Frequency, Chunks and Hesitations

A Usage-based Analysis of Chunking in English

Ulrike Schneider

NIHIN studies
New Ideas in Human Interaction

Albert-Ludwigs-Universität Freiburg
Universitätsbibliothek

Frequency, Chunks and Hesitations

A Usage-Based Analysis of Chunking in English

Inaugural-Dissertation
zur
Erlangung der Doktorwürde
der Philologischen Fakultät
der Albert-Ludwigs-Universität
Freiburg i. Br.

vorgelegt von
Ulrike Schneider
aus Darmstadt

SS 2014

Titel der eingereichten Fassung:
Frequency, Hesitations and Chunks – A Usage-based Study of Chunking in English

Erstgutachter: Prof. Dr. Benedikt Szmrecsanyi
Zweitgutachter: Prof. Dr. Peter Auer
Drittgutachter: Prof. Dr. Dr. h.c. Christian Mair

Vorsitzender des Promotionsausschusses
der Gemeinsamen Kommission der
Philologischen, Philosophischen und Wirtschafts-
und Verhaltenswissenschaftlichen Fakultät: Prof. Dr. Bernd Kortmann

Datum der Disputation: 10. Juli 2014

To Urs Lufft

Above all for encouraging me to be inquisitive.

Acknowledgments

My thanks go first and foremost to my supervisors, Benedikt Szmrecsanyi and Peter Auer, for the continuous support and guidance they have provided over the years. I am also grateful for the funding, resources, support and linguistic network provided by my graduate school, the DFG-funded "Research Training Group for Frequency Effects in Language" (DFG GRK 1625/1) at the University of Freiburg. I particularly want to thank the school's spokesperson Stefan Pfänder for always holding his protecting hand over us (in his own lovely words).

I greatly profited from being able to present my work at various conferences and workshops and receive valuable suggestions and comments. For this, I am particularly grateful to the linguistic communities at the Universities of Freiburg and Mainz and the many academic guests who visited us over the years.

A lot of this work would not have been possible without a knowledge of statistics and scripting languages. Thus I would like to thank Sascha Wolfer, Stefan Th. Gries, Harald Baayen, Volya Kapatsinski and Carolin Strobl for answering my various questions about R in general and Classification and Regression trees and random forests in particular. I also wish to thank the bwGrid for making their computational resources available to me and Janne Schulz for providing the most patient technical support.

Special thanks go to Matthias Eitelmann, Florian Dolberg and Sebastian Schmidt for reading large parts of this work at various drafting stages and offering valuable feedback and suggestions. I am also extremely grateful to Michael Schäfer for his calming reassurance during the last stages of the writing process.

Finally, a big personal 'Thank You' to my parents, my brothers and all my friends in Freiburg, Mainz and wherever else you are. You are my backbone and you kept me going.

Despite the help I received, any errors remain, of course, entirely my own.

Zusammenfassung in deutscher Sprache

Die vorliegende Studie befasst sich aus anwendungsbasierter Sicht mit der mentalen Verarbeitung von Mehrwortsequenzen. Dem liegt die Annahme zugrunde, dass Abfolgen von Wörtern durch häufige Wiederholung zu Routinen werden können. Diese sind dann schneller mental abrufbar als selten verwendete Kombinationen. Frequente Sequenzen wie *I don't know* oder *I'm trying* sollten daher schneller und flüssiger produziert werden können als selten verwendete Kombinationen wie *I don't recall* oder *I am attempting* oder gar nie zuvor verwendete Abfolgen.

Routinen wie *I don't know* werden auch als 'Chunks' bezeichnet. Im Fokus dieser Studie steht die Chunkingtheorie von Bybee (2007b), die Beschreibt, dass mit zunehmender Wiederholung mentale Verbindungen zwischen den Wörtern in einer Sequenz stärker werden. Chunking beginnt damit schon bei niederfrequenten Sequenzen und wird immer intensiver, je frequenter die Sequenz ist. Hoch frequente Kombinationen sind nach Bybee schließlich einfach als größere Einheit aus dem mentalen Lexikon abrufbar. Bybee geht zudem davon aus, dass der Grad der 'Gechunktheit' einer Sequenz entscheidend von deren absoluter Vorkommenshäufigkeit abhängt. Sie postuliert, dass Übergangswahrscheinlichkeiten und andere Maße der Kookkurrenz-Wahrscheinlichkeit wenn überhaupt nur eine untergeordnete Rolle spielen.

Die vorliegende Studie vergleicht auf Basis dieser und weiterer gebrauchsbasierter Theorien den Einfluss absoluter Vorkommenshäufigkeit und relativer Kookkurrenz-Maße auf die mentale Verarbeitung von Mehrwortsequenzen. Als Indikator werden dabei Zögerungssignale herangezogen. Sprecher sollten während der Artikulation von stark gechunkten Sequenzen keine Zögerungssignale benötigen und es vorziehen, Chunks ohne Unterbrechung zu artikulieren. Falls gefüllte sowie ungefüllte Pausen und Diskursmarker gebraucht werden um Zeit zu schaffen für die Bewältigung von Planungsschwierigkeiten, so sollten diese nicht innerhalb von starken Chunks platziert werden, sondern stattdessen zwischen Wörtern, die weniger stark mental verknüpft sind. So lässt sich beispielsweise erklären, dass in den folgenden Beispielen das hoch frequente *we've got* als ununterbrochene Einheit artikuliert wird, während *we've enjoyed*, eine sehr viel seltenere Konstruktion, unterbrochen wird.

(1) *you know* we've got (sw2331.A.s133)

(2) we've *uh [pause]* enjoyed (sw2316.A.s154)

Die Studie untersucht die Platzierung von Zögerungssignalen im Switchboard NXT Korpus. Bei diesem Korpus handelt es sich um Transkripte von Telefonkonversationen

im amerikanischen Englisch. Aus diesem Korpus wurden mehr als 11.000 gefüllte und ungefüllte Pausen sowie Diskursmarker, die im Kontext von Präpositionalphrasen und Satzanfangsstrukturen auftreten, extrahiert. Als Analyseinstrumente dienen bei dieser Untersuchung Classification and Regression Trees sowie Random Forests. Diese Regressionsmethoden teilen Daten anhand ihrer Charakteristika in immer kleinere Untergruppen ein und können so Rückschlüsse darüber zulassen, unter welchen Bedingungen verschiedene Sprecher die gleichen Entscheidungen treffen. Für eine solche Analyse sind diese neuen Verfahren besonders geeignet, da sie berücksichtigen können, dass den Sprechern mehr als nur zwei Optionen zur Platzierung von Zögerungssignalen zur Verfügung stehen.

Die Studie zeigt, dass Kookkurrenzfrequenzen und andere Maße der Kookkurrenz-Wahrscheinlichkeit in sieben der 14 ausgewählten Kontexte einen signifikanten Einfluss auf die Platzierung von Zögerungssignalen haben. Damit ist ein genereller Nachweis für die kognitive Realität von Chunking erbracht.

Im Detail zeigen die Ergebnisse, dass sich mit diesem Instrument gerade Chunking zwischen den Wörtern zu beiden Seiten der Präpositionalphrasengrenze nachweisen lässt und auch dass starkes Chunking in diesem Kontext sehr häufig vorkommt. Zudem findet sich Evidenz für bisher kaum berücksichtigte Phänomene: Die 'Chunking-inventare' verschiedener Sprecher einer Sprache überlappen zwar größtenteils – bedingt durch die Strukturen der Sprache – aber jeder Sprecher verfügt zusätzlich über Chunks, die sich aus seinen Lebensumständen ergeben. In dieser Datenbank traf das besonders auf Namen von Wohnorten (z.B. *Boise, Idaho*) und Arbeitgebern zu (z.B. *Richardson Symphony*). Zudem können Zögerungssignale am Satzanfang selbst zu einem Teil eines Chunks werden. So entstehen Elemente wie *and-uh*, die für längeres Zögern genutzt werden können.

Die Resultate untermauern die Hypothese, dass Chunking ein gradueller Prozess ist, der nicht nur hoch frequente Sequenzen betrifft. Es finden sich keine Anhaltspunkte, dass bei der Verarbeitung kategorisch zwischen Chunks und nicht-Chunks unterschieden wird. Stattdessen zeigt sich, dass Verknüpfungen mit zunehmender Frequenz gestärkt werden. Für Bybees These, dass die Sequenz als Ganzes gespeichert wird, finden sich keine eindeutigen Anzeichen. Im Gegenteil weisen die Regressionen nach, dass es teilweise auch am hoch frequenten Ende des Spektrums noch Unterschiede zwischen dem 'Anziehungsgrad' zwischen Wörtern gibt, was auf immer weiter ansteigende Attraktionsgrade hinweist.

Eine weitere zentrale Erkenntnis ist, dass absolute Kookkurrenzfrequenz nicht als bester Prädiktor abschneidet. Übergangswahrscheinlichkeiten und andere Kookkurrenz-

maße erweisen sich als ebenso gute Prädiktoren für Chunking wie absolute Frequenz und sind in einigen Kontexten sogar bessere Prädiktoren.

Contents

List of Tables

List of Figures

List of Tables

1 Introduction

The present study provides statistical evidence that frequency of use shapes the mental representation of multi-word expressions and, furthermore, that hesitations are the visible footprints of this process. After all, "language users are creatures of habit" (Szmrecsanyi 2006:1), who prefer to use the same set of structures over and over, thus developing skilled routines, which are no longer interrupted by hesitations.

Speakers rarely use the full extent of their creative potential (cf. Pawley and Syder 1983:193). Instead, they tend to resort to well-practised phrases and expressions (cf. Sinclair 1991:110). Thus, native speakers of English will opt for the greetings in (1) and (2) rather than aiming for the creative and novel ones in (3).

(1) Hi, how are you?

(2) Good morning! How are you?

(3) Pleasant start to the day! How is your health?

The fact that speakers prefer to use expressions they have encountered before (cf. Bybee 2010:53), process these faster (cf. Arnon and Snider 2010) and produce them more fluently than novel creative formations (cf. Pawley and Syder 1983; Erman 2007) suggests that recurrent sequences are mentally logged in some way and can therefore be accessed faster and more easily than uncommon or entirely new sequences.

One explanation for this phenomenon is that sequences such as *good morning* and *how are you* are routine behaviours. A routine develops quite generally and in any context (i.e. linguistic and non-linguistic) when, through frequent repetition, motor activities become 'chunked', that is welded into one longer sequence (cf. Langacker 2000:3-4). Thus, for example, the series of movements required for tying shoelaces or knitting can be practiced until they can be performed as one fast, precise sequence. In the same way, speakers are so practiced at using linguistic chunks that they no longer have to assemble *how, are* and *you* into a sentence, but can retrieve the entire sequence as a single unit from the mental lexicon. In fact, this principle is so prevalent that it has been estimated that up to 80 per cent of language output is made up of chunks and other sorts of pre-constructed units (cf. Altenberg 1998:102).

The placement of hesitations such as filled and unfilled pauses provides information about how practised or 'chunked' the sequences surrounding the hesitation are. These signals of planning problems are not scattered randomly throughout speech, but follow from the speaker's familiarity with the constructions he or she is using. If one word in a chunk evokes all others, speakers should not hesitate within a chunk. In other words,

chunks should not be interrupted by hesitations. As a result, the following hesitant versions of the chunks in (1) and (2) sound very unnatural.

(4) *how are *um* you?

(5) *good *uh* morning

As such, hesitations are a valuable, yet largely unexplored, source of evidence of the mental reality of chunks. The case studies presented in this book not only use hesitations to confirm the claim that frequency plays a central role in chunking, but also explore how these effects can best be modelled.

1.1 Scope

I define chunking as

> the process whereby a sequence of words become mutually ever more strongly represented. This means that in the mental network, the associations between their respective nodes become stronger until the sequence may eventually be retrievable as a single unit.

Thus a chunk is

> a mentally represented multi-word unit. This representation is either in the form of strong mutual activation or a combined node in the network.

This model builds on Bybee's (2007b) Linear Fusion Hypothesis which states that "items that are used together fuse together" (Bybee 2007b:316). Bybee hypothesises that there is a "sequential link" between items that are used together, which is strengthened through repeated use (Bybee 2007b:318, 319), so that the components of frequent sequences prime and automate each other (Bybee 2007b:316). According to Bybee, this process, also referred to as 'chunking', starts with the first encounter of a sequence and leads to the fusion of the sequence into a single unit which can be stored in memory (Bybee 2007b:324).

Bybee's model of the mind is a usage-based exemplar model (2006; 2010), which assumes that linguistic items are mentally connected on various levels and dimensions (Bybee 2010:22), thus forming a network of associations. The network is highly redundant, incorporating entries for both complex and simple items (Bybee 2010:24). Item representations are strengthened by repeated exposure to a phenomenon (Bybee 2010:19). Within this framework, Bybee defines chunking as sequential relations

developing between co-occurring words (Bybee 2010:25; 33-4), but also postulates that, from the first encounter, sequences are stored as wholes in memory. According to Bybee, chunking strength increases continually starting with the first time a sequence is encountered. As a consequence, even rarely-encountered combinations of words may be chunked (depending on how recent the only or last encounter). The difference between these weak chunks and more frequent ones is that in the case of weak chunks, the representations of the individual words are more easily accessible than the representation of the chunk while strong, frequent chunks are so much more easily accessible than their component parts that they can easily be accessed as wholes (cf. Bybee 2010:36). In other words, high-frequency chunks are complex units which can be processed as single items (cf. Bybee 2010:47; 52).

The conception of the mind as a network is not singular to Bybee's approach, though. Usage-based frameworks like Bybee's are located in the realm of cognitive linguistics (cf. Croft and Cruse 2004:291) and draw on cognitive concepts. Thus network models are also underlying cognitive grammars, such as Construction Grammar (Goldberg 1995:5; Fried and Östman 2004:12) and Radical Construction Grammar (Croft 2001) as well as other usage-based approaches such as that of Langacker (2000:6) who, in turn, draws on parallel distributed processing (cf. Rumelhart and McClelland 1986; McClelland and Rumelhart 1986).[1]

The approach taken here crucially differs from Bybee's in the way frequency effects are modelled. Bybee believes that the representation strength of larger units is directly correlated with absolute co-occurrence frequency of the corresponding parts. Thus, the more frequently a sequence is used, the more easily it can be accessed as a whole. The model applied here, on the other hand, assumes that the mind not only keeps track of absolute frequencies of co-occurrence, but also of more abstract probabilistic patterns of co-occurrence, such as, for example, the relative chance of two words appearing in sequence compared to how likely they are to occur in other combinations. Mental connections between words are modelled by means of measures of associations which reflect these probabilistic patterns. The absolute and the relative implementation of frequency may make very different predictions about chunking strengths. Based on absolute co-occurrence frequency, very un-unit-like pairs such as *and the* or *I just* are among those predicted to have the highest chunking strength, while transitional probability, a probabilistic measure, rates semantic units like *wind up* and *lack of* among

[1] Of course, even a network model offers nothing more than a schematic simplification of processing. Neurologically, linguistic units are "patterns of mental (ultimately neural) activation [...] not 'stored' in any particular location" (Kemmer and Barlow 2000:xii; see also Langacker 2000:6).

the strongest chunks[2]. The probabilistic approach fits well into Bybee's framework as it provides ratios of usage of the whole versus usage of the parts, which could be interpreted as the likelihood of access to the whole. Nevertheless, Bybee strongly rejects such approaches (2010:97-101). Stefanowitsch and Gries (2003), Wiechmann (2008), Ellis, Simpson-Vlach and Maynard (2008) as well as Kapatsinski and Radicke (2009), however, provide strong support for probabilistic implementations of frequency (see also Bod 2010).

In summary, I postulate that frequency of use has an effect on the processing of multi-word sequences. I aim to show that besides absolute co-occurrence frequencies chunking reflects relative frequencies and associations between words. The study thus builds on Bybee's as the central usage-based theory and aims to refine the model with respect to its assumptions concerning the influence of usage frequencies on the mental representation of chunks.

1.2 Terminology

Throughout the study, I will make use of the terms 'chunking' and 'chunk', despite the fact that the use of this terminology could be interpreted as suggesting a binary distinction between chunked and non-chunked units. My conception of chunking, however, is not binary but continuous: the term 'chunk' thus describes multi-word sequences along the upper end on the chunkiness scale which clearly display signs of mental representation as a unit. This choice of terminology was made in order to stay consistent with Bybee's work.

1.3 Objective & Methodology

It is the objective of this study to provide evidence of the psycholinguistic validity of frequency-based chunking in general and the mental representation of probabilistic relations in particular. Both objectives will be pursued with the help of empirical corpus-based analyses which model chunking on a two-word (or 'bigram') level.

Based on existing evidence that speakers are unlikely to interrupt mentally cohesive units with hesitations (cf. Goldman-Eisler 1968; Beattie and Butterworth 1979; Shriberg and Stolcke 1996; Kapatsinski 2005; Kapatsinski 2010), the placement of filled and unfilled pauses as well as a set of discourse markers is taken as indicators of the

[2] *And the* occurs 1,110 times in the Switchboard NXT corpus; *I just* occurs 795 times, *wind up* 25 times and *lack of* just 12 times. Yet the direct transitional probability of *and the* is only .04; that of *I just* is .02. The direct transitional probability of *wind up,* on the other hand, is .93 and that of *lack of* is 1.

chunkiness of sequences. In (6) and (7), for instance, *type of* and *I think* are regarded as rather chunky, based on their frequency and mutual attraction, while *of book* and *personally I* are not because their frequency and attraction are much lower. In other words, the former pairs are expected to have a far stronger mental representation than the latter. Consequently, both simple and probabilistic conceptions of chunking predict that speakers should be more likely to interrupt *of book* or *personally I* than *type of* or *I think*. The speakers' hesitation placement confirms this prediction and can thus be interpreted as evidence that chunks are cognitively real.

(6) type of *[pause]* book (sw3056.A.s36)

(7) *well* personally *you know* I think (sw2062.B.s2)

However, absolute co-occurrence frequency and probabilistic measures of association do not always make the same predictions about the chunkiness of multi-word units. The expression *in the olden days* exemplifies this. The complete sequence is comparatively infrequent, yet the archaic term *olden* can only be combined with *days* and *times*. Therefore, in relation to its total usage, *olden* frequently co-occurs with *days* and probabilistic measures would predict that the sequence is highly chunked, while based on frequency, associations between *olden* and *days* should be very weak. Therefore, analyses of hesitation placement can also be used to explicitly contrast absolute frequency and more complex, relative measures of association, and so offer valuable information on how chunking can best be modelled.

In this study, hesitation placement within 14 syntactically controlled contexts – six types of prepositional phrases and eight pre-verbal sentence-initial contexts – will be analysed. These contexts were selected because they are very common in speech and thus yield sufficient numbers of data-points to allow for large-scale analyses. Furthermore, they provide information concerning the relation between chunks and constituents, thus answering questions such as: do phrase boundaries and chunk boundaries coincide?

The influence of the absolute co-occurrence frequency of the words in the vicinity of a hesitation will explicitly be contrasted with four selected measures of association:

- direct transitional probability
- backwards transitional probability
- mutual information
- lexical gravity G

I hypothesise that the more complex a measure, i.e. the more information needed to calculate it, the better it should perform. Among the selected predictors, complexity increases as we move down the list:

- Co-occurrence frequency is one of the simplest predictors possible because it only reflects a single absolute count.
- Transitional probabilities additionally reflect knowledge about the frequency of one of the words in the pair.
- The mutual information score is calculated based on the frequency of the pair and the frequency of both component words.
- Lexical gravity G is additionally based on information about the number of other combinations the words can occur in.

With increasing complexity of calculation measures should reflect a broader knowledge about the distributions of words in language. As speakers experience these distributions in their day-to-day input, the more complex a measure, the more closely it approximates a speaker's experience and it may thus more accurately reflect how he processes speech. Consequently, the performance of the predictors should increase as we move down the list.

The data employed will be the Switchboard NXT corpus of American English consisting of 830,000 words of telephone conversations. The corpus is highly annotated and time-aligned, allowing not only for a reliable extraction of hesitations and unfilled pauses, but also for posing syntactic restrictions.

Data will be analysed with the help of Classification and Regression Trees ('CART trees'; Hothorn, Hornik and Zeileis 2006) as well as random forests (Hothorn, Hornik and Zeileis 2006; Strobl et al. 2007; 2008). These algorithms "grow" trees through recursive binary partitioning of the data, with the aim of creating ever purer "branches", i.e. subgroups of the data. While CART trees rely on a single tree per dataset, random forests grow thousands of trees using only a random selection of data points and predictors in each tree (cf. Strobl, Malley and Tutz 2009b). Such analyses have several advantages over generalised linear models and other regression approaches commonly applied in linguistics. CART trees and random forests can handle multinomial outcomes and complex interactions as well as collinear predictors and large numbers of predictors (cf. Tagliamonte and Baayen 2012). Despite this, the methodology has not yet been widely applied in linguistics. Thus the present studies also serve to highlight the advantages of recursive partitioning methods for linguistic theory building.

Corpus-based methods in cognitive and psycholinguistics have previously been challenged for not controlling for a range of factors (cf. Pickering and Branigan 1999:136) as well as for producing only coarse-grained results (cf. Glynn in Arppe et al. 2010:7). They are thus seen as tools for exploratory analyses, providing hypotheses which are in need of experimental corroboration (cf. Branigan et al. 1995:492). Gries (2005:386-7), however, argues that corpus studies are superior to experiments in two crucial respects:

- Experiments are conducted in an artificial setting; their results thus have a limited scope for generalisation. Corpora, on the other hand, cover different registers, subjects and levels of formality. Their results thus allow for much broader generalisations.
- The actual usage frequency of a construction can only be determined with the help of a corpus.

Zeschel (in Arppe et al. 2010:10) further argues that existing evidence of frequency effects in language learning and processing is so pervasive that these effects must be considered factual. Therefore "corpus-derived findings are not necessarily in need of additional experimental corroboration in order to qualify as relevant for cognitive research" (Arppe et al. 2010:10). Kemmer and Barlow (2000:xv) even argue that corpora are the ideal tool for usage-based analyses. This study builds on these pro-corpus arguments. The following studies will show that with an adequate statistical apparatus corpus data can not only validate experimental findings but can also provide evidence for broader generalisations as well as refine theory.

In summary, this work aims to further our understanding of the representation of speech in long-term memory and of the cognitive processes required for speech planning and production. In this respect, it is particularly targeted at providing new information on the effects which usage-frequency has on these processes. Considering its theoretical foundations, it is situated in the field of usage-based theories of human cognition and will mainly focus on Bybee's theory as it is the cornerstone of the discipline yet in need of empirical validation. This work also heavily draws on ideas and results from research on formulaic language and also on structural descriptions of hesitation placement. Analyses are thus set at a cross-roads, built on underpinnings from both cognitive linguistics and psycholinguistics with a focus on the empirical implementation of processing and making predictions about unseen data.

1.4 Structure of the Present Study

This study is structured as follows. Chapter 2 will provide a review of the relevant literature, thus presenting the theoretical underpinnings. It first introduces the general concept of frequency effects which usage-based models are based on and briefly details which assumptions about grammar, the lexicon as well as the overall mental representation of language such models must be based on. Chunking is then introduced as a particular kind of entrenchment.

Chunks, as defined here, however, are not the only multi-word unit which has been assumed to be cognitively represented. Therefore, I provide a typology of the kinds of multi-word units (MWUs) heretofore postulated in cognitive linguistics and psycholinguistics. Based on six parameters Section 2.2.1 will illustrate which assumptions about MWUs have been made in different disciplines and in which respects these are in accordance with or irreconcilable with my own assumptions. It will become clear that any definition of a MWU can only capture a subset of the spectrum of possibly represented units. Six exemplary approaches will then be evaluated in more detail. These are Pawley and Syder (1983) as an example of a theory drawing on formulaic language and an approach which investigates the placement of hesitations to draw conclusions about chunking, Sinclair (1991) and Biber et al. (1999) which both share a strong emphasis on corpus linguistic methodology and finally Bybee (2007b; 2007a; 2010; Beckner and Bybee 2009), Arnon and Snider (2010) as well as Kapatsinski and Radicke (2009), which focus on cognitive or psycholinguistic theory building. It is particularly the findings of the last group which this study aims to evaluate, i.e. the way in which multi-word frequency effects can best be modelled and how the resulting units are stored.

Finally, Chapter 2 also reviews studies on hesitation placement, drawing attention to the different factors which have been considered determinants of where speakers interrupt their utterance to hesitate. Particular attention will be given to the incapacity of syntactic factors to explain the existing variation and how usage-based and probabilistic approaches excel at doing so.

Chapter 3 will present the data and methodology used and gives an overview of the design of the following studies. The chapter further supplies information about the Switchboard NXT corpus, its annotations and how I extracted data from the corpus. Furthermore, the selected set of hesitations is described, paying particular attention to all assumptions made about hesitations, the reasons for including certain discourse markers and the coding of hesitations. The chapter finally describes the methodology used, listing the selected measures of associations as well as giving a detailed account of the workings of the regression method.

Chapters 4 and 5 describe empirical analyses of hesitation placement in two different contexts. Chapter 4 provides an analysis of prepositional phrases. The question at the focus of this chapter is whether the usage frequency of all word-pairs in the phrase as well as the statistical associations between the words in the phrase can explain why hesitations are placed where they are. In other words, do absolute and relative frequencies of co-occurrence explain why a speaker would say *you know on the spot* but *on the [pause] lookout*? The study further addresses whether chunking across the prepositional phrase boundary is possible or even common. The regression methods applied for this purpose are CART trees an random forests. Results will reveal that co-occurrence frequencies as well as probabilistic chances of co-occurrence have a strong influence on hesitation placement and consequently on speech planning. Hesitation placement is, to a significant degree, predictable from these usage-based factors and this is particularly true for the phrase boundary. The latter indicates that the words to the left and the right of the prepositional phrase boundary commonly form chunks.

In Chapter 5, the same methodology is applied to sentence beginnings. Hesitation placement before and within the verb phrase will be analysed. In light of the finding that chunking across the prepositional phrase boundary is a regular phenomenon, particular attention will be paid to subject-verb chunking. I will also address the absolute beginning of the sentence and investigate whether specific sentence-initial time-buying devices emerge. Results show that the majority of hesitations is uttered at or very near the sentence onset – speakers prefer to plan a sentence before they start uttering it. Very few hesitations are uttered as late as after the subject, let alone within the verb phrase. Yet I can show what few hesitations are moved to these positions mark word-pairs of below average frequency or cohesiveness. Finally, I find very strong evidence that hesitations themselves can become part of chunks. Frequent coordinating conjunctions, such as *and* and *but*, often merge with following pause fillers to form chunks like *and uh* and *but uh* which serve as longer time-buying devices.

In the discussion and conclusion in Chapter 6, I evaluate the advantages and disadvantages of the methodology applied in the two corpus-based studies. I will specifically point out how CART trees with their graphical representation of distributions in the data profit linguistic analyses of this sort. Furthermore, I interpret my findings in light of theory building and statistical modelling of chunks. I will conclude with a discussion of the the relevance of my findings in the light of Bybee's and other models of the mind and point out potential objects of further research.

This print also includes a number of appendices which primarily contain graphs providing more detailed information about numerous steps of the analyses. Wherever such information is available, reference will be made in the text. In addition, Appendix

P provides selected R scripts used for the generation of graphs and statistical calculations.

2 Frequency, Chunks & Hesitations

This chapter details the theoretical foundations of the present study. Section 2.1 contextualises chunking within the greater realm of frequency effects. Section 2.2 then compares the chunking approach taken in the present study to existing definitions of mentally represented multi-word sequences in the literature and explains in which respects these overlap or differ. This overview is followed by a more detailed discussion of approaches to the analysis of multi-word sequences which are relevant to my own approach. Section 2.3 is devoted to previous analyses of hesitation placement. It illustrates that hesitations have been found to predominantly occur at the boundaries of different units of encoding.

2.1 Frequency Effects

Central to usage-based work like this is the assumption that language use shapes the mental representation of language and consequently the system. Every instance of usage of a sound or word or construction is considered to be conditioned by previous experiences and in turn to influence instances to come (cf. Beckner et al. 2009:2; Bod 2010; Bybee 2006:730; Kemmer and Barlow 2000:ix). The more frequently a type or token is used, the stronger its impact.

This view of language as a "complex adaptive system" (Beckner et al. 2009:1) is irreconcilable with earlier structural and generative views as it violates their fundamental assumptions. This applies particularly to three dichotomies, which are traditionally at the heart of generative theories.

competence vs performance – This dichotomy poses a strict separation of knowledge (competence) from usage (performance). It is based on an interest in speakers' knowledge of a language and disinterest in their use of it, the latter being considered just an imperfect output of their knowledge. Any influence of performance on competence was initially not considered (cf. Chomsky 1965:4).

language vs other human cognitive abilities – Language, i.e. Universal Grammar, is considered to be innately given, represented in a module separate from other cognitive abilities. Learning a language means deducing from the input which structures are possible in the given language (cf. Chomsky 1965:32-3). Thus language is 'learned' in a fundamentally different way from other cognitive abilities.

grammar vs lexicon – A speaker's knowledge of the structure of his or her language (grammar) and his inventory of words (lexicon) are considered two separate entities (cf. Chomsky 1965:84).

Additionally, generative theories strive for minimum redundancy in the lexicon as well as in grammar (cf. e.g. Chomsky 1965:184), meaning that the lexicon is, per definition, as compact as possible, keeping entries small. Grammars furthermore strive for a minimum of symbols and rules. These theories, however, are "unable to represent the lexical or collocational dimension of syntagmatic structure" (Barlow 2000:317) and to reflect psychological reality (cf. Langacker 2000:2).

Usage-based models[3], on the other hand, argue that language structure is derived from "more general psychological capacities" (Langacker 2000:2; also "domain-general cognitive processes" Bybee 2010:1), thus no longer assigning language a special role among human cognitive abilities. Usage-based theories do not consider grammar to be an abstract set of rules (cf. Beckner et al. 2009:5) but instead propose that "grammar is the cognitive organization of one's experience with language" (Bybee 2006:711). In usage-based models,

> [t]he basic units of grammar are constructions, which are direct form-meaning pairings that range from the very specific (words or idioms) to the more general (passive construction, ditransitive construction), and from very small units (words with affixes, *walked*) to clause or even discourse-level units. (Beckner et al. 2009:5)

Constructions are cognitively represented in the form of exemplars or exemplar clouds, which are "rich memory representations" (Bybee 2010:14), containing structural, lexical, phonetic, semantic and contextual information. Usage-based approaches thus also do away with the remaining two distinctions, knowledge (i.e. competence) and usage (i.e. performance), far from being separate, are seen as depending on one another and grammar and the lexicon are considered "highly inter-twined" (Beckner et al. 2009:7). The resulting network of cognitive representations is not minimalist and economical (cf. Bybee and Hopper 2001:8), but highly redundant with separate entries for nested and overlapping units (cf. Bybee 2010:24). Thus it represents, for example, simple words like *car*, complex words like *car sharing*, constructions like *drive a car* and other multi-word units like *designated driver*.

This "emergent system" (Bybee 2007a:8) is necessarily dynamic. Every token of a linguistic unit which a speaker encounters influences cognitive representations (cf. Bybee 2007a:8, 2010:18; Langacker 2000:10-1), therefore "even adult grammars are not fixed

[3] A term coined by Langacker (1987).

and static but have the potential to change as experience changes" (Beckner et al. 2009:7; cf. also Bybee 2010:10). As a consequence, language use shapes the mental representation of language. The more often a language user encounters the same or a similar usage event, the stronger its impact on his mental representations.

Usage-based theories explain changes in this dynamic system by the aforementioned set of 'domain general cognitive processes' (cf. Bybee 2007a:8). Among these are three particularly linked to usage frequency, namely entrenchment, categorisation and analogy. The following sections describe these processes and list some exemplary effects they have on language acquisition, processing and change.

Entrenchment – The more often a motor activity is repeated, the more firmly it becomes entrenched, and therefore every repetition of an activity facilitates its execution. Through frequent repetition, the elements of a complex sequence can then coalesce into a "pre-packaged" unit (Bybee and McClelland 2005:384; cf. also Langacker 2000:3). This facilitation in execution is brought about through ever-increasing cognitive connections between the elements in the sequence, a process often referred to as *chunking* (cf. Bybee 2007b:316, 2010:7; Beckner et al. 2009:6). The more frequent an entrenched word or chunk, the more likely it will be affected by 'automatisation'; the gestures by which the words are articulated start to overlap, leading to 'phonetic reduction' (cf. Bybee 2006:720; 2010:75; Bybee and Hopper 2001:11; Bybee and McClelland 2005:382).

In morphosyntax, however, entrenchment has the opposite effect. Here it has a conserving effect because

> [h]igh-frequency structures become more entrenched in their morpho-syntactic structure and resist restructuring on the basis of productive patterns that might otherwise occur. (Bybee 2006:715)

Therefore high-frequency verbs in English have maintained their irregular past tense (e.g. *keep-kept*), while low-frequency verbs are regularising (e.g. *weep-weeped*; cf. Bybee 2006:715).

Through either repeated access as a unit, the conserving effect (cf. Bybee 2006:715) or phonetic reduction (cf. Bybee 2010:48), a structure can "become autonomous from etymologically related forms" (Bybee 2006:715) or "the construction that originally gave rise to it" (Bybee 2006:720). Thus contractions like *I'm* or *I'll* are autonomous of other pronoun-verb constructions, because they display more extreme phonetic reduction (cf. Bybee 2010:48). The *going to* future in English is also a strong case of autonomy, as it lost its original senses of motion and purpose and took on its current meaning of future (cf. Bybee 2006:19-20). The latter is an example of the most extreme form of autonomy,

namely grammaticalisation (cf. Bybee 2010:48), whereby an existing construction gives rise to "a new grammatical morpheme and a new construction" (Bybee 2006:719).

Eventually, frequent co-occurrence and combined usage of items may lead to reanalysis, meaning that the internal structure of the chunk is gradually lost (cf. Bybee 2007b:316) and words merge into a single constituent or change their grammatical category. In the present example, *going to* has come to be analysed as a single unit (cf. Bybee 2006:721).

Categorisation – Categorisation is a special kind of comparison, whereby a new form is compared against an established standard in order to determine whether it belongs to the same category (cf. Langacker 2000:4). Words or phrases are, for example, recognised by means of categorisation (cf. Bybee 2010:7).

Categories show prototype effects, which means that some members of the category are considered more central or prototypical than others (cf. Bybee 2010:18,79). Bybee (2006) along with Bybee and Eddington (2006) argue that this prototype or 'central' member is also the most frequent (Bybee 2006:727).

Utterances are categorised according to their form, meaning and the context in which they are encountered. Such a category of utterances is referred to as an exemplar (cf. Beckner et al. 2009:7). Exemplars can again be grouped and thus form the basis for the emergence of constructions (cf. Bybee 2006:718).

Analogy – Creative novel utterances can be formed in analogy to known and categorised utterances (cf. Bybee 2010:8), thereby "the more frequent of the members of a paradigm tends to serve as the basis of new analogical forms" (Bybee 2010:25). Analogy also plays a role in diachronic language change, where it mostly leads to the loss of an alternation in a paradigm, which is referred to as 'analogical levelling' (cf. Bybee 2010:66). For example, irregular past-tense forms are shifted over to the more frequent regular paradigm (*leapt* becoming *leaped*). Due to the conserving effect of entrenchment this happens earlier for low-frequency than for high-frequency verbs (cf. Bybee 2010:66).

Currently, cognitive linguists and psycholinguists are gathering evidence for the cognitive reality of frequency effects and their limits as well as which types of linguistic units are affected by them. There are indications that it is not only concrete surface structures like multi-morphemic words and multi-word sequences which can be affected, but also more abstract elements like constructions. Bybee and Hopper (2001) point out that these questions are tied together and that any evidence for frequency effects is simultaneously evidence for the cognitive reality of a particular kind of unit, because

> [l]inguistic material cannot accrue frequency effects unless the brain is keeping track of frequency in some way; frequency effects cannot be attributed to units unless they are items in storage that are affected by experience. (Bybee and Hopper 2001:9)

My own work focusses on only one kind of frequency effect, namely the entrenchment effect of chunking and the corresponding unit of mental representation, the chunk. The following section introduces chunking in more detail and illustrates how my approach ties in with previous work in the field.

2.2 Chunking

Chunking was defined above as the process whereby words in a sequence gradually become mutually more strongly connected. Associations between the words in the sequence are strengthened and the sequence may even receive its own holistic representation. Analyses of chunking generally focus on the role of usage frequency in this process. A chunk is thus a sequence of words, which, through repeated use, has become routinised and can therefore be recalled and produced with ease, (almost) like a single word.

This concept of a chunk and chunking is connected to a usage-based and frequency-oriented perspective. A large number of other terms are currently in use to describe mentally-represented units beyond a word in length. Terminology differs depending on the discipline and personal preferences. The following is a broad though non-exhaustive list of established labels:

- chunk (e.g. Bybee 2007a; Beckner and Bybee 2009)
- construction (e.g. Goldberg 1995; Croft 2001; Fillmore, Kay and O'Connor 2003)
- collocation (e.g. Mel'čuk 1998; Sinclair 1991)
- collostruction (e.g. Stefanowitsch and Gries 2003)
- formula (e.g. Ellis 2002; Simpson-Vlach and Ellis 2010)
- idiom (e.g. Sinclair 1991; Levelt 1992)
- lexical bundle (e.g. Biber and Conrad 2003)
- lexicalised sentence stems (e.g. Pawley and Syder 1983)
- memorised sentences (e.g. Pawley and Syder 1983)
- phrase (e.g. Simpson-Vlach and Ellis 2010)
- phrasal lexeme (e.g. Moon 1998)
- phraseme (e.g. Mel'čuk 1998)
- phraseologism (e.g. Gries 2008)
- prefab (structures) (e.g. Barlow 2000; Erman and Warren 2000; Erman 2007)
- recurrent word combinations (e.g. Altenberg 1998)
- recurrent word associations (e.g. Barlow 2000)
- single multi-word unit of meaning (e.g. Sinclair 1991)

Different terms may designate the same, partially overlapping, or entirely different concepts. The following section presents a typology of approaches by highlighting six distinctive aspects featuring in any definition of multi-word units. It details previous definitions of multi-word units (MWUs) and illustrates how the concept of chunking ties in with them. Section 2.2.2 then summarises definitions and findings from several studies which are most influential for this work. Wherever I refer to a specific concept or definition, I will use the terminology of the author; for general references, the theory-independent term MWU will be used.

2.2.1 A Typology of MWUs

Gries (2008) develops a taxonomy of different models of MWUs (for a similar scheme see Gibbs 2007). He finds that they can be characterised along the lines of six parameters which constitute typical cornerstones in any definition. In the following, my concept of chunks will be placed in the context of these parameters.

2.2.1.1 The Nature of the Elements Involved

MWUs can potentially form at any level of abstraction. For an analysis of English noun phrases, Bybee (2007b), for instance, lists the following four levels:

(8) Very specific: *my mother*, *my computer, the car, a problem, an idea*
Partially general: [*my* + NOUN] [POSS PRO + *mother*]
More general: [POSSESSIVE + NOUN]
Fully general: [DETERMINER + NOUN](Bybee 2007b:325)

Potentially, even further or mixed levels are conceivable (e.g. lemma level). Most concepts are, however, restricted to a particular level. Sinclair (1991), for example, defines "[c]ollocation in its purest sense" as "only the lexical co-occurrence of words" (Sinclair 1991:170), and thus operates on the first, "very specific", level. Ellis (2003), on the other hand, declares that chunking "operates at concrete and abstract levels" (Ellis 2003:76) and Gries (2008) defines phraseologisms as "the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic items of various kinds" (Gries 2008:6), including grammatical patterns, phrases and clauses (2008:5-7). Hence, he mixes different levels, requiring one specific element and permitting a more general group for any further element. Corpus-based studies based on so-called 'n-grams', i.e. word sequences of a specified length, work by definition on

the word level (cf. Gries 2008:16; for an example see also Simpson-Vlach and Ellis 2010).

