cook-Gumperz (1977) presents evidence that repetition is often used for instruction-giving in classrooms. Watson-Gegeo and Boggs (1977) show that Hawaiian children use repetition in contradiction routines; Ervin-Tripp (1979) notes that repetition is frequently used as a remedial tactic in turn-taking trouble in child discourse. Ochs (1979) argues that repetition is used as an attention-getting strategy and a means to achieve definiteness in children's interaction among themselves and with caretakers. Hatch, Peck, and Wagner-Gough (1979) examined repetition in children's acquisition of formulae and the subsequent impact on rule formulation. Goodwin (1983) describes what she calls 'aggravated partial-repeat correction formats' in conversations among urban black children age four to fourteen, where repeats are produced with challenging intonation. Erickson (1984: 141) demonstrates that black adolescents use repetition of pitch and phrases to engage listeners in a dialogue of "call and response."
3. Well-known, though rather irrelevant to the present study, is also *semantic priming*, which occurs when processing of a word is facilitated by just having processed a semantically related word. Thus, for instance, processing and recognition of the word *cat* is aided by just having been exposed to the word *dog* (cf. Meyer and Schvaneveldt 1971; cf. Cleland and Pickering 2003 for semantic enhancement in syntactic priming). Some researchers also distinguish *word-order priming* from other priming varieties (cf. Hartsuiker and Westenberg 2000).
4. Reprinted from Kathryn Bock, "Syntactic persistence in language production", *Cognitive Psychology* 18 (3), p. 361, Copyright ©1986, with permission from Elsevier.
5. Reprinted from Deborah Tannen, *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*, p. 73, Copyright 1989, with kind permission from Cambridge University Press.
6. Reprinted from Deborah Tannen, *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*, p. 77, Copyright 1989, with kind permission from Cambridge University Press.
7. SENTENCELENGTH and TTR will be controlled for differing string lengths of the options under analysis. For instance, a sentence containing a site of analytic comparison is in-

trinsically longer – by one word (*more*) – than the same sentence containing a site of synthetic comparison. This bias would have skewed results and was thus removed. This is also an issue for future marker choice (where BE GOING TO markers contain more material than WILL markers), genitive choice (where the *of*-genitive contains more material than the *s*-genitive), and infinitival vs. gerundial complementation (where the infinitive marker *to* only occurs with infinitival complementation).

8. The reason for this eclecticism is that otherwise, the number of observations would be too low for reliable statistical analysis.
9. Varbrul was designed in the 1970's, and has not been very substantially updated since. Varbrul is therefore antique in terms of software. Among the most serious limitations of the Varbrul package are the following: (i) the data need to be in a specific, somewhat archaic format for Varbrul to handle them, (ii) Varbrul can only deal with nominal variables, (iii) Varbrul cannot handle interactions between independent variables, (iv) Varbrul does not report R^2 (see Bayley and Young forthcoming for a review of the Varbrul package). The logistic regression module of SPSS has none of these limitations.
10. All of these figures are imaginary.
11. It is worth digressing briefly to discuss the importance of this measure. Not reporting variance explained (R^2) is somewhat convenient since it helps avoid questions such as whether the predictors studied – however significant they may be in isolation – are of substantial interest, or whether inclusion of *other* predictors would have enhanced the analyst's ability to explain linguistic variation. Take, for instance Poplack and Tagliamonte (1996), or almost any other study in the Varbrul tradition. The reader is bombarded with probabilistic weights, but nowhere is one told how successful the corresponding predictors are in actually explaining the linguistic variable studied. This usual omission in Varbrul research of course also has to do with the fact that the program does not report R^2, but it does report a goodness-of-fit measure (*log likelihood*), which is never reported either. In all, while reporting R^2 may make painfully clear inadequacies and omissions in terms of the overall explaining power of one's research design, reporting such a measure increases accountability.
12. For several reasons, it will not be feasible to report probabilistic weights for the type of logistic regression utilized and for the type of independent variables considered in the present study (cf. Pampel 2000: 23; Jaccard 2001: 3).
13. Statistical significance of odds ratios will normally be computed on the basis of the *Wald statistic*, which is the ratio of the unstandardized logit coefficients to their standard errors. For large coefficients, however, *Wald* becomes unreliable (cf. Menard 2002: 39), which is why large coefficients will be tested by specifying a model with and without the independent and testing the change in *Hosmer and Lemeshow's G* for significance.
14. This chapter will focus exclusively on comparatives. First, superlatives are less frequent than comparatives, which makes statistical analysis difficult. Second, most previous research on comparison in English has dealt primarily with comparatives, and as this is not a study on comparison specifically, this chapter too will content itself to research comparatives.
15. In non-standard English, there is a third possibility: double comparatives (e.g. *more nicer*). Due to this pattern's comparatively low text frequency (cf. Kytö and Romaine 1997: 329; Biber et al. 1999: 525), double comparison will not be considered in this chapter.

16. Disyllabic adjectives with final stress that are attested with analytic comparison in Leech and Culpeper's data include *acute, afraid, akin, aware, bizarre, compact, complete, correct, exact, extreme, intense, mature, obscure, polite, precise, profane, profound, remote, robust, secure, severe, sincere* (Leech and Culpeper 1997: 361).
17. Left out from this selection is a *horror aequi*-related variable, namely whether the presence of a token ending in *-er* in the immediate adjacency of an adjective influences the type of comparison the adjective will take. The reason is that there were only some 36 such patterns (of the type *an extremer winter* or *a rather costlier car*) in the whole database, which is a number too low for reliable statistical analysis.
18. In this context, mention should be made that there is a collinearity issue with LENGTH, MORPH, and STRESS especially in the CG and CSPAE (cf. the collinearity measures in Appendix B), which may distort results: (i) adjectives that end in *-y* or begin in *un-* tend to be longer than other adjectives, (ii) adjectives that are stressed on the last syllable also tend to be longer than other adjectives, and (iii) adjectives that end in *-y* or begin in *un-* tend not to be stressed on the final syllable. These correlations also obtain in the DS and in FRED, but there their magnitude is not a cause for concern.
19. The regression line might appear to some readers as improbably horizontal, given the distribution of the dots. Note, however, that the majority of dots sitting on the *x*-axis represent more than one speaker (often many more), to which the regression is of course sensitive. This will also have to be kept in mind for similar scatterplots in the chapters to follow. By illustration, consider speaker WESBE (FRED corpus): This speaker uses only one relevant (optional) analytic form and one relevant synthetic form – in this order. Thus, speaker WESBE never actually switches from synthetic comparison to analytic comparison and sits, in the righthand graph in Figure 6, on the *x*-axis (at $x = 50$).
20. CG: step χ^2 = 54.39, df = 8, $p < 0.001$; CSPAE: step χ^2 = 21.37, df = 8, p = 0.006; DS: step χ^2 = 46.19, df = 8, $p < 0.001$; FRED: step χ^2 = 107.05, df = 8, $p < 0.001$.
21. Figure 7 – exactly as similar graphs that will be presented in what follows – is based on 19 measuring points. These have been arrived at by dividing the observed textual distance between PREVIOUS and CURRENT into 20-tiles, i.e. into 20 equal groups (which have 19 cut-off points); the percentage of matches between PREVIOUS and CURRENT was then determined separately for each 20-tile.
22. * significant at $p < .05$, ** $p < .01$, *** $p < .005$.
23. ATRIGGER(75) shows whether the token *more* precedes CURRENT by 75–51 words; ATRIGGER(25) shows whether the token *more* precedes CURRENT by 25–6 words; ATRIGGER(5) shows whether the token *more* precedes CURRENT by less than 6 words.
24. Step χ^2 = 25.04, df = 9, p = 0.03.
25. This subset consisted of the following texts: LAN008–LAN014, NBL001, NBL003, NBL006, NBL007, NBL008, WES001, WES002, HEB001–HEB041, SAL001–SAL039, WAR001, DUR001–DUR003, LAN001–LAN007, KEN006–KEN008, KEN014, LND001, LND002, SFK011–SFK036, CON001–CON010, DEV001, SOM001–SOM014, DEN001–DEN004, GLA001–GLA007. These comprise ca. 1,300,000 words and thus approximately 55% of the entire FRED corpus; dialect areas included in the sample are the Hebrides, the Midlands, the North of England, Wales, the Southwest, and the Southeast.
26. When the *s*-genitive is used, the possessum cannot be determined by an article (**the man's the house*). Therefore, definite or indefinite articles determining the possessum