'Chunk' originally refers to any kind of entrenched sequence which is represented in memory (cf. Newell 1990:7). Consequently, it designates linguistic sequences at any level. Bybee adopts the term in this sense (2010:7) and notes that chunks form the basic components of "constructions, constituents and formulaic expressions" (2010:7). She thus starts out from the 'very specific' level (i.e. word sequences), but allows for conclusions about representation on more abstract 'general' levels (e.g. constructions).

The present study proceeds in a similar manner, working with surface-level sequences in order to draw conclusions about mental relations between individual words. Thus the representation of 'very specific' items is at the focus. Wherever indicated, conclusions about possible mental representations on a more abstract or 'general' level will be drawn.

2.2.1.2 The Number of Elements Involved

A second criterion, which features in some definitions, though often only for operational purposes, is the minimum number of elements the sequence must contain or a limit on the maximum number of elements in the sequence. As we are concerned with *multi*-word units, the minimum per se has to be two.

Corpus-based studies, particularly those dealing with large amounts of data often impose more strict limits in order to "keep data to a more manageable size" (Simpson-Vlach and Ellis 2010:6). Hence, Simpson-Vlach and Ellis (2010), limit the 'formulas' they analyse to three to six words in length. Biber, Conrad and Cortes (2004) even limit their interest to four-word sequences only.

For practical reasons, in the present study the size of n-grams is restricted to two. I thus operate on the basis of word pairs or 'bigrams'. This option was chosen because longer n-grams, per se, tend to be rarer in speech (cf. Bybee 2010:35) and corpora need to be very large to obtain reliable information about their distribution.

2.2.1.3 Frequency of the Sequence

More than any other parameter, the way frequency is implemented in a definition of a MWU reflects the kind of effect expected. Four commonly-used approaches can be distinguished (classification of the three initial approaches based on Arnon and Snider 2010:69):

(a) *Words-and-Rules Approach* (e.g. Chomsky 1965) – Only semantically non-transparent phrases, such as *rain cats and dogs*, are expected to be stored as units. Frequency effects are not expected in this case. Such approaches are therefore of minor interest in this work.

(b) *Frequency-Threshold Approach* (e.g. Erman and Warren 2000; Vogel Sosa and MacFarlane 2002; Wray 2002; Biber, Conrad and Cortes 2004; Simpson-Vlach and Ellis 2010) – Phrases with a frequency of occurrence beyond a certain threshold can be stored holistically. However, assumptions of holistic storage cannot generally be equated with expectations of threshold effects (see below). Inversely, in practice, not all studies using a threshold approach actually expect holistic storage; they merely use a threshold as a simplified way of reflecting stronger chunking (see (c)).

The following are examples of frequency-thresholds applied in the literature: 10 occurrences per million words (cf. Simpson-Vlach and Ellis 2010); 40 occurrences per million words (cf. Biber, Conrad and Cortes 2004); arbitrary cut-off point based on impressionistic frequency i.e. "if [...] extensions are felt to be common, they have been considered part of the prefab in question" (Erman and Warren 2000:40).

Arnon and Snider (2010) criticise threshold approaches, pointing out that they build on a logical fallacy (see discussion of Arnon and Snider 2010 in Section 2.2.2).

(c) *Continuous Approach* (e.g. Langacker 1987; Ellis, Simpson-Vlach and Maynard 2008; Arnon and Snider 2010) – Processing of the sequence is influenced by every repetition, meaning that the unit-like character of the string intensifies with frequency of use. While in some models this effects arises from strengthened connections between the components of the MWU, other models assume that MWUs are stored holistically from the first encounter. In the latter case, the effect is caused by an increased activation of the whole the more frequent it is (cf. e.g. Arnon and Snider 2010:69). For computational reasons, continuous approaches sometimes operationalise frequency in bins (cf. e.g. Ellis, Simpson-Vlach and Maynard 2008 who use three frequency bins). For a comparison of the predictions made by a frequency threshold approach and a continuous approach, see Kapatsinski and Radicke (2009), summarised in Section 2.2.2 below.

(d) *Probabilistic Approach* (e.g. Gries 2008; Hilpert 2013; Stefanowitsch and Gries 2003; Wiechmann 2008) – Chunking strength rises with the probability of co-occurrence. This can be operationalised through measures of association (e.g. Gries 2008). Hence, as in the continuous approach, chunking is considered a gradable phenomenon. Yet, contrary to the continuous approach, the chunking strength of low-frequency sequences can be high, depending on the probabilistic tendency of the words in the sequence to co-occur.

Most probabilistic approaches thus measure how likely a given word is to be used in a particular construction considering all contexts the word occurs in.

Wray (2002) advocates the opposite perspective, namely that a count of how often an idea is expressed with a particular sequence in relation to the total number of times the idea is expressed would be the best way of determining whether a "word string is the preferred way of expressing a given idea" (2002:30). She concedes, however, that there is as yet no corpus tagged in a way which allows this (2002:31; for a similar concept see Erman and Warren 2000:31).

There are also models which combine aspects from two or more of the above categories. Kapatsinski and Radicke (2009), for instance, argue that their data can best be explained by a model which assumes that mental connections between the components of an MWU rises throughout the frequency spectrum (continuous approach), but that extremely highly frequent sequences are stored holistically (threshold approach; for a more detailed account of Kapatsinski and Radicke 2009 see Section 2.2.2).

The present study models chunking strength as a gradable phenomenon. Importantly, it incorporates both absolute co-occurrence frequency (continuous approach) as well as probabilistic measures of association (probabilistic approach) in order to evaluate which approach better captures the given effects. In case both of these approaches are inadequate and chunking, in fact, happens abruptly, models are designed to capture this and indicate the threshold level. The basic assumption that I start out from is that chunks are stored in the form of strengthened connections between components. Due to their comparatively higher complexity, probabilistic measures of association are expected to reflect this more accurately than simple co-occurrence frequency. Results will be evaluated empirically in order to determine whether there is evidence that MWUs actually receive combined entries in the mental lexicon.

2.2.1.4 Distance between the Elements Involved

While n-gram-based approaches are, by definition, only concerned with immediately adjacent elements (cf. Simpson-Vlach and Ellis 2010), others allow material to intervene (cf. Gries 2008:5; Wray 2002:9). This also includes the possibility of any kind of expansion of the multi-word unit, be it at the left or right margin.

(9) run (dangerously) amok
(10)(quite) all right
(11)What (in fact) did you do?

In an n-gram-based approach, for instance, the full and the reduced version of (9) would not be picked up as instances of the same MWU. Other approaches, like Erman and Warren (2000), on the other hand, allow non-obligatory extensions of MWUs. So (10), including the adverb, is considered an extension of the MWU *all right* (2000:35). Most approaches permitting intervening elements also permit nested MWUs like *in fact* in (11) (example taken from Erman and Warren 2000:46; further examples of approaches which allow nesting are Altenberg 1998:115; Beckner and Bybee 2009:30; Bybee 2007b:319). Bybee (2007b) hypothesises that "optional elements have weaker sequential links than obligatory elements and thus looser constituency" (Bybee 2007b: 319).

As far as any possible distance between the elements in a MWU is concerned, the approach taken in my work exemplifies n-gram-based work. It only picks up chunking of immediately adjacent words. Should any words intervene, like *dangerously* in (9), the MWU is no longer recognised. Instead of classifying (9) as an instance of *run amok*, a bigram-based approach sees only the word-pairs *run dangerously* and *dangerously amok*, which could be chunks in their own right.

2.2.1.5 Lexical and Syntactic Flexibility of the Elements Involved

The 'time'-*away* construction (for a detailed analysis see Jackendoff 1997) best exemplifies different views on lexical and syntactic flexibility. In n-gram-based studies (e.g. Biber, Conrad and Cortes 2004; Simpson-Vlach and Ellis 2010), which generally operate on a strict surface level, the three tokens (12), (13) and (14) would (if each were of sufficient frequency) be picked up as separate types of MWUs (though in some interpretations (12) and (13) would be subsumed as tokens of the same MWU type, see Simpson-Vlach and Ellis 2010).

(12) dance the night away

(13) danced the night away

(14) while the day away

All other approaches, and particularly cognitive grammar and construction grammars, allow more flexibility. The latter would subsume (12) to (14) under one construction type of the form *X the 'time' away* (see also Erman 2007 for an explicit analysis of fixed versus semi-fixed slots in MWUs).

Occasionally, studies such as Erman and Warren (2000) approach the issue of flexibility from the opposite direction, specifying not the maximum flexibility permitted,

but the minimum stability required (for a similar approach see Gries 2008:5-6). Thus Erman and Warren require

> that for anything to be a prefab the choice of one word must determine, or at least definitely restrict, the choice of at least one other, normally adjacent, word[. This] excludes from consideration constructions [...] such as *make somebody do something*, [...] which contains only one lexically specified item *(somebody, do,* and *something* all represent unrestricted choices of words). (Erman and Warren 2000:32)

The basic set-up of my approach is that of an n-gram-based study. However, in addition to a quantitative word-form-based analysis, I qualitatively analyse the data in order to see whether findings at the specific level also apply to more general levels. This means that initially only surface-level word-pairs are extracted from the corpus. Hence any syntactic or lexical variation would be registered as different bigrams. For instance *night away* and *day away* in the above examples would be picked up as separate bigrams. The quantitative analysis would then consider the frequencies and hesitation pattern of each one separately. The additional qualitative analysis of the data, however, could reveal if they both showed similar frequencies and hesitation placement and should therefore be considered a combined MWU of the more flexible form 'time' *away*.

2.2.1.6 Semantic Unity and Semantic (Non-) Compositionality

Finally, definitions differ concerning the requirement of semantic unity and semantic non-compositionality. Some theories assume that sequences only converge into a MWU if the meaning of the whole is no longer deducible from the meaning of its parts, as for example Mel'čuk (1998) who defines a 'phraseme' as a "non-compositional lexical unit" (Mel'čuk 1998:24). Occasionally, a range of sense restrictions are assumed for MWUs, as in Moon (1998) who defines 'phrasal lexemes' as

> the sorts of item that for reasons of semantics, lexico-grammar, or pragmatics are regarded as holistic units rather than compositional strings. Such items include pure idioms, proverbs, similes, institutionalized metaphors, formulae, sayings, and various other kinds of institutionalized collocation. (Moon 1998:79)

More commonly, however, MWUs are believed to form when the words within them "function as a semantic unit, i.e. [...] have a sense just like a single morpheme or word" (Gries 2008:6). Beckner and Bybee explain how the notions of semantic unity and mental unity tie together:

> [B]ecause elements that are semantically related tend to occur together, most chunks are also semantically coherent and therefore considered to be constituents in most theories of grammar. (Beckner and Bybee 2009:31)

The authors however caution that this does not imply that chunks necessarily need to follow constituent structure, but that, contrarily, constituent structure emerges from chunking (Beckner and Bybee 2009:42; see also Section 2.2.2), implying that some constituent boundaries assumed by traditional grammars (cf. Huddleston and Pullum 2002) should be rejected. Biber et al. (2004) also point out that

> most lexical bundles do not represent a complete structural unit. [...] Instead, most lexical bundles bridge two structural units: they begin at a clause or phrase boundary, but the last words of the bundle are the first elements of a second structural unit. (Biber, Conrad and Cortes 2004:377)

N-gram-based corpus studies require neither semantic unity nor non-compositionality for a string to be considered a MWU. Though Biber et al. (2004) concede that

> frequency is only one measure of the extent to which a multi-word sequence is prefabricated; sequences with idiomatic meanings are usually rare but clearly prefabricated (Biber, Conrad and Cortes 2004:376).

This study takes a quantitative, statistical approach to the collection of evidence for the cognitive reality of chunks. It thus relies on objectively measurable criteria such as corpus frequencies, measures of association and hesitation placement. I make no assumptions about the role of semantics in chunking.

2.2.2 Previous Research on MWUs

As I have shown above, research into MWUs does not follow one school of thought. The umbrella-term 'MWU' is a vague and fuzzy concept, mostly due to the very nature of the subject: units form in various different shapes and studies generally centre on one particular type only. In order to arrive at a clearly delimited as well as empirically analysable field of research, I have narrowed down the type of MWU analysed in this study to immediately adjacent word-pairs. This section contains detailed discussions of the approaches which form the pillars of this work.

Pawley and Syder (1983) approach MWUs from the perspective of research on formulaic language (see also e.g. Ellis 2002; 2003; 2008; Ellis, Simpson-Vlach and Maynard 2008; Simpson-Vlach and Ellis 2010; Wray 2002). In this early model, usage frequency features implicitly only, in that sentence stems are described as "common" or

"conventionalised". The analysis is of importance to this study because it is one of the first to utilise hesitations as indicators of holistic storage.

Starting out from the observation that native speakers consistently make "natural and idiomatic" selections ("nativelike selection") and produce these "stretches of spontaneous connected discourse" fluently ("nativelike fluency", 1983:191), Pawley and Syder propose that native-like sentences cannot be produced based on Chomskyan syntactic rules alone. Interestingly, they hypothesise that *two* classes of units are necessary to explain native utterances: "memorised sentences" and "lexicalized sentence stems" (1983:205). According to the authors,

> [t]he terms refer to two distinct but interrelated classes of units, and it will be suggested that a store of these two unit types is among the additional ingredients required for native control (Pawley and Syder 1983:205).

A sentence stem is defined as "a unit of clause length or longer whose grammatical form and lexical content [are] wholly or largely fixed" (1983:191). Importantly, sentence stems are units belonging to 'competence' in Chomsky's sense. Being lexicalised, they are "part of the speech community's common dictionary" (1983:209). Like a word, a stem functions as a "conventional label for a conventional concept" (1983:209). Except for some completely fixed expressions, sentence stems allow for a restricted degree of "inflection" and "expansion" (1983:214-5; see also Sections 2.2.1.4 and 2.2.1.5 above). (15) to (17) are some examples of sentence stems listed by the authors (1983:213):

(15) Who do you think you are?

(16) P_i thinks the world of P_j.

(17) I think so.

Memorised sequences, on the other hand, are performance phenomena (1983:208-9). Native-like fluency, particularly, can only be explained by looking at individual speakers and their choices – reflected in memorised sequences – rather than by analysing "timeless knowledge shared by the members of a language community" (1983:209), which is reflected in sentence stems.

Not all memorised sequences are lexicalised, though (1983:209). They are defined as

> strings which the speaker or hearer is capable of consciously assembling or analysing, but which on most occasions of use are recalled as wholes or as automatically chained strings. (1983:205)

They are stored holistically in the mental lexicon (1983:218).

(18) Are you all right?

(19) I enjoyed every minute of it.

(20) There's nothing you can do about it now.

Memorised sequences include complete clauses and sentences like (18) to (20) (1983:206-7), but also phraseological units, defined as

> sequences which contain a nucleus of fixed lexical items standing in construction with one or more variable elements (often a grammatical inflection), the specification of the variables being necessary to complete the clause. (1983:205)

In one of the first approaches to bringing together research on MWUs and hesitations, Pawley and Syder (1983) analyse the placement of hesitations in spoken language. They find that speakers generally hesitate[4] at or near clause boundaries (1983:200), being mostly unable to encode more than a clause of eight to ten words at a time (1983:202). The authors therefore postulate that fluent stretches of speech, particularly when longer than one clause, are normally memorised sentence stems (1983:208). They consequently consider memorised stems the "normal building blocks of fluent spoken discourse" (1983:208).

Pawley and Syder's results are largely based on introspection and qualitative analyses of a small set of transcribed conversations. Later investigations have since come to similar conclusions, namely that a proportion of (spoken) language must consist of MWUs. Sinclair (1991:109), for instance, distinguishes between speech production by means of the open-choice principle and the idiom principle, postulating that a full understanding of the way language is composed affords both principles. He describes the former as a "'slot-and-filler' model" (Sinclair 1991:109) of language, where utterances are seen as a series of slots, each to be filled from the lexicon. "At each slot, virtually any word can occur" (Sinclair 1991:109), necessitating complex choices. Yet what Sinclair finds in large corpora is that words and phrases have a strong tendency to occur in particular grammatical constructions and semantic environments as well as in the environment of particular words (Sinclair 1991:112). Furthermore,

> [t]here is a broad tendency for frequent words, or frequent senses of words, to have less of a clear and independent meaning than less frequent words or senses. [... And] [n]ormal text is made up of the occurrence of frequent words, and the frequent senses of words. Hence normal text is largely delexicalized [...]. (Sinclair 1991:113)

---

[4] The authors explicitly consider unfilled pauses of unspecified lengths, decreases in the speed of articulation and hedges (Pawley and Syder 1983:200-1). The example transcripts also contain pause fillers which may also have been included in the analysis.

Based on this observation, he postulates

> that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices even though they might appear to be analysable into segments. (Sinclair 1991:110)

These allow the language user to generate utterances following the "idiom principle". Compared to the open choice principle which allows for the composition of utterances out of individual words, but requires many decisions, the idiom principle is based on "fewer and larger choices" (Sinclair 1991:113). During speech planning, the speaker can switch back and forth between the open-choice and the idiom principle.

Importantly, Sinclair neither sets out to delimit which units are planned according to the idiom principle nor does he coin a name for all 'idiom-principle units'. Instead, he emphasises the diversity, listing "collocations" (Sinclair 1991:115), "[i]dioms, proverbs, clichés, technical terms, jargon, expressions, phrasal verbs, and the like" (Sinclair 1991:111). He concludes that "when we have thoroughly pursued the patterns of co-occurrence of linguistic choices there will be little or no need for a separate residual grammar or lexicon" (Sinclair 1991:137).

Like Sinclair, Biber et al. (1999) take a strictly frequency-driven approach. Their method emphasises the usefulness of large-scale corpora for the identification of MWUs, which they term 'lexical bundles' and define as "words that show a statistical tendency to co-occur" in a register (Biber et al. 1999:989). Based on this definition, the authors conduct a large-scale study, extracting three- to six-word sequences from conversational and academic corpora. Shorter and longer sequences are rejected for practical reasons. The only restricting criteria are

- a minimum frequency of at least ten occurrences per million words for three- and four-word sequences and five times for five- and six-word sequences (Biber et al. 1999:992-3),
- occurrence of the bundle in at least five different texts to avoid picking up idiosyncrasies (Biber et al. 1999:992-3) and
- that bundles must not cut across a turn boundary or be interrupted by a punctuation mark (Biber et al. 1999:993).

The definition of a word follows the orthography of the corpus. Any further restrictions based on semantic or structural grounds are explicitly rejected (1999:990).

The authors find that, according to this definition, only 30 per cent of spoken language and 21 per cent of academic prose are composed of recurrent expressions (Biber et al. 1999:995). This is a very small proportion compared to, for example, Altenberg's 80-per-cent estimate (1998:102). Furthermore, spoken language contains

more ultra-high frequency bundles (≥200 occurrences per million words) than academic prose. The authors also claim that "recognizably trite or idiomatic expressions are typically not frequent" (Biber et al. 1999:998), yet show that frequent verb phrase expressions are often "relatively idiomatic" (Biber et al. 1999:1007).

Concerning the internal structure of lexical bundles, Biber et al. note that most bundles do not consist of a single, complete constituent, but tend to encompass material from several units (Biber et al. 1999:991, 999; see also Section 2.2.1.6). "Thus many bundles in conversation contain the beginning of a main clause, followed by the beginning of an embedded complement clause" (Biber et al. 1999:991), such as (21) and (22).

(21) I don't know why

(22) I thought that was

The "systematic patterns of use" (Biber et al. 1999:989) which the authors finally come up with present structural groupings and illustrate the bundles' conversational functions. Biber et al.'s work shows that a frequency-driven approach to language analysis, free of prior theoretical assumptions, can reveal meaningful patterns in language. However, as Biber and colleagues themselves point out in consecutive work co-occurrence frequencies are not explanatory, they merely "identif[y] patterns that must be explained" (Biber, Conrad and Cortes 2004:376). Biber et al. (1999) consider usage frequency both the driving force behind the formation of MWUs and the outcome of this process. While this may indeed be the case (see Bybee 2007a:18; Bybee 2010:53), it is unsatisfactory proof that there are really cognitive processes leading up to the use of lexical bundles. Such approaches could therefore be criticised as showing frequencies, but not really frequency effects.

Bybee (2007b, 2007a, 2010), Bybee and Scheibman (2007), Beckner and Bybee (2009), Arnon and Snider (2010), along with Kapatsinski and Radicke (2009) will exemplarily be detailed here as specimens of cognitive and psycholinguistic studies. In contrast to many other, predominantly cognitive, approaches (e.g. Langacker 1987, 2000; Goldberg 1995), which conceptualise MWUs on an abstract level, by allowing at least open slots and some degree of variation, the studies presented here conceptualise MWUs as surface-level sequences. Their results are thus particularly relevant in the context of my study, which also models MWUs as surface sequences. They furthermore focus on frequency and frequency-related probabilities as determinants of the strength and form of mental representation.

Bybee's sizeable body of work revolves around the question of how frequency of use affects language structure and how concrete instances of use relate to the general cognitive representation of language (Bybee 2010:12). She argues that "human beings are sensitive to recurring sequences of stimuli and record them in memory" (Bybee 2007b:323).

Bybee's conception of MWUs – 'chunks' in her terminology – is stated in her "Linear Fusion Hypothesis", which postulates that "items that are used together fuse together" (Bybee 2007b:316). Bybee explains that "items that are used together frequently will form tighter bonds than items that occur together less often" (Bybee 2007b:319), so that the items in a chunk prime and automate each other (Bybee 2007b: 316). Importantly, chunking in Bybee's sense is a gradual process whereby the strength of the sequential relation depends on the co-occurrence frequency of the sequence (Bybee 2010:25).

In the context of her exemplar model, Bybee conceptualises a chunk as a sequence which, through repeated use, has become automated and "can be accessed and executed as a unit" (Bybee 2007b:316). From the first encounter, the entire sequence is stored in memory. Frequency of use leads to a rise in representation strength, which Bybee describes as follows:

> Certainly words that have never been experienced together do not constitute a chunk, but otherwise there is a continuum from words that have been experienced together only once and fairly recently, which will constitute a weak chunk whose internal parts are stronger than the whole, to more frequent chunks such as *lend a hand* and *pick and choose* which are easily accessible as wholes while still maintaining connections to their parts. (Bybee 2010:36)

Thus there is always a competition between the parts and the whole, the more frequently the chunk is accessed, the weaker the parts and vice versa. Eventually, in very frequent chunks, the weakening of the parts may lead to reduced analysability and compositionality (Bybee 2010:52, see also Bybee 2007b:316). Furthermore, highly frequent chunks are processed as single items (Bybee 2010:47; 52). In Bybee's terms, both sequential links and holistic exemplars describe the strong unit-like behaviour of frequent sequences.

According to Bybee, it is possible for two smaller chunks to form a larger one (Bybee 2010:34). She, however, cautions that the longer a sequence, the more specific and less "useful" it is. This means that larger chunks usually occur less often than smaller ones (Bybee 2010:35). As well as this, the fact that the smaller chunks within the larger one are more frequent than the whole means that some relations in the chunk are stronger

than others. Within the larger chunk (23), for example, the 'sub-chunk' (24) is more frequent than both the whole and the sub-chunk (25).

(23) I don't know
(24) I don't
(25) don't know

In a study of vowel reduction, Bybee finds that "an adverb intervening between the subject and *don't* blocks vowel reduction [in *don't*], [...] but an adverb between *don't* and the verb does not" (Bybee 2010:44 based on Bybee and Scheibman 2007). She argues that this means that the cohesion between *I* and *don't* is stronger than between *don't* and *know* and concludes that there can be "varying degrees of cohesion and constituency" (Bybee 2010:44) within larger chunks.

According to Bybee, these differences in cohesion are "what gives language its hierarchical structure" (Bybee 2010:35). Thus the traditional notion that constituency leads to sequentiality is rejected. Bybee and colleagues hypothesise that the opposite is the case, i.e. that "sequentiality is more basic than hierarchy" (Bybee 2007b:326, see also Bybee 2010:136-8). Beckner and Bybee (2009:33-5), for example, argue that complex prepositions like *in spite of* are not only chunked, but also behave like constituents. Bybee furthermore maintains that contractions, as they occur between modals and *have* (e.g. *could've*), modals and *not* (e.g. *shouldn't*), subject pronouns or nouns and auxiliaries (e.g. *I'll, they're*), are proof of their behaviour as constituents (Bybee 2007b:327-8 see also summary in Bybee 2010:137).

Bybee and colleagues (e.g. Bybee 2002; 2007b:327-9; 2010:37-45; Bybee and Scheibman 2007) have furthermore repeatedly shown that phonetic reduction phenomena are affected by usage frequency. Importantly, reduction not only shows word frequency effects but also phrase frequency effects: "it is not just the frequency of the word that determines its degree of reduction, but rather the frequency of the word in the reducing environment" (Bybee 2010:37, based on Bybee 2007c). This means that reduction is primarily correlated to chunk frequency. For instance, Bybee and Scheibman (2007:298-9) in their study of *don't* reduction show that flapping only occurs when the subject is a pronoun and the strongest reduction in the vowel only occurs when the sequence is *I don't* – the most frequent chunk in the set.

A final point of note is the fact that Bybee is doubtful of probabilistic approaches in general and collostruction analysis (Stefanowitsch and Gries 2003) in particular (Bybee 2010:97; see also Bybee and Eddington 2006). Collostruction analysis determines the attraction of a particular lexeme (L) to a construction (C), by taking into account

- the frequency of L in C,
- the frequency of L in all other constructions,
- the frequency of C with lexemes other than L and
- the frequency of all other constructions with lexemes other than L (Stefanowitsch and Gries 2003:218).

The resulting measure of collostruction strength is probabilistic in the sense that it evaluates how likely L and C are to occur together given their distribution in the data. The more strongly observed frequencies of co-occurrence exceed expected co-occurrences, the stronger the collostruction. This means that high-frequency lexemes need to occur in a construction more often than low-frequency lexemes to receive the same collostruction strength rating.

Bybee argues that the frequency of L in C should be the most important factor determining collostruction strength "with perhaps the frequency relative to the overall frequency of the construction playing a role" (Bybee 2010:97). Her argument rests on the fact that a high-frequency and a low-frequency lexeme occurring equally often in a given construction do not receive the same collostruction strength rating. She holds that this method of calculation presupposes that the mind somehow "devalues" high-frequency lexemes in a construction and that collostruction analysis does not address by which "cognitive mechanism" speakers' minds do this (Bybee 2010:100-1). As other measures of association (such as lexical gravity G; see Section 3.3.1 below) also "devalue" frequency in this way, this point may apply to probabilistic approaches in general.

Finally, Bybee rejects the claim made in collostruction analysis that measures of association can reflect semantic relations. She holds that "[s]ince no semantic considerations go into the analysis, it seems plausible that no semantic analysis can emerge from it" (Bybee 2010:98).

Arnon and Snider (2010) also investigate how exactly frequency affects the mental representation of MWUs. Their focus is on the question of whether there is a frequency threshold at which MWUs are stored holistically. To this purpose, they analyse frequency effects in the comprehension of four-word phrases. In a decision task, adult native speakers of English are shown four-word phrases and asked to judge whether these are possible in English (Arnon and Snider 2010:70). The authors' design broaches a number of fundamental issues. First of all, they work with compositional, i.e. semantically transparent, phrases, such as (26) and (27) (2010:79-80; number in brackets indicate the frequency per million words).

(26) All over the place (21.45)

(27) Know what that is (6.25)

These are contrasted with low-frequency counterparts such as (28) and (29), which differ from them only in the last word and are controlled for sub-string frequency (i.e. the last words in a pair are always similarly frequent; the same goes for all two- and three-word combinations within pairs.)

(28) All over the city (0.65)

(29) Know what that was (1.05)

Arnon and Snider postulate that semantic non-compositionality is not required for frequency effects. Importantly, their choice of items shows that they expect effects on the concrete word-form level as opposed to effects on a lemmatised level or across constructions.

They furthermore contrast the threshold and continuous approaches (Arnon and Snider 2010:70). If the threshold approach is correct, frequency effects should only be observable for high frequency sequences and a model with two frequency bins (i.e. high and low) should suffice. If, however, every occurrence strengthens the activation of the sequence, then effects should be observable for strings of all frequencies and a continuous measure of frequency should be more apt to describe results (2010:70).

The study shows that participants' reaction times are significantly influenced by string frequency. They consistently respond faster to higher frequency items, irrespective of whether the high/low threshold is set at ten, five or one instance(s) per million words (Arnon and Snider 2010:72-4). This is not compatible with approaches postulating independent representations for high-frequency strings only and no representation of lower-frequency sequences (cf. e.g. Wray 2002). Based on these findings, Arnon and Snider proceed to demonstrate that frequency, if modelled as a continuous variable, is indeed a better predictor of reaction times than if modelled as a binary factor (2010:75). Thus the authors not only show that there are frequency effects in processing for compositional surface forms but also that the effect is gradient.

They point out that apart from their findings, there are theoretical arguments to reject threshold approaches. Arnon and Snider argue that "speakers cannot know a priori which phrases will become frequent enough to merit storage" (2010:77), so information would initially have to be stored for all phrases, only to be discarded for low-frequency strings later on, which the authors deem unlikely. In a similar vein, they argue against approaches which only permit idiosyncratic expressions to be included in the lexicon: as the child learner does not yet fully grasp which sequences are

semantically transparent and which are not, he is not fully able to treat them separately (Arnon and Snider 2010:77).

Finally, Kapatsinski and Radicke (2009) also address the issue of holistic storage. They conduct experiments wherein speakers need to detect the sequence /ʌp/ in words and MWUs. Theoretically, facilitated processing of frequent MWUs could be caused either by strengthened mutual activation between the words in the MWU ("distributed account" Kapatsinski and Radicke 2009:500) or through the representation of larger units in the lexicon ("localist account" Kapatsinski and Radicke 2009:500).

Results of earlier studies (by Morton and Long 1976 as well as Vogel Sosa and MacFarlane 2002) suggest a U-shaped effect of MWU frequency on accessibility of the words within the MWU: detectability of parts increases the more frequent the MWU, but decreases for "ultra-high-frequency" MWUs (Kapatsinski and Radicke 2009:503). Kapatsinski and Radicke argue that this curve could be explained by both the localist account and the distributed account:

Localist Explanation: Highly frequent sequences are stored in the mental lexicon as MWUs. For any sequence with a frequency below the threshold for storage in the lexicon, an increase in frequency of the phrase leads to improved predictability and therefore detectability of the words within it. In the case of high-frequency sequences, however, both the words and the sequence are represented in the lexicon which leads to "between-level competition during lexical access" (Kapatsinski and Radicke 2009:501). Hence in the case of highly frequent MWUs, competition from the phrase hinders detection of the parts. Importantly, the localist account differs from other frequency threshold approaches, which postulate combined storage for (highly) frequent MWUs, but do not predict any frequency effects for strings below the threshold for combined storage.

Distributed Explanation: MWUs are not separately represented in the lexicon. Instead, the more often the sequence is used the stronger the mutual activation between its parts. This means that the higher the frequency of the MWU, the more the words within it prime each other and consequently the more easily predictable and detectable they are. Detectability, however, also depends on "how surprising, and therefore salient, the occurrence of the word is" (Kapatsinski and Radicke 2009:504). This so-called 'surprisal' is inversely related to predictability: the more predictable a word, the less surprising it is and therefore the harder it is to detect. The combination of these two factors also predicts a U-shaped effect (Kapatsinski and Radicke 2009:503-4).

In order to test whether the proposed U-shaped effect emerges, the authors conduct two experiments. First, they select 'verb + *up'* collocations from Google and the British National Corpus which range across the entire frequency spectrum (e.g. *get up, sign up, line up, catch up, eke up*). For practical reasons, this set is grouped in seven frequency bins. The collocations are then used in 240 experimental sentences which are read out and recorded. The same number of control sentences not containing *up* is created. The recordings are played to 20 native speakers of English, who are asked to press a 'present' button as soon as they hear *up*. If they do not hear it, they have to press an 'absent' button after the sentence. Participants' reaction time and error rate are recorded.

As expected, results show a U-shaped curve. The higher the frequency of the verb-particle collocation, the faster participants detect it and the more accurate their responses, "except for the highest-frequency collocations. Detecting the particle is harder in highest-frequency verb-particle collocations than in less frequent collocations" (Kapatsinski and Radicke 2009:515). Furthermore, a quadratic function is significantly better at describing reaction times than a linear model (Kapatsinski and Radicke 2009:509).

As both accounts predict the same effect for /ʌp/-detection in MWUs, the first experiment confirms both accounts equally and does not help to discern which route of processing is taken. However, localist and distributed predictions for /ʌp/-detection in individual words (e.g. *puppy*) are conflictive. Therefore, Kapatsinski and Radicke claim that the pattern of frequency effects encountered for words must answer the question of whether larger units are stored locally.

Both theories assume that individual words are stored in the mental lexicon irrespective of their frequency. According to Kapatsinski and Radicke, for /ʌp/-detection in words, the distributed account predicts that the more frequent a word, the more the word's parts prime each other and parseability rises. Consequently there should be a positive correlation between word frequency and /ʌp/-detectability (Kapatsinski and Radicke 2009:504). The localist account, on the other hand, predicts that the more frequent the word, the more strongly it is activated as a whole, which should lead to increased competition between the parts and the whole as frequency increases. As a result, there should be a negative correlation between word frequency and /ʌp/-detectability (Kapatsinski and Radicke 2009:504).

Thus the experiment is repeated with stimuli in which /ʌp/ does not occur as a particle but inside other words (e.g. *cup, hiccups, upholstery;* Kapatsinski and Radicke 2009:506). In addition to word frequency, location of /ʌp/ within the word, length of

the word, whether /ʌp/ is a morphological or syllabic constituent, stress and duration are included in the analysis (Kapatsinski and Radicke 2009:511). Results reveal that /ʌp/ is harder to detect the more frequent the word, particularly if /ʌp/ is not in word-initial position (Kapatsinski and Radicke 2009:514-5). Importantly, /ʌp/-detection in highly frequent words does not deviate from the pattern.

In summary, the authors' results are consistent with the localist hypothesis.

> They indicate that the highest-frequency phrases are stored in memory as lexical units but they also suggest that a phrase needs to be extremely frequent to be stored in the lexicon. (Kapatsinski and Radicke 2009:516)

In summary, all of the approaches detailed in this section investigate co-occurrence patterns of words in speech. Pawley and Syder (1983) point out that the great fluency with which native speakers can produce utterances and the idiomatic choices native speakers make can only be explained by the postulation that not every sentences is creatively assembled word by word, but that speakers must have an inventory of units longer than the word mentally available which they can use to create utterances. A particularly interesting aspect of their study in the context of my own approach is the finding that speakers utter the sequences that are supposed to be mentally stored more fluently than other stretches of speech. The memorised sentences and lexicalised sentence stems postulated by Pawley and Syder are, however, defined introspectively and are in need of empirical corroboration.

Sinclair (1991) confirms the first of Pawley and Syder's hypotheses that there are stretches longer than the word which a speaker has stored in memory and that speakers can alternate between using these larger building blocks and using individual words to form utterances. It is Sinclair's contribution to point out that we should not rely on introspection, but on co-occurrence patterns in large-scale corpora to determine which words together constitute a unit.

Biber et al. (1999) conduct such a large-scale corpus study extracting frequent three- to six-word sequences. Theirs is a typical example of an n-gram-based approach which strictly operates on a surface, word-form level, not allowing for variation or open slots. Interestingly, they find that frequently occurring stretches in speech and writing often encompass parts of several constituents and do not necessarily begin or end at constituent boundaries. The methodology applied in this study will be based on Biber and colleagues' yet go beyond in two respects: Firstly, Biber et al. impose an arbitrary frequency threshold, discarding all sequences which are less frequent, which leads to the loss of a lot of material. Thus, more complete information about the building blocks of language can be obtained if all sequence types are extracted and coded for their token

frequency. Secondly, Biber et al. obtain knowledge about which sequences occur frequently in both written and spoken language yet without including other factors, like fluency or reading pace, they cannot draw conclusions about the role these patterns play in processing or why they emerge.

It is noteworthy that both Pawley and Syder (1983) and Biber et al. (1999) draw attention to the fact that we may find idiosyncratic variation. Not all speakers make the same experiences or have the same preferences, which may lead to different patterns of co-occurrence in their output. Biber et al. control for this by only taking sequences into account which occur in at least five different texts. It will be interesting to see whether we find indications for large differences in speakers' preferences in the following studies and in how far these can be problematic for corpus studies analysing the speech of many speakers at once.

The work by Bybee and colleagues is particularly important in this context because Bybee interprets her findings in the context of an elaborate model of the mind. In her exemplar model, entire sequences receive a holistic representation in the mind at the very first encounter. Any encounter of the unit from then on strengthens its representation. Usage is further expected to strengthen the 'sequential links' between the elements themselves, so that words that are often used together "prime[...] or automate" each other (Bybee 2007b:316).

Kapatsinski and Radicke (2009) explicitly contrast two models of the mind. According to the first, the distributed account, increases in usage frequency of a sequence lead to stronger mental associations between the units in the sequence, which would be in line with a concept of strengthened sequential links between words. Yet, according to the localist account, the sequence receives its own representation in memory. Kapatsinski and Radicke interpret the results of their experiments as evidence for the localist account. The localist model for which they find evidence, however, differs from Bybee's in that only high-frequency sequences are expected to be stored holistically. Words in sequences below the threshold for holistic storage behave as predicted by the distributed account: associations between them are strengthened with increasing co-occurrence frequency (Kapatsinski and Radicke 2009:518).

Arnon and Snider (2010) also contrast a threshold approach which predicts only storage of entire sequences above a certain frequency and a continuous approach which predicts that all sequences are stored and get gradually more entrenched the more frequently they are encountered, which means that the level of activation between the words rises. Their experiment provides evidence for the continuous approach.

Thus this overview of different approaches showed that research into co-occurrence patterns can reveal meaningful building blocks which speakers have recourse to when

constructing utterances. The postulation of such MWUs can, in turn, explain a variety of effects found in language change and language processing. The studies detailed here, however, either define MWUs introspectively or rely on absolute co-occurrence frequencies and not on relative measures of the association between two words. It is highly relevant to see whether the implementation of these measures improves model building. Finally, while all studies are based on the assumption that MWUs are cognitively represented in some way, there is not yet one model of the mind which can conclusively explain all effects.

2.3 Hesitation Placement

The following analyses rest on the assumption that hesitations are not scattered throughout speech at random; instead they are located at the boundaries of units of encoding. This section provides an overview of previous findings concerning regularities in hesitation placement. It illustrates which phonological and syntactic units have been proposed as units of encoding and thus as determinants of placement and comes to the conclusion that structural units alone cannot fully explain where speakers stop to hesitate. I furthermore present evidence that chunks and other multi-word units can serve as an explanation for speakers' choice of interruption point.

The section is structured according to the different units proposed as determinants of placement: intonation units, syntactic constituents and MWUs. A further separation of studies according to their research object, i.e. the type of hesitation, will be forgone as, despite displaying certain differences in preference (i.e. some hesitations are, for example, rather placed at sentence boundaries while others are preferred at phrase boundaries), there are overwhelming consistencies across all types of hesitations concerning where they are *not* placed, i.e. they tend to occur at the boundaries between rather than within units.

2.3.1 Hesitation Placement Depending on Intonation Units

Studies by Boomer (1965) as well as Clark and Fox Tree (2002) which concentrate on phonemic clauses as the basic units of encoding generally distinguish between hesitation placement at unit boundaries, after the first word and later in the unit. They compare how often hesitations are placed at each of these locations in order to discern regularities.

Boomer (1965) analyses interviews of 16 American men and finds that filled and unfilled pauses are preferentially placed after the first word in a phonemic clause, irrelevant of the length of the clause (1965:151-3). He argues that this proves that language planning takes place at the level of the phonemic clause and that pauses are most likely to occur after the first word because at this point "at least a preliminary decision has been made concerning [the clause's] structure", but it is the point "before the lexical choices have been finally made" (Boomer 1965:156). According to him, "[h]igh information lexical words" (Boomer 1965:155) are placed later in the phonemic clause.

Clark and Fox Tree (2002) argue that claims about preferred hesitation locations need to be made based on relativised frequencies, as each phrase has only one initial planning point and one after the first word, but many within it. The authors claim that most

planning is required at the boundaries of intonation units, less after the first word and least at later points within the intonation unit (Clark and Fox Tree 2002:94). Comparing the actual number of *uh* and *um* occurring in each of these three locations in the London-Lund corpus to the number of opportunities of occurrence in a given location, they find that filled pauses are most likely to occur at intonation unit boundaries (43 per 1,000 opportunities), followed by after the first word (27 per 1,000 opportunities) and within the intonation unit (13 per 1,000 opportunities; Clark and Fox Tree 2002:94). Furthermore, fillers are more likely to be followed by a pause within an intonation unit than at the boundary (Clark and Fox Tree 2002:95). Clark and Fox Tree account for the differences in co-occurrence with pauses by arguing that fillers are less needed at unit boundaries because it is acceptable for speakers to pause at boundaries. Within a unit, the "local importance" (Clark and Fox Tree 2002:97), or disruptiveness, of a pause is greater. Speakers therefore feel a greater need to mark it with a filler.[5]

In summary, Clark and Fox Tree's results indicate that hesitations are preferentially placed at the boundaries of intonation units. Their assessment that listeners expect pauses to be placed at unit boundaries and that speakers mark diverging behaviour with fillers corroborates this finding. The approach, however, cannot explain observed variation, particularly between placement at the boundary versus after the first word.

2.3.2 Hesitation Placement Depending on Constituents

The majority of studies (cf. Maclay and Osgood 1959; Goldman-Eisler 1968; Cook 1971; Clark and Clark 1977; Shriberg 1994; Clark and Wasow 1998; Biber et al. 1999; Bortfeld et al. 2001; Schilperoord and Verhagen 2006) take the word, phrase or clause as the unit of language planning and study hesitation placement in relation to these units and their boundaries. Planning points emphasised are mostly sentence and phrase boundaries. Furthermore, the question of whether content or function words are stronger attractors of hesitations receives some attention. Where not indicated otherwise, studies make no reference to a particular model of language planning.

Maclay and Osgood (1959) analyse the distribution of filled and unfilled pauses in a 50,000 word sample of conference contributions. They select 16 frequent phrase types ('frames') of different length, i.e. six noun phrases (e.g. 'Determiner Noun' and 'Determiner Adverb Adjective Noun'), eight prepositional phrases (e.g. 'Preposition Verb-ing Noun' and 'Preposition Determiner Adjective Noun') and two verb phrases ('Verb(fin) Verb(inf)' and 'Modal Verb(fin) Verb(inf)') and record the placement of filled and unfilled pauses within them (Maclay and Osgood 1959:31-2). They find that both

[5] For a discussion of further aspects of Clark and Fox Tree 2002 see Section 3.1.3.1.

types of pauses are significantly more likely to occur before content words than before function words (Maclay and Osgood 1959:32-3) and that about half the filled and unfilled pauses occur at phrase boundaries (Maclay and Osgood 1959:33). Maclay and Osgood conclude that these placement preferences result from speakers starting to utter a constituent before they have made the most difficult lexical choices within it (Maclay and Osgood 1959:41).

Based on Boomer's (1965) and Maclay and Osgood's (1959) results, Clark and Clark (1977) introduce a typology of potential planning points within constituents. Their three major planning points are

a. Grammatical junctures
b. Other constituent boundaries
c. Before the first content word within a constituent (Clark and Clark 1977:267-8)

According to Clark and Clark, at the first point, speakers stop to plan the sentence outline and its first constituent. At the second point, speakers stop to plan the next constituent. The last point corresponds to the moment where "speakers have committed themselves to the syntactic form of the constituent [...], but before they have planned the precise words to fill it out" (1977:268).

The predominance of these planning points has been confirmed by other studies (cf. Goldman-Eisler 1968, Bortfeld et al. 2001, Shriberg 1994 and Biber et al. 1999). Goldman-Eisler (1968:95) reports that between 47 and 61 per cent of unfilled pauses in her data of cartoon descriptions[6] occur at phrase boundaries and Bortfeld et al. (2001:138) find that in their experimental setting 39.6% of fillers are placed at phrase boundaries.

Shriberg (1994:149-51) shows that in Switchboard about 43% of filled pauses occur in sentence-initial position. When relativised according to the chance of occurring sentence-initially versus medially, filled pauses are in fact found to be four times more likely to occur sentence-initially than sentence-medially. Shriberg adds that there is a significant, but weak, correlation between filler type and sentence position: *um* is slightly more likely to occur in sentence-initial position while *uh* occurs more frequently in sentence-medial position (Shriberg 1994:154; see also Swerts 1998:490).

---

[6] Participants were asked to describe cartoons and to summarise their morale. The task was very specific and included precise instructions such as "formulating the general point, meaning, or moral of the story in as concise a form as you can" (Goldman-Eisler 1961:165) and "stick to the first reasonable version, and then keep repeating the same wording" "until I stop you" (which meant six repetitions each; Goldman-Eisler 1961:165). Other studies have shown that such "lexical suppression" and lack of options may lead to *more* filled pauses (Christenfeld 1994:198; for further critique of Goldman-Eisler's study see O'Connell and Kowal 2004). Whether this may also have an effect on the *placement* of pauses is yet unknown.

Biber et al. (1999) in turn claim that it is unfilled pauses which have the greatest propensity to occur "at major boundaries between syntactic units" in their corpus of British and American conversations, while filled pauses additionally "occur at lesser or medial syntactic boundaries [...], such as before the beginning of dependent clauses and coordinate constructions" (Biber et al. 1999:1054; see also 1060 and Swerts 1998:489).

Cook's (1971) findings confirm the penchant of hesitations to occur sentence initially. In eleven interviews conducted at the university of Aberdeen, he finds significantly more filled pauses than would be expected "before pronouns, conjunctions, 'well', 'yes', 'no'" (Cook 1971:138) – all words typically occurring in sentence-initial position (see also Biber et al. 1999:1057-8 on repetitions of pronouns and conjunctions). These findings aside, he notes that filled pauses also frequently occur before the second and third word in a clause (1971:138), which emphasises the relevance of the third position in Clark and Clark's typology.

Schilperoord and Verhagen (2006:145), on the other hand, find that, in a corpus of dictated letters by Dutch lawyers, unfilled pauses are predominantly placed *after* determiners and conjunctions, which are typically the first and second words in a clause. This is confirmed by findings from Fox, Hayashi and Jasperson (1996), who analyse repetitions in prepositional phrases in English and Japanese and find that English speakers often repeat the preposition and article preceding the noun, apparently to delay the production of the latter. In Japanese which has postpositions rather than prepositions, speakers "do not use recycling to delay the production of nouns" (Fox, Hayashi and Jasperson 1996:205), because they have no material available before the first content word in the phrase.

Altogether, the results presented in this section so far are largely homogenous and tie in with the results from studies taking the intonation unit as the level of language planning. The predominant location of hesitations is generally described to be at the boundary between larger units, followed by the position preceding the first content word within a unit, thus confirming Clark and Clark's typology. This has been interpreted as speakers planning at the phrase or clause level, occasionally starting their utterance, however, before they have finalised the lexical choices.

The hesitation pattern mostly found is that pauses at boundaries are more likely to be unfilled, while pauses within units are more likely to be filled[7]. This pattern confirms Clark and Fox Tree's (2002) conclusion that pausing is more acceptable at the boundary between syntactic or intonation units, so that speakers feel less need to mark any disruptions at boundaries with fillers.

---

[7] cf. however Maclay and Osgood (1959:33-4), who find that the filled pause is more likely to occur at phrase boundaries and before function words than the unfilled pause.

Clark and Wasow's (1998) "commit-and-restore model" of speech production, which accounts for the location of repetitions and repeat rates of different parts of speech, offers some explanations for the predominance of repetitions at constituent boundaries which can be generalised to other types of disfluencies.

The authors find that in the Switchboard corpus, certain function words, such as conjunctions, determiners and pronouns, show far higher repeat rates than content words (Clark and Wasow 1998:210-1). As function words are generally far more frequent than content words, the difference in repeat rates could result from the fact that the former are more easily accessible. Being more easily retrievable, function words can be repeated until the rest of the sentence has been fully constructed (cf. Clark and Wasow 1998:206, 210). Yet the authors show that the accessibility hypothesis cannot account for the fact that *I'm*, *at*, *had* and *because* (repeat rates: 56.1, 16.3, 8.9, 3.1 per thousand) are far more likely to be repeated than equally frequent *out* and *them* (repeat rates: 1.2 and 0.5 per thousand; Clark and Wasow 1998:211). Clark and Wasow interpret this finding as evidence that speakers plan speech "one major constituent at a time" (Clark and Wasow 1998:204), because the most-repeated content words were those which typically occur at the left edge of phrases and larger constituents. Indeed, more detailed investigations show that the same personal pronoun is more likely to be repeated occurring at the left edge of a constituent than elsewhere (Clark and Wasow 1998:216).

Importantly for the present context, Clark and Wasow also make claims about the role of the complexity of constituents. The more complex the constituent, the more likely it will commence with a hesitation. They test this hypothesis by means of analysing 500 fluent NPs and 500 NPs containing repetitions in Switchboard, comparing restart rates in complex constituents, simple constituents and fixed phrases and find that speakers are more likely to produce (30) than (31) or (32) (examples taken from Clark and Wasow 1998:212, 235).

(30) *Complex constituent*: the the time we were there at the warehouse

(31) *Simple constituent*: the the diesels

(32) *Fixed phrase*: a a lot of

The authors attribute differences in repeat rates to complexity; the greater the grammatical weight of a constituent, the more difficult it is to plan (Clark and Wasow 1998:205). The fact that it is predominantly the first word in the constituent that is repeated is interpreted as evidence that major constituents, irrelevant of their length, are the basic units of speech planning (Clark and Wasow 1998:204). Low repeat rates in fixed phrases must thus mean that they are the least complex. The authors, in fact, argue

that these are no longer fully compositional and possibly even planned as units (Clark and Wasow 1998:212).

In summary, the studies presented in this and the previous section provide evidence that speakers hesitate in predictable locations, which must result from the units in which speech is planned. None of the presented theories, however, can conclusively account for why speakers sometimes prefer to hesitate at constituent boundaries and sometimes within them. Explanations pointing to the complexity of the upcoming constituent (e.g. Clark and Wasow 1998; see also Goldman-Eisler 1968:60-9) cannot account for hesitations in different locations in constituents of equal complexity.

2.3.3 Hesitation Placement Depending on Usage-Based and Probabilistic Factors

As early as 1954, Lounsbury recognised that "statistical uncertainty" could explain hesitation placement. He claims that due to syntactic restrictions and cultural preferences, "certain message events co-occur more often than others" (Lounsbury 1954:96), leading to "habits of varying strength" (1954:96), reflected in different transitional probabilities[8]. He hypothesises:

> *Hypothesis 1*: Hesitation pauses correspond to the points of highest statistical uncertainty in the sequencing of units of any given order. (Lounsbury 1954:99)

He claims that the lower the transitional probability at any given point, the longer the pause[9] at that point (1954:98).

> *Hypothesis 2*: Hesitation pauses and points of high statistical uncertainty correspond to the beginning of units of encoding. (Lounsbury 1954:100)

> *Hypothesis 3*: Hesitation pauses and points of high statistical uncertainty frequently do not fall at the points where immediate-constituent analysis would establish boundaries between higher-order linguistic units or where syntactic junctures or 'facultative pauses' would occur. (Lounsbury 1954:100)

This early psycholinguistic concept of units of encoding which are characterised by high transitional probabilities (or at least transitional probabilities higher within them

---

[8] Note that according to Lounsbury, habits form on at least three levels, a semantic, a grammatical and a motor skill level. Hesitations, however, are supposed to only reflect units on the semantic level of encoding (Lounsbury 1954:98).

[9] Lounsbury never explicitly defines which kinds of pauses – filled or unfilled – he considers. He analyses pauses as the "latency" (1954:98) period between two events and explains that in speech "[t]here are pauses *and* perhaps quite a bit of hemming and hawing" (1954:98; emphasis added), which leads to the conclusion that his theories are meant to apply to unfilled pauses only.

than between them and their surrounding context) and which do not necessarily correspond to phrases or larger constituents is very much in line with a probabilistic concept of chunking.

Lounsbury himself did not yet have the computational possibilities to provide evidence for his hypotheses, which have been taken as starting points in studies to follow. These studies can roughly be divided into three groups: *early work* from the pre-computer era with limited statistical possibilities and small data sets, *later studies* applying advanced empirical methods and, finally, studies analysing hesitations as a means of underpinning usage-based theories of MWUs, where the quantitative procedure is often not at the focus.

Early Work

Goldman-Eisler's (1968) seminal work belongs to the early group. She uses a Cloze procedure to estimate transitional probabilities, whereby participants are given the preceding context of a word and are asked to guess the word. The faster participants guess the word, the higher the transitional probability (1968:34). Based on the transitional probabilities thus obtained, she concludes that pauses occur where transitional probability is low. Most importantly, she notes that "[t]he relationship[...] was not reciprocal. Forty-six of the 75 words of low predictability (p=0 or 0.10) were not preceded by pauses but by fluent speech" (Goldman-Eisler 1968:37).

Based on further experiments, she adds that speech production is influenced by "the probability structure of subsequent speech" (corresponding to direct transitional probability) as well as of "preceding speech" (corresponding to backwards transitional probability; Goldman-Eisler 1968:43; see also Section 3.3.1). Goldman-Eisler eventually concludes that a part of language must be prefabricated:

> The conception of ready-made sentence schemata, models of sentences or modules implies that they are selected in one piece so to speak, that they are not constructed from individual lexical elements – and this would account for the fluency of speakers irrespective of their complexity, in the same way as efficiency in mass production is a matter of use of prefabricated units. (Goldman-Eisler 1968:128)

While Goldman-Eisler's experimental work lacks the statistical sophistication to be considered evidence of probabilistic influences on hesitation placement in today's sense, her findings on direct transitional probability are confirmed by other Cloze-based studies (cf. e.g. Beattie and Butterworth 1979), corpus-based studies (cf. e.g. Kapatsinski 2005) and by n-gram-based language models (cf. e.g. Shriberg and Stolcke 1996).

Later Studies

Several studies show that low frequency as a form of low probability affects speakers' hesitation behaviour: infrequent words are more likely to be repeated or preceded by hesitations than highly frequent words (cf. Biber et al. 1999:1059; Beattie and Butterworth 1979:208; Kapatsinski 2010). Kapatsinski points out that "a high-frequency word is a more cohesive unit" (2010:72) and comes to mind faster than a low-frequency word (2010:85), its production being more automatic and "less susceptible to conscious control" (2010:102). Therefore, even when length is held constant, low frequency words are more likely to be interrupted in a repair sequence (2010:83) in the Switchboard corpus.

Stolcke and Shriberg (1996) develop an n-gram-based language model in which filled pauses, repetitions and self-corrections are modelled like words. For this purpose they use the Switchboard corpus. They find that particularly at constituent boundaries, "the [filled pause] itself is the best predictor of the following word, not the context preceding the [filled pause]" (Stolcke and Shriberg 1996:3). They conclude that filled pauses "correlate strongly with certain lexical choices or syntactic structures" (Stolcke and Shriberg 1996:3).

While these studies make use of powerful empirical tools, they do not aim to uncover the cognitive processes underlying their results. In other words, they only analyse the data 'en masse' and are not interested in unveiling which linguistic structures are associated with particular probabilistic properties.

Theories of MWUs

There is very little work which explicitly brings together usage-based theories of MWUs and the placement of hesitations. Apart from Pawley and Syder's (1983) study introduced in Section 2.2.2, only Bybee (2007b) and Erman (2007) investigate the placement and length of pauses in relation to MWUs. In terms of design, these studies are the exact opposite of those studies described above, where the placement of hesitations was at the focus and refined statistical procedures were employed. Here, the cognitive formation of MWUs is the prime interest and hesitations are employed as (occasionally rather exemplary) evidence of this process.

Erman (2007) investigates the occurrence of filled and unfilled pauses in relation to 'prefabs', which are defined as MWUs with a "conventional meaning" and "restricted exchangeability", which "means that at least one member of the prefab cannot be replaced by a synonymous item without causing a change of meaning or function and/or idiomaticity" (2007:33). Erman explicitly does not take frequency into account. She notes that "there are numerous frequent word combinations that do not satisfy our criteria for 'prefabhood'" (Erman 2007:48). These are therefore considered sequences of

"open" (i.e. non-prefabricated) slots. This means that (33) and (34) are prefabs in Erman's sense, while (35) and (36) are not (examples taken from Erman 2007:33-4, 48)

(33) choose words

(34) Am I (ever) glad!

(35) open the window

(36) watch the telly

Erman codes all words in 30,000-word extracts from both the Bergen Corpus of London Teenager Language (COLT) and the London-Lund corpus (2007:27). She finds more pauses in non-prefabricated language (88.7%) than in prefabricated language (11.3%; 2007:41). Furthermore, pauses within prefabricated sequences were "considerably shorter" than those in non-prefabricated language (2007:41). There is, however, no difference in pausing behaviour before "fixed" and "semi-fixed" slots within prefabs (2007:43-4).

Due to Erman's highly subjective definition of prefabs and the impressionistic pause marking the study is based on (2007:38), it can hardly be considered a probabilistic account of pause placement and should rather be considered a rough indication that pause placement is not unrelated to MWUs.

In her work on the relation between sequentiality and constituent structure, the general point of which has already been introduced in Section 2.2.2, Bybee (2007b) analyses the formation of MWUs in the English noun phrase. For this purpose, she selects eleven highly frequent nouns, listed in (37), from the Switchboard corpus and analyses their preceding and following context and the pauses within this context (2007b:320).

(37) husband, mother, computer, movie, school, car, house, money, idea, class, problem

Bybee finds that pauses are more likely to follow nouns than to precede them (2007b: 320-1). Out of 7,870 nouns only one per cent is preceded by a pause while 34 per cent are followed by pauses. Bybee also finds that 'within-NP'-combinations of the type 'X +Noun', are more frequent than 'cross-phrasal' combinations of the type 'Noun+X'. From this frequency distribution and the placement of pauses, Bybee concludes that "the co-occurrence patterns for X+N are stronger than for N+X" (2007b:320-1). Thus she finds bonds between words within a traditional constituent to be stronger than bonds between words belonging to different constituents. Yet (38) and (39) show some instances where this is apparently not the case (examples taken from Bybee 2007b:322).

(38) lot/amount(s) of money

(39) to/in class/school

Of + Noun (particularly *money*) occurs frequently in Bybee's data. As *of* is not part of the noun phrase, this could point to strong bonds *across* phrase boundaries. Yet, Sinclair (1991:85-6) argues that in such cases we do not find a head (*lot*) and a postmodifying prepositional phrase. Instead the second noun (*money*) is the head, modified by *lot of*. According to this analysis, *of* and *money* can be considered to occur within the same constituent thus supporting Bybee's hypothesis (2007b:322). High rates of *and* and relative *that* following the noun (2007b:323), however, cannot be explained in this way.

Bybee's highly interesting claims that constituent structure is secondary to co-occurrence frequencies and that this can be deduced from pause placement in speech are still in need of further empirical support. We particularly need more data which 'zooms in' on types like (38) in order to show whether in these cases co-occurrence frequency is 'meaningful', in the sense that it reflects cognitively represented MWUs (which would therefore not be interrupted by pauses).

In summary, Goldman-Eisler (1968), Beattie and Butterworth (1979), Shriberg and Stolcke (1996), Biber et al. (1999), Kapatsinski (2005), Bybee (2007b), Erman (2007) and Kapatsinski (2010) illustrate that there is an influence of word and co-occurrence frequencies as well as transitional probabilities on hesitation placement.

However, results from Kapatsinski (2005) indicate that this may not always be the case. In an analysis of the influence of probabilistic factors on the extent of recycle in repetitions, Kapatsinski finds that in 92% of cases where a repair is "initiated within three words from a clause boundary" (Kapatsinski 2005:482), speakers start their repeat from the clause boundary (2005:483). This could be due to extremely low transitional probabilities at clause boundaries. Yet this is not the case in Kapatsinski's data. He finds that "speakers do tend to recycle to the nearest constituent boundary rather than a transition with lower frequency or probability by default" (Kapatsinski 2005:491). He finally concludes that such structural factors are so strong that speakers "ignore differences in frequencies and probabilities between words in a sentence, unless they are larger than a particular value" (Kapatsinski 2005:491). As none of the above-cited studies explicitly take the correlation between co-occurrence frequencies or transitional probabilities and constituent boundaries into account, it remains unclear whether this is also true for other types of hesitations, like filled and unfilled pauses. Furthermore, the exact way in which probabilistic and structural factors interact has yet to be analysed.

3 Data & Methodology

This section provides the methodological spine of the present study. It will first illustrate the make-up and coding of the Switchboard NXT corpus, which serves as the data base. It will then proceed to introduce the set of hesitations at the focus of the study as well as the software employed. Finally, the chapter will explain the predictors (i.e. independent variables) and the empirical methodology of the following studies.

3.1 Data

3.1.1 The Switchboard NXT Corpus

Switchboard NXT (*NXT Switchboard Corpus Public Release 2008*) as used here is a subset of the larger Switchboard corpus (cf. Godfrey, Holliman and McDaniel 1992), which is a 2.9-million-word sample of spoken American English, more precisely of telephone conversations between previously unacquainted adults representing all dialect areas of the United States (cf. Godfrey and Holliman 1997). It is the single most widely used corpus for studies of hesitation phenomena (cf. e.g. Acton 2011; Bell et al. 2003; Clark and Wasow 1998; Kapatsinski 2010; Shriberg 1994; Shriberg 1996; Shriberg and Stolcke 1996; Stolcke and Shriberg 1996; Tily et al. 2009). Switchboard has also been extensively used for the study of frequency effects (cf. e.g. Arnon and Snider 2010; Bybee 2007b; 2010; Gregory et al. 1999).

The original Switchboard corpus was collected by Texas Instruments in 1990-1 (cf. Godfrey and Holliman 1997). It consists of 2,430 conversations ranging between 1.5 and ten minutes in length, averaging six minutes each and totalling 240 hours (cf. Bell et al. 2003:1003; Calhoun et al. 2010:390). 543 speakers (302 male and 241 female) participated, of which 50 speakers were declared "target speakers" and were selected at least 25 times (cf. Godfrey, Holliman and McDaniel 1992:I-517-9). They thus contributed sixty to ninety minutes each. The remaining speakers participated in one to twenty calls (cf. Godfrey, Holliman and McDaniel 1992:I-519), though attention was paid so that "(1) no two speakers would converse together more than once and (2) no one spoke more than once on a given topic" (cf. Godfrey and Holliman 1997).

Prior to recording, participants were given a list of 70 predetermined conversation topics from which they chose a variety of topics they were interested in (cf. Calhoun et al. 2010:390), ranging from AIDS and air pollution to pets and vacation spots[10]. They were then paired by a computer-operated system with another participant who shared

[10] For a complete list of topics see http://www.ldc.upenn.edu/Catalog/docs/LDC97S62/topic_tab.csv.

one of their interests. The system then played a recording which introduced the conversation topic (cf. Godfrey and Holliman 1997; Godfrey, Holliman and McDaniel 1992:I-517-8). Speakers were given as much unrecorded time to introduce themselves as they wished. Furthermore, they were told that they were free to end the conversation at any time (cf. Godfrey, Holliman and McDaniel 1992:I-518).

The corpus has been transcribed, corrected and annotated several times. The NXT version, which will be used in this study, is based on two orthographic transcripts. One is the transcript released in the LDC's *Treebank3* (Marcus et al. 1999). This is an improved version of the *Switchboard-1 Release 2* transcript (cf. Godfrey and Holliman 1997), which in turn is based on the original 1993 Switchboard release (cf. Calhoun et al. 2010:391). Treebank3 only contains 1,126 of the total of 2,430 conversations in Switchboard (cf. Calhoun et al. 2010:391-2). 650 of the conversations included in Treebank3 were syntactically parsed. 642 of these, comprising a total of ca. 830,000 words, form the basis for Switchboard NXT. The remaining eight conversations had to be excluded due to technical difficulties (cf. Calhoun et al. 2010:394).

The second transcript included in Switchboard NXT is the so-called *MS-State* transcript (cf. Deshmukh et al. 1998), a manually-corrected and time-aligned transcript based on the Treebank3 release, produced by the Institute for Signal and Information Processing at Mississippi State University. While the MS-State transcript is generally more accurate than the Treebank3 version, both were included in the NXT release. This was necessary due to the extensive annotations available for each version (cf. Calhoun et al. 2010:392). Both transcripts were aligned at the word level. However, due to differences between the transcripts, partly caused by different transcription conventions (e.g. *don't* (MS State) vs. *do n't* (Treebank3)), but also by actual differences in the transcripts, 0.5% of Treebank3 words and 2.2% of MS-State words remain unaligned (cf. Calhoun et al. 2010:413). A simple transfer of all annotations from one orthographic transcript to another would have resulted in some inconsistent links, hence the transcripts were both included and aligned by attributing all words with unique IDs (cf. Calhoun et al. 2010:392-3).

Despite the MS-State transcript being more accurate, I opted for using the Treebank3 transcript for the present study, because the syntactic information is linked to the Treebank3 transcript only. Figure 3.1 shows all corpus and annotation layers and their relations. The layers used in the present study are given in black. The following sections will provide further information about the terminals and syntax layers of annotation, which are the two layers used here.

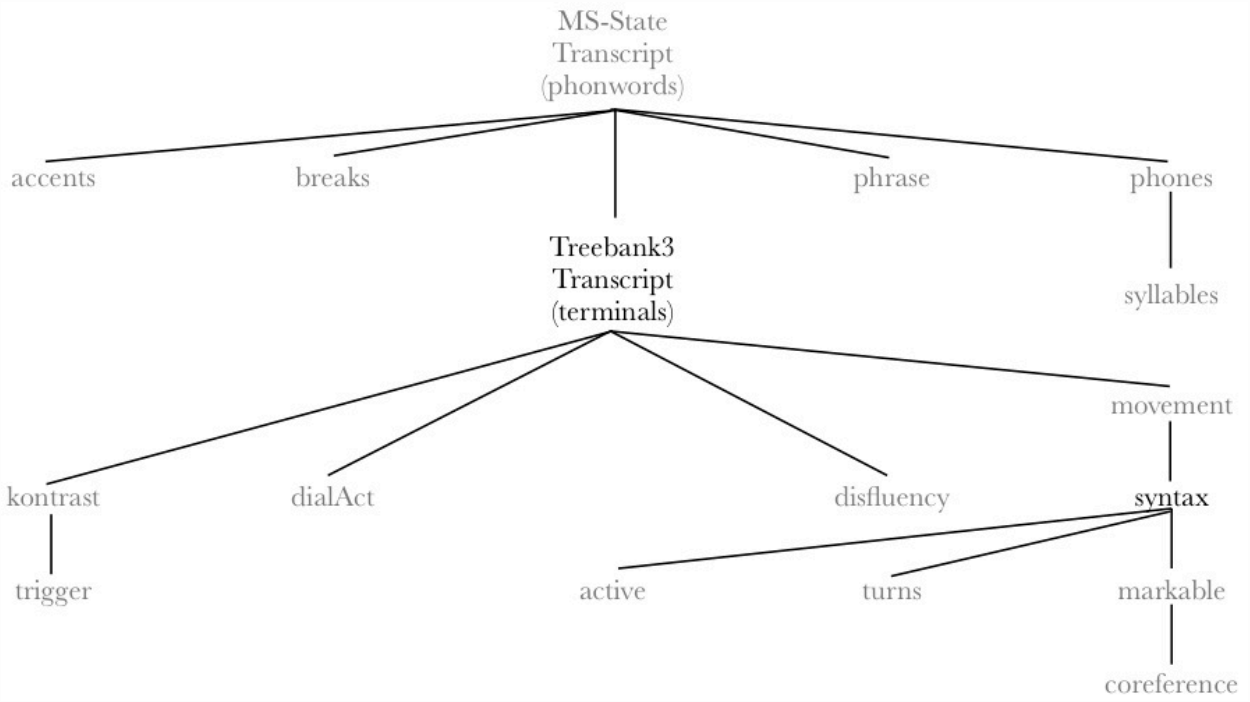


Figure 3.1: Layers of the NXT-format Switchboard Corpus (NXT Switchboard Corpus Public Release 2008)

3.1.1.1 The Terminals Layer

The terminals layer in Switchboard NXT is based on the Treebank3 transcript of the corpus (cf. Calhoun et al. 2010:393), but contains information beyond the orthographic transcript. Most importantly, it links every word in the Treebank3 transcript to its equivalent in the MS-State transcript and also to any other layer of the corpus (see Figure 3.1). The layer also contains start and end times of words as well as part-of-speech (POS) information. This has been slightly adapted from the Treebank3 transcript, which was annotated with the Penn treebank tagset, a condensed version of the Brown corpus' tagset (cf. Calhoun et al. 2010:393-4; Marcus, Marcinkiewicz and Santorini 1993:314).

One POS-tag worth mentioning is *UH*, which is used to mark a variety of elements which are not part of the emergent syntactic structure and (mostly) carry little or no propositional meaning. This includes the following groups: fillers (including *oh* and *huh*), exclamations (e.g. *wow*, *god*), *yes* and *no* answers (including *yeah*, *nope* etc.), greetings and adieus (e.g. *hello*, *bye-bye*), continuers and assessors (mostly one-word turns, such as *exactly* and *uh-huh*) and discourse markers (e.g. *well* and *like*; *Addendum to the POS Tagging Guidelines*; adapted from Santorini 1990).

For the present project, this category is both too narrowly and too widely defined. It is too widely defined, because it subsumes non-sentence elements belonging to many different parts of speech. While the original *Dysfluency Annotation Stylebook for the*

Switchboard Corpus (Meteer and Taylor 1995) demands labelling filled pauses, discourse markers and so forth with an individual letter so that they can be distinguished, the *Addendum to the Guidelines,* which has been applied in the tagging of Switchboard NXT, does not do so (cf. Meteer and Taylor 1995:5; *Addendum to the POS Tagging Guidelines*:2). Hence they can no longer be distinguished by the tag and appear as one large group. On the other hand, the tag is too narrowly defined, as it does not apply to the phrasal discourse markers *you know* and *I mean* (*Addendum to the POS Tagging Guidelines*:2). For these reasons, the *UH*-tag could not be used to retrieve the entire group of hesitations to be investigated (see Section 3.1.2). It proved useful, however, for distinguishing between the discourse markers *like* and *well* and instances of *like* and *well* as adverbs, verbs and nouns.

3.1.1.2 The Syntax Layer

The syntactic bracketing information for Switchboard NXT was drawn from the Penn Treebank Project and based on the Treebank3 transcript of the corpus (cf. Calhoun et al. 2010:393). The parsing of the Penn Treebank, in turn, is a hand-corrected version of the output of Fidditch, a deterministic parser (cf. Marcus, Marcinkiewicz and Santorini 1993:320; Hindle 1994).

Fidditch was designed to provide a theory-independent annotation of syntactic structures in written and spoken English (cf. Hindle 1994:104, 108). It constructs hierarchical trees annotated with structural traces representing deep-structure relations (1994:109, 111).

When Fidditch was used for the parsing of the Penn Treebank, some of its original categories were collapsed (cf. Marcus et al. 1993:320 and see Appendix B for a comparison of the different tagsets), resulting in a simplified, flat and relatively theory-independent annotation (cf. Marcus, Marcinkiewicz and Santorini 1993:321).

Despite the theory-independent bracketing system aimed at, a subset of Fidditch null elements was retained in the Penn Treebank because it was considered "the easiest mechanism to include information about predicate-argument structure" (Marcus, Marcinkiewicz and Santorini 1993:321). The conversion of the corpus annotations to Switchboard NXT lead to some further changes, most notably the inclusion of disfluency information and the shift of null elements to other layers.

3.1.1.3 Definition of a Word in Switchboard NXT

The word is defined as follows in Switchboard NXT:

- All elements separated by spaces are considered separate words. Consequently, complex prepositions are treated as sequences of prepositions, or prepositions and nouns etc. Non-hyphenated compounds, such as *nursing home* or *family member*, are treated as separate words in the corpus.
- Contracted forms such as *I'll* and *you're* are considered separate words. *Don't* is rendered as *do* + *n't*. Genitive *'s* is tagged as an individual word.
- Acronyms and alphabetisms are not treated the same: acronyms such as *AIDS* are treated as one word while alphabetisms, such as *IBM*, are treated as several words.

3.1.1.4 Definition of a Sentence in Switchboard NXT

As described above, the syntax layer of annotation in Switchboard NXT was adapted from the parsing of the Treebank3 transcript. Treebank3 is segmented into so-called 'slash units', which Calhoun et al. (2010:393) describe as "sentence-like chunks" or "utterance unit[s]". These were parsed as the highest-level S-brackets in Switchboard NXT (cf. Calhoun et al. 2010:393). The sentence count on all further levels of annotation in Switchboard NXT follows from this segmentation.[11]

Examples (39) to (41), annotated as sw2102.B.s70, sw2102.B.s71 and sw2102.B.s72 in the corpus, were consecutively uttered by a single speaker.

(39) (S (INTJ right))

(40) (S (INTJ well)
(SBAR (WHADVP when)
(S (NP my kids) (VP were (ADJP little) (ADVP)))) (NP I) (VP did (VP have (NP a T V set))))

(41) (S and (NP I) (VP did (VP watch (NP (NP (NP a lot) (PP of (NP Sesame Street))) and (NP (NP a lot) (PP of (NP Electric Company)))) (ADVP as well)))))

These examples illustrate that slash units were defined on strictly syntactic grounds; at any point where all syntactic brackets are closed, a new unit begins. This means all non-embedded clauses and a small number of extra-clausal elements are treated as

[11] The sentence codes given throughout this work are derived from Treebank3; the MS-State transcript's count occasionally differs.

separate units (cf. also Calhoun et al. 2010:393). The discourse markers *well*, *like*, *you know* and *I mean*, which will be extracted for the following analyses, do not receive a separate S-bracket and are consequently not treated as separate units on the syntax layer.

Interestingly, in Switchboard NXT another element which could be interpreted as a sentence marker is found on the terminals layer of annotation. According to the punctuation marks ("punc") annotated on this layer, (39) to (41) constitute a single unit, shown here as (42).

(42) Right well when my kids were little I did have a TV set and I did watch a lot of Sesame Street and a lot of Electric Company as well. (sw2102.B.s70-72)

For the following analyses, a sentence was defined as a parse unit on the syntax layer of annotation, meaning that (38) to (40) are treated as three sentences. This choice was made for two reasons. Firstly, it is unclear on what grounds (i.e. prosodic, semantic or otherwise) syntactic parse units were merged to form larger 'punc' units on the terminals layer. Secondly, automatic search scripts can more reliably recognise when a sequence of words makes a syntactic sentence than when it makes a 'punc unit' due to the fact that each word has been annotated with a unique code which contains the syntactic sentence count (e.g. sw2102.B.*s70*_6). This code, also called the 'nite id', functions as a link to other layers of annotation which means that, for example, information concerning the syntactic function of a word is obtainable via this code.

3.1.2 Hesitations: Definitions & Retrieval

I define hesitations as elements which are not part of the emergent syntactic structure, do not contribute to the propositional meaning of the utterance and are related to planning problems (see also Fox Tree 1995:709). This separates hesitations from interjections and non-linguistic material because only hesitations result from planning problems, yet it permits a wide variety of hesitation devices. Not all of these will be considered in the following study. Analyses will be restricted to hesitations which constitute an interruption of the speech flow, but where the utterance is continued immediately afterwards without recycling or altering parts of the message. This applies

to filled and unfilled pauses, like those in (43), as well as to a small set of discourse markers[12], like, for example, *you know* in (44).

(43) I play *uhm [pause]* once a week in the in the summer

(44) So when Alice and I go round we always *you know* borrow a book for it 's uh like our own library really

Other types of disfluencies were excluded for methodological reasons. First of all, phonetic lengthenings, which can be used as time-buying devices, were ignored because these cannot be searched automatically in the corpus. Secondly, repetitions and self-corrections including any subgroups thereof (e.g. false starts, deletions, substitutions) were also excluded. This choice will be illustrated by means of the repetition in (45) and the self-correction in (46).

(45) Play it back *to the to the* world

(46) So I suppose that means that *they could one could [pause]* take it over pretty quickly[13]

In (45), the speaker utters *play it back to the* fluently and then repeats *to the* before he produces *world*. Thus we have an interruption point after *the* and additionally a retraction point, i.e. the beginning of the prepositional phrase. The situation is similar in (46), where the speaker interrupts his utterance after the first instance of *could* in order to retrace back to the subject, which he replaces. (The following pause constitutes another interruption point.) By contrast, to describe the location of hesitations such as the pause and filler in (43) and the discourse marker in (44) only the interruption point is necessary. The speaker does not retrace, so there is no retraction point. Consequently, disfluencies with retraction and without retraction do not easily lend themselves to analysis in a single framework. This is why the present study is limited to unfilled pauses, the filled pauses *uh* and *um*, and the discourse markers *well*, *like*, *you know* and *I mean*.

---

[12] I have chosen to use the terms 'filled pause' and 'discourse marker', because these are the most popular labels (Schourup 1999:228). However, the concepts are also known by a variety of other names. Filled pauses are also referred to as automatisms (Fehringer and Fry 2007) or fillers (Vasilescu, Candea and Adda-Decker 2005) and have been variously classified as asides (Clark 2004), editing terms (Heeman and Allen 1999, Levelt 1983), hesitations (Biber et al. 1999), interjections (Clark and Clark 1977), performance additions (Clark and Fox Tree 2002), suspension devices (Clark 1996) and time-buying devices (Fehringer and Fry 2007). Discourse markers, in turn, are also referred to as discourse particles (Schourup 1985), editing expressions (Bortfeld et al. 2001), editing terms (Heeman and Allen 1999), meaningless particles (Hosman 1989) parenthetic phrases (Crystal 1988), pragmatic expressions (Erman 1987) or smallwords (Gilquin 2008).

[13] Examples taken from the British component of the International Corpus of English (ICE-GB). They are in order: <ICE-GB:S1A-025 #222:1:A>, <ICE-GB:S1A-025 #326:1:B>, <ICE-GB:S1A-018 #250:1:A>, <ICE-GB:S1A-023 #310:1:B>. Italics are my own and the transcription of pauses was changed.

While the research question and the set-up of the study call for selectiveness in the choice of hesitations, it allows for more flexibility where the causes for hesitations are concerned. Goldman-Eisler, for instance, defines three kinds of decisions required for speech production. Thus, according to her, hesitations can arise from the following (Goldman-Eisler 1968:32):

- Semantic choice – Planning the content of the utterance.
- Syntactic choice – Planning the outline of the syntactic structure.
- Lexical choice – Selecting the actual words.