phrase of an *of*-genitive were not included in the count in order not to skew results.

27. This only lists the intralinguistic factors. If extralinguistic factors are included as well, the third and fourth most important predictors in FRED are whether or not a speaker is from the Midlands or from Wales, respectively.
28. CSAE: step $\chi^2 = 65.08$, df = 10, $p < 0.001$; FRED: step $\chi^2 = 117.58$, df = 10, $p < 0.001$.
29. * significant at $p < .05$, ** $p < .01$, *** $p < .005$.
30. The graph on the CSAE is somewhat bumpier than the graph on FRED because the number of observations is lower in the CSAE. Thus, statistical outliers do not cancel themselves out so easily in the CSAE.
31. Of course, sometimes there is no optionality; in conditional protases, for instance, prescriptive tradition bans usage of WILL (cf. Comrie 1985; also see below, section 2).
32. It is also well-known, though irrelevant for the present study, that WILL is massively more frequent in written language than in spoken language (for instance, Berglund 1997, 1999b, 2000b; Biber et al. 1999; Mair 1997b; Martin and Weltens 1973; Wekker 1976).
33. *Will*, too, can have non-future-marking homonyms (e.g. *his last will*). Due to these homonyms' negligible text frequency, however, no attempt was made to remove them manually.
34. Note that if CURRENT takes a BE GOING TO based future marker, the token *going* does not count.
35. The following figures supplement Table 16. For simplicity, only estimates significant at $p < .05$ are displayed.
36. There are fewer BE GOING TO → WILL switches than WILL → BE GOING TO switches in the CG, CSAE, and in FRED. Thus, in these corpora, BE GOING TO is more persistent than WILL. The opposite is true for the DS, while the two forms are exactly equally persistent in the CSPAE.
37. CG: step $\chi^2 = 7{,}990.86$, df = 13, $p < 0.001$; CSPAE: step $\chi^2 = 6{,}488.79$, df = 13, $p < 0.001$; DS: step $\chi^2 = 5{,}301.80$, df = 13, $p < 0.001$; CSAE: step $\chi^2 = 124.24$, df = 13, $p < 0.001$; FRED: $\chi^2 = 166.31$, df = 13, $p < 0.001$.
38. * significant at $p < .05$, ** $p < .01$, *** $p < .005$. Note that the good fits cannot be due merely to the large datasets – the CSPAE dataset has the best fit, yet it only has less than half the size of the CG and DS datasets.
39. One exception is the 5-word threshold in FRED, which is associated with a significant exp(b) value of greater than 1.
40. The three distance thresholds in GO-TRIGGER were conflated into a single dichotomy, namely whether or not there is such a trigger in a context of 75 words prior to CURRENT (coded 0 for such a trigger not present, and 1 for present).
41. The phenomenon will be referred to as 'particle placement,' thus avoiding the theoretical assumptions associated with the also common term 'particle movement.'
42. This phrase structure analysis follows Radford (1988: 91–100).
43. Another related factor is the number of subsequent mentions of the direct object and textual distance to the next mention: Chen (1986) has claimed that the more often the referent of the direct object of the construction is mentioned in the *subsequent* discourse, the greater the probability for *V+Part+NP* patterning. Along the same lines, Chen (1986) has argued that the earlier the referent of the direct object is mentioned in the subsequent discourse, the greater the preference for *V+NP+Part* patterning. However, these two factors will not be considered here for two reasons: for one thing, Gries (2003a: Table 6)

shows that they have a weak explanatory value. Second, while Chen's factors may make sense for written texts, they are – from a theoretical perspective – doubtful when applied to conversational data, which proceeds on a temporal axis rather than on a left-right one.

44. This subset consisted of the following texts: LAN008–LAN014, NBL001, NBL003, NBL006, NBL007, NBL008, WES001, WES002, HEB001–HEB041, SAL001–SAL039, WAR001, DUR001–DUR003, LAN001–LAN007, KEN006–KEN008, KEN014, LND001, LND002, SFK011–SFK033. These comprise approximately 1,000,000 words and thus 40% of the entire FRED corpus; dialect areas included in the sample are the Hebrides, the Midlands, the North of England, and the Southeast.
45. I am indebted to Stefan Th. Gries and Anatol Stefanowitsch for giving me access to the complete list. The p values which their list consists of were transformed mathematically into a 0–100 scale as follows: p values of verbs that Gries and Stefanowitsch found to have a preference for *V+Part+NP* were subtracted from $+2$; p values of verbs that have a preference for *V+NP+Part* were multiplied by -1. Subsequently, all values were multiplied by $+50$.
46. FRED: step $\chi^2 = 44.86$, df = 10, $p < 0.001$; CSAE: step $\chi^2 = 21.7$, df = 10, $p = 0.017$.
47. * significant at $p < .05$, ** $p < .01$, *** $p < .005$.
48. It should be noted, however, that Gries' R^2 values (for instance, those reported in Gries 2003b) are substantially higher than the ones reported here. There are some likely reasons why Gries' models account for more variance in particle placement: first, Gries' book-length analysis of particle placement is more fine-grained than the one presented here. Second, he analyzes a corpus of written and spoken Standard British English; this study's data consist of conversational American English and dialect speech in the British Isles. It is conceivable that in this study's data, particle placement simply varies more freely, or is also determined by other factors.
49. The reason is that software cannot distinguish, for example, between a *V+N* pattern (e.g. *I hate bowling*) and a genuine *V+ger.* pattern (e.g. *I hate saying that*), unless the gerund is tagged as a gerund.
50. This subset consisted of the following texts: LAN008 - LAN014, NBL001, NBL003, NBL006, NBL007, NBL008, WES001, WES002, HEB001–HEB041, SAL001–SAL 039, WAR001, DUR001–DUR003, LAN001–LAN007, KEN006–KEN008, KEN014, LND001, LND002, SFK011–SFK036, CON001–CON010, DEV001, SOM001–SOM014, DEN001–DEN004, GLA001–GLA007. These comprise ca. 1,300,000 words and thus 55% of the entire FRED corpus; dialect areas included in the sample are the Hebrides, the Midlands, the North of England, Wales, the Southwest, and the Southeast.
51. This list has been adapted from Quirk and Greenbaum (1993: 46–47) and Kolln (1994: 89–90).
52. The *-ing* form is not in this list because it is already covered by the independent variable ING-HORRORAEQUI.
53. Tokens that start in <th>, such as *the*, were not counted.
54. * significant at $p < .05$, ** $p < .01$, *** $p < .005$.
55. The following figures supplement Table 27. For simplicity, only estimates significant at $p < .05$ are displayed.
56. The following figures supplement Table 27. For simplicity, only estimates significant at $p < .05$ are displayed.

57. The following figures supplement Table 27.
58. CG: step χ^2 = 197.28, df = 15, $p < 0.001$; CSPAE: step χ^2 = 64.61, df = 15, $p < 0.001$; DS: step χ^2 = 137.49, df = 15, $p < 0.001$; CSAE: step χ^2 = 35.75, df = 15, $p = 0.002$; FRED: step χ^2 = 56.77, df = 15, $p < 0.001$.
59. * significant at $p < .05$, ** $p < .01$, *** $p < .005$.
60. Averages are not weighted by corpus size.
61. Comparing shares of explained variance across different data sources and models requires, at a minimum and among other things, that the number of predictors in each model be the same throughout. However, the models presented in the course of this study were tailored to the respective alternation studied, and thus the models differed in the kind and number of predictors included. The idea behind Figure 21 is to remedy this discrepancy by recalculating the logistic regression models presented earlier on the basis of three very basic terms only: PREVIOUS, TEXTDIST, and TEXTDIST $*$ PREVIOUS (hence, df = 3). Inclusion of TEXTDIST and TEXTDIST $*$ PREVIOUS is meant to control, for instance, for the fact that optional comparatives are less frequent than future markers, which is why persistence is weaker in comparison choice. Including textual distance controls for this distorting factor. The models underlying Figure 21 are therefore comparably basic α-persistence models; what is displayed is the average Nagelkerke R^2, i.e. the share of variance – per alternation – that is explained collectively by PREVIOUS, TEXTDIST, and TEXTDIST $*$ PREVIOUS.
62. In all, four regression runs (CG, CSPAE, DS, FRED) were conducted for comparison strategy choice, five for future marker choice (CG, CSPAE, DS, CSAE, FRED), two for both particle placement and genitive choice (CSAE and FRED), and five for complementation strategy choice (CG, CSPAE, DS, CSAE, FRED).
63. This means that in comparison strategy choice, the CSPAE ($N = 219$) and FRED ($N = 170$) datasets for comparison choice had to be omitted; in complementation strategy choice, the CSAE dataset ($N = 102$) had to be omitted.
64. This puts aside the CSAE dataset on genitive choice, where persistence lasts for more than 1,000 minutes which very much looks like a freak estimate.
65. Due to the relative infrequency of comparatives, SAMETURN and SAMESPEAKER were not included in the four regression runs on comparison strategy choice.
66. Along these lines, it seems worth mentioning that TTR appears to be a better and more potent determinant of α-persistence than SENTENCELENGTH.
67. Readers should rest assured that the author is fully aware of the fact that any serious linguist would have looked at more context anyway. (1) is somewhat curtailed for expository reasons.
68. $p < 0.005$, χ^2 = 37.1, df = 1.
69. $p < 0.001$, χ^2 = 71.7, df = 1.
70. This difference is significant at the 0.001 level (χ^2 = 13.24, df = 1).

References

Primary sources (corpora)

Corpus of Spoken American English (CSAE)
2000 Compiled by John Du Bois, Wallace Chafe, Charles Meyer, and Sandra Thompson. Distributed by the Linguistic Data Consortium.
[Installments 1 and 2: c. 166,000 words]

Corpus of Spoken Professional American English (CSPAE)
1999 Compiled by Michael Barlow. Distributed by athelstan. http://www.athel.com/cspa.html.
[2m words]

Freiburg English Dialect Corpus (FRED)
2005 Compiled by Bernd Kortmann, Susanne Wagner, Lukas Pietsch, Tanja Herrmann, and collaborators. Freiburg: English Department, University of Freiburg. http://www.anglistik.uni-freiburg.de/institut/lskortmann/FRED/.
[2.4m words]

The British National Corpus (BNC)
2000 World Edition. BNC Consortium/Oxford University Computing Services.
[100m words in total; context-governed component (CG): 6m words, demographically sampled component (DS): 4m words]

Secondary sources

Abbi, Anvita
1985 Reduplicative structures: A phenomenon of the South Asian linguistic area. In *For Gordon H. Fairbanks*, Veneeta Acson and Richard Leed (eds.), 159–171. (Oceanic Linguistics Special publication 20.) Honolulu: University of Hawaii Press.

Altenberg, Bengt
1982 *The Genitive v. the Of-Construction. A Study of Syntactic Variation in 17th Century English*. Malmö: CWK Gleerup.

Anderwald, Lieselotte, and Bernd Kortmann
2002 Typology and dialectology: A programmatic sketch. In *Present Day Dialectology*, Vol. 1, Johannes Berns and Jaap van Marle (eds.), 159–171. Berlin/New York: Mouton de Gruyter.

Aston, Guy, and Lou Burnard
1998 *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Barber, Charles
1964 *Linguistic Change in Present-Day English*. Edinburgh: Oliver and Boyd.

Bauer, Laurie
1994 *Watching English Change: An Introduction to The Study of Linguistic Change in Standard Englishes in The Twentieth Century*. London: Longman.

Bayley, Robert, and Richard Young
forthcoming VARBRUL: A special case of logistic regression. In *A Handbook of Computation Techniques in Linguistics,* Dennis Preston and Robert Bayley (eds.). Amsterdam/Philadelphia: Benjamins.