Most hesitations, and particularly discourse markers, can also serve other mostly rhythmic and interactional functions (cf. Biber et al. 1999:1054; Boomer and Dittmann 1962:216; Clark 2004; Clark and Clark 1977:267; Deese 1984:112; Kowal and O'Connell 1993; Mukherjee 2007:580; Tottie and Svalduz 2009). And many pauses are, of course, caused by the physical needs of speaking, such as breathing (cf. Goldman-Eisler 1968:24-5).

Some studies (e.g. Erman 2007:32) therefore limit their interest to hesitations caused by making lexical choices. I reject this procedure, firstly because it is prone to circularities. Based on subjective criteria, the researcher excludes large numbers of hesitations and finds then that the remaining set conforms to his theory. Secondly, and most importantly, such subjective distinctions are not necessary in a study of hesitation placement and chunking. My hypothesis that highly cohesive word pairs are not interrupted should hold for the insertion of any extra-syntactic material. This means, no matter what the cause of *uh* or the function of *like*, if the mind is responsive to frequencies and probabilistic tendencies, speakers should not interrupt chunks with them.

The following sections detail how the different categories of hesitations were defined and explain the motivation behind these definitions.

3.1.2.1 Unfilled Pauses

Unfilled pauses are not marked in the corpus, but can be calculated from the given start and end times of words. Minimum and maximum pause lengths were adopted. The minimum threshold is based on Goldman-Eisler who argues that

> breaks in phonation of less than 0.25 sec [should] not [be] considered as discontinuities. This might mean loss of some data, but it ensures the clear

> separation of hesitation pauses from phonetic stoppages. (Goldman-Eisler 1968:12)[14]

The minimum pause length was therefore set to 0.2 seconds. Pauses of 0.2 seconds and above were included irrespective of whether the speaker breathed in the interval or whether he or she maintained or reduced the level of articulatory tension. This generalisation could be made because speakers should tend to articulate chunked sequences without interruptions and should prefer to breathe before or after them.

As it can be assumed that a speaker in a telephone conversation would be unlikely to pause for longer than one second without filling that pause with some kind of floor-holding or hesitation device, unless interrupted by the other speaker or external factors such as other people in the room or the door bell, all unfilled pauses of more than one second in length were also excluded[15].

Furthermore, unfilled pauses at utterance boundaries were excluded, because these cannot be faithfully attributed to one of the speakers. In total, the corpus contains 43,855 unfilled pauses of between 0.2 and 1 second in length which are not located at utterance boundaries.

3.1.2.2 Filled Pauses

Filled pauses are easily retrievable from the corpus because they are consistently transcribed as *uh* and *um* and POS-tagged as *UH*[16]. In this way, all 17,423 cases of *uh* and all 3628 cases of *um* were selected. Of course, discourse markers may also serve as pause fillers, yet they will be referred to as a separate group throughout this study. This choice was made based on the outcome of the study described in Section 3.1.3.1 below.

---

[14] This lower limit has been taken and adjusted, like, for example, in Boomer's (1965:150) and Holmes' (1988:327) threshold of 200 ms. See also Ford and Thompson (1996:146), who claim that only pauses of at least 300 ms are noticeable to the listener. Goldman-Eisler's threshold has also been claimed to be too high (O'Connell and Kowal 2004:465, based on Hieke, Kowal and O'Connell 1983) or not useful at all, because "any arbitrary definition of speech pause in terms of duration alone violates certain underlying linguistic and psychological realities" (Boomer and Dittmann 1962:219).

[15] cf. also Jefferson's (1989) "standard maximum tolerance" of one second.

[16] Note that Shriberg (1994:42-3) describes a number of errors in the transcription of filled pauses, particularly that many were missed by transcribers. She, however, used the original Switchboard release (Godfrey, Holliman and McDaniel 1992), while the present study is based on a much later, improved transcript, where it can be trusted that such mistakes have been, at least partly, corrected. Should some instances of fillers still be missing from transcripts, this should not alter any results of the present study, as this merely alters the frequency of fillers, but not their placement.

3.1.2.3 Discourse Markers

Discourse markers are a group of expressions derived from several word classes. The groups' two central characteristics are syntactic optionality and a lack of propositional meaning. In other words,

> they have a core meaning which is procedural, not conceptual, and their more specific interpretation is 'negotiated' by the context, both linguistic and conceptual. (Fraser 1999:931)

Among a wealth of other functions, it has been shown that some discourse markers can be used to mark ongoing lexical and content search or to announce repair sequences (cf. Jucker 1993:447*;* Fung and Carter 2007:418; Müller 2005:189). These functions, which correspond to those of hesitations, have been most frequently described for the discourse markers *well*, *like*, *you know* and *I mean*.

Jucker (1993) defines *well* as a "signpost", which indicates that "the addressee has to reconstruct the background against which he can process the upcoming utterance" (1993:438). One of the core meanings of *well,* according to Jucker, is signalling delay (1993:438, 447). Fung and Carter (2007:415, 423-4) add that *well* and *I mean* can be used as time-buying devices. The authors also believe that discourse markers may serve several functions at once (Fung and Carter 2007:414; see also Gilquin 2008:125). A single instance of *well* could thus serve as both a time-buying device and as a marker of a topic shift.

Müller (2005) finds that *well* is frequently used to indicate "that the speaker is searching for the right phrase" (Müller 2005:109) while *you know* is used to mark lexical or content search and *like* can be used when "searching for the appropriate expression" (2005: 208). According to her, where discourse markers are used as time-buying devices, they are characteristically accompanied by filled and unfilled pauses and other hesitations (2005:109, 158, 208). Finally, Levey (2006) confirms Müller's claim that *like* can be used as a marker of disfluency and finds that "*like* occurs with self-repairs and the elaboration of preceding remarks" (Levey 2006:426).

Discourse markers are sometimes included in hesitation studies. Clark and Wasow, for instance, include discourse markers such as *I mean* and *you know* among the group of editing expressions they analyse (Clark and Wasow 1998:201; see also Clark and Clark 1977:270; Clark 1996:262-3). They do not, however, equate them with filled pauses, which they consider lesser disruptions in speech than discourse markers (Clark and Wasow 1998:220).

Highly interesting in the context of the present study are the findings that discourse markers like *you know* can be used instead of the pause fillers *uh* and *um* in American English (Tottie and Svalduz 2009) and Schiffrin's observation that "several markers –

y'know, I mean, oh, like – can occur quite freely within a sentence at locations which are very difficult to define syntactically" (Schiffrin 1987:32).

For the given reasons, *well*, *like*, *you know* and *I mean* are included in this study of hesitation placement. *Well* and *like* are easily retrievable from the corpus because in Switchboard NXT single-word discourse markers are marked with the POS-tag *UH* (see also Section 3.1.1.1 for more details). The corpus contains a total of 5,364 cases of *well* and 1,488 cases of *like*. Y*ou know* and *I mean,* on the other hand, are not tagged *UH*. I therefore searched for all cases where *you* immediately preceded *know* and *I* immediately preceded *mean* and where the sequence was syntactically parsed as *(S (NP*, which excludes cases like *Do you know*. In this way, a total of 8,810 instances of *you know* and 2,024 instances of *I mean* where retrieved. Due to the limitations of the corpus' annotations, it was not possible to automatically distinguish any more accurately between the discourse markers and instances of these phrases with propositional meaning. The latter are highly unlikely to occur in the specific syntactic contexts selected for the studies in Chapters 4 and 5, though.

3.1.3 Hesitation Coding

A large proportion of all hesitations do not occur alone. Even excluding cases like (46), where pauses occur in the context of repair sequences, more than a third of hesitations occur in a 'cluster'[17], as is the case in (47).

(47) *uh [pause]* by the supreme court (sw3509.B.s88)

Speakers use clusters like *you know uh [pause]*, *well [pause]* or *uh uh*. Due to the large number of combinations, including each of them as a separate predictor in the analysis would mean raising the number of hesitation types from seven (*[pause], uh, um, like, well, you know, I mean*) to over eighty, all with very low token frequencies. Excluding all of these cases, on the other hand, would mean a great loss of valuable data. Therefore, grouping clusters to end up with low type and high token frequencies would be the best option. For small datasets, however, even a model with seven classes of hesitations would be too fine-grained. A scheme which also groups the clusters with the individually occurring hesitations is preferable. The following three-group schema would be ideal in terms of group size.

Group 'pause' – unfilled pauses unaccompanied by other hesitations

[17] Extrapolation based on results from the prepositional phrase data presented in Chapter 3. Out of 6,155 hesitations in this dataset, 2,293 occur in clusters of two or more (1,025 clusters in total).

Group 'u' – the fillers *uh* and *um* and all clusters which consist of one or more instances of *uh/um* and optional pauses, irrespective of their order (e.g. *uh um*, *[pause] uh uh*, *uh [pause] um*)

Group 'dm' – the discourse markers and all clusters which consist of one or more discourse markers, optional fillers and optional pauses (e.g. *[pause] I mean*, *you know like uh*, *I mean [pause] well*)

However, regrouping of this sort bears the risk of lumping together elements with fundamentally different behaviour and functions. As shown in Section 3.1.2.3, discourse markers are generally considered to also fulfil other functions besides hesitating. Therefore, *you know* and *like* might fulfil different functions and consequently be placed at different locations within an utterance. According to my hypothesis, discourse markers should not be placed within chunks, irrespective of their function, yet it should still be tested whether their behaviour is generally sufficiently similar to warrant subsuming them under one category.

The same is true for filled pauses. There is a long-running debate concerning whether filled pauses carry meaning (cf. Maclay and Osgood 1959; Howell and Young 1991; Christenfeld 1994; Clark 1996; Clark and Wasow 1998; Clark and Fox Tree 2002; Schilperoord and Verhagen 2006; Mukherjee 2007; Corley and Stewart 2008) and whether *uh* and *um* might even have different meanings. Clark and Fox Tree (2002) claim that *uh* and *um* are substantially different elements, used to announce minor and major delays respectively. This raises the question whether *uh* and *um,* as well as their various combinations with pauses and also with each other, can be subsumed under one category. If they behaved fundamentally differently concerning their placement and their tendency to combine with pauses, treating them as members of the same category of hesitations would not adequately represent their use[18].

In order to establish whether the envisaged regrouping scheme is warranted for *uh/um* as well as for the discourse markers, I will repeat some of Clark and Fox Tree's analyses. First, however, I will provide a summary of their argument and the most important counter arguments.

[18] See also Acton (2011:1) who conducts a sociolinguistic study and claims that "while *um* and *uh* share a great deal in the way of interpretation, association, and usage, they are far from perfect substitutes".

3.1.3.1 Excursion: Do *uh* and *um* Have Different Meanings?

Clark and Fox Tree argue that *uh* and *um* are two *distinct* English words, namely interjections used to express different meanings in spoken language. They argue that *uh* and *um* are planned like other words (2002:75). According to the "filler-as-word hypothesis" (2002:79)

> [*u*]*h* and *um* are interjections whose basic meanings are these:
>
> *Uh*: 'Used to announce the initiation [...] of what is expected to be a minor delay in speaking.'
>
> *Um*: 'Used to announce the initiation [...] of what is expected to be a major delay in speaking.' (Clark and Fox Tree 2002:79)

They present evidence that in the London-Lund corpus

- *um* is followed by filled and unfilled pauses more often than *uh*.
- *um* is followed by longer unfilled pauses than *uh*.
- *um* is more likely to be preceded by filled and unfilled pauses than *uh*.
- unfilled pauses before *um* tend to be slightly longer than before *uh* (2002:82-6).

O'Connell and Kowal (2005) refute this hypothesis, arguing that

> *uh* and *um* fail, respectively, to be reliably predictive of minor and major delays, [...] they are not even typically followed by silent pauses [and] do not fit the uses characteristic of interjections. (O'Connell and Kowal 2005:560)

They criticise the use of the London-Lund corpus for its impressionistic pause marking, which, they claim, only allows the study of *perceived* pause length (Clark and Fox Tree 2002:81). O'Connell and Kowal test Clark and Fox Tree's hypotheses using a corpus of TV and radio interviews with then Senator Hilary Clinton, arguing that

> [i]f indeed the conventional usage of *uh* and *um* signals an upcoming minor or major delay, then professional speakers should arguably be the most expert in using this device to their purposes. (O'Connell and Kowal 2005:560)

They find that in their data, only 20% of *uh* and 40% of *um* are followed by unfilled pauses at all. O'Connell and Kowal point out that this means that the prediction that a delay is to follow is false in 80% and 60% of cases respectively and that therefore "*uh* and *um* are poor perceptual cues to silent pauses for the listener" (2005:562)[19].

[19] Note that Clark and Fox Tree's argument actually refers to *uh* and *um* being signals of *delay*, which they define as "any combination of pauses and fillers" (Clark and Fox Tree 2002), while O'Connell and Kowal only test whether they are signals that silent pauses will follow.

Concerning the mean length of pauses following *uh* and *um*, O'Connell and Kowal find that the length ranges are so similar "as to exclude the notion of any reliable perceptual difference between the silent pauses after *uh* and the silent pauses after *um*" (2005:564).

Disregarding the primary hypothesis that *uh* and *um* are English words, i.e. interjections, neither Clark and Fox Tree (2002) nor O'Connell and Kowal (2005) really underpin nor undermine the indirect hypothesis that *uh* and *um* are different elements used in different situations or for different purposes (whatever these may be). Both studies suffer from an unfortunate choice of corpora, the former, because a corpus with non-acoustically measured pause marking was chosen and the latter because of its choice of professionally trained speakers who generally aim for an uninterrupted fluent delivery and to avoid fillers altogether.

In order to establish whether the envisaged regrouping scheme is warranted for *uh*/*um* as well as for the discourse markers, some of Clark and Fox Tree's analyses are repeated and extended to the use of discourse markers.

First, the claim that *um* is more likely to be preceded and followed by pauses than *uh* is investigated. For this purpose, it was counted how often fillers and discourse markers co-occur with pauses in the context of six selected prepositional phrase types in the Switchboard NXT corpus. Analyses are based on a total of 3,742 unfilled pauses, 1,562 filled pauses and 644 discourse markers (including only 17 tokens of *well* and 10 of *I mean;* for more details about the dataset see Section 4.2). Table 3.1 below shows the results.

	absolute	**as percentage of filler/DM**	**as percentage of pauses**
[pause] uh	260	18.5%	6.9%
[pause] um	39	25.3%	1.0%
uh [pause]	481	34.2%	12.9%
um [pause]	83	53.9%	2.2%
[pause] you know	152	35.9%	4.1%
[pause] like	30	15.5%	0.8%
you know [pause]	42	9.9%	1.1%
like [pause]	14	7.2%	0.4%

Table 3.1: Co-occurrences of fillers and discourse markers (DM) with pauses

Both *uh* and *um* tend to be followed rather than preceded by pauses. While *um* is used slightly more frequently with pauses, differences between the two fillers are not statistically significant (based on a 2x2 chi-square test). Discourse markers, by contrast,

tend to be preceded by pauses (*well* and *I mean* are ignored here because of their low frequency). In-group differences are again not significant. Differences between fillers and discourse markers, however, are highly significant (p<.001, based on a 2x2 chi-square test).

These results refute Clark and Fox Tree's (2002:82, 84) claim that *um* is followed and preceded more often by pauses than *uh* is. Instead both fillers show uniform tendencies in their usage, especially if compared to the usage of discourse markers.

In the following step, it is determined whether the mean length of pauses before and after the two fillers differ, as claimed by Clark and Fox Tree (2002). The same is repeated for the discourse markers. Figure 3.2 shows the results.

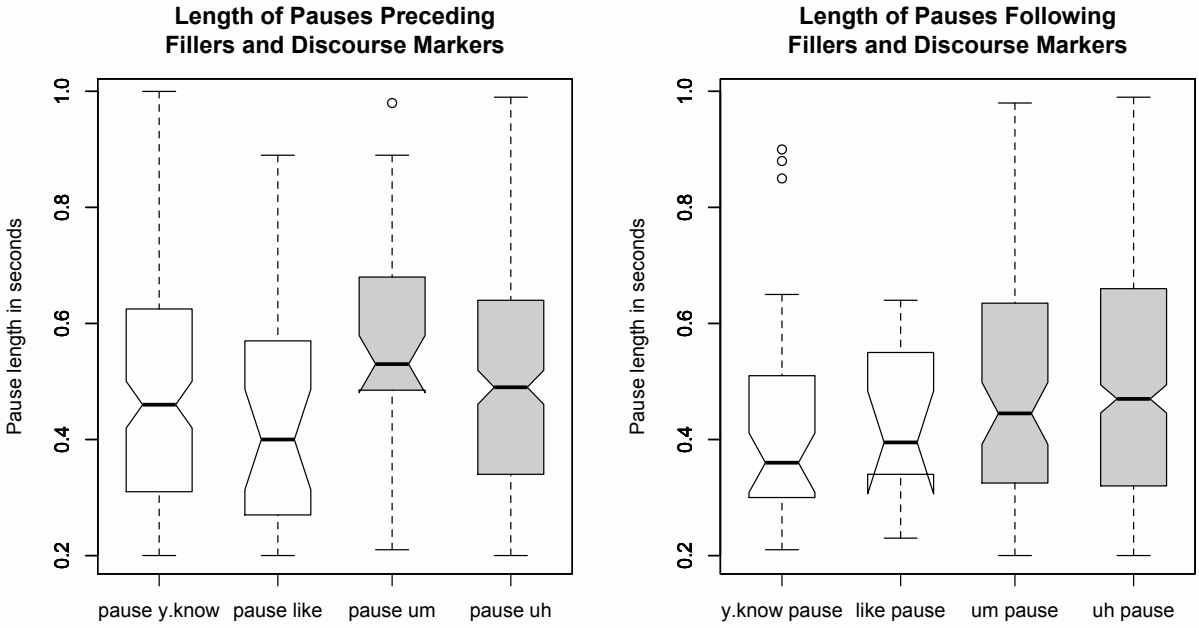


Figure 3.2: Distribution of pause lengths before and after uh, um, you know ('y.know') and like.

Mean pause duration before *uh* is 0.51 seconds, while it is 0.56 seconds before *um*. This difference does not reach significance (p<.1, based on Wilcoxon rank-sum test; Field, Miles and Field 2012:655, chosen here because the data is not normally distributed). The two grey-shaded boxes in the left plot in Figure 3.2 visualise how similar in length pauses before *um* ('pause um') and those before *uh* ('pause uh') are. The fact that the notches in the boxes overlap indicates that the median pause lengths do not differ significantly (Chambers et al. 1983:62).

Mean pause duration after *uh* and *um* is 0.5 and 0.49 seconds respectively, which is not a significant difference (based on Wilcoxon rank-sum test). The grey-shaded boxes in the right panel in Figure 3.2 indicate how similar the medians are.

Preceding the discourse markers, mean pause lengths are 0.49 (preceding *you know*) and 0.44 (preceding *like*; difference non-significant, based on Wilcoxon rank-sum test). Following the discourse markers, mean pause lengths are 0.42 (following *you know*) and 0.43 (following *like*; difference non-significant, based on Wilcoxon rank-sum test).

When we compare fillers and discourse markers as groups, we find that pauses preceding fillers are significantly longer than those preceding discourse markers ($p<.05$, based on Wilcoxon rank-sum test) and also that pauses following fillers are significantly longer than those following discourse markers ($p<.01$, based on Wilcoxon rank-sum test).

These findings refute Clark and Fox Tree's hypotheses. It appears that *uh* and *um* are generally combined with pauses in very similar ways. It is therefore unlikely that their meanings differ. Instead, fillers behave as one group whose usage contrasts with that of discourse markers.

It remains to be clarified whether speakers use fillers and discourse markers in the same or in different syntactic contexts. To establish whether this is the case, a 19x7 table is drawn, with the seven hesitation elements as rows (i.e. *uh, um, well, like, you know, I mean*) and all possible positions of occurrence in the different prepositional phrases as columns (e.g. *hesitation* Prep N; Prep *hesitation* N; see Table 4.4). Only fillers and discourse markers which do not occur in a cluster are included. The resulting table captures the distribution pattern for each hesitation type.

Based on this table, a distance matrix was generated (based on the product-moment correlation *r*; cf. Gries 2009b:313; R-package used: *amap*, Lucas 2010)[20] and finally a cluster dendrogram was drawn (based on the ward amalgamation rule; cf. Gries 2009b: 317-8)[21].

The resulting dendrogram, shown in Figure 3.3, clusters *uh* and *um* together and contrasts them with the discourse markers and unfilled pauses (*well* and *I mean* are separated because of their extremely low frequencies). This indicates that *uh* and *um* have a similar distribution pattern in prepositional phrases. If speakers prefer to place *uh*

---

[20] The more common Euclidean distance was not applicable in this case, because it is "based on the spatial distance between vectors" (Gries 2009b:315), which means that in this case it would have separated highly frequent elements (e.g. unfilled pauses) from less frequent elements (e.g. *well*), irrespective of their distribution in the prepositional phrase structures. The product-moment correlation, on the other hand, is "based on the similarity of the curvature of the vectors" (Gries 2009b:315). It groups hesitation elements together which show similar patterns of distribution in the prepositional phrase structures, irrespective of their overall high or low frequency.

[21] Ward "joins those elements whose joining increases the error sum of squares least" (Gries 2009b:317), which generates comparatively small clusters and is of broad applicability (cf. Gries 2009b:317).

before the noun in a particular phrase type, they tend to place *um* in the same position. The same can be said for the discourse markers *like* and *you know*.

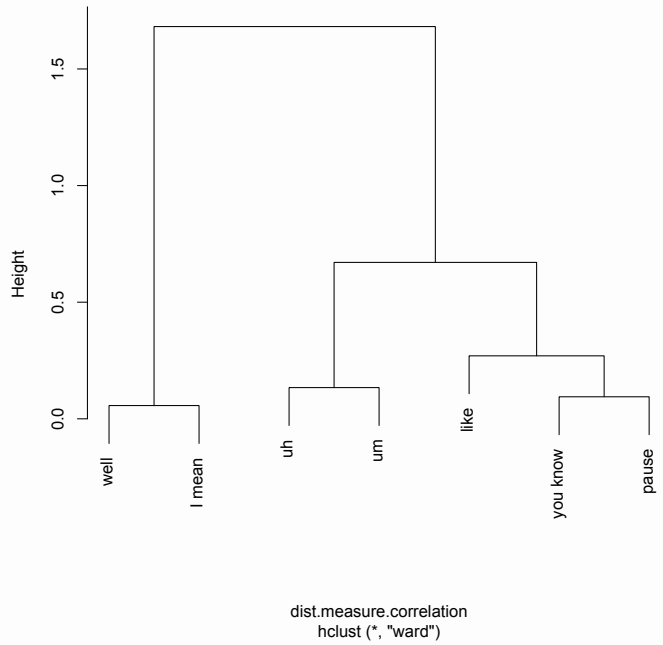


Figure 3.3: Cluster dendrogram, based on distribution patterns of hesitations in prepositional phrases

These results confirm that the fillers *uh* and *um* may be handled as a single group in an analysis and need not be included as separate items. I further take these results as evidence that the included discourse markers can be subsumed under one label.

As a final step of the analysis, four further categories are added. In this way, we can establish whether hesitations occurring in clusters can be grouped with the individually-occurring fillers and discourse markers as planned. The new categories are

cluster u – Provides information about the syntactic distribution of clusters in the prepositional-phrase dataset; contains the distribution of clusters consisting of *uh* and *um* in combination with each other and/or with pauses.

all u – Combines *uh*, *um* and *cluster u*.

cluster dm – Provides information about the syntactic distribution of clusters in the prepositional-phrase dataset, contains clusters consisting of discourse markers occurring in combination with each other and/or with filled and unfilled pauses.

all dm – Combines *well, like, I mean, you know* and *cluster dm*.

Based on this expanded data-set, a second dendrogram is constructed. All technical parameters are kept stable. I merely excluded the data for *well* and *I mean* to emphasise that the other categories are not clustered the way they are because they contrast with the low-frequency cluster. The resulting dendrogram, shown in Figure 3.4, indicates that

- *Uh* and *um* have a similar syntactic distribution in prepositional phrases.
- Instances of *uh* and *um* occurring in combination with each other or with unfilled pauses ('cluster u') are distributed in prepositional phrases like individually-occurring fillers.
- The category 'all u' adequately represents the three groups *uh*, *um* and 'cluster u'.
- The same holds true for the discourse markers *like* and *you know* and the larger category 'all dm'.

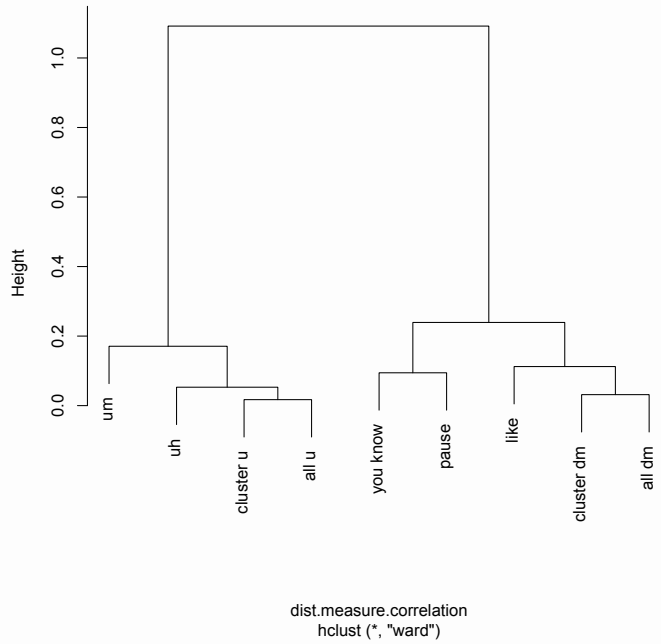


Figure 3.4: Cluster dendrogram, based on distribution patterns of hesitations and hesitation combinations in prepositional phrases

Results indicate that the data can be regrouped as planned above. Hence only three hesitation types remain, namely 'pause' (individually occurring pauses), 'u' (the fillers *uh* and *um*, potentially co-occurring with each other and/or pauses) and 'dm' (the discourse markers *like*, *well*, *you know* and *I mean*, potentially co-occurring with each other and/or

pauses). The last group does not adequately represent the behaviour of *well* and *I mean*, which is not problematic as they are extremely infrequent in the data employed here.

However, it should be noted here that the apparent similarity in the behaviour of *uh* and *um* and their combinability with pauses might be the result of imprecise transcriptions – a question which can neither be verified or falsified without the help of the original sound files. If it were the case that transcribers did not consistently distinguish between *uh* and *um*, this would lead to the same result, namely that it would be best to subsume *uh* and *um* under one category.

3.1.4 The Number of Hesitations in Spoken Language

Relying on hesitations as indicators of chunking means that only a proportion of material from a corpus can be analysed, namely those contexts which are disfluent. This section briefly lists how pervasive a phenomenon hesitations actually are in spoken language and thus illustrates that the proportion of corpus data which can be analysed in studies restricted to disfluent contexts is considerable.

Table 3.2 presents results from previous studies on the frequency of hesitations. Differences in estimates partly arise from incongruities in the set of hesitations analysed. However, a number of factors have been shown to influence the number of hesitations used. These include sociolinguistic factors like gender and age (cf. Bortfeld et al. 2001; Feldstein, Brenner and Jaffe 1963; Shriberg 1994; Shriberg 1996; Tottie and Svalduz 2009) as well as the conversational setting (cf. Feldstein, Brenner and Jaffe 1963; O'Connell, Kowal and Ageneau 2005; O'Connell and Kowal 2005; Shriberg 1996) and the topic (cf. Schachter et al. 1991). Interestingly, longer or more complex sentences do not appear to lead to higher hesitation rates (cf. Goldman-Eisler 1968; Shriberg 1994; Fehringer and Fry 2007).

---

[22] Goldman-Eisler's results are based on a very specific task which includes such precise instructions as "formulating the general point, meaning, or moral of the story [in a cartoon] in as concise a form as you can" (1961:165). The instructions furthermore include the command to "stick to the first reasonable version, and then keep repeating the same wording" "until I stop you" (which meant six repetitions each; 1961:165). In Goldman-Eisler's table of results, the maximum total length of a description (including speech, pauses and other hesitations) is only 79.9 seconds and the shortest length of speech periods in a summary is as short as 2.9 seconds (1961:167). I therefore assume that results are only based on the initial descriptions and summaries.

Note also that Goldman-Eisler (1968) summarises the results of this study as follows: "Most of the group, however, paused between 40% and 50% of their total speaking time" (1968:18). This description corresponds to the same results once the initial pauses are excluded.

Other studies have shown that such "lexical suppression" and lack of options may lead to more filled pauses (cf. Christenfeld 1994:198; for further critique of Goldman-Eisler's study see O'Connell and Kowal 2004).

Study	Rate per 100 Words	Hesitations	Data
Acton 2011: 3	1.14 0.54 *uh* 0.6 *um*	filled pauses	Corpus of American speed dating conversations
Acton 2011: 6	3.05	filled pauses	Switchboard
Biber et al. 1999:1054	1.3 0.8 *uh/er* 0.5 *um/erm*	filled pauses	Conversational subcorpora of the Longman Spoken and Written English Corpus, which comprise British and American English
Biber et al. 1999:1054	1.9	unfilled pauses	Conversational subcorpora of the Longman Spoken and Written English Corpus, which comprise British and American English
Bortfeld et al. 2001:131	5.97	repetitions, self-corrections, filled pauses, editing expressions	Corpus of speakers discussing tangrams and photos of children.
Fehringer and Fry 2007:46, 52	5.11	filled pauses	Bilingual speakers speaking their L1 (German and English speakers)
Fehringer and Fry 2007:46, 52	6.45	filled pauses	Bilingual speakers speaking their L2 (German and English speakers)
Fox Tree 1995:710	6	repetitions, self-corrections, filled pauses	Based on a re-evaluation of findings from a range of previous studies by other authors.
Goldman-Eisler 1961:167	over 50 % of speaking time	unfilled pauses	Based on experiments; value given is a ratio of the time spent pausing versus the time spent uttering *fluent* speech, the latter being defined as the time spent uttering words excluding all "irrelevant vocal productions, i.e. noise, such as repetitions of the same words or other obvious forms of marking time vocally" (1961:167). A large proportion of the pause time is actually the delay before speakers begin their first utterances (1961:168).[22]
Maclay and Osgood 1959:34	10.97 (3.87 filled pauses)	repetitions, self-corrections, filled pauses, unfilled pauses	Speech of conference participants
Tottie and Svalduz 2009	0.65 0.32 *uh* 0.35 *um*	filled pauses	Corpus of Spoken American English from the University of Santa Barbara (CSAE)
Tottie and Svalduz 2009	1.45 0.85 *er* 0.6 *erm*	filled pauses	Spoken component of the British National Corpus

Table 3.2 Frequency of hesitations in spoken English

3.2 Software

The Switchboard NXT corpus was specially developed to be used with the *NITE XML Toolkit* (cf. Calhoun et al. 2010:388-9). The NITE XML software package can be used for data retrieval and to graphically display structural relationships (cf. Calhoun et al. 2010:389-90; for more information see Carletta et al. 2009; "Readme SWBD Queries"; "Readme Tools")

The toolkit, however, proved to have some limitations in its functions. Furthermore, a number of bugs have not been resolved yet and the output format is very restricted. Therefore, it was abandoned for the present project and instead *R* (R Development Core Team 2009) was used. *R* has the advantage of being able to retrieve, analyse and graphically display results in the same environment.

In order to make all relevant information accessible with *R*, the *terminals* and *syntax* annotation layers had to be combined and reformatted. The focus was on visualising sentences as word sequences, i.e. non-hierarchically. To improve readability, only relevant information was maintained and other markup, like, for example, links to the other transcript, was deleted. To facilitate the linking process and to be able to refer to quotes from the corpus consistently, a unique ID was created for every word. It consists of the file number (e.g. sw2005), the speaker (A or B) and the sentence and word counts as given in the Treebank3 transcript (resulting in e.g. sw2005.A.s2_8).

Where not indicated otherwise, I developed the scripts myself, based on instructions and examples in Baayen (2008), Tagliamonte and Baayen (2012), Gries (2009a, 2009b) and Field, Miles and Field (2012). Appendix P lists some scripts in the order of their use in this book. Wherever additional R packages were used for graphics or statistics, these are quoted and listed in the bibliography.

3.3 Methodology

My hypotheses will be tested in a regression analysis. Technically, statistical regression assesses regularities in the data by comparing actual outcomes to a predicted pattern of outcomes. An algorithm deduces how far the behaviour of one variable – the dependent one – is governed by the behaviour of others – the predictors. In the following analyses, the dependent variable is always the location where a hesitation is placed and the predictors are mostly frequencies and frequency-based measures of association. So, for example, whether a hesitation is placed before or after a preposition is analysed by means of looking at the relations holding between the preposition and the words surrounding it. This means that whether a speaker prefers (48) or (49) should depend on the frequency of *going to* and *to school* and the associations holding between the words in these pairs.

(48) ?going *uh* to school
(49) ?going to *uh* school

According to my hypothesis, the more strongly the two words attract or the more easily one word in the pair can be guessed given the other, the stronger the bond is between them and the less likely the speaker is to interrupt the sequence with a hesitation. This design is based on a set of fundamental assumptions:

It is not possible to predict whether or not a speaker needs to hesitate – As Goldman-Eisler (1968) points out, the relationship between transitional probability and hesitation placement is not reciprocal. Sequences with lower transitional probabilities are more likely to be interrupted by hesitations than those with higher transitional probabilities, but this does not imply that every sequence with a very low transitional probability is interrupted by a hesitation – after all, most speech is uttered fluently. Therefore, a corpus-based study of chunking cannot attempt to predict whether a speaker needs to hesitate or not. It should merely be predictable *where* a speaker stops to hesitate.

Hesitations should be deleted prior to the calculation of transitional probabilities etc. – I will calculate co-occurrence frequencies and measures of association based on a version of the corpus in which all relevant hesitations have been removed. This step must be taken in order to obtain bigrams which reflect the actual transitions of interest. In (48) and (49), for example, I am interested in how often *going* and *to* occur together (and, in a later step, how much they attract each other). Consequently, bigrams calculated on the basis of a 'cleaned-up' hesitation corpus provide far more information than bigrams like *going uh* retrieved from a standard corpus.

Hesitation placement should be analysed in structurally similar settings – Notwithstanding the aim to extract general regularities and tendencies, it is important to limit the scope of the analysis to specific syntactic environments, as frequencies, transitional probabilities and the like may not be comparable across the board. First of all, word classes have been claimed to hold characteristic relations to the words in their surroundings. Nouns, for example, tend to be preceded by determiners, and it has been claimed that they therefore form tighter bonds to their left context than their right (cf. Bybee 2007b:318–323), whereas verbs are often followed by a restricted set of prepositions and consequently may form tighter bonds to their right.

Secondly, structuralist studies of hesitation placement have shown that hesitations are preferentially placed at phrase boundaries or before the first content word in a constituent (cf. Maclay and Osgood 1959; Goldman-Eisler 1968; Clark and Clark 1977; Shriberg 1994; Biber et al. 1999; Bortfeld et al. 2001). This preference may be a frequency effect, resulting from low co-occurrence frequencies of words to the left and right of the phrase boundary. Yet it may also illustrate that speech is planned constituent by constituent and that in speech planning frequency merely interacts with structural factors. Therefore, it is well imaginable that the relation between the words before and after a phrase boundary is of prime importance and that relations between words within the phrase might have a much weaker influence on hesitation placement. For these reasons, comparing associations between entirely different word-pairs is of limited usefulness. The optimal solution is to compare only stretches of speech that consist of a set sequence of parts of speech.

Measures of association should be compared and combined – Predictions about the associations between words based on absolute co-occurrence frequency and on probabilistic measures of association can diverge considerably. Even different measures of association do not always make the same predictions. Therefore, it is important to compare them. Additionally, it may be the case that a combination of factors best predicts chunking. Consequently, a statistical model which can take into account interactions is indispensable.

Based on these assumptions, I employ a multinomial and multivariate regression analysis, 'multivariate' meaning that it considers many predictors at once and 'multinomial' meaning that it can deal with more than two distinct outcomes (cf. Field, Miles and Field 2012:922). This is important, because in contexts like (50), the speaker has available to him four possible 'slots' to hesitate: before *by*, before *the*, before *supreme* and before *court*.

(50) *uh [pause]* by the supreme court (sw3509.B.s88)

In my analysis, each hesitant transition constitutes one data point and the characteristics of the word-pairs in its surrounding are its attributes.

3.3.1 Measures of Association

The following section gives an overview of the selected measures of association, their computation and the highest-ranked bigrams according to each measure and explains how they evaluate the degree of attraction in a bigram. The list of measures applied is by no means exhaustive (cf. Wiechmann 2008 for a more comprehensive list). Measures were selected for their current popularity or, in the case of Lexical Gravity G, for their innovativeness.

All calculations are based on a POS-tagged version of Switchboard NXT, in which all relevant hesitations have been removed (n=ca. 780,000). Furthermore, all predictors are calculated on a bigram level. For the present purposes, a bigram is defined as two consecutive words not crossing sentence boundaries. The definition of a word follows the formatting of the corpus (see Section 3.1.1.3). Clitics are considered separate words, cliticised negations are represented as *n't_RB*. POS-Tags were considered to be part of the word.

Finally, the word and bigram frequencies obtained are not normalised or lemmatised[23], but are given as they appear in the corpus. Measures are calculated on the basis of these absolute frequencies. All analyses, calculations and statistics were conducted in R. Where not indicated otherwise, I developed the scripts.

Bigram Frequency – Bigram frequency measures how often two words occur in a corpus in a specific order. Bigram frequency is commonly used as a simple measure of chunking strength, assuming that combined usage leads to stronger associative ties between words (cf. e.g. Bybee 2007b). For the extraction of bigrams from the corpus, the script in Gries (2009a:122-3) was adapted.

There are 180,266 bigram types in the corpus, the vast majority of which (127,499) are hapax legomena. The most frequent bigrams are listed in (51). Table A.2 in the Appendix provides a legend to the abbreviations used as POS-tags in Switchboard NXT.

23 See Kapatsinski (2010:81), who argues, based on a statistical analysis, that surface frequency is a "better predictor of interruption" than lemmatised frequency.

(51) it_PRP 's_BES (6,717)
that_DT 's_BES (4,943)
do_VBP n't_RB (4,689)
i_PRP do_VBP (3,629)
and_CC i_PRP (3,566)
i_PRP think_VBP (3,434)
i_PRP 'm_VBP (3,078)
i_PRP i_PRP (2,628)
in_IN the_DT (2,436)
a_DT lot_NN (2,193)

Direct Transitional Probability – Direct transitional probability (TPD) measures how likely the first word is to be followed by the second (cf. Kapatsinski 2005:6–7). It is unidirectional in that it only looks from the first word to the second and not vice versa. The measure is also known as "Conditional Bigram (Probability)" (cf. e.g. Gregory et al. 1999) and "Forward Bigram Probability" (cf. e.g. Tily et al. 2009). It is defined as the frequency of the bigram divided by the frequency of the first word in the bigram (cf. Kapatsinski 2005:6-7).

$$TPD = \frac{F(Bigr.)}{F(w_1)}$$

Direct transitional probabilities based on the Switchboard NXT corpus range from one to almost zero. Median TPD is 0.007 (mean: 0.11; but the data is not normally distributed). Interestingly, there is a large group (n=9,292) of bigrams with a TPD of one. These are mostly cases where both the first word and the bigram are hapax legomena, though some are more frequent, as shown in (52) (numbers in brackets are bigram frequency and TPD).