Behaghel, Otto
1909/1910 Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern [Correlations between weight and ordering of sentential constituents]. *Indogermanische Forschungen* 25: 110–142.

Berglund, Ylva
1997 Future in present-day English: Corpus-based evidence on the rivalry of expressions. *ICAME Journal* 21: 7–20.

1999a Exploiting a large spoken corpus: An end-user's way to the BNC. *International Journal of Corpus Linguistics* 4 (1): 29–52.

1999b Utilising present-day English corpora: A case study concerning expressions of future. *ICAME Journal* 24: 25–63.

2000a *Gonna* and *going to* in the spoken component of the British National Corpus. In *Corpus Linguistics and Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20)*, Christian Mair and Marianne Hundt (eds.), 35–49. Amsterdam/Atlanta: Rodopi.

2000b "You're gonna, you're not going to": A corpus-based study of colligation and collocation patterns of the *(BE) going to* construction in present-day spoken British English. In *PALC'99: Practical Applications in Language Corpora: Papers from The 2. International Conference at the University of Lódz, 15–18 April 1999*, Barbara Lewandowska-Tomaszcyk and Patrick James Melia (eds.), 161–192. Frankfurt a.M.: Peter Lang.

Biber, Douglas, Susan Conrad, and Randi Reppen
1998 *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan
1999 *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Binnick, Robert
1971 *Will* and *Be Going To*. In *Papers from the Seventh Regional Meeting of the Chicago Linguistics Society*. Chicago: Chicago Linguistics Society.

Blanken, Gerhard, Jürgen Dittmann, and Claus-W. Wallesch
1992 Studies on the "speechless man". The case of speech automatisms. In *Prehistory, History and Historiography of Language, Speech and Linguistic Theory*, Bela Brogyanyi (ed.), 339–358. Amsterdam/Philadelphia: Benjamins.

Bock, Kathryn
1986 Syntactic persistence in language production. *Cognitive Psychology* 18: 355–387.

1989 Closed-class immanence in sentence production. *Cognition* 31: 163–186.

1990 Structure in language: Creating form in talk. *American Psychologist* 45: 1221–1236.

Bock, Kathryn, and Zenzi Griffin

2000 The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General* 129 (2): 177–192.

Bock, Kathryn, and Anthony Kroch

1989 The isolability of syntactic processing. In *Linguistic Structure in Language Processing*, Greg Carlsen and Michael Tanenhaus (eds.), 157–196. Dordrecht: Kluwer.

Bock, Kathryn, and Helga Loebell

1990 Framing sentences. *Cognition* 35: 1–39.

Bolinger, Dwight

1961 Syntactic blends and other matters. *Language* 37: 366–381.

1968 *Aspects of Language*. New York: Harcourt.

1971 *The Phrasal Verb in English*. Cambridge, MA: Harvard University Press.

Bortfeld, Heather, Silvia Leon, Jonathan Bloom, Michael Schober, and Susan Brennan

2001 Disfluency rates in conversation: Effects of age, relationship, topic, role and gender. *Language and Speech* 44: 123–147.

Boyce, Suzanne, Catherine Browman, and Louis Goldstein

1987 Lexical organization and Welsh consonant mutations. *Journal of Memory and Language* 26: 419–452.

Boyland, Joyce, and John Andersen

1998 Evidence that priming is long-lasting. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Branigan, Holly, Martin Pickering, and Alexandra Cleland

1999 Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin and Review* 6 (4): 635–640.

2000a Syntactic coordination in dialogue. *Cognition* 75: 813–825.

Branigan, Holly, Martin Pickering, Simon Liversedge, Andrew Stewart, and Thomas Urbach

1995 Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research* 24: 489–507.

Branigan, Holly, Martin Pickering, Andrew Stewart, and Janet McLean

2000b Syntactic priming in spoken production: Linguistic and temporal interference. *Memory and Cognition* 28 (8): 1297–1302.

Braun, Albert

1982 *Studien zur Syntax und Morphologie der Steigerungsformen im Englischen* [Studies in the syntax and morphology of comparatives in English]. (Schweizer Anglistische Studien 110.) Bern: Francke.

Brennan, Susan, and Herbert Clark

1996 Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22 (6): 1482–1493.

Brugmann, Karl

1909 Das Wesen der lautlichen Dissimilationen [The nature of sound dissimilations]. *Abhandlungen der philologisch-historischen Klasse der königlich-sächsischen Gesellschaft der Wissenschaften* 27: 141–178.

Cedergren, Henrietta, and David Sankoff
1974 Variable rules: Performance as a statistical reflection of competence. *Language* 50: 333–355.

Chang, Franklin, Gary Dell, Kathryn Bock, and Zenzi Griffin
2000 Structural priming as implicit learning: A comparison of models of sentence production. *Journal of Psycholinguistic Research* 29 (2): 217–229.

Chen, Ping
1986 Discourse and particle movement in English. *Studies in Language* 10 (1): 79–95.

Chomsky, Noam
1965 *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
1970 Remarks on nominalization. In *Readings in English Transformational Grammar,* Roderick Jacobs and Peter Rosenbaum (eds.), 232–286. Waltham: Ginn and Company.
1986 *Knowledge of Language. Its Nature, Origin, and Use*. New York: Praeger.

Cleland, Alexandra, and Martin Pickering
2003 The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language* 49: 214–230.

Close, R.A
1988 The future in English. In *Kernprobleme der englischen Grammatik: Sprachliche Fakten und ihre Vermittlung*, Wolf-Dietrich Bald (ed.), 51–66. München: Langenscheidt-Longman.

Cohen, Laurent, and Stanislas Dehaene
1998 Competition between past and present: Assessment and interpretation of verbal perseverations. *Brain* 121: 1641–1659.

Comrie, Bernard
1982 Future time reference in the conditional protasis. *Australian Journal of Linguistics* 2: 143–152.
1985 *Tense*. Cambridge: Cambridge University Press.

Cook-Gumperz, Jenny
1977 Situated instructions: Language socialization of school age children. In *Child Discourse*, Susan Ervin-Tripp and Claudia Mitchell-Kernan (eds.), 103–121. New York: Academic Press.

Curme, George
1931 *A Grammar of the English Language*. Vol. 2: Syntax. D. C. Heath and Company.

Dahl, Lisa
1971 The *s*-genitive with non-personal nouns in modern English journalistic style. *Neuphilologische Mitteilungen* 72: 140–172.

Danchev, Andrei, and Merja Kytö
1994 The construction *be going to* + infinitive in Early Modern English. In *Studies in Early Modern English*, Dieter Kastovsky (ed.), 59–77. Berlin/New York: Mouton de Gruyter.

Danchev, Andrei, A. Pavlova, M. Nalchadjan, and O. Zlatareva

1965 The construction *going to* + inf. in modern English. *Zeitschrift für Anglistik und Amerikanistik* 13: 375–386.

Declerck, Renaat

1991 *Tense in English: Its Structure and Use in Discourse*. London/New York: Routledge.

Dehé, Nicole, Ray Jackendorff, Andrew McIntyre, and Silke Urban (eds.)

2002 *Verb-Particle Explorations*. (Interface Explorations 1.) Berlin/New York: Mouton de Gruyter.

Dell, Gary

1986 A spreading-activation theory of retrieval in sentence production. *Psychological Review* 93 (3): 283–321.

Dixon, Robert

1991 *A New Approach to English Grammar, On Semantic Principles*. Oxford: Oxford University Press.

Drews, Etta

1996 Morphological priming. *Language and Cognitive Processes* 11: 629–634.

Duffley, Patrick

1999 The use of the infinitive and the *-ing* after verbs denoting the beginning, middle and end of an event. *Folia Linguistica* 33: 295–331.

Duranti, Alessandro, and Elinor Ochs

1979 Left-dislocation in Italian conversation. In *Discourse and Syntax*, Talmy Givón (ed.), 377–416. New York: Academic Press.

Erickson, Frederick

1984 Rhetoric, anecdote, and rhapsody: Coherence strategies in a conversation among black American adolescents. In *Coherence in Spoken and Written Discourse*, Deborah Tannen (ed.), 81–154. Norwood, NJ: Ablex.

Ervin-Tripp, Susan

1979 Children's verbal turn-taking. In *Developmental Pragmatics*, Elinor Ochs and Bambi Schieffelin (eds.), 391–414. New York: Academic Press

Estival, Dominique

1985 Syntactic priming of the passive in English. *Text* 5: 7–21.

Fanego, Teresa

1996a The development of gerunds as objects of subject-control verbs in English (1400–1760). *Diachronica* 13: 29–62.

1996b On the historical development of English retrospective verbs. *Neuphilologische Mitteilungen* 97: 71-79.

1997 On patterns of complementation with verbs of effort. *English Studies* 78: 60–67.

Fraser, Bruce

1965 An examination of the verb-particle construction in English. Ph. D. diss., Massachusetts Institute of Technology.

1974 Review of Dwight Bolinger, *The Phrasal Verb in English*. *Language* 50 (3): 568–575.

Freed, Alice

1979 *The Semantics of English Aspectual Complementation*. Dordrecht: Reidel.

Friederici, Angela, Herbert Schriefers, and Ulman Lindenberger

1998 Differential age effects on semantic and syntactic priming. *International Journal of Behavioral Development* 22 (4): 813–845.

Garrod, Simon, and Martin Pickering

2004 Why is conversation so easy? *Trends in Cognitive Science* 8 (1): 8–11.

Giles, Howard

1980 Accommodation theory: Some new directions. In *Aspects of Linguistic Behaviour: A Festschrift in Honour of Robert Le Page*, S. De Silva (ed.), 105–136. York: University of York Press.

Givón, Talmy

1993 *English Grammar. A Function-Based Introduction*. Amsterdam/Philadelphia: Benjamins.

Goodwin, Marjorie Harness

1983 Aggravated correction and disagreement in children's conversations. *Journal of Pragmatics* 7: 657–677.

Gramley, Stephan

1980 Infinitive and gerund complements with the verbs *begin* and *start*. *Arbeiten aus Anglistik und Amerikanistik* 5: 159–186.

Gries, Stefan Th.

1999 Particle movement: A cognitive and functional approach. *Cognitive Linguistics* 10: 105–145.

2002 The influence of processing on syntactic variation: Particle placement in English. In *Verb-Particle Explorations*, Nicole Dehé, Ray Jackendorff, Andrew McIntyre, and Silke Urban (eds.), 269–288. Berlin/New York: Mouton de Gruyter.

2003a Grammatical variation in English: A question of 'structure vs. function'? In *Determinants of Grammatical Variation in English*, Günter Rohdenburg and Britta Mondorf (eds.), 155–174. Berlin/New York: Mouton de Gruyter.

2003b *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London/New York: Continuum Press.

2005 Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research* 34 (4): 365–399.

Gries, Stefan Th., and Anatol Stefanowitsch

2004 Extending collustructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9 (1): 97–129.

Haegeman, Liliane

1983 Be going to, gaan, and aller: Some observations on the expression of future time. *International Review of Applied Linguistics* 11 (2): 155–157.

1989 *Be going to* and *will*: A pragmatic account. *Journal of Linguistics* 25 (2): 291–317.

Hall, R.M.R., and Beatrice Hall

1970 A note on *will* vs. *going to*. *Linguistic Inquiry* 1: 138–139.

Halliday, Michael

1991 Corpus studies and probabilistic grammar. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Karin Aijmer and Bengt Altenberg (eds.), 30–43. London: Longman.

Halliday, Michael, and Ruqaiya Hasan
1976 *Cohesion in English*. London: Longman.

Hartsuiker, Robert, and Casper Westenberg
2000 Word order priming in written and spoken sentence production. *Cognition* 75: B27–B39.

Hasher, Lynn, and Rose Zacks
1988 Working memory, comprehension, and aging: A review and a new view. In *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 22, G. Bower (ed.), 193–225. New York: Academic Press.

Hatch, Evelyn, Sabrina Peck, and Judy Wagner-Gough
1979 A look at process in child second-language acquisition. In *Developmental Pragmatics*, Elinor Ochs and Bambi Schieffelin (eds.), 269–278. New York: Academic Press.

Hawkins, John
1994 *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.
1999 Processing complexity and filler-gap dependencies across grammars. *Language* 75: 244–285.

Hertzog, Christopher
1991 Aging, information processing speed, and intelligence. In *Annual Review of Gerontology and Geriatrics*, Vol. 11, K. W. Schaie (ed.), 55–79. New York: Springer.

Hopper, Paul
1998 Emergent grammar. In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, Michael Tomasello (ed.), 155–176. Mahwah, NJ/London: Erlbaum Associates.

Hudson, Richard
1984 *Word Grammar*. Oxford: Blackwell.
1997 Inherent variability and linguistic theory. *Cognitive Linguistics* 8 (2): 73–108.

Hundt, Marianne
1997 Has BrE been catching up with AmE over the past thirty years? In *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora*, Magnus Ljung (ed.), 135–151. Amsterdam: Rodopi.
1998 *New Zealand English Grammar – Fact or Fiction? A Corpus-Based Study in Morphosyntactic Variation*. Amsterdam/Philadelphia: Benjamins.

Jaccard, James
2001 *Interaction Effects in Logistic Regression*. (Quantitative Applications in the Social Sciences 135.) Thousand Oaks: Sage Publications.

Jefferson, Gail
1972 Side sequences. In *Studies in Social Interaction*, David Sudnow (ed.), 294–338. New York: Free Press.