(52) ca_MD n't_RB (741, 1)
wo_MD n't_RB (132, 1)
willing_JJ to_TO (51, 1)
civil_NNP war_NNP (16, 1)
et_FW cetera_FW (15, 1)
wind_VBP up_RP (14, 1)
according_VBG to_IN (12, 1)
lack_NN of_IN (12, 1)

In the case of *Civil War*, we see that the POS-tagging marks all words in multi-word proper names as proper names. Occasionally, high transitional probabilities and high MI scores result from this practice.

Backwards Transitional Probability – Backwards Transitional Probability (TPB) is another unidirectional measure of association. It measures how likely the second word is to be preceded by the first (cf. Kapatsinski 2005:6–7). Backwards Transitional Probability is defined as the frequency of the bigram divided by the frequency of the second word in the bigram (cf. Kapatsinski 2005:6-7). It is sometimes known as "Reverse Conditional

Bigram (Probability)" (cf. e.g. Gregory et al. 1999) or "Backward Bigram" (e.g. Tily et al. 2009).

$$TPB = \frac{F(Bigr.)}{F(w_2)}$$

Direct transitional probabilities also range from one to almost zero with 10,403 bigrams receiving a backwards transitional probability of one. The median is 0.007 (mean: 0.12). (53) lists some examples of bigrams which have a backwards transitional probability of one (numbers in brackets are bigram frequency and TPB).

(53) i_PRP 'm_VBP (3078, 1)
new_NNP york_NNP (72, 1)
i_PRP suppose_VBP (52, 1)
san_NNP antonio_NNP (51, 1)
united_NNP states_NNP (16, 1)
oh_UH dear_UH (15, 1)
air_NN conditioning_NN (13, 1)
the_DT longest_JJS (8, 1)

Mutual Information Score – The mutual information score (MI) assesses how strongly the two words in a bigram attract by calculating how much more often they occur together than would be expected by chance (cf Manning and Schütze 1999; Oakes 1998). It is a bidirectional measure of association because, unlike the transitional probabilities, it takes associations from left to right as well as from right to left into account. My calculation follows the formula used by Wiechmann (2008:264–265). First, the product of the frequencies of the first and second words is divided by the number of words in the corpus. The bigram frequency is then divided by the result of the first quotient. Finally, a logarithm is computed (here to the base of two).

$$MI = \log \left(\frac{F(Bigr.)}{\frac{F(w_1) \times F(w_2)}{\sum words}} \right)$$

MI 'rewards' unexpectedly frequent bigrams and 'punishes' those which consist of highly frequent words yet occur more rarely than expected. Consequently, if two bigrams occur with the same frequency, it favours the one which is less likely to occur by chance. The highest MI score is awarded where the bigram itself as well as both component words are hapax legomena.

In the present corpus, MI takes on values between -8.43 and 19.57. These values group around a mean of 4.68 (standard deviation: 4.08). The positive value of the mean indicates that the patterning is more consistent than a random pattern (cf. Gries,

personal communication, 2010). (54) shows ten examples of bigrams which receive the highest MI of 19.573.

(54) aesthetically_RB pleasing_JJ
beef_NNP bourguignonne_NNP
berlin_NNP wall_NNP
bit_RB faddish_JJ
collapsible_JJ sailboat_NN
glove_NN compartment_NN
humongous_JJ cactus_NN
juvenile_JJ delinquents_NNS
loa-_VBD loaned_VBD
self-discipline_NN which_IN

Lexical Gravity G – Lexical gravity G (G) is a relatively new measure of attraction which assesses how likely among all *possible* combinations of words the combination in a given bigram is (cf. Daudaravičius and Marcinkevičienė 2004). Like the mutual information score it is bidirectional, yet it differs from MI and transitional probabilities in one crucial respect. The latter operate based on an "assumption of complete independence" (Gries and Mukherjee 2010:3), meaning that they ignore the fact that semantic and syntactic factors restrict the possible combinations of words. Lexical gravity G reflects these restrictions. Therefore, there is no linear correlation between G and MI. G is comparatively complex to calculate (the logarithm is calculated to a base of two):

$$G = \log\left(\frac{F(Bigr.) \times type\ freq\ after\ w_1}{F(w_1)}\right) + \log\left(\frac{F(Bigr.) \times type\ freq\ before\ w_2}{F(w_2)}\right)$$

For the present corpus, G scores range from -13.03 to 16.37 (mean: -1.96; standard deviation: 2.19). In contrast to MI, the fact that the mean is negative does not provide information about the degree of randomness in the distribution (cf. Gries, personal communication, 2010). (55) lists the ten bigrams which receive the highest G scores (values in brackets are G).

(55) and_CC i_PRP (16.37)
of_IN the_DT (16.06)
in_IN the_DT (15.93)
and_CC and_CC (15.23)
it_PRP 's_BES (15.08)
and_CC then_RB (14.93)
that_DT 's_BES (14.89)
and_CC it_PRP (14.66)
i_PRP do_VBP (14.39)
it_PRP was_VBD (14.37)

Gries (2010) shows that in a cluster analysis of corpus registers G clearly outperforms MI, leading him to suggest that

> the corpus-linguistic approach to collocational statistics should maybe be reconsidered, to move away from the nearly 30 [...] measures that only

> include token frequencies to one that also includes type frequencies. (Gries 2010:11)

He concedes, however, that G may not be the "ultimate solution" and that the formula may be improved by including the distribution of type frequencies (Gries 2010:11-2). Moreover, G correlates strongly with log bigram frequency.

3.3.2 Other Predictors

The following further predictors will be included in the analyses:

Word Frequency – Word frequency is clearly not a measure of association. However, word frequencies can show whether an apparent effect of chunking is in fact caused by the frequency of only one word in the bigram (cf. Biber et al. 1999; Kapatsinski 2010; Stolcke and Shriberg 1996). For example, a hesitation placed before word Y could be placed there because Y is more strongly associated with the following Z than with the preceding X or simply because Y is infrequent.

Word frequencies were extracted from the corpus with the help of the script provided in Gries (*2009a*:106-7). There are 21,975 word types in the corpus, 9,618 of which are hapax legomena.

Hesitation type – This predictor is included because filled pauses, unfilled pauses and discourse markers may be placed differently. Thus, variation in placement could actually be merely an effect of hesitation type.

The coding distinguishes between the categories 'pause' (unfilled pauses), 'u' (filled pauses and clusters) and 'dm' (discourse markers and clusters; see Section 3.1.3).

3.3.3 Nonparametric Regression: Recursive Partitioning

All analyses in the present study are based on classification and regression trees (CART), also called 'conditional inference trees' (cf. Hothorn, Hornik and Zeileis 2006) and on Random Forests. These are data-driven, nonparametric regression approaches (cf. Strobl, Malley and Tutz 2009b:325), which are more suitable for handling the present data than other multivariate regression analyses because they have a number of qualities which other approaches lack (cf. Tagliamonte and Baayen 2012:159, 170):

They can handle multinomial outcomes. – The present study asks where in a structure the hesitation is expected. As most of the selected contexts are more than two words long, speakers mostly have more than two placement options. In (53) and (54), for instance,

the speaker can pause either before *going*, before *to* and before *school*. Binomial regressions, can only handle hypotheses of the kind 'Does the speaker hesitate before or within the phrase?', which would substantially limit the scope of the analysis.

They can deal with collinear predictors (cf. Tagliamonte and Baayen 2012:161). – This matters because all measures of association were eventually derived from frequencies, which leads to some covariation.

They can be applied to unbalanced designs (cf. Tagliamonte and Baayen 2012:171). – Hesitations are not evenly distributed throughout a structure, but instead often preferentially occur in one particular spot.

They can cope with complex interactions[24] (cf. Tagliamonte and Baayen 2012:171).

CART trees and random forests will be run in R (R Development Core Team 2009). I will proceed to explain their application and point out further advantages and the approaches' limitations. For a condensed description of the methodology see also Schneider (2014) and for a detailed analysis of the exemplary dataset used here see Section 4.4.6.

3.3.3.1 CART Trees

The basic mechanism underlying CART trees is recursive binary partitioning of the data, following an algorithm which increasingly purifies the resulting subgroups so that 'leaves' become more "homogenous with respect to the levels of the response variable" (Tagliamonte and Baayen 2012:159; cf. also Therneau and Atkinson 1997:6). For every split, the algorithm determines the predictor and splitting point best suited to reduce impurity (cf. Baayen 2008:149; Strobl, Malley and Tutz 2009b:327). This process is repeated until one of three possible stop criteria is reached.

> (a) a given threshold for the minimum number of observations left in a node is reached or
>
> (b) a given threshold for the minimum change in the impurity measure is not met any more by any variable.
>
> Recent classification tree algorithms also provide statistical stopping criteria that incorporate the distribution of the splitting criterion [...]. (Strobl, Malley and Tutz 2009b:327)

[24] For problems handling perfect interactions see Strobl, Malley and Tutz (2009a:28), where the authors concede that in "a perfectly symmetric, artificial XOR problem, a tree would indeed not find a cutpoint to start with".

The binary splits result in a tree with branches ending in terminal 'leaf nodes'. These are "non-overlapping subsets that jointly comprise the full data set" (Baayen 2008:149). The number of terminal nodes reflects the complexity of a model (cf. Hothorn, Hornik and Zeileis 2006:665). Interactions are represented in the tree where a predictor appears in only one of the two branches resulting from a previous split (cf. Baayen 2008:154; Strobl, Malley and Tutz 2009a:11). Even non-linear and non-monotone associations (cf. Strobl, Malley and Tutz 2009b:325) can be captured and visualised in the leaves of the tree.

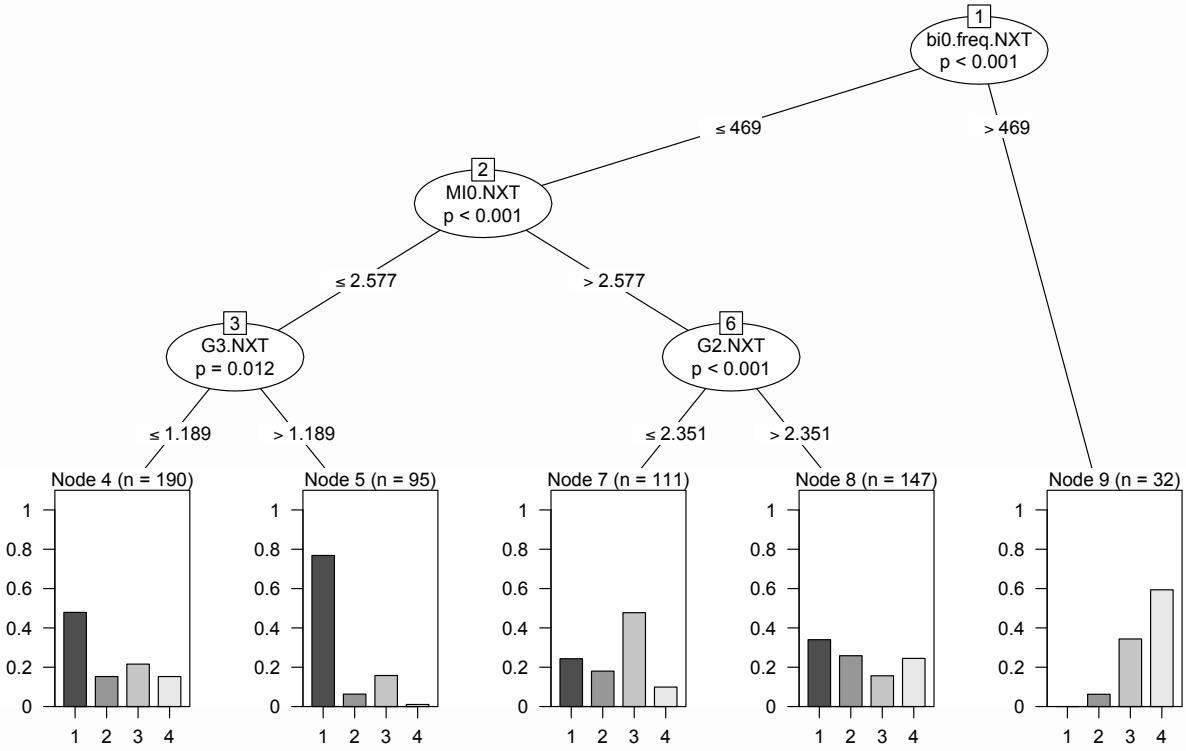
The CART mechanism used in this chapter is the *ctree* package for R (Hothorn, Hornik and Zeileis 2006). For some time, another package, namely *rpart* (Therneau et al. 2011; for detailed descriptions see also Therneau and Atkinson 1997 and Atkinson and Therneau 2000), has been described as "the de-facto standard in open-source recursive partitioning software" (Hothorn, Hornik and Zeileis 2006:663). Yet *ctree* excels *rpart* in several respects.

- Most notably, *ctree* implements the 'statistical stopping criteria' cited above, which ensure that the resulting trees do not adapt too closely to the data, which would result in overfitting (cf. Strobl, Malley and Tutz 2009a:10). Results from an overfitted model reflect the variation in the particular sample, but no longer adequately represent the behaviour of the larger population the sample was taken from (cf. Strobl, Malley and Tutz 2009b:327-8).
- *ctree* ensures that variable selection is not biased towards predictors with many possible splits, which was the case with previous algorithms (cf. Strobl, Malley and Tutz 2009b:342; Hothorn, Hornik and Zeileis 2006:664). It has been shown that this results in more accurate data partitioning (cf. Hothorn, Hornik and Zeileis 2006:667).
- The graphical output of *ctree* offers a lot more visual information than that of *rpart*.

Figure 3.5 shows an exemplary CART tree. The algorithm was provided with all predictors listed in Sections 3.3.1 and 3.3.2 in order to predict hesitation placement in a four-word phrase. At each split, the algorithm aims for splitting the set into groups with more homogenous hesitation behaviour. All nodes (i.e. splits and leaves) are numbered from one to nine. At each split, the predictor and splitting point are listed. At the first split, for instance, the algorithm selected the predictor 'bi0.freq.NXT' (i.e. the combined frequency of the words to the left and the right of the phrase boundary) as the splitting criterion and a value of 469 as the splitting point. Not all predictors are necessarily

represented in a tree. The algorithm selects only those which, at any given point, are the best possible splitting criteria (cf. Strobl, Malley and Tutz 2009b:327).

The bar graphs at each terminal node show the distribution of outcomes in the leaf, in this case, how many hesitations were placed in the different positions in the phrase. The highest bar in a leaf indicates the model's prediction for this leaf. So, under the conditions created by Splits 1, 2 and 3, the model predicts that all hesitations will occur at Position 1 (i.e. before the first word in the phrase). Thus, in Node 4, roughly 50% of data-points are assessed correctly, while the remaining 50% are misclassified (because they actually occur in positions other than 1). Misclassification rates across the leaves of a single tree can differ considerably.



List of Abbreviations

		Word Frequencies		**Bigram Measures**	
w.freq	Word Frequency	w0	Word Preceding the Preposition	bi0	X + Preposition
bi.freq	Bigram Frequency	w1	Preposition	bi1	Preposition + Determiner
TPD	Direct Transitional Probability	w2	Determiner	bi2	Determiner + Adjective
TPB	Backwards Transitional Probability	w3	Adjective	bi3	Adjective + Noun
MI	Mutual Information Score	w4	Noun		
G	Lexical Gravity G				

Figure 3.5: Exemplary CART tree

It is worth noting that the fact that a tree is grown at all shows that there are effects in the data. If there are none, *ctree* does not grow a tree because it refuses to create splits which do not lead to a reduction of the noise in the data (cf. Baayen 2008:150).

To assess the quality of the model, the total number of correct and false predictions is compared to a baseline model "that simply predicts the most likely realization for all data points" (Baayen 2008:153) in a chi-square test of significance. In this case, the preferential placement of hesitations is before the first word (i.e. in Position 1). Hence the baseline model would predict that all hesitations occur at Position 1.

Furthermore, to be more cautious, individual cells are compared. This means that we ask whether the number of correct classifications in the *ctree* model significantly exceeds those of the baseline model and, reversely, whether the number of misclassifications of the *ctree* model is significantly lower than that of the baseline model. For this purpose, the residuals of both models are compared. If the value of the residuals exceeds 2, the two models' performance can be considered statistically significantly different[25].

Additionally, the grouping generated by the tree, i.e. the content of the individual leaves, can be analysed qualitatively to reveal linguistic commonalities among structures with similar frequencies and hesitation behaviour.

Nevertheless, Hothorn, Hornik and Zeileis (2006) warn that CART trees also have certain disadvantages.

> Since a key reason for the popularity of tree based methods stems from their ability to represent the estimated regression relationship in an intuitive way, interpretations drawn from regression trees must be taken with a grain of salt. (Hothorn, Hornik and Zeileis 2006:671)

The routine relies on a single tree only and is therefore unstable to small changes in the data, so that a small number of changed values or differences in the number of data points can lead to a different partitioning of the data (cf. Strobl, Malley and Tutz 2009b: 330). Finally, all splits in the tree are only locally optimal, as a

> variable and cutpoint are chosen with respect to the impurity reduction they can achieve in a given node defined by all previous splits, but regardless of all splits yet to come (Strobl et al. 2009b:333).

[25] I thank Sascha Wolfer for suggesting this method of model comparison which provides the best way known to me to compare the two models.

3.3.3.2 Random Forests

Problems caused by relying on a single tree can be solved by using an entire forest of differently generated trees. This possibility is offered by random forests, implemented in R with the *cforest* command in the *party* package (Hothorn et al. 2006; Strobl et al. 2007; 2008)[26]. Random forests generate an ensemble of trees (cf. Strobl, Malley and Tutz 2009b:331), each based on a random subsample (without replacement) of data-points and predictors (cf. Strobl, Malley and Tutz 2009b:332-3). Tree growth is unstopped and the finished trees are not pruned (cf. Strobl, Malley and Tutz 2009b:331). Each tree then gets to 'vote' on the most likely response for a given data-point (cf. Strobl, Malley and Tutz 2009b:334; Tagliamonte and Baayen 2012:161). *cforest* in particular uses a voting scheme which averages observation weights and is therefore more precise than averaging predictions directly (cf. R Documentation [help function] for cforest {party}). In this way, "ensemble methods utilize the fact that classification trees are unstable but, on average, produce the right prediction" (Strobl, Malley and Tutz 2009b:332).

The use of a random selection of predictors for every split creates conditions which are as diverse as possible. In this way, splits emerge which may not have been the locally optimal splits had all predictors been considered, but which eventually lead to a better overall result (cf. Strobl, Malley and Tutz 2009b:333). Furthermore, predictors which might never appear in a single tree, because they are always marginally outperformed by another correlated predictor, have a chance to perform, allowing for an objective comparison of their predictive power.

It is important to keep in mind that, unlike CART trees, random forests are a 'black-box' tool, meaning that the resulting model cannot be graphically displayed or analysed. Neither is it possible to determine the best-performing tree in the forest (cf. Strobl, Malley and Tutz 2009b:333). As a consequence, complex interactions, while grasped by the model, cannot be visualised.

To grow an optimal forest, first the ideal number of predictors to be considered at every split ('mtry') and the best number of trees ('ntree') need to be determined. The more trees grown, the more reliable the result. And the more predictors in the dataset, the more trees need to be grown and the more predictors need to be considered for every split (cf. Strobl, Malley and Tutz 2009b:343). Strobl, Malley and Tutz (2009a, Supplement:3) state that "[t]he square root of the number of variables is often suggested as a default value for mtry".

---

[26] Procedure and scripts for random forests used throughout this study are based on Strobl Malley and Tutz (2009a, Supplement), Tagliamonte and Baayen (2012) and Shih (2011). Depending on their size, random forests, and especially the corresponding variable importance measures can be extremely computationally intensive. I gratefully acknowledge the bwGRiD making their computational resources available to me and the administrators providing useful help and guidance.

In the prepositional phrases analysed in Chapter 4, the number of predictors considered ranges from 14 to 26 and in the analyses of sentence-initial sequences in Chapter 5, it ranges from 8 to 26. The number of predictors most frequently considered is 20. To create uniform conditions, mtry was set to five in all cases, which is also the default setting in *cforest* (cf. R Documentation [help function] for cforest {party}). This choice was confirmed in a test with 200 forests for the structure 'Preposition Determiner Adjective Noun', where forests with mtry=3 performed worse than those with mtry=5[27].

Concerning forest size, authors agree that the larger the better (cf. Goldstein, Polley and Briggs 2011:20; Genuer, Poggi and Tuleau 2008:16-7; Shih 2011:3). In order to determine which forest size suffices for the present purposes, 1,000 trial forests were grown for each of the 14 datasets. For each set, 100 forests of each 100, 500, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000 and 8,000 trees were grown (mtry=5). The left pane in Figure 3.6 shows the results for an exemplary dataset.

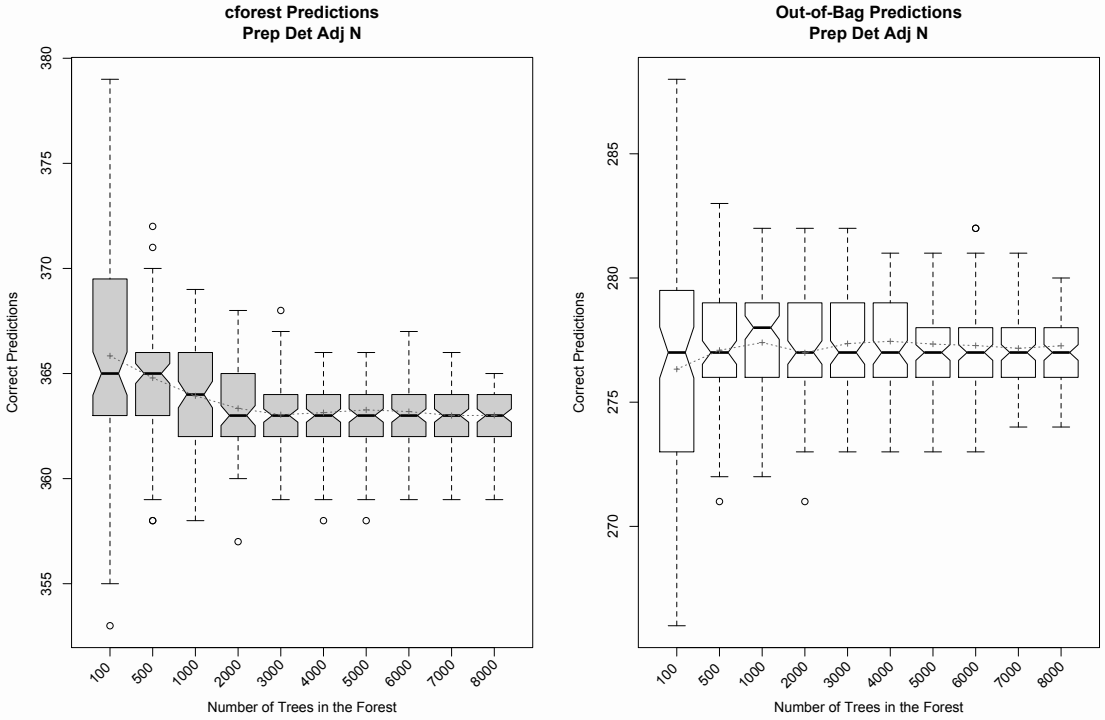


Figure 3.6: Number of correct predictions for an exemplary dataset based on differently-sized forests

While the forests with the highest predictive accuracy can always be grown with the smallest forests (ntree=100), these small forests still show a lot of variation in predictive accuracy. Variation decreases with forest size. In all of the prepositional-phrase datasets, means are stable from about 3,000 trees onwards and variation does not decrease

[27] The procedure was restricted to these two options because, in practice, mtry other than 3 or 5 are rarely seen.

significantly if more than 3,000 trees are used (see Appendix D). Therefore, in all prepositional phrase datasets, ntree was set to 3,000. The situation was slightly different in the case of the datasets of sentence-initial sequences. Here, results were stable from 2,000 trees on and some analyses even crashed when much larger forests were used (indicated by missing values in the graphs in Appendices F to M). Therefore, slightly smaller forests, of only 2,000 trees will be employed in Chapter 5.

Finally, random forests are truly random; hence two runs will produce different results. They can only be replicated exactly if a 'random seed' is set, which controls the generation of all 'random' elements (cf. Strobl, Malley and Tutz 2009b:343). Wherever *cforest* results will be reported, I will also give the seed.

Model performance is evaluated in the same way as the performance of CART trees. However, cforests "come with their own built-in test sample: the out-of-bag observations" ('OOB'; Strobl, Malley and Tutz 2009b:341), which are those observations from the original dataset which were not included in the learning samples for the trees (cf. Strobl, Malley and Tutz 2009b:335). The authors claim that the forest's estimate of error rate is "naive and overoptimistic" (Strobl, Malley and Tutz 2009b:335) and that the out-of-bag predictions offer a more conservative and therefore realistic estimate (cf. Strobl, Malley and Tutz 2009b:335). Consequently, I also compare out-of-bag results to the baseline model.

Results reported for *cforest* as well as for the out-of-bag observations are always derived from the same seed, meaning that they stem from the same model. The model chosen for reporting is always one where the number of correct predictions and the number of correct out-of-bag predictions are as close as possible to both the mean and median performance of models of the respective size.[28]

In order to assess whether model predictions are solid, thus allowing for generalisations, cross-validation is often employed (cf. Field, Miles and Field 2012:916). Cross-validation involves fitting a model to one subset of the data and then using it on another, non-overlapping subset and comparing accuracy (cf. Bortz 2005:454). This is effectively how out-of-bag predictions are generated in the present data-set. Combined with the fact that the performance of an average is reported, this effectively renders cross-validation redundant.

The predictive power of individual predictors is ranked by variable importance scores, generated with the *varimp* command (Figure 3.7 shows an exemplary ranking of scores). The *varimp* function artificially turns predictors into non-significant predictors and measures by how much prediction accuracy decreases (cf. Tagliamonte and Baayen

[28] Both, means and medians, can be seen in Figure 3.6: the median is indicated by the thick black horizontal lines in the middle of the boxes and means are shown as crosses, connected by a dashed line.

2012:160). If a predictor only causes noise in the data, prediction accuracy actually increases after permutation of the predictor. Generally, the importance of inconsequential predictors is close to zero. As a consequence

> [a]ll variables with importance that is negative, zero, or positive but with a value that lies in the same range as the negative values can be excluded from further exploration. (Strobl, Malley and Tutz 2009b:343)

The dashed vertical lines in the variable importance graphs indicate this range (see Figure 3.7). Note that "*varimp* cannot (yet) handle missing values" (R Documentation [help function] for cforest {party}).

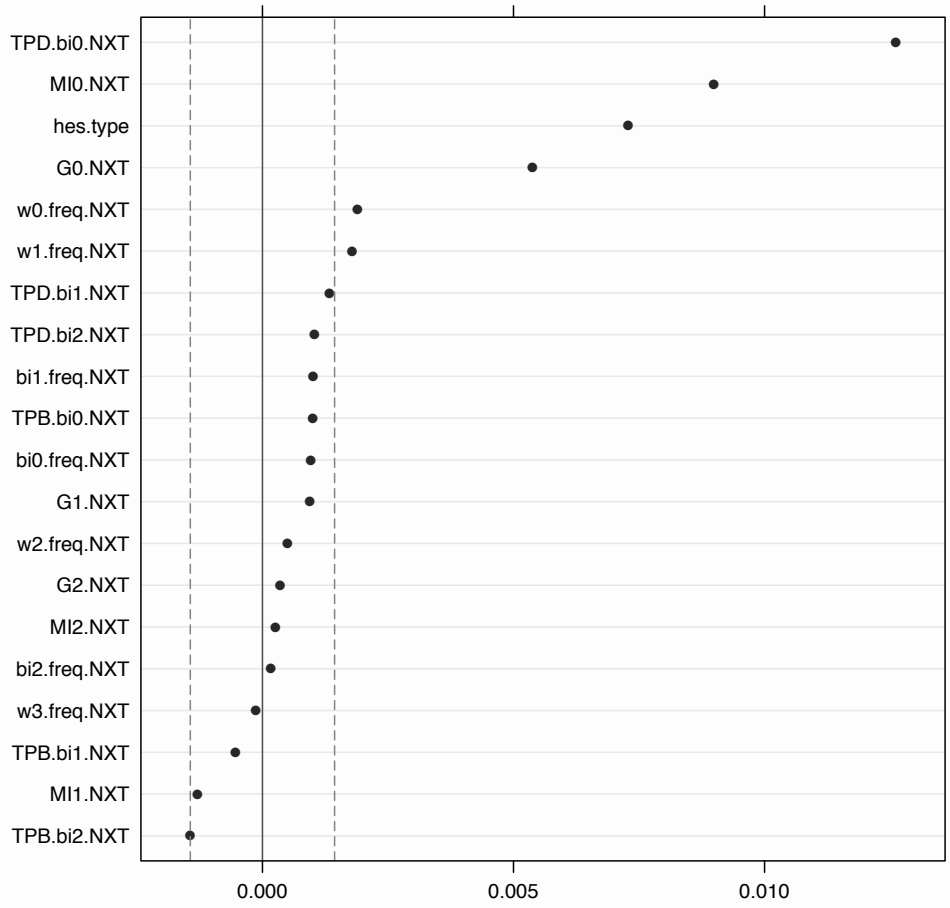


Figure 3.7: Variable importance of predictors for an exemplary dataset

Varimp scores neither provide information about the directionality of the effect nor about interactions. The score each predictor is assigned cannot be interpreted or compared across studies. It can only serve as a relative ranking of predictors within the same model (cf. Strobl, Malley and Tutz 2009b:336)[29].

In the present case, however, all models are provided with the same types of predictors. Each individual model not only considers the frequency of each bigram in the selected structure, but also all transitional probabilities etc. If scores in one of the models are particularly low, then this will affect all types of predictors. In this particular case, it is therefore possible to combine variable importance scores from several models in order to evaluate which type of predictor (e.g. bigram frequency, the mutual information score etc.) performs best overall. The more scores from different models deviate, the larger the standard deviation will be. This ensures that large inter-model differences will not lead to false assumptions about different effect sizes, but instead will lead to the cautious interpretation that no significant differences can be discerned.

While some authors suggest combining random forests with logistic regression methods (cf. e.g. Tagliamonte and Baayen 2012), this is not an option for multinomial data, as the shortcomings of binary logistic regression pertain. Therefore, this study opts for combining single trees with random forests.

[29] Theoretically, *varimp* offers the possibility to adjust "for correlations between predictor variables", so that "[t]he resulting variable importance score is conditional in the sense of beta coefficients in regression models" (R Documentation [help function] for cforest {party}). This option is often reported in the relevant literature (cf. Tagliamonte and Baayen 2012:178; Shih 2011:3). Interestingly, Strobl, Malley and Tutz (2009a:31, Supplement:5), who argue that permutation importance artificially prefers correlated predictors if correlations are not included in *varimp*, do not use this adjustment. Despite some predictors being clearly correlated, this option was not chosen in this study, mainly because it was not computationally feasible, even on the high-powered bwGRiD machines. However, application of this setting can also be rejected on theoretical grounds as I am not only interested in investigating the strength of the frequency effect, but also in the 'cognitive reality' of different measures of association. Working with residuals instead of the measures themselves skews their importance ranking (Kapatsinski, personal communication, 2011).

4 Hesitation Placement in Prepositional Phrases

This chapter deals with hesitation placement in prepositional phrases; specifically with the effect that two-word string frequency has on speakers' choice of where to interrupt the speech flow to place a hesitation. Variation in placement in phrases of different length and complexity, such as in (56) a. and b. and (57) a. b. and c., will be investigated[30].

(56) a. *uh [pause]* on Monday (sw2611.A.s6)
b. on *uh* Stalin (sw4858.A.s33)
(57) a. *you know* on the spot (sw4765.B.s22)
b. on *[pause]* the person (sw4133.B.s48)
c. on the *[pause]* lookout (sw4184.B.s153)

While the structure in examples (56) a. and b. as well as in (57) a., b. and c. above is identical, speakers still chose a different hesitation pattern. In the examples labelled a., the hesitation is placed before the phrase, while in b. and c. it occurs at different positions within the phrase. Structural explanations cannot account for this kind of variation, which is by no means limited to prepositional phrases.

Analyses will be based on two-word sequences, so-called bigrams. The focus is on relations holding across the phrase boundary and their effect on hesitation placement as well as on investigating whether the mind only keeps track of absolute co-occurrence frequency or whether it also monitors relative chance of co-occurrence, which is here approximated by means of four frequency-based probabilistic measures of association.

Results indicate that two-word string frequency and probabilistic relations between the words in the pair have a profound influence on speech production. When speakers run into planning trouble and need to stop the speech flow to hesitate, they are significantly more likely to interrupt the pair of words in the phrase which is the least frequent or least attracted than any other pair. Reversely, strongly attracted pairs are significantly less likely to be interrupted than other two-word strings.

These effects are strongest at the prepositional phrase boundary, warranting the conclusions that chunks in violation of constituent boundaries are common and that speakers hesitate most commonly at the prepositional phrase boundary and deviate from this pattern if the words to the left and the right of the boundary are strongly attracted.

---

[30] Examples given throughout this work are mostly not complete sentences, but only relevant excerpts. Where, instead of the hesitation placement, the structure is at the focus, hesitations are often removed for better legibility.

4.1 Background & Previous Research

Quantitative analyses require comparatively large datasets, which makes highly frequent constructions their best subjects. Prepositional phrases are such common structures. Even rather complex types of prepositional phrases, such as, for example, 'Preposition Determiner Adjective Noun', are still frequent enough to yield a sufficient number of data-points for statistical analyses. Additionally, a number of other factors make prepositional phrases interesting objects for a study on chunking.

No sentence boundary – A good proportion of prepositional phrases does not occur sentence-initially, so an analysis of hesitation placement can be restricted to sentence-medial hesitations. This bears the advantage that the left context can always be included in the analysis meaning that wherever the hesitation occurs in the phrase, we can calculate the attractions between the words occurring before and after it. In sentence-initial contexts, this is not possible because the first word in the sentence has no left context and consequently neither do sentence-initial hesitations.

Relations across the phrase boundary – According to Bybee (2007b:326), words which frequently occur in sequence develop into a syntactic constituent. Consequently, attractions between words within a phrase should be stronger than attractions between words at different sides of a phrase boundary and hesitations should therefore predominantly be placed at phrase boundaries. Studies which find that speakers have a tendency to place hesitations at constituent boundaries (cf. Goldman-Eisler 1968:95; Swerts 1998:489-90; Biber et al. 1999:1054; Bortfeld et al. 2001:138; Kapatsinski 2005:482-3) could be considered evidence that this is in fact the case. Bybee (2007b: 327-330), on the other hand, also finds some chunks which violate the boundaries of traditional constituent structure, which, in turn, should lead to hesitations being placed within the phrase. Such frequency effects might explain why not *all* hesitations in the studies by Goldman-Eisler and others were placed at phrase boundaries.

Other approaches, however, suggest that attractions between words have no more than a minor influence on hesitation placement near major constituent boundaries. Kapatsinski (2005:482-3, 491), for example, finds that in 92% of repairs started up to three words after a clause boundary, speakers recycle back to the boundary, irrespective of all but the strongest attractions between the words in the surrounding context (see also Fox and Jasperson 1995). On the one hand, such findings may indicate that the preference for recycling back to clause boundaries is not an effect of the co-occurrence frequency of the words to both sides of the clause boundary, but an independent effect, possibly resulting from clauses being structural basic units of planning (cf. Power 1986) or speakers' desire to produce constituents in an uninterrupted flow (cf. Clark and

Wasow 1998). On the other hand, clauses may actually be usage-based units (cf. Bybee 2007b) with clause boundaries marking the boundaries of larger chunks or constructions, which Kapatsinski's bigram-based approach did not capture.

From Kapatsinski's (2005) results it appears that phrase boundaries which do not coincide with constituent boundaries are not such strong attractors of retraction. It will therefore be interesting to see whether the effects described by Kapatsinski can also be found for minor constituent boundaries and for the set of hesitations selected here, which does not include repetitions.

Chunking across the prepositional phrase boundary – Prepositional phrases can fulfil a range of syntactic roles, functioning as postmodifiers of noun phrases, as adverbials or as complements of verbs and adjectives. The strength of the relation between the preposition and its preceding context may vary considerably depending on the function of the phrase. It has been claimed that when functioning as complement of a verb or adjective, the preposition is more strongly related to its left than to its right context, because the former determines the choice of preposition (cf. Quirk et al. 1985:657). So, in the sequence *sorry for that*, the connection in the pair *sorry for* is supposed to be stronger than that in the pair *for that*. According to Altenberg (1998:110), such verb-complement combinations are more likely to form an MWU than combinations of other phrase types, which are not as closely related. Thus varying syntactic functions are said to correlate with stronger or weaker relations holding across the prepositional phrase boundary and thus with a different degree of 'chunkiness'. This should be measurable in the attraction between words right and left of the boundary and should also be reflected in speakers' hesitation patterns.

Pullum and Huddleston (2002) emphasise that some of the functions that prepositions fulfil are grammaticalised. They list such grammaticalised uses of prepositions as forming the passive (see (58)) or genitive constructions (see (59); examples taken from Pullum and Huddleston 2002:601). Prepositions can also enter into a wide range of further (semi-)idiomatic constructions such as those listed in (60) (cf. Pullum and Huddleston 2002:617-626). The authors refer to any preposition which has become an obligatory part of at least one grammatical construction as a "grammaticised preposition" (Pullum and Huddleston 2002:647). In this sense, they name *of* "the most highly grammaticised of all prepositions" (Pullum and Huddleston 2002:658).

(58) He was interviewed by the police.

(59) They were mourning the death of their king.

(60) for example, on the spot, out of, by means of, at last, for free

Some of these constructions, like passive 'BE verb-ed by X', are very abstract. Specific instantiations are mostly rare and few verbs predominantly occur in the passive voice and form a strong attraction to the construction. Other constructions, by contrast, are fixed, e.g. *in charge of* (cf. Pullum and Huddleston 2002:618). In the latter case, *charge* and *of* are likely to form a cohesive unit. I expect fixed constructions to be chunked and consequently not to be interrupted by hesitations. It will be very interesting to see whether hesitation patterns will allow for conclusions about chunking on a more abstract level, such as, for example, 'Quantifier + *of*'. Because *of* constructions are so common, they are the most likely candidates for analysis of their chunk status on concrete as well as abstract levels.

Other researchers have also commented particularly on the grammaticalised uses of *of* and the many idiomatic expressions it enters. Sinclair (1991:85) states that, in nominal groups, it is the function of *of* "to introduce a second noun as a potential headword". He claims that in many cases, such as (61), "the second noun [...] appears to be the most salient" (Sinclair 1991:85) and goes so far as to argue that in (62) and (63), the second noun is the headword (Sinclair 1991:86; examples taken from Sinclair 1991:85-6).

(61) this kind of problem

(62) one of my oldest friends

(63) a lot of the houses

This claim is confirmed by findings from Clark and Wasow (1998:212), who show that "fixed expressions" like *a lot of* are treated much more like single units than other noun phrases in the sense that *a* in *a lot of* is much more rarely repeated than *a* in other noun phrases. This finding also provides evidence that hesitations can serve as indicators of chunking in these contexts.

In her aforementioned study of the placement of pauses in noun phrases (see Section 2.3.3), Bybee (2007b:320-2) notes that the prepositions *of*, *to* and *in* are in the list of 19 items which most commonly precede her set of selected nouns. Unfortunately, she provides no information on whether the noun-preposition combinations are likely to be interrupted by pauses or whether the pause is instead placed after the noun in these cases (which is Bybee's indicator for chunkiness; Bybee 2007b:320).