Jespersen, Otto
1909 *A Modern English Grammar on Historical Principles*. Vol. 7: Syntax. Copenhagen: Munksgaard.

Johnstone, Barbara
1984 Repeating yourself: Discourse paraphrase and the generation of language. In *Proceedings of the Eastern States Conference on Linguistics*, 250–259. Columbus, OH: Department of Linguistics, Ohio State University.

Jucker, Andreas
1993 The genitive versus the *of*-construction in newspaper language. In *The Noun Phrase in English. Its Structure and Variability*, Andreas Jucker (ed.), 121–136. Heidelberg: Carl Winter.

Keenan, Elinor
1977 Making it last: Repetition in children's discourse. In *Child Discourse*, Susan Ervin-Tripp and Claudia Mitchell-Kernan (eds.), 125–138. New York: Academic Press.

Kemper, Susane
1992 Language and aging. In *The Handbook of Aging and Cognition*, Fergus Craik and Timothy Salthouse (eds.), 213–270. Hillsdale, NJ: Erlbaum.

Kempley, S.T., and John Morton
1982 The effects of priming with regularly and irregularly related words in auditory word recognition. *British Journal of Psychology* 73: 441–445.

Kempson, Ruth
1977 *Semantic Theory*. Cambridge: Cambridge University Press.

Kennedy, Arthur
1920 *The Modern English Verb-Adverb Combination*. Stanford: Stanford University Press.

Kjellmer, Goran
1998 On contraction in modern English. *Studia Neophilologica* 69: 155–186.

Kolln, Martha
1994 *Understanding English Grammar*. 4th ed. New York: MacMillan.

Konopka, Agnieszka, and Kathryn Bock
in press Structural persistence from idiom production.

Kortmann, Bernd
2002 New prospects for the study of dialect syntax: Impetus from syntactic theory and language typology. In *Syntactic Microvariation*, Sief Barbiers, Leonie Cornips, and Susanne van der Kleij (eds.), 185–213. Amsterdam: Meertens Instituut.

2003 Comparative English dialect grammar: A typological approach. In *Fifty Years of English Studies in Spain (1952:2002). A Commemorative Volume*, Ignacio Palacios, María José López Couso, Patricia Fra, and Elena Seoane (eds.), 65–83. Santiago de Compostela: University of Santiago.

2004 *Do* as a tense and aspect marker in varieties of English. In *Dialectology meets Typology: Dialect Grammar from a Cross-Linguistic Perspective*, Bernd Kortmann (ed.), 245–275. Berlin/New York: Mouton de Gruyter.

Krug, Manfred
2000 *Emerging English Modals: A Corpus-Based Study of Grammaticalization*. Berlin/New York: Mouton de Gruyter.

Kruisinga, Etsko, and P.A. Erades
1953 *An English Grammar*. Vol. 1. 8th ed. Groningen: Noordhoff.

Kuryłowicz, Jerzy
1964 *The Inflectional Categories of Indo-European*. Heidelberg: Carl Winter.

Kytö, Merja
1990 *Shall* or *will*? Choice of the variant form in early modern English, British and American. In *Historical Linguistics 1987: Papers from the 8th International Conference on Historical Linguistics*, Henning Andersen and Konrad Koerner (eds.), 275–88. Amsterdam/Philadelphia: Benjamins.

Kytö, Merja, and Susan Romaine
1997 Competing forms of adjective comparison in modern English: What could be *more quicker* and *easier* and *more effective*? In *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, Terttu Nevalainen and Leena Kahlas-Tarkka (eds.), 329–352. Amsterdam: Rodopi.

Labov, William
1966a The linguistic variable as a structural unit. *Washington Linguistics Review* 3: 4–22.
1966b *The social stratification of English in New York City*. Washington D.C.: Center for Applied Linguistics.
1969 Contraction, deletion, and inherent variability of the English copula. *Language* 45: 715–762.

Langacker, Ronald
1987 *Foundations of Cognitive Grammar*. Volume 1: Theoretical prerequisites. Stanford: Stanford University Press

Laver, Gary, and Deborah Burke
1993 Why do semantic priming effects increase in old age? A meta-analysis. *Psychology and Aging* 8: 34–43.

Leech, Geoffrey, and Jonathan Culpeper
1997 The comparison of adjectives in recent British English. In *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, Terttu Nevalainen and Leena Kahlas-Tarkka (eds.), 125–132. Amsterdam: Rodopi.

Levelt, Willem
1983 Monitoring and self-repair in speech. *Cognition* 14: 41–104.
1989 *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt, Willem, and Stephanie Kelter
1982 Surface form and memory in question answering. *Cognitive Psychology* 14: 78–106.

Levin, Samuel
1982 Are figures of thought figures of speech? In *Contemporary Perceptions of Language: Interdisciplinary Dimensions*, Heidi Byrnes (ed.), 112–123. Washington D.C.: Georgetown University Press.

Lindquist, Hans
2000 *Livelier* or *more lively*: Syntactic and contextual factors influencing the comparison of disyllabic adjectives. In *Corpora Galore: Analyses and Techniques in Describing English*, John Kirk (ed.), 125–132. Amsterdam: Rodopi.

Mair, Christian

1997a Parallel corpora: A real-time approach to the study of language change in progress. In *Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora*, Magnus Ljung (ed.), 195–209. Amsterdam: Rodopi.

1997b The spread of the *going-to*-future in written English: A corpus-based investigation into language change in progress. In *Language History and Linguistic Modelling*, Raymond Hickey and Stanislaw Puppel (eds.), 1537–1543. Berlin/New York: Mouton de Gruyter.

2002 Three changing patterns of verb complementation in Late Modern English: A real-time study based on matching text corpora. *English Language and Linguistics* 6 (1): 105–131.

2003 Gerundial complements after *begin* and *start*: Grammatical and sociolinguistic factors, and how they work against each other. In *Determinants of Grammatical Variation in English*, Günter Rohdenburg and Britta Mondorf (eds.), 329–345. Berlin/New York: Mouton de Gruyter.

2004 Corpus linguistics and grammaticalisation theory: Beyond statistics and frequency? In *Corpus Approaches to Grammaticalisation in English*, Christian Mair and Hans Lindquist (eds.), 121–150. Amsterdam/Phildalephia: Benjamins.

Martin, Willy, and Jan Weltens

1973 A frequency-note on the expression of futurity in English. *Zeitschrift für Anglistik und Amerikanistik* 21: 289–98.

Matthews, Robert

1979 Are the grammatical sentences of a language a recursive set? *Synthese* 40: 209–224.

McKone, Elinor

1995 Short-term implicit memory for words and non-words. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21: 1108–1126.

Menard, Scott

2002 *Applied Logistic Regression Analysis*. (Quantitative Applications in the Social Sciences 106.) 2d ed. Thousand Oaks: Sage Publications.

Merritt, Marilyn

1982 Repeats and reformulations in primary classrooms as windows of the nature of talk engagement. *Discourse Processes* 5: 127–145.

Meyer, David, and Roger Schvaneveldt

1971 Facilitation in recognizing pairs of words: Evidence of dependence between retrieval operations. *Journal of Experimental Psychology* 90: 227–234.

Miller, Jim, and Regina Weinert

1998 *Spontaneous Spoken Language. Syntax and Discourse*. Oxford: Clarendon Press.

Milsark, Gary

1988 Re: Doubl-ing. *Linguistic Inquiry* 3: 542–549.

Mitchell, Bruce

1985 *Old English*. Vol. 1. Oxford: Clarendon Press.

Mondorf, Britta

2000 *Wider-ranging* vs. *more old-fashioned* views on comparative formation in adjectival compounds/derivatives. In *Proceedings of the Anglistentag 1999*, Bernhard Reitz and Sigrid Rieuwerts (eds.), 35–44. Trier: Wissenschaftlicher Verlag Trier.

2002 The effect of prepositional complements on the choice of synthetic or analytic comparatives. In *Perspectives on Prepositions*, Hubert Cuyckens and Günter Radden (eds.), 65–78. Tübingen: Niemeyer.

2003 Support for *more*-support. In *Determinants of Grammatical Variation in English*, Günter Rohdenburg and Britta Mondorf (eds.), 251–304. Berlin/New York: Mouton de Gruyter.

Nicol, Janet

1996 Syntactic priming. *Language and cognitive processes* 11: 675–679.

Nicolle, Steve

1997 A relevance-theoretic account of *be going to*. *Linguistics* 33: 355–377.

Ochs, Elinor

1979 Planned and unplanned discourse. In *Discourse and Syntax,* Talmy Givón (ed.), 51–80. New York: Academic Press.

Orwin, Robert

1994 Evaluating coding decisions. In *The Handbook of Research Synthesis*, Harris Cooper and Larry Hedges (eds.), 139–162. New York: Russell Sage Foundation.

Osselton, Noel

1988 Thematic genitives. In *An Historic Tongue: Studies in English Linguistics in Memory of Barbara Strang*, Graham Nixon and John Honey (eds.), 138–144. London: Routledge.

Palmer, Frank

1974 *The English Verb*. London: Longman.

Pampel, Fred

2000 *Logistic Regression. A Primer*. (Quantitative Applications in the Social Sciences 132.) Thousand Oaks: Sage Publications.

Pickering, Martin, and Holly Branigan

1998 The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language* 39: 633–651.

Pickering, Martin, Holly Branigan, Alexandra Cleland, and Andrew Stewart

2000 Activation of syntactic information during language production. *Journal of Psycholinguistic Research* 29 (2): 205–216.

Pickering, Martin, and Simon Garrod

2004 Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27: 169–226.

Polanyi, Livia

1979 False starts can be true. In *Proceedings of the 4th Annuary Meeting of the Berkeley Linguistics Society*, 628–639. Berkeley: Linguistics Society.

Poplack, Shana
1980 The notion of the plural in Puerto Rican English: Competing constraints on (s) deletion. In *Locating Language in Time and Space*, William Labov (ed.), 55–68. New York: Academic Press.

Poplack, Shana, and Sali Tagliamonte
1993 The zero-marked verb: Testing the creole hypothesis. *Journal of Pidgin and Creole Languages* 8: 171–206.
1996 Nothing in context: Variation, grammaticization and past time marking in Nigerian Pidgin English. In *Changing Meanings, Changing Functions*, Philip Baker and Anand Syea (eds.), 71–94. Westminster: University Press.

Postma, Albert
2000 Detection of errors during speech production: A review of speech monitoring models. *Cognition* 77: 97–131.

Potter, Mary, and Linda Lombardi
1998 Syntactic priming in immediate recall of sentences. *Journal of Memory and Language* 28: 265–282.

Potter, Simeon
1969 *Changing English*. London: Deutsch.

Pound, Luise
1901 *The Comparison of Adjectives in English in the XV and the XVI Century*. (Anglistische Forschungen 7.) Heidelberg: Carl Winter.

Poutsma, Hendrik
1914 *A Grammar of Late Modern English*. Groningen: Noordhoff.

Pullum, Geoffrey, and Arnold Zwicky
1999 Gerund participles and head-complement inflection conditions. In *The Clause in English: In Honour of Rodney Huddleston*, Peter Collins and David Lee (eds.), 251–271. Amsterdam/Philadelphia: Benjamins.

Quirk, Randolph
1974 *The Linguist and the English Language*. London: Arnold.

Quirk, Randolph, and Sidney Greenbaum
1993 *A University Grammar of English*. Harlow: Longman.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik
1985 *A Comprehensive Grammar of the English Language*. London: Longman.

Raab-Fischer, Roswitha
1995 Löst der Genitiv die *of*-Phrase ab? Eine korpusgestützte Studie zum Sprachwandel im heutigen Englisch [Is the genitive replacing the *of*-construction? A corpus-based study of language change in Modern English]. *Zeitschrift für Anglistik und Amerikanistik* 43: 123–132.

Radford, Andrew
1988 *Transformational Grammar. A First Course*. Cambridge: Cambridge University Press.

Rastle, Kathleen and Deborah Burke
1996 Priming the tip of the tongue: Effects of prior processing on word retrieval in young and older adults. *Journal of Memory and Language* 35: 586–605.

Rohdenburg, Günter
1995 On the replacement of finite complement clauses by infinitives in English. *English Studies* 76: 367–388.

1996 Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7: 149–182.

1998 Syntactic complexity and the variable use of *to be* in 16th to 18th century English. *Arbeiten aus Anglistik und Amerikanistik* 23: 199–228.

2003 Cognitive complexity and horror aequi as factors determining the use of interrogative clause linkers in English. In *Determinants of Grammatical Variation in English*, Günter Rohdenburg and Britta Mondorf (eds.), 205–249. Berlin/New York: Mouton de Gruyter.

Rohdenburg, Günter, and Britta Mondorf (eds.)
2003 *Determinants of Grammatical Variation in English.* Berlin/New York: Mouton de Gruyter.

Rohdenburg, Günter, and Julia Schlüter
2000 Determinanten grammatischer Variation im Früh- und Spätneuenglischen [Determinants of grammatical variation in Early and Late Modern English]. *Sprachwissenschaft* 25 (4): 443–496.

Rohr, Anny
1929 Die Steigerung des neuenglischen Eigenschaftswortes im 17. und 18. Jahrhundert mit Ausblicken auf den Sprachgebrauch der Gegenwart [On the comparison of the adjective in Modern English during the 17th and 18th century, with a perspective on language use today]. Ph. D. diss., University of Giessen.