Beckner and Bybee (2009:41) find that out of 6,254 tokens of *in spite* in the COCA, 6,241 are followed by *of*. Thus the direct transitional probability from *in spite* to *of* is 99.5%. A qualitative diachronic analysis furthermore shows that associations between *in spite of* and the meaning of *spite* have weakened over time to a degree that *in spite of* today is no longer transparent (cf. Beckner and Bybee 2009:38). The authors conclude that their findings undermine the standing definition that there is a constituent boundary

before *of* and that instead *in spite of* constitutes a constituent (Beckner and Bybee 2009:41).

Vogel Soza and MacFarlane (2002) conduct a reaction time study in order to test whether highly frequent 'X+*of*' combinations are stored holistically. They selected 24 utterances containing *of* from the Switchboard corpus and grouped them into four frequency bins (see Table 4.1). The recordings were then played to participants who were told to press a key as soon as they heard *of*. A response was considered correct if the key was pressed within 1700ms after the onset of *of*. Results show that the percentage of correct responses declines the more frequent the *of*-bigram (see Table 4.1). In the two highest frequency bins, participants only responded correctly in 38% and 37% of cases respectively. Interestingly, reaction times (also shown in Table 4.1) did not change gradually, but stayed within the same range until the highest frequency bin, where reaction was highly significantly ($p<.001$) slower than in all other bins combined (cf. Vogel Sosa and MacFarlane 2002:232-3).

The authors conclude that the highest-frequency pairs are chunked in the sense that they are stored holistically. The whole has become autonomous from its constituent parts and therefore access to the parts is no longer necessary to process the whole (Vogel Sosa and MacFarlane 2002:234).

Group	1	2	3	4
Bigram frequency	1 - 99	100 - 299	300 - 799	>= 800
Bigrams	sense of piece of sums of each of example of colleague of	care of because of kinds of bit of any of much of	couple of part of most of all of think of type of	kind of lot of one of out of sort of some of
Percentage of correct responses	60%	47%	38%	37%
Average reaction time	779.2 ms	773.8 ms	775.9 ms	956.6 ms

Table 4.1: Rate of detection of 'of' in a word-monitoring test (cf. Vogel Sosa and MacFarlane 2002)

In light of this evidence, it is highly expectable that in certain constructions prepositions form a chunk with the preceding word. Evidence of such chunks which violate traditional phrase boundaries is particularly strong for *of*. I hypothesise that these proposed chunks will be detectable in the present data, meaning I expect speakers to

place hesitations elsewhere than the phrase boundary if the pair bridging the boundary is chunked. Furthermore, the present approach models chunking in a very concrete, word-form level. It will be interesting to see whether the approach allows for conclusions concerning whether some constructions, e.g. 'Quantifier + *of*', are represented on a more abstract level.

Embedded phrases – Prepositional phrases typically contain a noun phrase as their prepositional complement, which allows for an analysis of whether the boundary of the embedded noun phrase is reflected in the frequency and attraction of word-pairs in the prepositional phrase and whether it has an effect on the placement of hesitations. This is particularly interesting in light of Bybee's (2007b:321-2) finding that some preposition-noun combinations, such as *of money* and *to/in school*, are very frequent and might therefore form chunks across the noun phrase boundary. Furthermore, the noun phrase can range in complexity from the simplest phrase type (just a noun) to more complex structures such as 'Determiner Adjective Noun', which allows for a comparison of frequency effects in simple and expanded phrases.

Results are comparable – Finally, hesitation placement in prepositional phrases has been researched elsewhere. A basic analysis of hesitation placement in prepositional phrases, for example, can be found in Maclay and Osgood (1959:30-34). The authors were among the first to investigate whether there is a system to the distribution of filled and unfilled pauses in spontaneous speech. For their analysis, Maclay and Osgood select 16 phrase types, half of which are prepositional phrases. Table 4.2 provides their prepositional phrase results; numbers for filled and unfilled pauses are added up and the position after the noun is ignored.

The analysis is restricted to a comparison of occurrence before function versus content words and placement within the phrase versus at phrase boundaries. The authors find that hesitations are significantly more likely to occur before content words than before function words and that 53% of pauses are placed at phrase boundaries (Maclay and Osgood 1959:32-3). However, if we count only the prepositional phrase transitions listed in the table below, this figure drops to 30.6%. Maclay and Osgood conclude that

> [i]t is as if we had available at some level of encoding a 'pool' of heavily practiced, tightly integrated word and phrase units, but selection from this pool requires simultaneous lexical and grammatical determinants. (Maclay and Osgood 1959:41)

Precisely these units are at the focus of the following analysis of hesitation placement in prepositional phrases.

Phrase Type	before Prep	before Det	before Adj1	before Adj2	before V-ing	before N1	before N2	Total
Prep N	81					144		225
Prep Det N	90	66				101		257
Prep N N	8					20	6	34
Prep Det N N	7	7				9	7	30
Prep Adj N	7		18			13		38
Prep Adj Adj N	4		7	4		5		20
Prep Det Adj N	28	24	45			28		125
Prep V-ing N	6				14	6		26
	231	97	70	4	14	326	13	755

Table 4.2: Excerpt of results from Maclay and Osgood 1959:31-2: Placement of filled and unfilled pauses in prepositional phrases

4.2 Data & Predictors

4.2.1 Selection of Phrase Types

For my analysis, hesitations occurring in the context of a restricted set of prepositional phrase types were extracted from the corpus. A set of simple and therefore frequently-used phrase types was preferable because it allows for a comparative analysis of hesitation placement in similar contexts. Maclay and Osgood's (1959) scheme contains such a selection of basic types. These were therefore adopted. However, only very few hesitations occurred in the context of the structures 'Preposition Adjective Adjective Noun' and 'Preposition Verb-ing Noun' in Switchboard NXT, so these structures were excluded. Table 4.3 below shows the selected set of phrase types and the number of hesitations or hesitation clusters occurring in each type.

Phrase Type	**Example**	**n**
Prep N	at home (sw3586.A.s110)	1,231
Prep Det N	to the park (sw3324.B.s82)	1,440
Prep N N	before spring break (sw2092.A.s186)	553
Prep Det N N	in the winter time (sw3124.B.s109)	494
Prep Adj N	with low mileage (sw2299.B.s72)	431
Prep Det Adj N	in the low forties (sw3377.B.s104)	575
	Total	4,724

Table 4.3: Types of prepositional phrases

4.2.2 Retrieval Procedure & Definitions

The data was gathered as follows. First, by means of searching the syntactic annotations (see Section 3.1.1.2), hesitations and hesitation clusters[31] occurring in the context of prepositional phrases were selected from a list of all hesitations used in Switchboard NXT. This set was then narrowed down to those hesitations placed within or directly preceding any prepositional phrase beginning in one of the sequences listed in Table 4.3.

Hesitations were selected for analysis, irrespective of how or whether the phrase continued after the sequence. Attention was paid, however, to formulate all search queries in a way which guaranteed that the structures 'Preposition Noun' and

[31] For a definition of hesitations and hesitation clusters see Section 3.1.2 above.

'Preposition Determiner Noun' were not followed by a noun (which would, in effect, have made 'Preposition Noun Noun' and 'Preposition Determiner Noun Noun' their subgroups). Furthermore, hesitations occurring after the final noun were not taken into consideration as hesitation placement in this position is likely to be determined by the planning demands imposed by the structure which follows.

The syntactic and part-of-speech tagging was only checked manually for a small, randomly chosen sample of sentences in Switchboard. It was found to be very accurate considering that the data consists not just of spoken language but of disfluent speech. Nonetheless tagging and parsing errors which occurred could not be corrected due to the vast amount of data to be searched and handled. I make the assumption that any noise potentially caused in this way will be balanced out by the large number of data-points.

Importantly, the tagger and the parser recognised phrasal verbs, i.e. they recognised that the preposition in these constructions functions as an adverbial particle and that the phrase is consequently not a prepositional phrase. (64) and (65) show the tagger and parser output in these cases. Phrasal verbs are therefore categorically excluded from analysis.

(64) (VP bring_VB (PRT up_RP) (NP painting_NN))
(65) (VP set_VB (NP a_DT siren_NN) (PRT off_RP))

Due to the setup of the corpus, some definitions are rather Switchboard NXT-specific:

- Due to the fact that in Switchboard NXT any sequence separated by a space from another sequence is treated as a separate word, complex prepositions are not tagged as such, but as individual words. Consequently, these are not picked up by my search query.
- As the letters in alphabetisms are separated by spaces in Switchboard, these are tagged as sequences of one-letter nouns (see also Section 3.1.1.3).
- All elements in titles and proper names are tagged as proper names. Therefore, *Pink Floyd*, *New Hampshire* and *muscular dystrophy* are all coded as sequences of proper names.
- Prepositions and coordinating conjunctions are coded with the same part-of-speech tag (see Appendix A), which may have resulted in my heuristics also picking up conjunctions. Yet, as the double precaution was taken to rely on both the syntactic parsing and the part-of-speech tagging, inaccurate hits of this sort are extremely rare.

Crucially, in the resulting dataset each hesitation or hesitation cluster constitutes a data-point, *not* each token of a phrase type. This makes it possible to handle phrases containing more than one hesitation, as is the case in (66).

(66) from [*uh*] industrial [*pause*] areas (sw2094.A.s9)

Cases like (66) are, in fact, counted as two data-points; one instance of *uh* occurring between the preposition and the noun and one instance of a pause occurring between the two nouns. Of course, regression models are able to predict only one of the hesitations in (66) correctly, as the environment considered for each hesitation is exactly the same. It would have been possible to avoid this issue by removing all hesitation tokens which occurred in structures such as (66) from the dataset. This would, however, have resulted in two serious shortcomings, a) a loss of about ten percent of data and b) a severe reduction in the scope of possible conclusions. Claims like 'hesitations *never* occur between structures with properties X, Y and Z' would no longer be warranted had such a substantial amount of data been deleted.

Finally, in order to be able to analyse relations holing across the phrase boundary, the word before the phrase was always extracted and included in the analysis, too. This led to the exclusion of all cases where the prepositional phrase occurred sentence-initially, because in these cases no word before the preposition existed.

4.2.3 Distribution of Hesitations

In total, the dataset consists of 3,742 individually-occurring hesitations and 982 hesitation clusters, adding up to 4,724 data-points (for a definition of hesitation clusters and reasons for combining clusters and individual hesitations in a single analysis, see Sections 3.1.2 and 3.1.3). Table 4.4 and Figure 4.1 illustrate that the placement of hesitations in the analysed data is extremely varied. Hesitations occur in all kinds of transitions; there is no position where speakers never hesitate. Only the position before the second content word in longer noun phrases is dispreferred.

Across the board, the position at the prepositional phrase boundary is the most popular (44.7%), followed by placement before the first content word (38.5%). A total of 2,611 hesitations occurs within the phrase. This means that on average a hesitation has a 44.7% chance of being placed at the prepositional phrase boundary compared to a 55.3% chance of being placed within the prepositional phrase. This result lies somewhere in between the rate of hesitations occurring at phrase boundaries found in Maclay and Osgood's set of prepositional phrases (1959; 30.6%; see above) and the rates reported by Bortfeld et al. (2001:138; 39.6%) and Goldman-Eisler (1968:95;

47-61%). Goldman-Eisler's higher rate may be due to the artificiality of her dataset. Deviations from Maclay and Osgood as well as Bortfeld et al. may arise from the fact that neither of these studies considers discourse markers, which show a propensity to occur at or near the phrase boundary; the present dataset does consider these, which increases the hesitation rate at the prepositional phrase boundary. In fact, hardly any discourse markers occur after the first transition within the prepositional phrase, while filled and unfilled pauses are used at all transitions – in some phrase types even more frequently before the first content word than elsewhere.

Phrase Type	**before Prep**	**before Det**	**before Adj**	**before N**	**before N**	**Total**
Prep N	596			635		1,231
Prep Det N	847	301		292		1,440
Prep N N	165			316	72	553
Prep Det N N	142	107		183	62	494
Prep Adj N	122		251	58		431
Prep Det Adj N	241	95	143	96		575
	2,113	503	394	1,580	134	4,724

Table 4.4: Distribution of hesitations across the six prepositional phrase types

Overall, the placement pattern confirms Clark and Clark's (1977:267-8) hypothesis that positions before the phrase and before the first content word in the phrase are preferred for hesitation placement.

Finally, it appears that the preferred hesitation placement in structures which do not contain a determiner is the place before the first content word, whereas preference generally shifts to the prepositional phrase boundary in structures with a determiner. It is, however, not possible to determine what exactly causes this shift without a more in-depth look at the data, which will follow in the next sections.

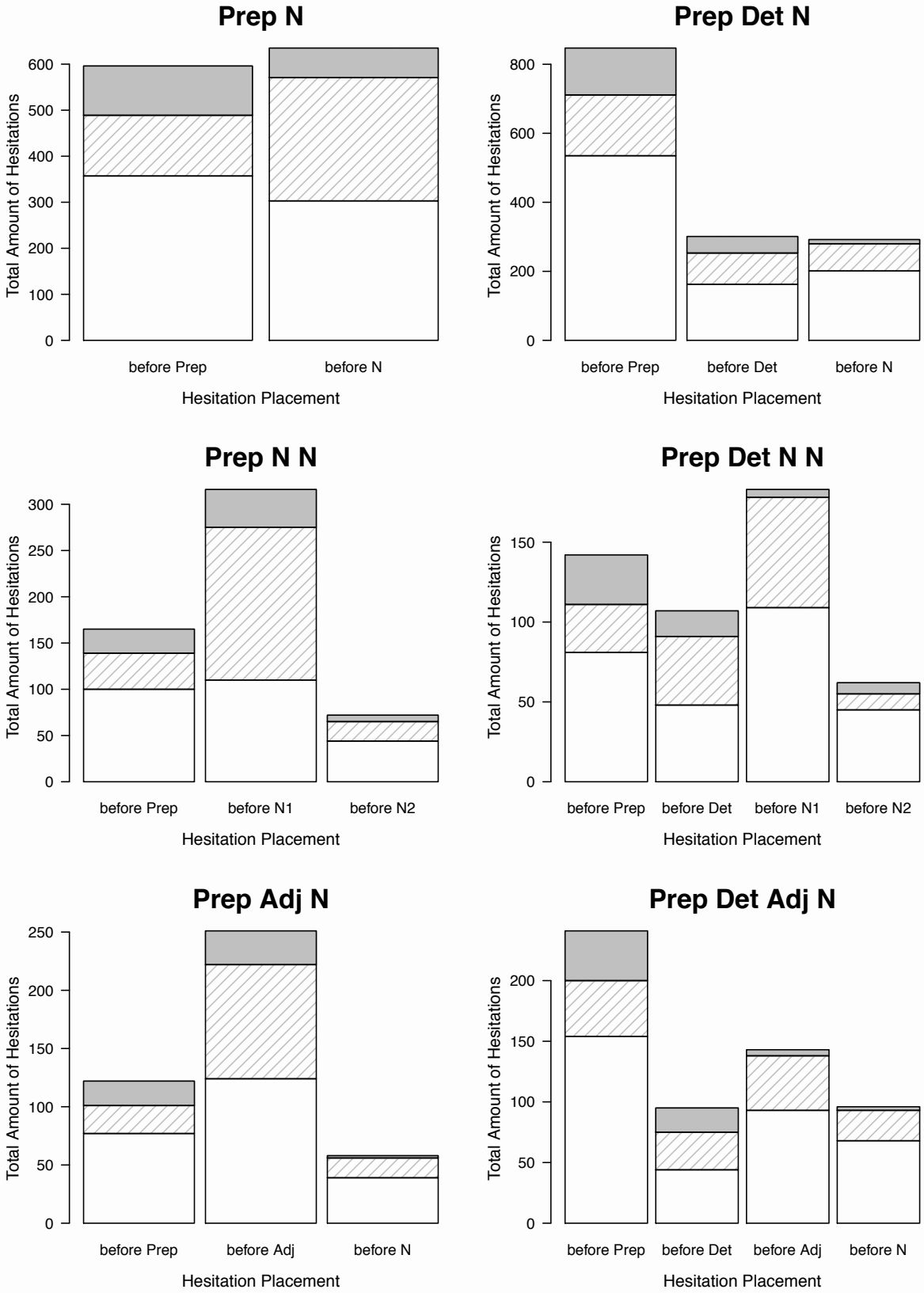


Figure 4.1: Distribution of hesitations across prepositional phrase types. White bars indicate unfilled pauses, ruled bars indicate filled pauses and grey bars indicate discourse markers.

4.2.4 Predictors

Analyses will draw on the following set of predictors, which are explained in more detail in Section 3.3.1. They are listed in order of increasing complexity, i.e. increasing amounts of information needed for their calculation. Theoretically, the more information is combined in a measure, the more 'knowledgeable' it should be. Consequently, performance as predictors of chunkiness should increase as we move down the list.

Bigram Frequency (bi.freq) – The absolute frequency of the word-pair (or bigram) in the Switchboard NXT corpus. Usage frequency is a simple measure of association. Usage-based theories expect that the more frequently a speaker uses a structure the more strongly it becomes entrenched. Consequently, we expect speakers to be less likely to insert hesitations into high-frequency pairs than into lower frequency pairs.

Direct Transitional Probability (TPD) – Each word has collocates which are likely to follow it, but there are also words which are very unlikely to follow. Direct (or forward) transitional probability indicates on a scale from zero to one how likely it is that the second word in a pair will come up given the first. The higher the transitional probability, the 'chunkier' the sequence should be.

Backwards Transitional Probability (TPB) – Backwards transitional probability is equivalent to its direct brother, only it measures how likely it is that the first word in the pair will *precede* the second.

Mutual Information Score (MI) – While the two previous measures were unidirectional in the sense that they only consider attractions in one direction (either forwards or backwards), the MI score is bidirectional and combines these two views. In Switchboard NXT, MI scores range from about -8.5 to 20. Sometimes pairs receiving a negative score are referred to as words that repel. Generally, the higher the score, the chunkier the pair should be.

Lexical Gravity G (G) – G is also bidirectional, though more complex than MI. In contrast to all previous measures of association, it takes into account that language is not a lottery, i.e. that syntax and semantics restrict the combinability of words. Not all theoretically possible combinations of words actually occur in language. Therefore, G measures how likely the combination AB is to occur given all *actually* observed combinations into which word A enters and all actual collocates of word B. In Switchboard NXT, G ranges from roughly -13 to 16.5. Again the higher the score, the

chunkier the pair should behave, and the words in negative-rated pairs are sometimes referred to as repellent.

Two further predictors are included in the analysis:

Word frequency (w.freq) – Word frequency is a control factor included in order to see whether low-frequency words attract hesitations irrespective of the larger context.

Hesitation Type (hes.type) – From the analysis of different hesitation types described in Section 3.1.3.1 we know that filled pauses, unfilled pauses and discourse markers do not always behave in the same way. Thus this predictor allows the analyses to distinguish between

- unfilled pauses (*pause*),
- filled pauses as well as combinations of filled and unfilled pauses (*u*) and
- discourse markers, including combinations of discourse markers and filled or unfilled pauses (*dm*)

This predictor is the only one in the set which is not numerical, but categorical.

4.2.5 Frequency-based Characteristics of all Transitions

In terms of frequency, transitional probability and the like, a typical 'Preposition Determiner' sequence differs considerably from a typical 'Noun Noun' combination. This section provides a brief characterisation of each type of transition or word-pair in the dataset. The aim is to illustrate characteristics of the dataset. Therefore, mean values and the standard deviation (sd) were only calculated for those pairs which actually occur in the set. This means that, for instance, not all 'Determiner Noun' combinations were taken into consideration, but only the set of 1,934 which is included in the present selection of data. For the sake of better comparability, separate tables and graphs for each transition were forgone and values are instead shown in comprehensive Table 4.5 as well as Figures C.1 to C.6 in the Appendix.

	n		Freq.	TPD	TPB	MI	G
X+Prep	**4,724**	mean	94.30	0.17	0.01	2.70	3.34
		sd	259.73	0.27	0.036	3.00	4.92
Prep+N	**1,784**	mean	26.17	0.005	0.26	4.20	0.05
		sd	78.71	0.03	0.30	2.37	3.93
Prep+Det	**2,509**	mean	735.91	0.13	0.04	2.28	10.45
		sd	857.86	0.11	0.04	0.85	4.38
Det+Adj	**575**	mean	67.06	0.004	0.31	3.54	3.49
		sd	132.88	0.01	0.27	1.80	4.52
Prep+ Adj	**431**	mean	5.67	0.002	0.13	3.06	-1.17
		sd	11.80	0.007	0.23	2.39	2.64
Det+N	**1,934**	mean	68.23	0.005	0.33	3.70	2.53
		sd	280.03	0.02	0.30	1.83	4.31
Adj+N	**1,006**	mean	10.55	0.15	0.14	9.04	0.36
		sd	35.05	0.25	0.23	3.50	3.05
N+N	**1,047**	mean	8.99	0.28	0.29	11.42	0.29
		sd	24.17	0.33	0.35	3.98	2.72

Table 4.5 Mean values and standard deviation (sd) of frequency, direct transitional probability (TPD), backwards transitional probability (TPB), mutual information score (MI) and (G) of all transition types in the dataset

4.2.5.1 X & Preposition

Linguistically, this is the most diverse group, because it is less narrowly defined than the other pairs. The X in this type can be any part of speech as long as it precedes the preposition. In this way, this group comprises all types of word-pairs which bridge the prepositional phrase boundary. The large variety of different combinations is reflected in the broad distribution of G values. For this group, lexical gravity G ranges from -10.1 (*kind like*) to 14.3 (*lot of*).

The group contains one highly frequent pair, namely *lot of* (absolute frequency in the corpus: 1,723; included in the dataset 91 times), and some with a direct or backwards transitional probability of one. (67) shows some examples of pairs with such a high transitional probability. The 100% transitional probability indicates that in the Switchboard NXT corpus the first word in these pairs is always followed by the second.

(67) compensate for, referring to, shortage of, most of

Backwards transitional probabilities are very low for this type of transition, owing to the fact that prepositions can fulfil a range of functions and occur in many different constructions. Thus while *compensate* is invariably followed by *for*, the preposition *for* will be used in a great range of other constructions, leading to a high direct transitional probability but a very low backwards transitional probability.

4.2.5.2 Preposition & Noun

This type of word-pair shows the characteristic pattern of function word content word sequences – low direct and high backwards transitional probabilities (compare Figures C.2 and C.3 in the Appendix) – which derives from the fact that function words can typically be followed by a large range of content or other function words, while content words like nouns are typically preceded by one of a much smaller set of function words. The mean backwards transitional probability for this group is 0.26, indicating that on average each noun is preceded by the same preposition in a quarter of all cases it is used. In 148 cases, the backwards transitional probability for this type of combination is as high as 100%. (68) lists some of these cases.

(68) from scratch, of periodicals, in Kingsport, at Amherst

The single outlier (see the plot for 'Preposition Noun' in Figure C.2 in the Appendix) with a direct transitional probability of one is *concerning recycling*.

4.2.5.3 Preposition & Determiner

This is the only type of pair in the dataset where both elements are function words. Both prepositions and determiners are small groups of highly frequent words, which, when put in sequence, do only have a limited number of combinatorial possibilities. This leads to this group's exceptionally high mean frequency and lexical gravity G which distinguish it from all other kinds of combinations in the dataset. It contains both the pair with the highest frequency (*in the*; absolute pair frequency in the corpus: 2,436) and

the highest G in the dataset (*of the*; lexical gravity G=16.1). Of course, these pairs carry little semantic content and probabilistically could simply co-occur by chance due to the high frequency of both determiners and prepositions. This is reflected in the comparatively low MI values.

4.2.5.4 Determiner & Adjective; Determiner & Noun

These groups are so similar in terms of their characteristics that they will be treated together. Both show the typical pattern of transitional probabilities found in most function word content word sequences (see also Section 5.2.5.2). There are 33 cases of backwards transitional probabilities of 100% among the 'Determiner + Adjective' combinations and 138 among the 'Determiner + Noun' combinations (i.e. cases where the adjective or noun is always preceded by the same word, namely a specific determiner). (69) and (70) show some examples.

(69) each succeeding, a manual, the verbal, a handcrafted, my bridal

(70) our employee, a tractor, the fast-food, the bulls, my bushes

Most of these cases reach this high transitional probability because the adjective or noun is a hapax legomenon and consequently only occurs in this single combination in the corpus. In the remaining cases, the content word is rare – the maximum corpus frequency being eight. The most frequent members of the groups are *a little* (corpus frequency: 657) and *a lot* (corpus frequency: 2,193). Of all pairs in these groups, *each succeeding* and *each assembly* receive the highest MI scores (11.9 and 11.3 respectively).

4.2.5.5 Preposition & Adjective

This type of combination has the lowest mean direct transitional probability of all bigram types in the dataset. The highest direct transitional probability reached in this group is a very low 0.08 (*past few*). This reflects the fact that prepositions can be followed by many different words (commonly determiners and nouns) and that adjectives, being optional elements in noun phrases, are unlikely to follow any particular preposition.

This also leads to members of this group being typically infrequent (highest corpus frequency: 75; *in high*) and having a low backwards transitional probability. There are, however, 22 outliers with a backwards transitional probability of one (see the fourth box plot in Figure C.3 in the Appendix). A selection of these outliers is shown in (71).

(71) by subsidized, as nonfeeling, with quarterly, of grilled

4.2.5.6 Adjective & Noun; Noun & Noun

These types will also be described together because they share most of the same characteristics. They are the only combinations of two content words in the dataset and consequently differ from the rest particularly in that they both, on average, have relatively high forward and backwards transitional probabilities and a far higher mean MI than any other group. Many combinations are, however, hapax legomena occurring only once in Switchboard NXT. This is the case for 33% of 'Adjective + Noun' combinations and for 53.4% of 'Noun + Noun' combinations.

Direct transitional probability reaches 100% for 54 tokens of adjective combinations and 116 tokens of noun combinations, and for 46 tokens of adjective combinations and 142 tokens of noun combinations backwards transitional probability reaches 100% (see Figures C.2 and C.3 in the Appendix). (72) and (73) show examples which have a direct transitional probability of 100% and (74) and (75) show examples which have a backwards transitional probability of 100%

(72) unleaded gas, rubbery cornstarch, childbearing years

(73) Vatican City, raspberry puree, Horseshoe Bay

(74) long layovers, Christian authors, judicial mishap, funny amendments

(75) television cameras, child molestation, carbon monoxide

In many cases, these exceptionally high transitional probabilities are reached because the bigrams in question are hapaxes. MI is very sensitive to hapaxes – scores are highest when both words in a pair are hapaxes, and the resulting combination necessarily only occurs once, too. Lexical gravity G, on the other hand, correlates strongly with the frequency of the word-pair and is therefore low for these groups.

Among the 'Noun + Noun' combinations there are many proper names and titles of books and films because all words in a title are tagged as proper names. Some examples are listed in (76).

(76) Carl Albrecht, Pink Floyd, Camp Goddard, East Fork, Captain Kirk

4.3 Previously Suggested Factors

This section addresses Lounsbury's (1954) first hypothesis concerning speech production in which he postulates that

> [h]esitation pauses correspond to the points of highest statistical uncertainty in the sequencing of units of any given order. (High statistical uncertainty = high transitional entropy.) (Lounsbury 1954:99)

According to this hypothesis, variation in the present data should be attributable to differences in transitional probability; hesitations should always be placed where transitional probability is lowest. This line of reasoning differs from the chunking hypothesis at the heart of this work; the arguments resulting from the two theories represent 'inverted' viewpoints.

The postulation of chunks leads to negative predictions, i.e. predictions about where hesitations are *not* placed. In other words, the assumption that chunking strength rises with transitional probability, and that the higher the chunking strength the less likely speakers are to interrupt a chunk, leads to the hypothesis that a speaker in need of some extra planning time will be *least* likely to interrupt the speech flow where transitional probability is *high* or *highest*. The focus is thus on the high-frequency or strong attraction range.

Lounsbury's way of reasoning, on the other hand, leads to positive predictions, i.e. assumptions concerning where speakers will *most* likely position a hesitation. This line of argumentation focusses on the low-frequency or low-attraction range of pairs. Predictions based on these two approaches are not contradictory but complement each other because the point of least attraction is also the least chunky in a given context.

Before proceeding to test Lounsbury's hypothesis, it is important to draw attention to the fact that his hypotheses[32] do not perfectly fit the methodology applied here. First of all, while he describes spontaneous speech as interspersed with "pauses and perhaps quite a bit of hemming and hawing" (Lounsbury 1954:98), his concept of pauses only encompasses the "latency" between two events (Lounsbury 1954:98) – thus presumably no filled pauses. It definitely does not include discourse markers. Secondly, Lounsbury may possibly refer to transitional probability calculated based on a larger context (Lounsbury 1954:93)[33]. Finally, Lounsbury's hypotheses are obviously only based on transitional probabilities. As the present study has further measures of attraction at its disposal, his hypotheses will be tested on all of them.

---

[32] For a brief discussion of his further hypotheses see Section 2.3.3 above.

[33] Transitional probability, as calculated here, takes into account a context of one word. It asks 'Given this word, how likely is a specific second one to follow?'. The same logic can be applied to larger contexts, thus asking 'Given two words, how likely is a specific third to follow?' and so on.

Lounsbury's hypothesis is confirmed if there is a measurable statistical tendency to place hesitations where the transitional probability or the score of any other measure of association is lowest. To this purpose, Table 4.6 shows the distribution of association strengths in the present dataset. Columns four to seven ('Distribution of lowest values') show how often a specific transition is the one with the weakest association in the phrase. The first row, for example, reads as follows: in the 'Preposition Noun' dataset, in 359 cases, the 'X+Preposition' pair was less frequent than the 'Preposition + Noun' pair and in 734 cases the 'Preposition + Noun' pair was the least frequent. The number of data-points considered is sometimes considerably lower than the absolute number of data-points available for a specific phrase type, due to the fact that only phrases with a single lowest point could be included in the analysis. This means that phrase types where two transitions received the same score were excluded. This was mostly the case where several bigrams in a phrase occurred only once.

It can be observed that different measures of association make very different claims about the data. A bigram type which scores low on one scale does not necessarily also receive low scores by other measures of association. Consequently, the 'weakest link', most frequently receiving the lowest score in the phrase, differs from measure to measure. If we look at 'Preposition Noun' at the top of the table again, we can see that according to backwards transitional probability, the word-pair bridging the phrase boundary ('Position 1') is most frequently the least attracted, while according to the other measures, it is the 'Preposition + Noun' pair which is more often the least attracted.

Associations holding across the prepositional phrase boundary are not generally the weakest. Only backwards transitional probability and the mutual information score tend to rate the pairs bridging the phrase boundary as less cohesive than transitions within the phrase. This is in accordance with findings presented in the previous section (see Table 4.5). It is a first indication that chunking may violate traditionally assumed constituent boundaries.

Results do not confirm Lounsbury's hypothesis in its strongest form. Hesitations are not placed at the "point of highest statistical uncertainty" (Lounsbury 1954:99) throughout. They do, however, suggest that a weaker version of the hypothesis is warranted. While not *all* hesitations are found at the point with the weakest attractions in the phrase, there are significantly more hesitations at this place than expected by chance.

Phrase Type	Measure	No. of data-points	Distribution of lowest values				No. of hes. at lowest position	Percent at lowest pos.	Significance level
			Position 1	Position 2	Position 3	Position 4			
Prep N	**Frequency**	1,093	359	734			673	61.57%	p<.001
	TPD	1,230	180	1,050			737	59.92%	p<.001
	TPB	1,231	1,120	111			633	51.42%	non-sig.
	MI	1,220	805	415			780	63.93%	p<.001
	G	1,231	406	825			737	59.87%	p<.001
Prep Det N	**Frequency**	1,332	703	17	612		649	48.72%	p<.001
	TPD	1,437	195	102	1,140		465	32.36%	below chance
	TPB	1,437	1,152	264	25		769	53.51%	p<.001
	MI	1,441	696	577	168		701	48.65%	p<.001
	G	1,441	751	50	640		696	48.30%	p<.001
Prep N N	**Frequency**	271	48	133	90		132	48.71%	p<.001
	TPD	546	23	519	4		319	58.42%	p<.001
	TPB	524	426	45	53		191	36.45%	non-sig.
	MI	547	305	239	3		258	47.17%	p<.001
	G	554	93	362	99		282	50.90%	p<.001
Prep Det N N	**Frequency**	313	57	0	67	189	86	27.48%	non-sig.
	TPD	486	26	9	433	18	192	39.51%	p<.001
	TPB	493	355	68	10	60	146	29.61%	p<.05
	MI	493	201	216	29	0	174	35.29%	p<.001
	G	493	105	1	191	196	158	32.05%	p<.001
Prep Adj N	**Frequency**	230	45	84	101		109	47.39%	p<.001
	TPD	430	21	382	27		256	59.53%	p<.001
	TPB	418	302	53	63		138	33.01%	below chance
	MI	428	192	231	5		222	51.87%	p<.001
	G	430	67	273	90		231	53.72%	p<.001
Prep Det Adj N	**Frequency**	406	89	1	27	289	149	36.70%	p<.001
	TPD	571	57	9	424	81	190	33.27%	p<.001
	TPB	575	402	76	10	87	235	40.87%	p<.001
	MI	575	254	247	74	0	224	38.96%	p<.001
	G	575	151	7	115	302	219	38.09%	p<.001

Table 4.6: Number and percentage of hesitations placed at the transitions with the lowest cohesion per phrase according to bigram frequency (Frequency), direct transitional probability (TPD), backwards transitional probability (TPB), mutual information score (MI) and lexical gravity G (G).

4.4 Analyses by Structure

This section forms the backbone of this chapter. It details the statistical analyses of the six selected prepositional phrase types. Based on the predictors listed in Section 4.2 above and introduced in more detail in Sections 3.3.1 and 3.3.2, hesitation placement and consequently chunking in the prepositional phrase contexts will be analysed. The regression methods employed for this purpose are CART trees and random forests which divide groups of data-points into ever more homogenous sub-groups based on each data-point's attributes. Thus, these analyses will, for example, separate phrases starting in low-frequency word-pairs from those beginning with high-frequency pairs if the internal algorithm has determined that hesitation behaviour in the former group differs significantly from that in the latter group. The final division of the data between the terminal sub-groups (also called leaves or nodes) can be further analysed both quantitatively and qualitatively.

Analyses provided here focus on the quantitative evaluation of whether hesitation behaviour in the individual nodes is sufficiently homogenous to be considered evidence of frequency effects in the data. In each subsection, I will also point out whether there are nodes in which one construction dominates or which splits allow for conclusions about the role of phrase boundaries; these are particularly interesting in the context of the focal questions of this chapter and the overall hypotheses of the study.

Sections 4.5 to 4.7 will be based on these analyses and findings. These sections will pay attention to particular aspects and phenomena as well as taking a meta perspective, not focussing on one phrase type but combining all six. Investigations conducted in these sections will often be of a more qualitative nature.

4.4.1 Preposition Noun

At 1,231 tokens, the 'Preposition Noun' dataset is one of the largest. As described in Section 4.2.2 above, this dataset encompasses hesitations occurring in the context of a preposition followed by either a singular or plural noun or proper name. All cases where the noun was followed by another noun were excluded, because they are already represented in the structure 'Preposition Noun Noun'.

The 'Preposition Noun' sample contains a diverse set of structures, ranging from fixed expressions and discourse markers, such as *of course* and *in terms*, to specific spacial relations, such as *to Hawaii* and *in Illinois.* In total, there are 234 proper names in the noun position of this phrase type. First or family names are surprisingly rare, however. Most proper names are major American cities, as well as names of states and continents, occurring with a variety of prepositions with spacial meanings, such as in (77) to (79).

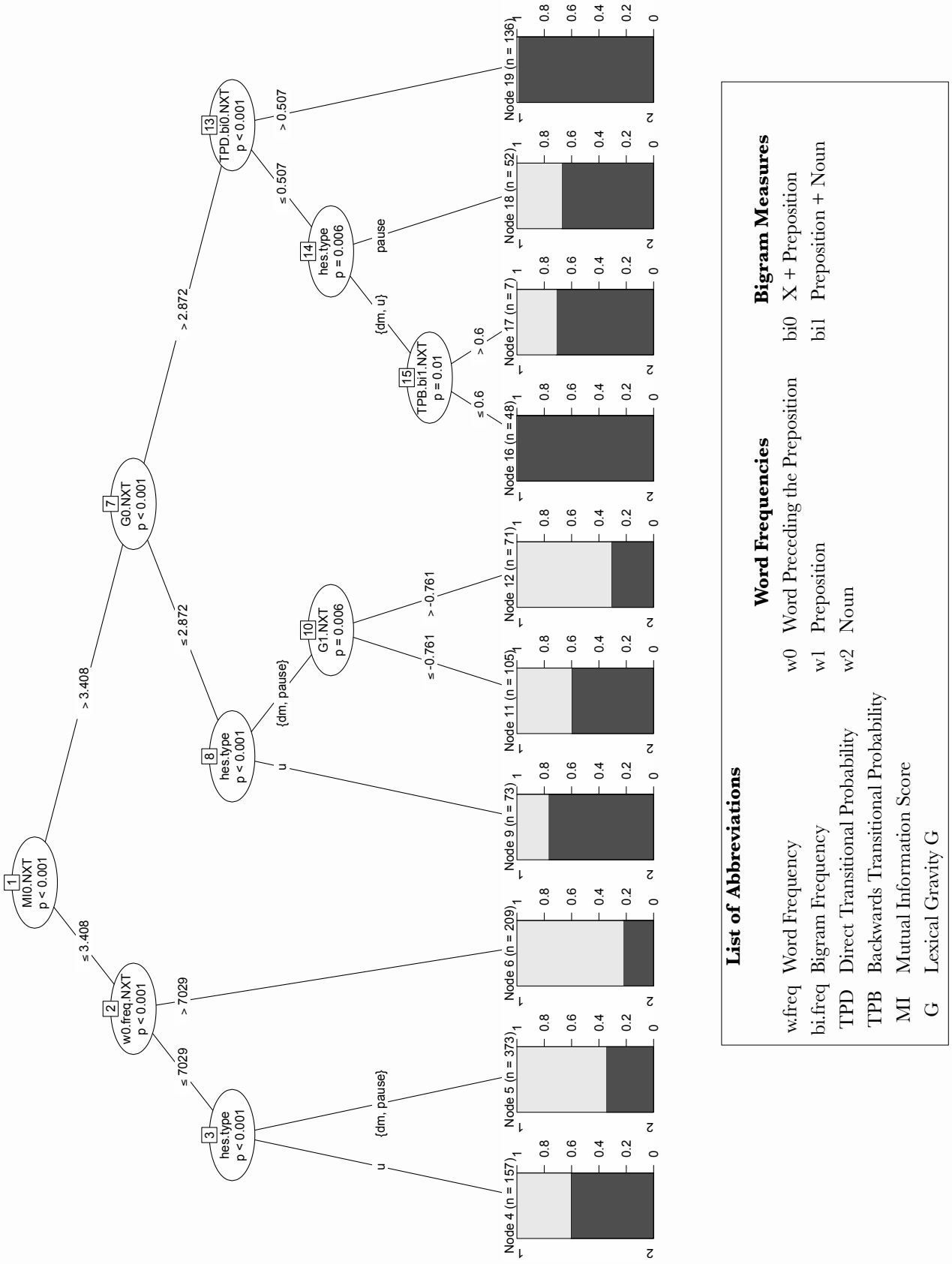


Figure 4.2: Ctree results for the structure 'Preposition Noun'. Labels at terminal node bar graphs (here: 1, light and 2, dark) indicate hesitation position before the corresponding words (w1=Preposition; w2=Noun).

(77) my family is from *[pause] uh [pause]* Europe (sw2259.A.s34)

(78)I saw him in a huge stadium in *[pause] uh* Philadelphia (sw2020.A.s140)

(79) we just send money *[pause]* to Washington (sw4080.A.s57)

For the analysis of frequency effects, a CART tree was fitted to the data with the help of the *ctree* function in *R*. As Figure 4.2 illustrates, *ctree* grows a very complex tree. It generates nine splits resulting in ten terminal leaves.

The effects seen in the tree are as expected; the stronger the attraction between two words, the less likely the pair is to be interrupted by hesitations and vice versa. This is most clearly evident in Node 19. In this group, there are strong attractions between the two words bridging the phrase boundary (i.e. the preposition and the word preceding it), evidenced by a high mutual information score (MI0 in Split 1), a high lexical gravity G (G0 in Split 7) and a high direct transitional probability (TPD.bi0 in Split 13). As a result, we find that only a single hesitation out of 136 is placed before the preposition.

The predominant outcome in a leaf signifies the leaf's prediction for all data-points it contains. In Node 4, for instance, the predominant behaviour is to place hesitations before the noun (Position 2). Therefore, the model's prediction is that all hesitations in this group are placed before the noun. Thus, for this leaf, model predictions are confirmed in 60.5% of cases and contradicted in the remaining 39.5%. In this way, and across all leaves, the tree predicts a total of 892 hesitations correctly, which corresponds to a misclassification rate of 27.54%. Whether this outcome suffices to warrant the claim that the predictors significantly improve model predictions, and thus that there are frequency effects in the data, is tested by comparing this result to that of a baseline model. This model works without predictors and simply presumes that hesitations are always placed at the same position, namely where they are placed in the majority of cases – in this case before the noun. Here, the baseline model predicts only 635 data-points correctly (misclassification rate: 48.42%). A chi-square test reveals that the *ctree* model offers a very highly significant improvement over the baseline model ($p<.001$; residuals: 10.2, -10.53). Table 4.7 provides a comparison of the performance of the model to the actual distribution in the data.

		Model Predictions		
	Hesitation Position	**pre Prep (1)**	**pre N (2)**	**Total**
Actual Distribution	**pre Prep (1)**	455	141	596
	pre N (2)	198	437	635
	Total	653	578	1,231

Table 4.7: Performance of ctree model for 'Preposition Noun'

The model performs equally well at predicting hesitations in both positions, yet not all terminal leaves are equally pure. Nodes 16 and 19 in particular stand out because they only contain hesitations before the noun. Node 16 is dominated by cases where the prepositional phrase functions as a prepositional object following a verb (31 out of 48 cases), such as (80) and (81).

(80) *you know* I was thinking about *uh [pause]* importance (sw2062.A.s15)

(81) without even having gone to *you know* school (sw2708.A.s74)

Node 19, on the other hand, is characterised by quantifying expressions, consisting of a quantifier (mostly *lot* and *lots*) followed by *of* and by hedges such as *kind of* and *sorts of*, see (82) and (83).

(82) there are a lot of *[pause]* jobs (sw2790.A.s108)

(83) because they do all kinds of *[pause] uh* gardening (sw2785.A.s70)

Most of the predictors selected by the model relate to the 'X+Preposition' bigram, indicating that the relation between the word-pair bridging the phrase boundary has a strong effect on hesitation placement. The type of hesitation also plays an important role, as evidenced by its being selected in all three major branches of the tree. If the hesitation is *uh* or *um,* or a cluster of filled and unfilled pauses (hesitation type 'u'), then the chance of the hesitation occurring before the noun is increased.

In order to obtain more stable information about frequency effects in the data and the importance of predictors, a random forest is grown. Forest results, unfortunately, cannot be displayed graphically, because forests are made up of 3,000 trees, all of which are different as they are based on random subsets of predictors and data-points (see also Section 3.3.3.2). Instead, Table 4.8 shows how the forest classifies the data-points. At a misclassification rate of merely 17.3%, the forest's performance is excellent and very highly significantly exceeds the performance of the baseline model ($p<.001$; residuals: 15.2, -15.69).

		Model Predictions		
	Hesitation Position	**pre Prep (1)**	**pre N (2)**	**Total**
Actual Distribution	**pre Prep (1)**	504	92	596
	pre N (2)	121	514	635
	Total	625	606	1,231

Table 4.8: Performance of cforest model for 'Preposition Noun'; ntree=3,000, mtry=5, seed=405

The model's excellent performance may, however, result from overfitting to the data. Overfitting in this case means that a forest generates too many terminal leaves which are too specifically tailored to a unique dataset resulting in claims which cannot be generalised. One way to avoid this, is to make use of so-called 'out-of-bag data'. As each tree in the forest is generated based on a random subset of data-points, there is always the corresponding subset of data unseen by the tree – the out-of-bag data. If we split this data up exactly as determined by the splits in the tree, we can see how well the model fares with new data. The resulting out-of-bag predictions provide a more conservative result (see Table 4.9). The misclassification rate among the out-of-bag prediction rises to 30.14% – still a very highly significant improvement over the baseline model ($p<.001$; residuals: 8.93, -9.22).

		Model Predictions		
	Hesitation Position	**pre Prep (1)**	**pre N (2)**	**Total**
Actual Distribution	**pre Prep (1)**	433	163	596
	pre N (2)	208	427	635
	Total	641	590	1,231

Table 4.9: Performance of cforest out-of-bag predictions for 'Preposition Noun'; ntree=3,000, mtry=5, seed=405

In summary, the fact that the *ctree* function grows a tree at all indicates that there is a frequency effect in the data. Tokens with similar frequency attributes display more homogenous hesitation behaviour than the entire set. A comparison of the model's number of correct predictions to the baseline model's shows that this frequency effect is significant, i.e. not likely due to chance. As individual trees rely on locally optimal splits which may not lead to the globally best solution, a random forest is grown to validate the results and provide more conclusive evidence concerning the importance of predictors. Forests, however, might over-fit to the data. Hence results of the out-of-bag sample are also provided. Both the general forest results and the out-of-bag results confirm the *ctree* prediction. There is a stable frequency effect in the data ($p<.001$). The more frequent a sequence or the more attracted the elements within it, the less likely it is to be interrupted by hesitations.

The variable importance measures shown in Figure 4.3, which can be calculated based on the forest results, show that the integration of the prepositional phrase into the sentence (i.e. measures relating to Bigram 0) influences hesitation placement far more strongly than the internal cohesion of the phrase (i.e. measures relating to Bigram 1). The mutual information score and direct transitional probability of the 'X+Preposition'

bigram are by far the best performing predictors. Finally, it should be noted that the type of hesitation used is also a highly significant predictor. If the hesitation type is *u*, the chance of the hesitation occurring before the noun increases. Consequently, the strong effect observed is not exclusively a frequency effect.

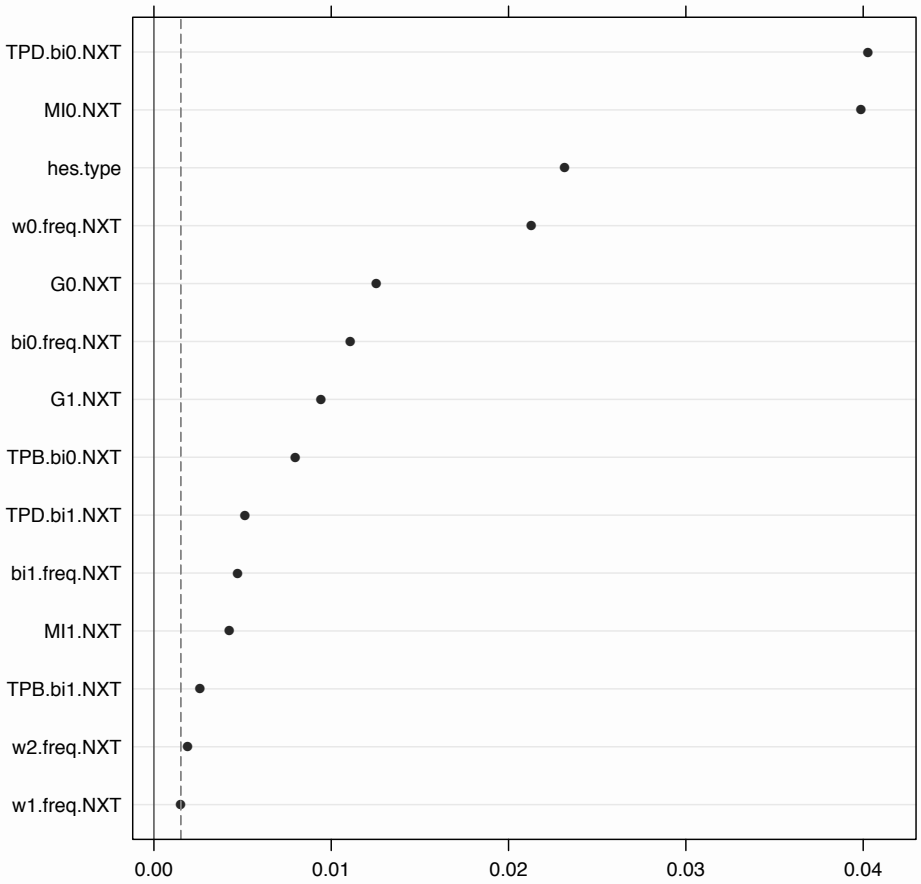


Figure 4.3: Variable importance of predictors for 'Preposition Noun' (mtry=5, ntree=3,000, seed=405, OOB=false, results from R version 2.13.1).

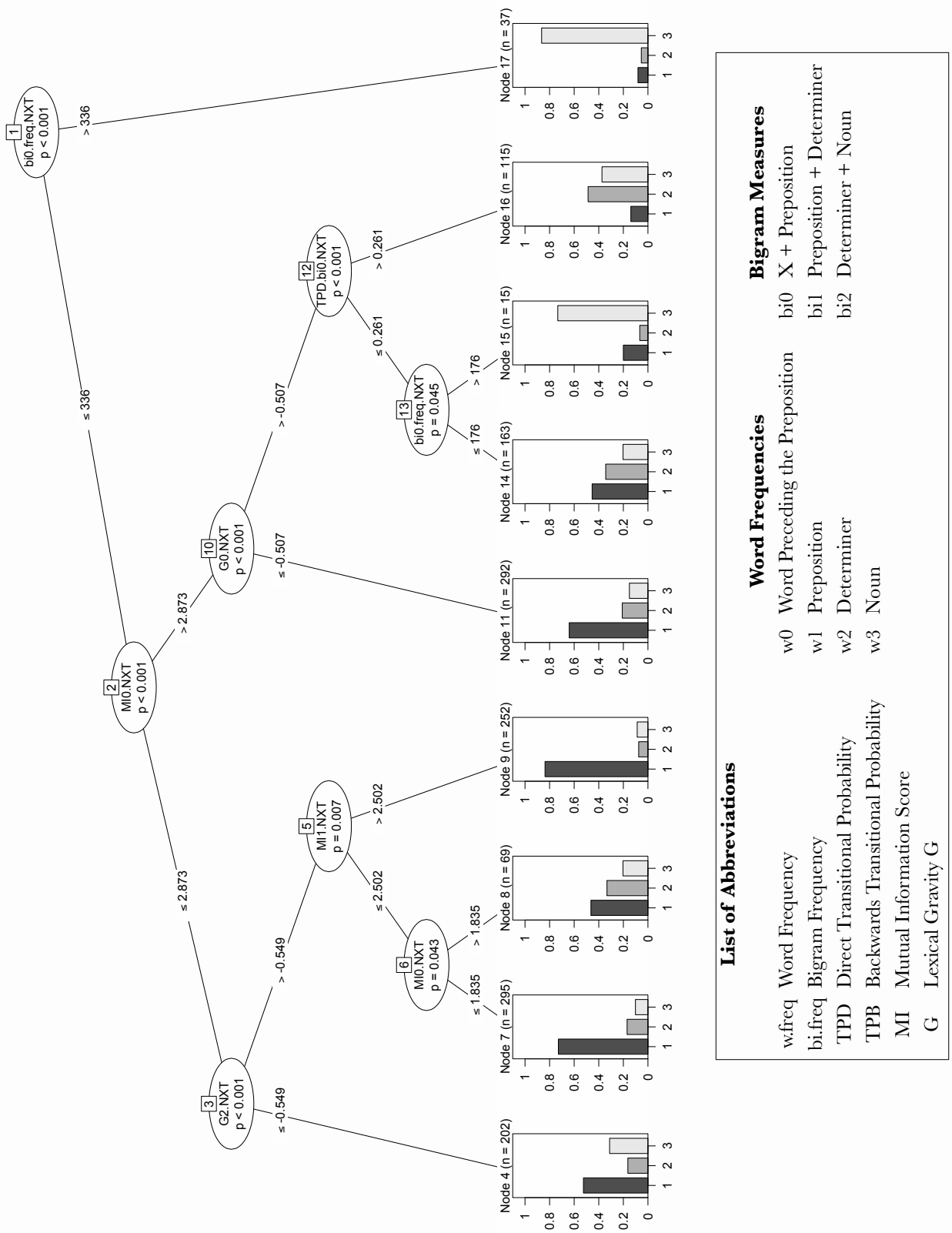


Figure 4.4: Ctree results for the structure 'Preposition Determiner Noun'. Labels at terminal node bar graphs (here: 1, dark; 2, medium and 3, light) indicate hesitation position before the corresponding words (w1=Preposition; w2=Determiner; w3=Noun).

4.4.2 Preposition Determiner Noun

The 'Preposition Determiner Noun' set is the largest (n=1,440)[34] in the present data. As with the last group, sequences of 'Preposition Determiner Noun' which were followed by another noun were excluded as these are subsumed under the structure 'Preposition Determiner Noun Noun' described below. While proper names were technically permitted in all structures, here, the determiner preceding the noun makes proper names extremely rare. (84) to (86) show some exemplary structures from the set.

(84) and you had to solve the mystery during *[pause]* the dinner (sw2627.B.s118)

(85) it would be great to have some of those *[pause]* organizations (sw2558.B.s20)

(86) with the Taurus or something similar in *uh* that regard (sw2336.B.s10)

A *ctree* with the usual predictors is grown on the data. It is highly complex, using eight splits to create nine terminal nodes. Figure 4.4 shows the tree and Table 4.10 lists the predictions of the model. The baseline model would only predict hesitations before the preposition in this case, thereby predicting hesitation placement correctly in 847 cases; this corresponds to a misclassification rate of 41.18%. At a misclassification rate of 35.83% (924 correct predictions), the tree performs highly significantly better than the baseline model ($p<.001$; residuals: 2.65, -3.16), indicating that there are frequency effects in the data.

		Model Predictions			
	Hesitation Position	**pre Prep (1)**	**pre Det (2)**	**pre N (3)**	**Total**
Actual Distribution	**pre Prep (1)**	825	16	6	847
	pre Det (2)	242	56	3	301
	pre N (3)	206	43	43	292
	Total	1,273	115	52	1,440

Table 4.10: Performance of ctree model for 'Preposition Determiner Noun'

Again, the make-up of the terminal nodes confirms my hypotheses. Strongly attracted pairs are more chunked than more weakly attracted or repellent word-pairs in the sense that the cohesive pairs are less likely to be interrupted by hesitations. Node 9 in particular exemplifies this. In this node, the preposition and the word preceding it do not strongly attract each other indicated by a low bigram frequency (bi0.freq in Split 1) and a low mutual information score (MI0 in Split 2); on the other hand, both phrase-internal

[34] Of originally 1,441 data-points, one had to be deleted, because of missing values, which varimp cannot handle.

bigrams show stronger cohesion than in the neighbouring nodes, reflected by the decisions in Splits 3 and 5. As a result, I expect hesitations to be placed at the prepositional phrase boundary rather than within the phrase and this is indeed what happens. Only 16.27% of hesitations are placed within the phrase, making this a very homogenous node.

Moreover, splits show that relations across the prepositional phrase boundary are crucial determinants of hesitation placement in these phrases. The first split indicates that once the 'X+Preposition' sequence is used with a certain currency, no other factors come into play and hesitations are almost exclusively placed before the noun. Sceptics might argue, though, that this owes to the fact that, at only 37 tokens, the resulting Node 17 is rather small and is therefore difficult to split into further parts as the minimum number of tokens per terminal node is seven. A closer look at Node 17, however, reveals that it is already linguistically highly homogeneous, consisting mostly of cases where the 'X+Preposition' bigram consists of a quantifying expression followed by *of*, such as in (85) above.

Splits 2, 10 and 12 (leading up to Node 16) further indicate that even if the word-pair bridging the prepositional phrase boundary is not frequent but instead strongly attracted, hesitations within it are dispreferred. In fact, a comparison of the two branches of the tree resulting from Split 2 reveals that relations within the prepositional phrase come into play only if the pair embracing the prepositional phrase boundary is of low frequency and low mutual attraction, which cements the important role of the prepositional phrase boundary for hesitation placement, even in phrases beyond two words in length.

Another leaf of interest is Node 9, where the weak attractions in the 'X+Preposition' bigram are often brought about by repetitions and self-corrections; their hiatus is between the preposition and the word preceding it, such as in (87), or by conjunctions preceding the phrase, as shown in (88). There are also a number of interesting cases in this node where the preposition is preceded by a particle or another preposition; (89) and (90) illustrate these cases.

(87) actually I think of it as a *[pause]* as a car (sw4728.A.s22)

(88) and *uh [pause]* in the movie (sw2160.A.s153)

(89) you say that you grew up *uh* in the sixties (sw2366.A.s14)

(90) in fact they were all in *uh* over the weekend for Easter (sw4785.A.s119; here the context indicates that the speaker means to say that everybody came home over the Easter weekend.)

Finally, Nodes 15 and 17 show very homogenous behaviour. They both consist to about 80% of tokens of hesitations occurring before the noun. Both clusters are dominated by structures beginning with 'Quantifier+*of*', such as (91).

(91) we get <u>all of</u> this *[pause]* information (sw2028.A.s215)

Next, a random forest is grown, which somewhat over-optimistically reduces the misclassification rate to 28.13% (see Table 4.11). This result, of course, represents a highly significant improvement over the baseline model ($p<.001$; residuals: 6.46, -7.72). The more conservative result based on the out-of-bag predictions (see Table 4.12) likewise highly significantly exceeds that of the baseline model (misclassification rate: 35.56%; $p<.001$; residuals: 2.78, -3.33). This shows there clearly is a frequency effect in the data.

		Model Predictions			
	Hesitation Position	**pre Prep (1)**	**pre Det (2)**	**pre N (3)**	**Total**
Actual Distribution	**pre Prep (1)**	826	10	11	847
	pre Det (2)	186	109	6	301
	pre N (3)	177	15	100	292
	Total	1,189	134	117	1,440

Table 4.11: Performance of cforest model for 'Preposition Determiner Noun', ntree=3,000, mtry=5, seed=1,282

		Model Predictions			
	Hesitation Position	**pre Prep (1)**	**pre Det (2)**	**pre N (3)**	**Total**
Actual Distribution	**pre Prep (1)**	803	23	21	847
	pre Det (2)	226	58	17	301
	pre N (3)	199	26	67	292
	Total	1,228	107	105	1,440

Table 4.12: Performance of cforest out-of-bag predictions for 'Preposition Determiner Noun', ntree=3,000, mtry=5, seed=1,282

The variable importance scores (Figure 4.5) confirm the high predictive value of all factors relating to the prepositional phrase boundary. The scores confirm the tree's assessment that the mutual information score and bigram frequency of the 'X + Preposition' pair, are successful predictors of hesitation placement. Yet the diagram reveals that direct transitional probability is actually more predictive than suggested by the *ctree*. Surprisingly, some word frequencies which were never chosen as a splitting criterion in the individual tree are rated highly predictive.

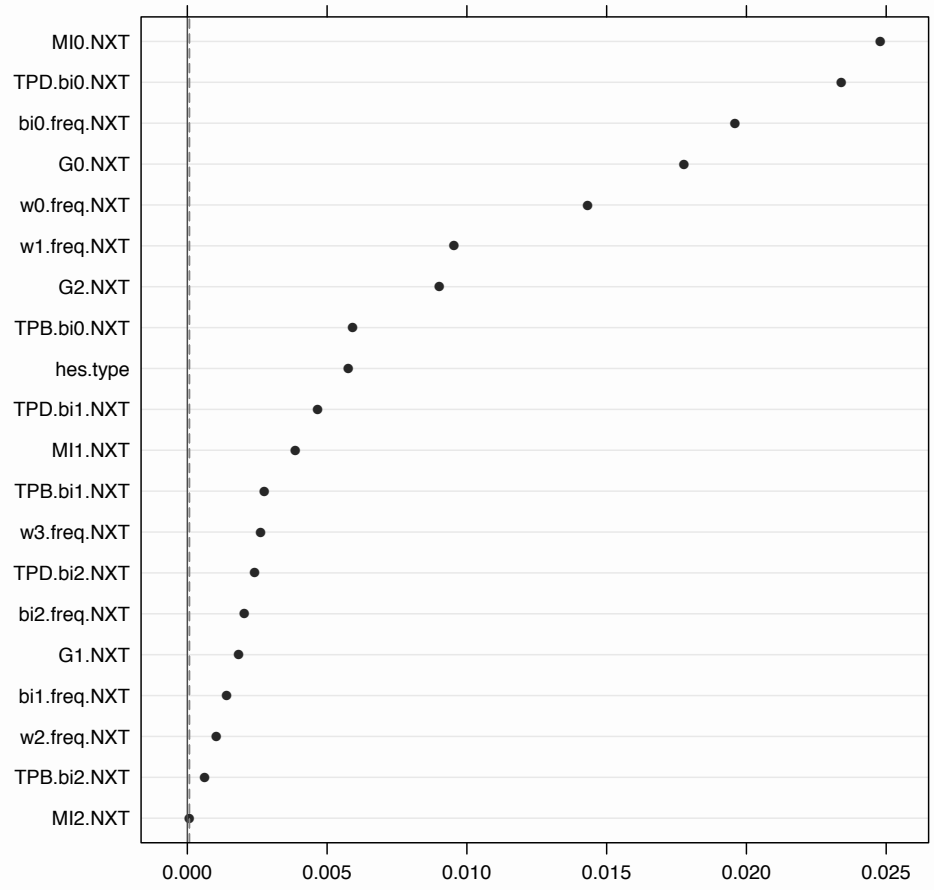


Figure 4.5: Variable importance of predictors for 'Preposition Determiner Noun' (mtry=5, ntree=3,000, seed=1,282, OOB=false, results from R version 2.13.1).

4.4.3 Preposition Noun Noun

The 'Preposition Noun Noun' subset is characterised by proper names. Around 200 of the 553 tokens contain at least one proper name as one of the nouns[35]. In most of these cases, both nouns form a composite proper name. Names of people, geographical locations (like cities, streets and lakes), institutions and sports clubs are among the most frequent. There are six characteristic ways in which these names can be built (in the POS tagging, both parts of such compositional proper names receive a proper name tag):

a. Sequences of two proper names, e.g. *Robin Williams, Saudi Arabia, Susan Sarandon*
b. A city followed by its state, e.g. *Scottsdale, Arizona; Boise, Idaho; Omaha, Nebraska*
c. A proper name followed or preceded by a geographical feature, e.g. *Lake Bonham, Kessler Park, Cape Hatteras*
d. A proper name followed or preceded by a noun or adjective, e.g. *Pink Floyd, New Hampshire, Camp Goddard, Wycliff Bible, Purdue University*
e. Compounds of two common nouns, e.g. *Central Park, American Express, White Rock, South Bend*
f. Abbreviations, e.g. *LA, UC*

The *ctree* model fitted to this data does not perform as well as the previous ones. At 346 correct classifications and a misclassification rate of 37.43%, it first appears to significantly outperform the baseline model (316 correct classifications, misclassification rate: 42.86%) at the $p<.01$ level. However, the low residuals of 1.69 and -1.95 suggest that significance is not reached.

The tree, as shown in Figure 4.6, partitions the data six times, creating seven terminal nodes. The initial and most important split is made based on the predictor 'hesitation type' and Splits 6 and 9 are based on word frequencies, which are not linked to multi-word frequency effects. Consequently, even if overall model performance had reached significance, it would not be indicative of chunking, but of other effects.

[35] Due to a scripting bug in the preparatory stages of the analysis, the letter before an <a> was deleted in all titles (which were printed in capitals in the original version of the corpus), leading to forms such as *aakenings* ('AWAKENINGS'), which were then counted as hapax legomena. While this affects only eight data-points in all other data sets combined, it affects ten data-points in this set.

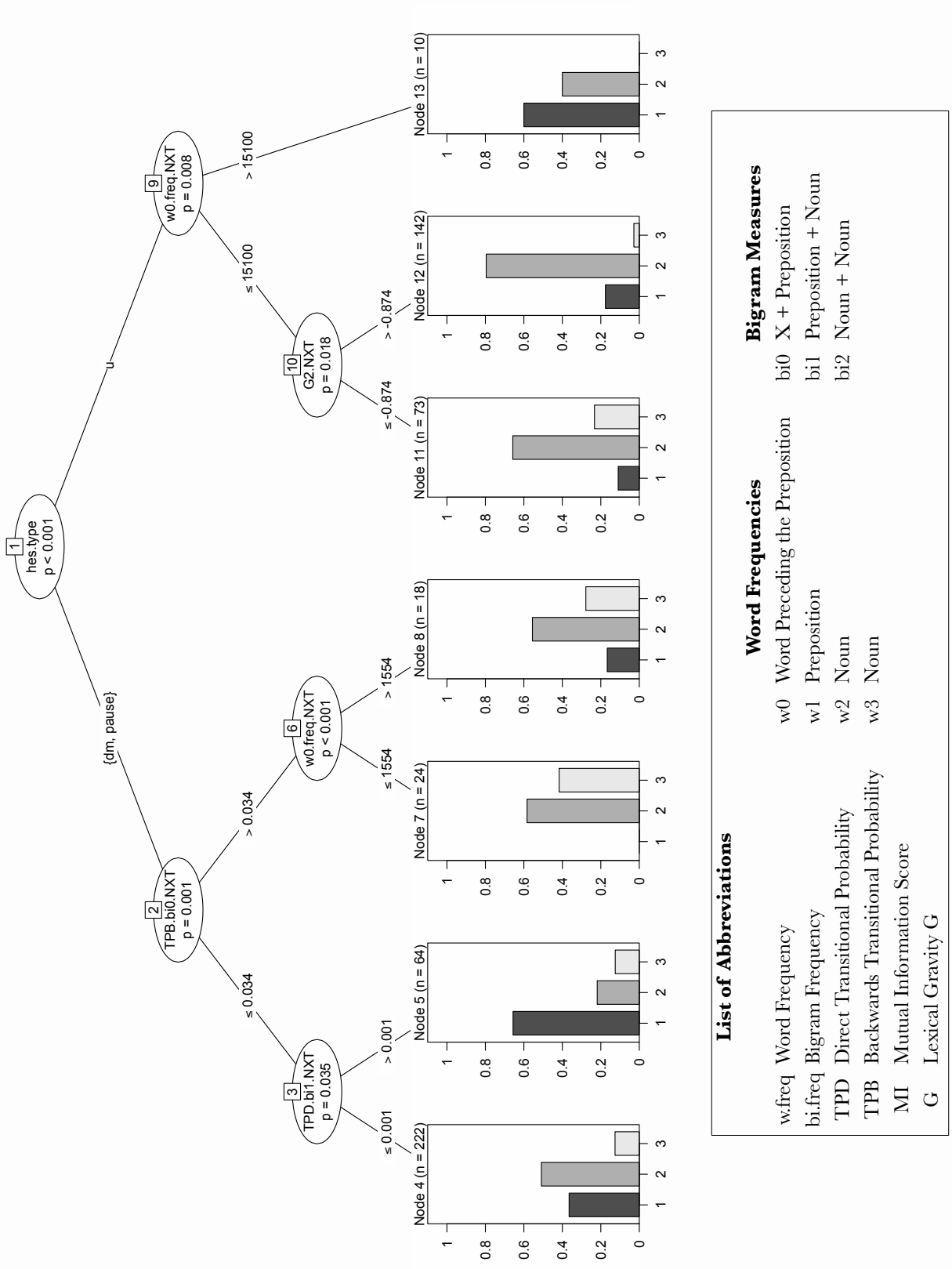


Figure 4.6: Ctree results for the structure 'Preposition Noun Noun'. Labels at terminal node bar graphs (here: 1, dark; 2, medium and 3, light) indicate hesitation position before the corresponding words (w1=Preposition; w2=Noun1; w3=Noun2).

		Model Predictions			
	Hesitation Position	**pre Prep (1)**	**pre N (2)**	**pre N (3)**	**Total**
Actual Distribution	**pre Prep (1)**	48	117	0	165
	pre N (2)	18	298	0	316
	pre N (3)	8	64	0	72
	Total	74	479	0	553

Table 4.13: Performance of ctree model for 'Preposition Noun Noun'

While most terminal leaves are very heterogeneous due to the poor performance, Node 12 is highly homogenous. The misclassification rate in Node 12 is only 20.42%, i.e. far below average. In this node, the vast majority of hesitations is placed before the first noun and very few are placed before the second noun. In neighbouring Node 11, the proportion of hesitations interrupting the 'Noun + Noun' sequence is a lot higher (compare the bars at Position 3 in Nodes 11 and 12). Split 10 explains that this effect a) is caused by different attractions between the two nouns and b) is in the predicted direction. The stronger the attraction between the two nouns – here indicated by the lexical gravity G (G2) value – the fewer hesitations occur between them. Interestingly, most of the two-word names fall into the higher-attraction category. Thus while Node 11 contains very few data-points in which the 'Noun + Noun' pair is made up of two-word names, these make up just over 50% of the tokens in Node 12.

I will now move on to Nodes 7 and 13. From a technical point of view, these nodes are not very interesting because they are neither particularly large nor homogenous, i.e. the dominant outcome only makes up around 60% of the data-points within these nodes. From a linguistic point of view, however, these are highly noteworthy due to the fact that they only contain two out of three possible outcome types.

In Node 7, no hesitations are placed before the preposition, which is of low frequency and strongly attracted to the word preceding it. This is indicated by a high backwards transitional probability (Split 2) and low word frequency (Split 6). In almost half of the tokens in this node, the preposition is part of a prepositional object, namely *talking about* and some form of GO + *to*; *[as] far as* is also prominent in this node.

In Node 13, on the other hand, no hesitations occur before the second noun. The group contains principally phrases ending in two-word proper names such as *San Salvador* and *Southeast Asia*.

Next, a *cforest* is grown. At 402 correctly classified tokens (misclassification rate: 27.31%, see Table 4.14), the forest's performance very highly significantly exceeds that of the baseline model ($p<.001$; residuals: 4.84, -5.59). This is, however, due to overfitting indicated by the much poorer performance on out-of-bag data (see Table 4.15). Here, the forest performs worse than the individual tree and consequently does not

significantly perform above baseline either (residuals: 0.73, -0.84). The variable importance scores (see Figure 4.7) again confirm the *ctree* result. What little effect we see, is mostly due not to frequency but to hesitation type.

		Model Predictions			
	Hesitation Position	**pre Prep (1)**	**pre N (2)**	**pre N (3)**	**Total**
Actual Distribution	**pre Prep (1)**	100	65	0	165
	pre N (2)	14	302	0	316
	pre N (3)	12	60	0	72
	Total	126	427	0	553

Table 4.14: Performance of cforest model for 'Preposition Noun Noun', ntree=3,000, mtry=5, seed=134

		Model Predictions			
	Hesitation Position	**pre Prep (1)**	**pre N (2)**	**pre N (3)**	**Total**
Actual Distribution	**pre Prep (1)**	53	112	0	165
	pre N (2)	40	276	0	316
	pre N (3)	16	56	0	72
	Total	109	444	0	553

Table 4.15: Performance of cforest out-of-bag predictions for 'Preposition Noun Noun', ntree=3,000, mtry=5, seed=134

Both, the individual tree and the forest fail to determine conditions under which hesitations are predominantly placed before the second noun. Such cases are rare, but do, in fact, constitute 13% of the 'Preposition Noun Noun' dataset. Instead, both models hugely overestimate the number of hesitations occurring before the first noun, predicting between 77.2% and 86.6% of hesitations to be placed in this position, while only 57.1% are.

The analysis of this dataset presumably suffers from the fact that it contains so many complex proper names which are often highly specific. Frequency effects for such compounds are not measurable across speakers, because they are highly dependent on a person's environment. One speaker's employer and hometown are presumably chunked for him or her, but may be completely novel terms to somebody else. Such ideolectal variation cannot be grasped by a corpus analysis.

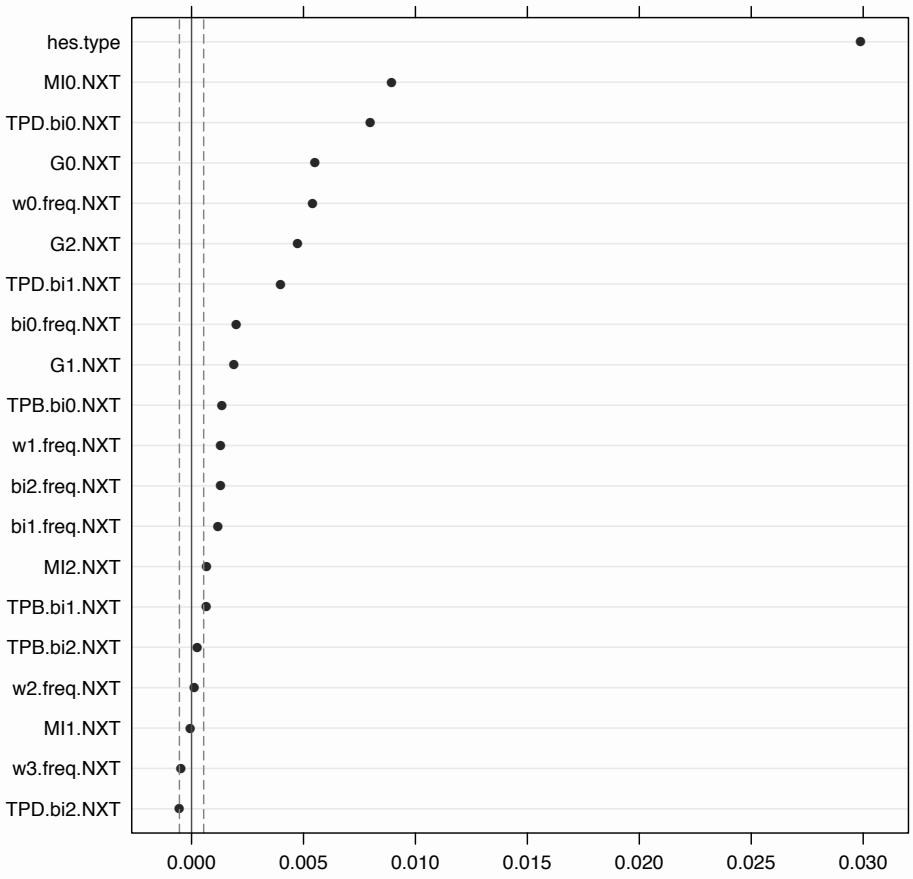


Figure 4.7: Variable importance of predictors for 'Preposition Noun Noun' (mtry=5, ntree=3,000, seed=134, OOB=false, results from R version 2.13.1).

4.4.4 Preposition Determiner Noun Noun

Despite the fact that the 'Preposition Determiner Noun Noun' structure also permits complex nouns, proper names are not as much of a problem here as they were for the analysis of the previous structure. Only about one fifth of the tokens in this set contain a proper name within the phrase (i.e. excluding cases where the proper name precedes the prepositional phrase). Also, names are not as specific as those in 'Preposition Noun Noun'. Here, they often relate to current affairs, such as *United States, Soviet Union* and *Persian Gulf,* or to well-known brands, such as *Honda* and *Audi,* which are less likely to differ substantially in different speakers' use. Nevertheless, expressions in this set are complex and often rare, such as in (92) and (93).

(92) they talk about the genealogy *[pause]* of the family tree (sw3825.A.s64)

(93) had always *you know* been raised on this *you know* emancipation proclamation (sw2253.B.s71)

Out of 500 tokens originally, six had to be deleted, because they contained missing values, which *varimp* cannot handle. The baseline model for this structure performs extremely poorly, classifying only 183 tokens correctly (misclassification rate: 62.96%). A *ctree*'s performance (see Figure 4.8 and Table 4.16) exceeds this rate. It classifies 218 tokens correctly (misclassification rate: 55.87%), which the chi-square test classifies as a significant improvement ($p<.01$). Yet, the residuals of 2.59 and -1.98 reveal that the difference is only marginal.

		Model Predictions				
	Hesitation Position	**pre Prep (1)**	**pre Det (2)**	**pre N (3)**	**pre N (4)**	**Total**
Actual Distribution	**pre Prep (1)**	129	0	13	0	142
	pre Det (2)	55	0	52	0	107
	pre N (3)	94	0	89	0	183
	pre N (4)	32	0	30	0	62
	Total	310	0	184	0	494

Table 4.16: Performance of ctree model for 'Preposition Determiner Noun Noun'

The tree is rather simple with only three splits and four terminal nodes and fails to predict hesitations in two of the possible positions. In line with all previous trees and forests, it emphasises the predictive value of measures relating to the word-pair bridging the prepositional phrase boundary (Bigram 0), particularly its mutual information score (see Split 1). In accordance with previous trees which included a split based on hesitation type, this tree separates filled pauses (hesitation type 'u') from the other hesitations.