Rosenbach, Anette
2003 Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In *Determinants of Grammatical Variation in English*, Günter Rohdenburg and Britta Mondorf (eds.), 379–412. Berlin/New York: Mouton de Gruyter.

2005 Animacy versus weight as determinants of grammatical variation in English. *Language* 81: 613–644.

Ross, John
1972 Doubl-ing. *Linguistic Inquiry* 3: 61–86.

Sacks, Harvey
1971 Unpublished Lecture Notes.

Sacks, Harvey, Emanuel Schegloff, and Gail Jefferson
1974 A simplest systematics for the organisation of turn-taking in conversation. *Language* 50: 696–735.

Saffran, Eleanor, and Nadine Martin
1997 Effects of structural priming on sentence production in aphasics. *Language and Cognitive Processes* 12: 877–882.

Sankoff, David
1998 Sociolinguistics and syntactic variation. In *Linguistics: The Cambridge Survey*, Vol. 4, Frederick Newmeyer (ed.), 140–161. Cambridge: Cambridge University Press

Sankoff, David, and Suzanne Laberge

1978 Statistical dependence among successive occurrences of a variable in discourse. In *Linguistic Variation: Models and Methods*, David Sankoff (ed.), 119–126. New York: Academic Press.

Sankoff, David, and William Labov

1979 On the use of variable rules. *Language in Society* 8: 189–222.

Schenkein, Jim

1980 A taxonomy for repeating action sequences in natural conversation. In *Language Production*, Vol. 1, Brian Butterworth (ed.), 21–47. New York: Academic Press.

Scherre, Maria, and Anthony Naro

1991 Marking in discourse: "Birds of a feather". *Language Variation and Change* 3: 23–32.

Schiffrin, Deborah

1982 *Cohesion in Discourse: The Role of Non-Adjacent Paraphrase*. (Working Papers in Sociolinguistics 97.) Austin: Southwest Educational Development Laboratory.

1987 *Discourse Markers*. Cambridge: Cambridge University Press.

Shepherd, Susan

1985 On the functional development of repetition in Antiguan Creole morphology, syntax, and discourse. In *Historical Semantics / Historical Word Formation*, Jacek Fisiak (ed.), 533–545. Berlin/New York: Mouton de Gruyter.

Silverstein, Michael

1976 Hierarchy of features and ergativity. In *Grammatical Categories in Australian Languages*, Robert Dixon (ed.), 112–171. Canberra: Australian Institute of Aboriginal Studies.

Stefanowitsch, Anatol

2003 Constructional semantics as a limit to grammatical alternation: The two genitives of English. In *Determinants of Grammatical Variation in English*, Günter Rohdenburg and Britta Mondorf (eds.), 413–441. Berlin/New York: Mouton de Gruyter.

Strang, Barbara

1968 *Modern English Structure*. 2d ed. London: Arnold.

Sweet, Henry

1892 *A New English Grammar: Logical and Historical*. Oxford: Clarendon Press.

Szmrecsanyi, Benedikt

2003 *Be going to* versus *will/shall*: Does syntax matter? *Journal of English Linguistics* 31 (4): 295–323.

2004 On operationalizing syntactic complexity. In *Le Poids des Mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Louvain-la-Neuve, March 10–12, 2004*, Vol. 2, Gérard Purnelle, Cédrick Fairon, and Anne Dister (eds.), 1032–1039. Louvain-la-Neuve: Presses universitaires de Louvain.

2005 Language users as creatures of habit: A corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory* 1 (1): 113–149.

Tanenhaus, Michael, Helen Flanigan, and Mark Seidenberg

1980 Orthographic and phonological activation in auditory and visual word recognition. *Memory and Cognition* 8: 513–520.

Tannen, Deborah

1982 Oral and literate strategies in spoken and written narratives. *Language* 58: 1–21.

1987 Repetition in conversation: Toward a poetics of talk. *Language* 63: 574–605.

1989 *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*. Cambridge: Cambridge University Press.

Taylor, John

1989 Possessive genitives in English. *Linguistics* 27: 663–686.

Tottie, Gunnel

2002 Non-categorical differences between American and British English: Some corpus evidence. In *Studies in Mid-Atlantic English*, Marko Modiano (ed.), 37–58. Gävle: University of Gävle Press.

Trudgill, Peter

1984 Standard English in England. In *Language in the British Isles*, Peter Trudgill (ed.), 32–44. Cambridge: Cambridge University Press.

Tyler, Lorraine, and William Marslen-Wilson

1977 The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior* 16: 683–692.

Van Dongen, Wilhelmus

1919 *He puts on his hat* and *He puts his hat on*. *Neophilologus* 4: 322–353.

Vosberg, Uwe

2003 The role of extractions and horror aequi in the evolution of *-ing*-complements in modern English. In *Determinants of Grammatical Variation in English*, Günter Rohdenburg and Britta Mondorf (eds.), 305–327. Berlin/New York: Mouton de Gruyter.

Řeřicha, Václav

1987 Notes on infinitival and *-ing* complements of the verbs *begin* and *start*. *Philologica Pragensia* 30: 129–132.

Wasow, Thomas

1997 Remarks on grammatical weight. *Language Variation and Change* 9: 81–105.

Watson-Gegeo, Karen Ann, and Stephen Boggs

1977 From verbal play to talk story: The role of routines in speech events among Hawaiian children. In *Child Discourse*, Susan Ervin-Tripp and Claudia Mitchell-Kernan (eds.), 57–90. New York: Academic Press.

Weiner, Judith, and William Labov

1983 Constraints on the agentless passive. *Journal of Linguistics* 19: 29–58.

Wekker, Herman

1976 *The Expression of Future Time in Contemporary British English*. Amsterdam: North Holland.

Wheeldon, Linda, and Stephen Monsell

1992 The locus of repetition priming of spoken word production. *Quarterly Journal of Experimental Psychology* 44A (4): 723–761.

Wheeldon, Linda, and Mark Smith
2003 Phrase structure priming: A short lived effect. *Language and Cognitive Processes* 18 (4): 431–442.

Wierzbicka, Anna
1998 The semantics of English causative constructions in a universal-typological perspective. In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, Michael Tomasello (ed.), 113–153. Mahwah, NJ/London: Erlbaum Associates.

Wright, Barton, and Merrill Garret
1984 Lexical decision in sentences: Effects of syntactic structure. *Memory and Cognition* 12: 31–45.

Zurif, Edgar, David Swinney, Penny Prather, Arthur Wingfield, and Hirma Brownell
1995 The allocation of memory resources during sentence comprehension: Evidence from the elderly. *Journal of Psycholinguistic Research* 24: 165–182.

Zwitserlood, Pienie
1996 Form priming. *Language and Cognitive Processes* 11: 589–596.

Index

BMN.06