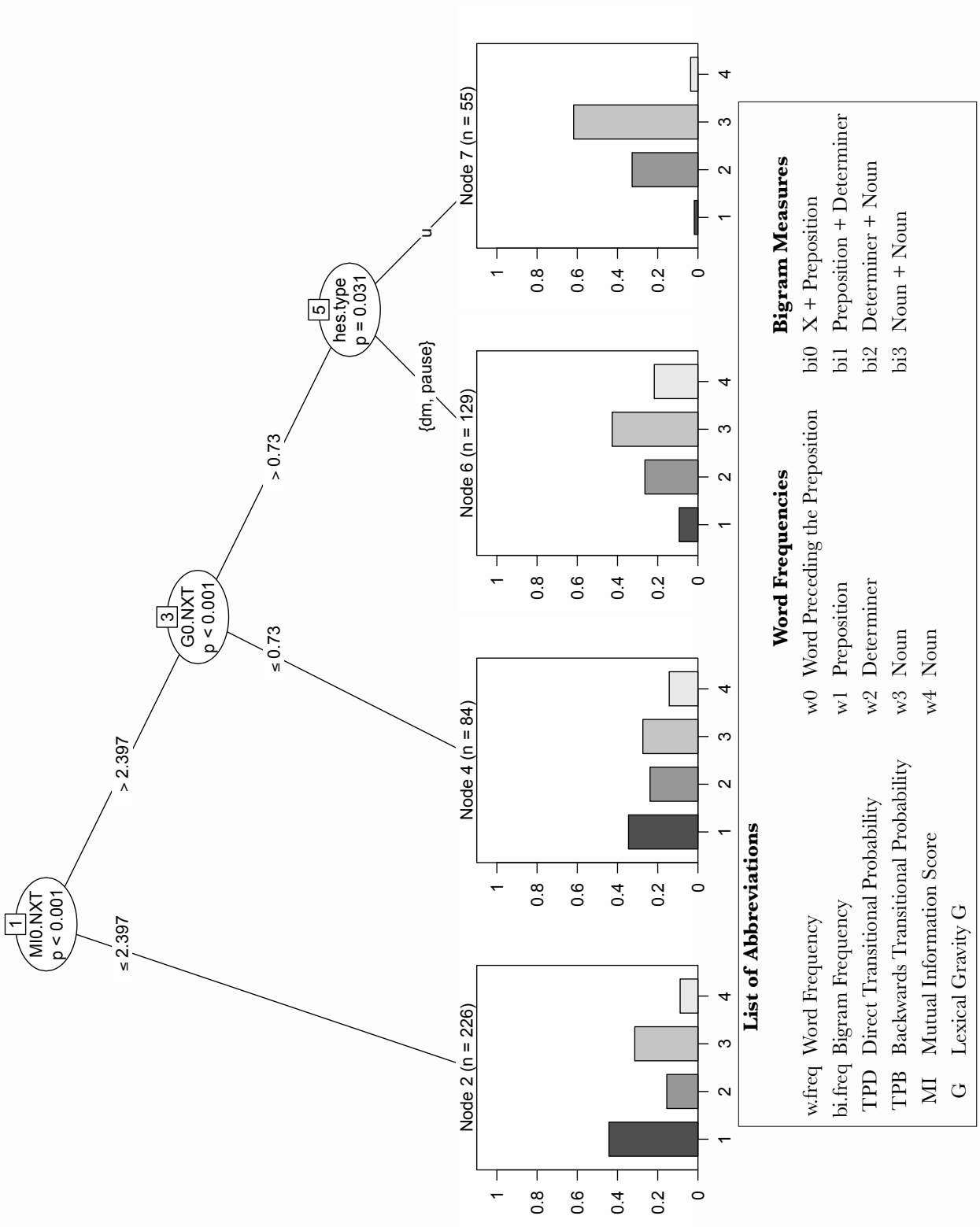


Figure 4.8: Ctree results for the structure 'Preposition Determiner Noun Noun'. Labels at terminal node bar graphs (here: 1, 2, 3 and 4) indicate hesitation position before the corresponding words (w1=Preposition; w2=Determiner; w3=Noun1; w4=Noun2).

As a result of the tree's poor performance, leaves are very noisy. The only leaf of interest is Node 7, which contains almost no hesitations before the preposition or the second noun. In this node, the pair bridging the prepositional phrase boundary is characterised by strong internal cohesion, evidenced by a high mutual information score (Split 1) and a comparatively high lexical gravity G (Split 3). This cohesion is displayed by a type of structure which has already been shown in other trees to be strongly cohesive and unlikely to be interrupted by hesitations, namely 'Quantifier+*of*'. This node further contains many cases where the prepositional phrase is, in fact, a prepositional object following a verb, such as in (94) and (95).

(94) *uh* I want to talk about *um* our family budget (sw2782.A.s5)

(95) they've got to stop worrying about *uh [pause]* the *uh* religious *[pause] uh* overtones (sw2383.B.s34)

Nodes 6 and 7 both show that, as expected, hesitations are not placed before the preposition if it forms a cohesive unit with the word preceding it and furthermore that, in these cases, hesitations are predominantly shifted to the position before the first noun. This tendency seems to be particularly strong for filled pauses (i.e. hesitation type 'u', see Split 5).

In addition to the individual *ctree*, a *cforest* is fitted. Forest results are again over-optimistic (misclassification rate: 34.62%, see Table 4.17), suggesting a very highly significantly better performance than the baseline model ($p<.001$; residuals: 10.35, -7.94). The out-of-bag predictions, however, confirm the results of the individual tree, predicting the exact same number of 218 tokens correctly – though in a different distribution (compare Tables 4.16 and 4.18).

The variable importance scores (see Figure 4.9) confirm the high predictive value of the three predictors chosen by *ctree*. Additionally, the frequency of the word preceding the preposition is ranked very highly.

		Model Predictions				
	Hesitation Position	**pre Prep (1)**	**pre Det (2)**	**pre N (3)**	**pre N (4)**	**Total**
Actual Distribution	**pre Prep (1)**	120	2	20	0	142
	pre Det (2)	22	44	41	0	107
	pre N (3)	27	2	154	0	183
	pre N (4)	16	4	37	5	62
	Total	185	52	252	5	494

Table 4.17: Performance of cforest model for 'Preposition Determiner Noun Noun', ntree=3,000, mtry=5, seed=561

		Model Predictions				
	Hesitation Position	**pre Prep (1)**	**pre Det (2)**	**pre N (3)**	**pre N (4)**	**Total**
Actual Distribution	**pre Prep (1)**	85	9	48	0	142
	pre Det (2)	31	8	68	0	107
	pre N (3)	44	15	124	0	183
	pre N (4)	16	3	42	1	62
	Total	176	35	282	1	494

Table 4.18: Performance of cforest out-of-bag predictions for 'Preposition Determiner Noun Noun', ntree=3,000, mtry=5, seed=561

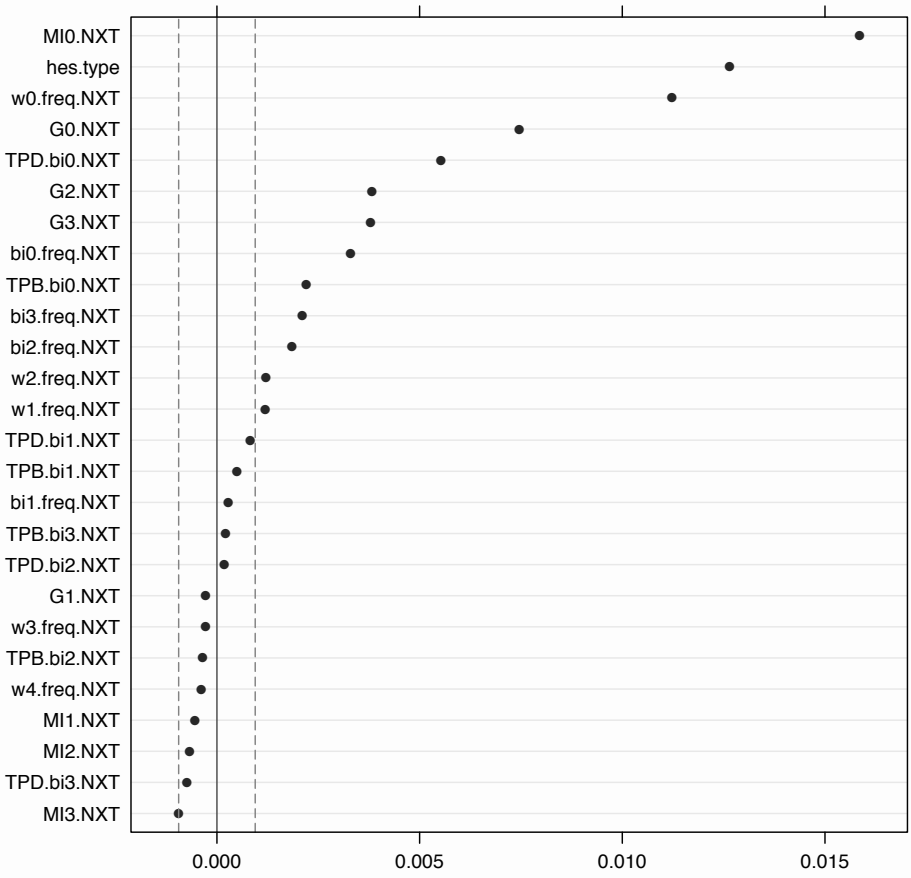


Figure 4.9: Variable importance of predictors for 'Preposition Determiner Noun Noun' (mtry=5, ntree=3,000, seed=561, OOB=false, results from R version 2.13.1).

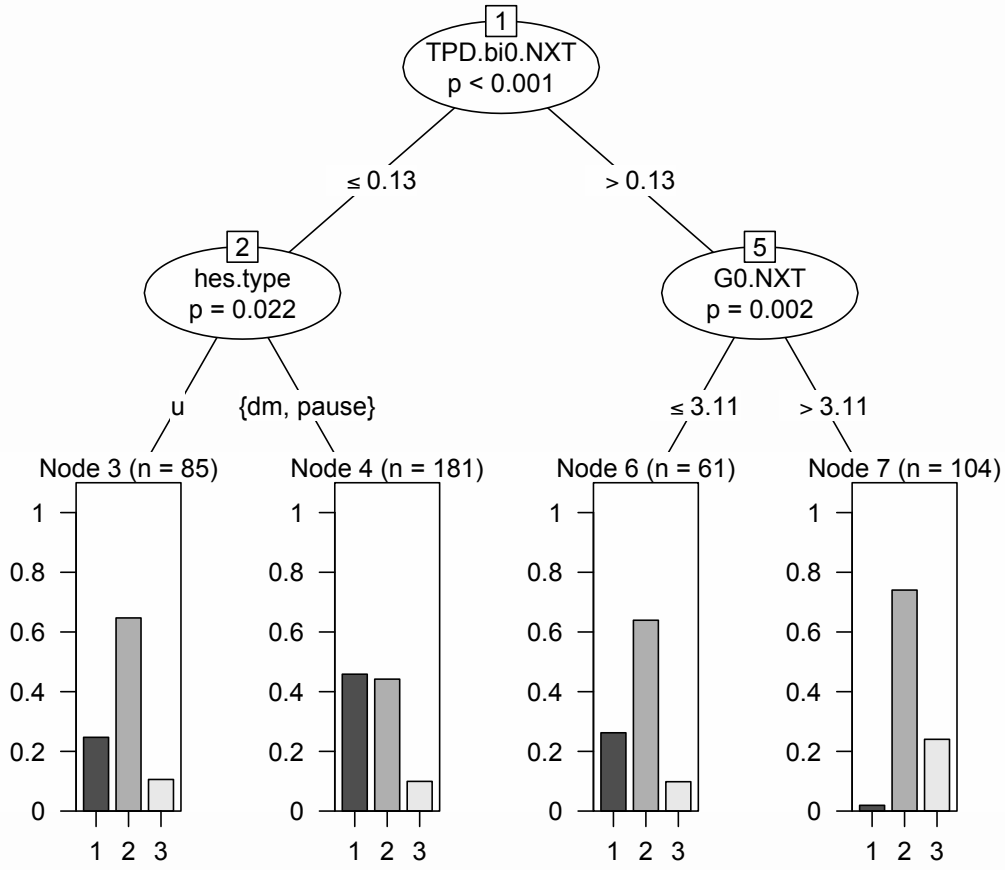


Figure 4.10: Ctree results for the structure 'Preposition Adjective Noun'. Labels at terminal node bar graphs (here: 1, dark; 2, medium and 3, light) indicate hesitation position before the corresponding words (w1=Preposition; w2=Adjective; w3=Noun).

List of Abbreviations

		Word Frequencies		**Bigram Measures**	
w.freq	Word Frequency	w0	Word Preceding the Preposition	bi0	X + Preposition
bi.freq	Bigram Frequency	w1	Preposition	bi1	Preposition + Adjective
TPD	Direct Transitional Probability	w2	Adjective	bi2	Adjective + Noun
TPB	Backwards Transitional Probability	w3	Noun		
MI	Mutual Information Score				
G	Lexical Gravity G				

4.4.5 Preposition Adjective Noun

This set has 432 tokens of which one had to be deleted due to missing values. The dataset is not characterised by one specific type of structure and contains extremely few proper names. (96) to (98) show some exemplary structures from the set.

(96) I'm even *[pause]* in worse shape (sw4626.B.s25)

(97) there's a couple of *uh [pause]* tall buildings (sw2938.B.s14)

(98) ours doesn't start until *uh [pause]* next week (sw2334.A.s245)

		Model Predictions			
	Hesitation Position	**pre Prep (1)**	**pre Adj (2)**	**pre N (3)**	**Total**
Actual Distribution	**pre Prep (1)**	83	39	0	122
	pre Adj (2)	80	171	0	251
	pre N (3)	18	40	0	58
	Total	181	250	0	431

Table 4.19: Performance of ctree model for 'Preposition Adjective Noun'

Of the remaining 431 data-points, the baseline model classifies 251 tokens correctly, corresponding to a misclassification rate of 41.76%. The *ctree* model, shown in Figure 4.10, hardly exceeds this, predicting only 254 data-points correctly (see Table 4.19). This increase of three is, of course, not a significant improvement. Performance is so poor because the predominant outcome in three out of all four terminal leaves is hesitation placement before the adjective. The tree thus fails to find conditions under which hesitations are predominantly placed before the noun. Additionally, in Node 4, the only node with a different prediction, placement before the preposition is just marginally more common than placement before the adjective, leading to considerable noise.

Only the last of the four terminal nodes (i.e. Node 7) in this small tree is noteworthy, because it achieves a misclassification rate of only 25.96% and contains only two hesitations before the determiner. Like many of the homogenous nodes in other trees, this one is characterised by a strong cohesion between the preposition and the word preceding it, evidenced by a high direct transitional probability (Split 1) and a high lexical gravity G (Split 5). Thus it again evidences the important role of the phrase boundary and the fact that effects are in the predicted direction; the higher the association between words, the less likely hesitations between them (compare hesitation rates before the preposition (Position 1) in Nodes 6 (low G) and 7 (high G)). Furthermore, this no-hesitation condition is again provided by 'Quantifier+*of*' expressions and hedges such as *kind of* and *sorts of.*

(99) I refinished a couple of *[pause]* old *uh* dressers (sw4936.B.s3)

(100) they tried to put me in some kind of *[pause]* immobilized walker (sw2433.A.s303)

A *cforest* performs better than the *ctree*, predicting 306 tokens correctly (misclassification rate: 29%), which very highly significantly exceeds the performance of the baseline model ($p<.001$; residuals: 3.47, -4.1). The out-of-bag control set, however, performs even worse than the baseline model. Its misclassification rate does not exceed 42.69% (247 correct predictions). This high uncertainty is also visible in the many negative values among the variable importance scores (see Figure 4.11), causing most of the predictors to be classified as non-significant. The three predictors chosen by *ctree* are among the significant ones. Yet lexical gravity G is again outperformed by the mutual information score. Overall, all measures relating to word-pairs within the prepositional phrase (Bigrams 1 and 2) are classified as non-significant and only measures relating to the pair bracketing the prepositional phrase boundary (Bigram 0) are considered to have significant predictive value.

		Model Predictions			
	Hesitation Position	**pre Prep (1)**	**pre Adj (2)**	**pre N (3)**	**Total**
Actual Distribution	**pre Prep (1)**	63	58	1	122
	pre Adj (2)	9	242	0	251
	pre N (3)	3	54	1	58
	Total	75	354	2	431

Table 4.20: Performance of cforest model for 'Preposition Adjective Noun', ntree=3,000, mtry=5, seed=1,069

		Model Predictions			
	Hesitation Position	**pre Prep (1)**	**pre Adj (2)**	**pre N (3)**	**Total**
Actual Distribution	**pre Prep (1)**	21	100	1	122
	pre Adj (2)	25	226	0	251
	pre N (3)	3	55	0	58
	Total	49	381	1	431

Table 4.21: Performance of cforest out-of-bag predictions for 'Preposition Adjective Noun', ntree=3,000, mtry=5, seed=1,069

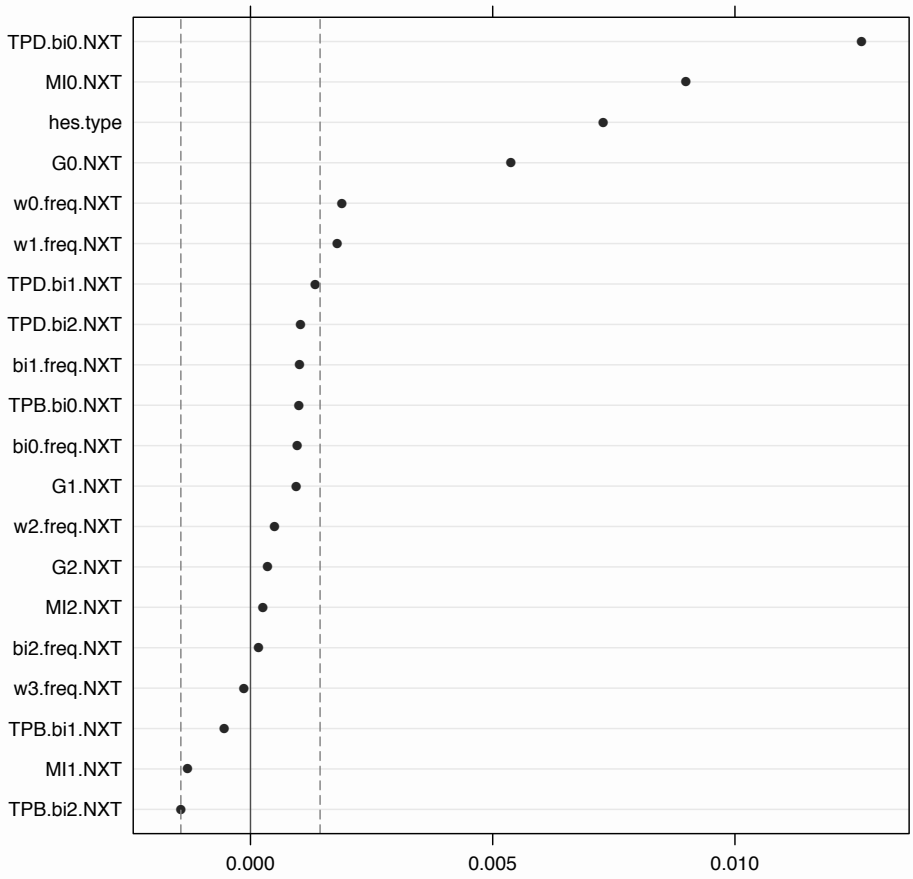


Figure 4.11: Variable importance of predictors for 'Preposition Adjective Noun' (mtry=5, ntree=3,000, seed=1,069, OOB=false, results from R version 2.13.1).

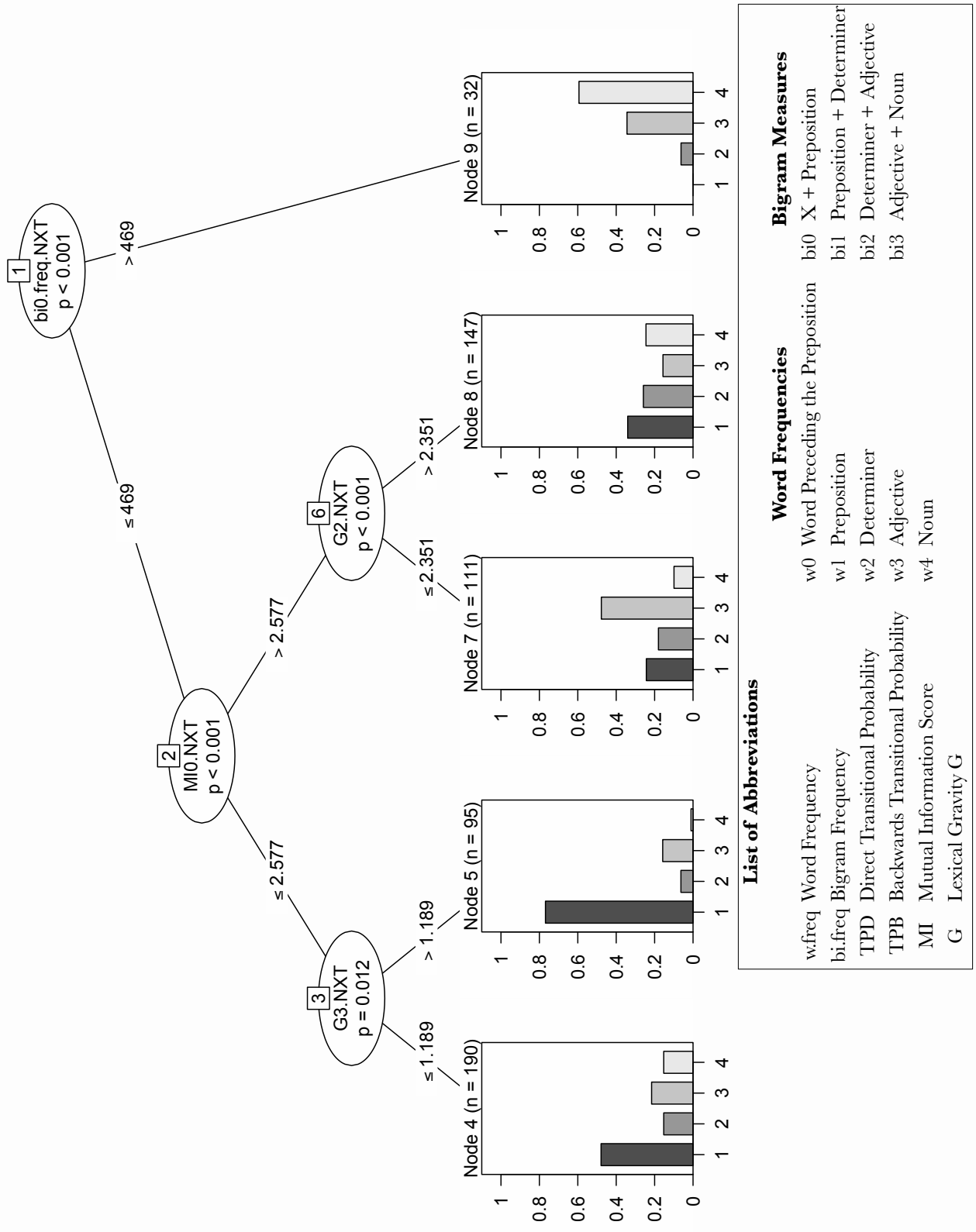


Figure 4.12: Ctree results for the structure 'Preposition Determiner Adjective Noun'. Labels at terminal node bar graphs (here: 1, 2, 3 and 4) indicate hesitation position before the corresponding words (w1=Preposition; w2=Determiner; w3=Adjective; w4=Noun).

4.4.6 Preposition Determiner Adjective Noun

The last of the prepositional phrase sets is 'Preposition Determiner Adjective Noun', which is structurally closest to 'Preposition Adjective Noun'. The outcome of the *ctree* model for this structure can be seen in Figure 4.12 and Table 4.22. Out of a total of 575 data points, the tree predicts 286 correctly, corresponding to a misclassification rate of 50.26%. In this case, the baseline model predicts 241 outcomes correctly (misclassification rate: 58.09%). Hence the *ctree* model preforms very highly significantly better than the baseline model ($p<.001$; residuals: 2.9, -2.46). Table 4.22 shows that the model overestimates the likelihood of a hesitation appearing before the preposition. This leads to the fact that terminal leaves for which the model predicts hesitations before the preposition (Position 1) are generally noisy.

		Model Predictions				
	Hesitation Position	**pre Prep (1)**	**pre Det (2)**	**pre Adj (3)**	**pre N (4)**	**Total**
Actual Distribution	**pre Prep (1)**	214	0	27	0	241
	pre Det (2)	73	0	20	2	95
	pre Adj (3)	79	0	53	11	143
	pre N (4)	66	0	11	19	96
	Total	432	0	111	32	575

Table 4.22: Performance of ctree model for 'Preposition Determiner Adjective Noun'

A comparison of nodes illustrates the direction of effects and the role of the phrase boundary. In Nodes 7 to 9, the words to the left and the right of the prepositional phrase boundary (Bigram 0) either co-occur frequently (Node 9, due to Split 1) or are strongly attracted (Nodes 7 and 8, due to Split 2). If we compare these nodes to Nodes 3 and 4, we see that the former contain far fewer hesitations at the prepositional phrase boundary (Position 1). Thus we can conclude that the more frequent the pair bridging the prepositional phrase boundary or the more attracted the words within it, the less likely it is to be interrupted by hesitations.

Furthermore, comparing Node 4 to Node 5 as well as Node 7 to Node 8 reveals the effects of phrase-internal cohesion. Node 5 displays internal cohesion between the adjective and the noun. In juxtaposition to Node 4, we see that there are far fewer hesitations at Position 4 in Node 5 than in Node 4; the attraction between the adjective and the noun means that the pair is far less likely to be interrupted than its more weakly attracted counterpart in Node 4. Interestingly, hesitations in Node 5 are not pushed to neighbouring transitions, i.e. the lack of hesitations before the noun (Position 4) does not lead to an increase in hesitations before the adjective (compare proportion of hesitations at Position 3 in Nodes 4 and 5). Instead, cohesion between the adjective and the noun

appears to 'push' hesitations right to the prepositional phrase boundary (compare proportion of hesitations at Position 1 in Nodes 4 and 5).

Differences between Nodes 7 and 8 are similar except the attraction between the determiner and the adjective in Node 8 (see Split 6) does not only result in an increase of hesitations at the prepositional phrase boundary, but also in a shift of hesitations to neighbouring positions (compare proportion of hesitations in Positions 2 and 4 in Nodes 7 and 8).

Linguistically, Node 5, which at a misclassification rate of 23.16% is the most homogenous, is characterised by highly frequent conventionalised expressions with strong internal cohesion, such as (101) to (104), which 'push' hesitations out of the prepositional phrase.

(101) in the long (run/term) (5x)
(102) (for/in) a long time (4x)
(103) on the other hand (3x)
(104) for the most part (4x)

Furthermore, Node 9 is of interest, as it contains no hesitations before the preposition. It is almost entirely comprised of phrases where the 'X+Preposition' bigram is *one of* (26 tokens), such as (105) and (106). The remaining tokens are instances of *out of, lot of* and *kind of*.

(105) one of the first *you know* choices (sw2521.B.s143)
(106) one of the top *uh [pause]* people (sw3658.B.s47)

The corresponding random forest predicts 363 of the 575 outcomes correctly, corresponding to a misclassification rate of 36.87%, a very good result compared to the baseline model (misclassification rate= 58.09%). The chi-square test shows that *cforest* performs very highly significantly better than the baseline model ($p<.001$; residuals: 7.86, -6.68). Table 4.23 shows the performance of the forest in more detail. The corresponding out-of-bag predictions are correct in 277 cases, corresponding to a misclassification rate of 51.83%. This is a far more conservative result, though still a significant improvement over the baseline model ($p<.01$, residuals: 2.38, -2.02).

		Model Predictions				
	Hesitation Position	**pre Prep (1)**	**pre Det (2)**	**pre Adj (3)**	**pre N (4)**	**Total**
Actual Distribution	**pre Prep (1)**	232	1	6	2	241
	pre Det (2)	55	21	17	2	95
	pre Adj (3)	57	0	80	6	143
	pre N (4)	47	3	16	30	96
	Total	391	25	119	40	575

Table 4.23: Performance of cforest for 'Preposition Determiner Adjective Noun', ntree=3,000, mtry=5, seed=63

		Model Predictions				
	Hesitation Position	**pre Prep (1)**	**pre Det (2)**	**pre Adj (3)**	**pre N (4)**	**Total**
Actual Distribution	**pre Prep (1)**	211	4	24	2	241
	pre Det (2)	67	3	20	5	95
	pre Adj (3)	80	5	46	12	143
	pre N (4)	55	4	20	17	96
	Total	413	16	110	36	575

Table 4.24: Performance of cforest out-of-bag predictions for 'Preposition Determiner Adjective Noun', ntree=3,000, mtry=5, seed=63

In line with the single tree analysed above, the variable importance ranking in Figure 4.13 shows that measures relating to the word-pair bridging the phrase boundary (Bigram 0) have the strongest effect on hesitation placement. Contrary to what the *ctree* suggests, the mutual information score actually outperforms bigram in this context (i.e. MI0 is the highest-ranking predictor, followed by bi0.freq). Furthermore, the role of some word frequencies is far more influential than suggested by the *ctree* model.

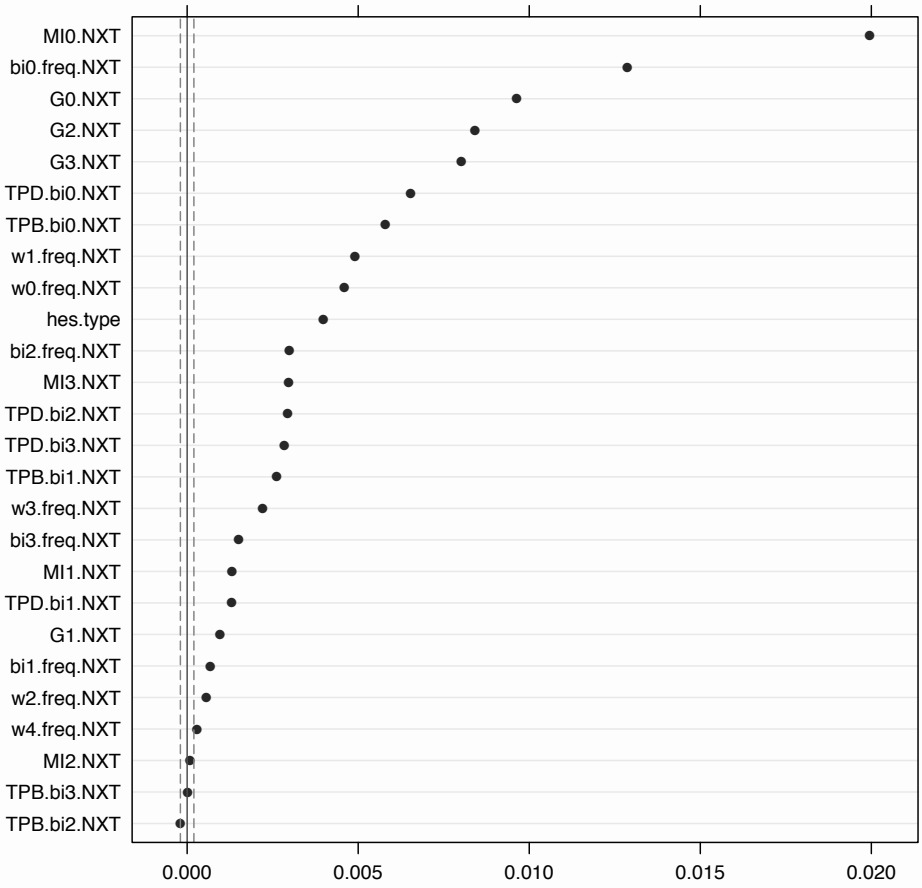


Figure 4.13: Variable importance of predictors for 'Preposition Determiner Adjective Noun' (mtry=5, ntree=3,000, seed=63, OOB=false, results from R version 2.13.1).

4.4.7 Summary

The previous sections provided analyses of hesitation placement in prepositional phrases. A grand total of 4,724 filled and unfilled pauses as well as discourse markers and hesitation clusters, such as *uh [pause] well* were analysed.

On average, speakers preferred to hesitate at the prepositional phrase boundary. Across all datasets, 2,113 hesitations (44.7%) were placed in this position. The second most common point to pause was before the first content word in the phrase; speakers hesitated there in 38.5% of cases. The boundary of the embedded noun phrase, which, of course, often coincides with the position before the first content word, was disfluent in 36.1% of cases. In phrases containing a second content word, the position before the second content word was the least popular to hesitate; only 14% of hesitations uttered in a phrase containing two content words was placed before the second content word.

Discourse markers were generally more likely to be placed at the prepositional phrase boundary than filled or unfilled pauses. Nevertheless, all types of hesitation devices are used at any transition. Thus some discourse markers are even used before the second content word in longer noun phrases.

These distributions were analysed with the help of Classification and Regression Trees (CART trees) and random forests (see Section 3.3.3). These regression tools were provided with information about the frequency of all words and word-pairs in the data as well as with several measures quantifying the associations between the two words in a pair. Analyses were conducted separately for each phrase type.

This summary of results focusses on the overall performance of the regressions models in Sections 4.4.1 to 4.4.6 and on similarities between the trees, their splits and terminal nodes. It excludes an evaluation of the performance of the predictors on which the models were based. This will be provided in detail in Section 4.5.

Tables 4.25 to 4.27 summarise the results for all prepositional phrase datasets. Due to the stepwise procedure, there are three separate results for each structure.

CART Trees – Table 4.25 provides the results of analyses based on CART trees, i.e. a single tree per analysis, a procedure whose use is sometimes cautioned against as it is considered unstable (cf. Strobl, Malley and Tutz 2009a:13). Model performance can be considered significant where *p*-values reach significance and the value of the residuals reaches or exceeds two. The first line in the table reads as follows: A CART tree fitted to the dataset of hesitations occurring in prepositional phrases of the type 'Preposition Noun' classifies 72.5% of data points correctly, corresponding to a misclassification rate of 27.5%; this represents a highly significant improvement when compared to the corresponding baseline model, thus indicating very highly significant frequency effects. These effects are confirmed by the test's residuals; the value of both highly exceeds two,

showing that the number of correct predictions highly significantly exceeds those of the baseline model, whilst the number of misclassifications is highly significantly lower.

In this way, highly significant frequency effects can be observed for hesitation placement in the structures 'Preposition Noun', 'Preposition Determiner Noun' and 'Preposition Determiner Adjective Noun'. In contrast, results for 'Preposition Noun Noun' and 'Preposition Determiner Noun Noun', which initially appear significant (p<. 01), do not stand up to scrutiny, as the residuals barely reach two.

Phrase	**MCR**	**Sig. level**	**Residuals**	
Prep N	27.5%	p<.001	10.2	-10.53
Prep Det N	35.8%	p<.001	2.65	-3.16
Prep N N	37.4%	p<.01	1.69	-1.95
Prep Det N N	55.9%	p<.01	2.59	-1.98
Prep Adj N	41.1%	non-sig.	-	-
Prep Det Adj N	50.3%	p<.001	2.9	-2.46

Table 4.25: Performance of the CART trees. Given are misclassification rates (MCR), p-values based on chi-square tests and the residuals of the chi-square tests.

Random forests – Table 4.26 lists results obtained with random forest analyses based on 3,000 individual trees. These are considered to be more reliable than individual trees (cf. Strobl, Malley and Tutz 2009a:15), yet may bear some risk of overfitting to the data (cf. Strobl, Malley and Tutz 2009a:19). For the present data, forest analyses provide very highly significant results for all six models.

Phrase	**MCR**	**Sig. level**	**Residuals**	
Prep N	17.3%	p<.001	15.2	-15.69
Prep Det N	28.1%	p<.001	6.46	-7.72
Prep N N	27.3%	p<.001	4.84	-5.59
Prep Det N N	34.6%	p<.001	10.35	-7.94
Prep Adj N	29.0%	p<.001	3.47	-4.1
Prep Det Adj N	36.9%	p<.001	7.86	-6.68

Table 4.26: Performance of random forests. Given are misclassification rates (MCR), p-values based on chi-square tests and the residuals of the chi-square tests.

Out-of-bag data – Table 4.27 summarises results based on random forest out-of-bag data, which purportedly provide the most conservative – because cross-validated – results (cf. Strobl, Malley and Tutz 2009a:19). In the present case, out-of-bag results are in line with CART results.

Phrase	**MCR**	**Sig. level**	**Residuals**	
Prep N	30.1%	p<.001	8.93	-9.22
Prep Det N	35.6%	p<.001	2.78	-3.33
Prep N N	40.5%	non-sig.	-	-
Prep Det N N	55.9%	p<.01	2.59	-1.98
Prep Adj N	42.7%	non-sig.	-	-
Prep Det Adj N	51.8%	p<.01	2.38	-2.02

Table 4.27: Performance of the random forests' out-of-bag sets. Given are misclassification rates (MCR), p-values based on chi-square tests and the residuals of the chi-square tests.

In summary, all analyses yield significant results for hesitation placement in the phrase types 'Preposition Noun', 'Preposition Determiner Noun' and 'Preposition Determiner Adjective Noun'. For three structures, there are highly significant influences of the selected measures of association on hesitation placement. Thus, the stronger the association between two words the less likely a hesitation will be placed between them.

The three structures in which no significant effects emerge should not be considered counterexamples. In the case of the phrase type 'Preposition Noun Noun', in particular, it appears that results point to a factor so far neglected in chunking studies, namely idiolectal variation. All speakers in Switchboard were given as much unrecorded time to introduce themselves as they wished, yet in many of the conversations speakers mention hometowns, employers and schools during the recorded time. The phrase type 'Preposition Noun Noun', in particular, is rich in descriptions such as *Sunnyvale, California,* see (107), or *Richardson Symphony,* see (108). These sequences are likely to be very frequent in the speech of some individuals, yet few are frequent in the corpus. Therefore, they are likely to be chunked for the individual speaker who uses them, but far too rare across speakers for their cohesion to be captured by association measures

(107) Hi I'm Nevin from *[pause]* Sunnyvale, California (sw4792.A.s1)

(108) I grew up next door to *uh [pause]* the Richardson Symphony (sw3152.B.s24)

Furthermore, there are some decisions which recur across many *ctree* (CART) analyses.

Firstly, whenever the mutual information score is chosen as a predictor, the selected splitting point lies somewhere in the range between 2.4 and 3.4 – a narrow band considering the mutual information score's full range of -8.43 to 19.57 (mean: 4.68; standard deviation: 4.08). The results of all of these splits are in line with my hypothesis; the higher the MI score of a pair, the less likely speakers are to place hesitations between the words in the pair.

Table 4.28 provides a list of exemplary 'X+Preposition' pairs to illustrate which kinds of pairs are described by these values. The results of the trees indicate that pairs like *seems like* and *extend beyond*, which fall into the high-MI category (third row), on average, are less likely to be interrupted than pairs in the low-MI range (first row), such as *and about* or *available in*. Pairs with an MI between 2.4 and 3.4, shown in the grey-shaded row, are sometimes assigned to the low-MI range and at other times to the high-MI range, depending on the exact splitting points chosen by the *ctrees*.

MI Range	Selected Bigrams
min. – 2.4	*or of* (-6.09); *of about* (-3.58); *and about* (-2.71); *and in* (-1.31); *but with* (-0.98); *and without* (0.03); *and during* (0.09); *house in* (0.95); *go from* (1.01); *bit from* (1.68); *planning for* (1.99); *available in* (2.24)
2.4 – 3.4	*use[Verb] for* (2.44); *sure about* (2.62); *more of* (2.65); *all of* (2.96); *guilty of* (3.05); *teaching in* (3.22); *money for* (3.34)
3.4 – max.	*painted with* (3.61); *one of* (4.10); *participating in* (5.30); *lot of* (5.47); *seems like* (6.72); *deviate from* (8.79); *mandatory upon* (1(

Table 4.28: Word-pairs at exemplary MI score levels.

Secondly, all lexical gravity G values chosen fall into two ranges, one from -0.9 to 0.7 and another from 1.2 to 3.1. The full range of G is -13.03 to 16.37 (mean: -1.96; standard deviation: 2.19). None of the other predictors produce such narrow ranges. All splits according to lexical gravity G indicate that the higher the G score of a pair, the less likely a speaker is to interrupt it to hesitate.

Table 4.29 provides a list of exemplary 'X+Preposition' pairs to illustrate which kinds of pairs are described by these values. For easier comparison of the MI score and lexical gravity G, the selected pairs are the same as in Table 4.28. The two ranges in which splits typically lie are shaded in grey. The splits in the *ctrees* indicate that pairs with a G score of less than -0.9, like *planning for* or *painted with*, tend to be disfluent more often than any of the pairs receiving a higher score. Conversely, pairs rated 3.1 or higher, like *more*

of and *lot of*, tend to have a greater chance of being uttered fluently than pairs receiving a lower G score.

G Range	Selected Bigrams
min. – -0.9	*or of* (-5.29); *planning for* (-3.19); *painted with* (-2.23); *of about* (-1.73); *deviate from* (-1.46)
-0.9 – 0.7	*mandatory upon* (-0.89); *directed towards* (-0.33); *extend beyond* (-0.29); *available in* (-0.14)
0.7 – 1.2	*participating in* (0.78); *bit from* (0.85); *and without* (0.94); *sure about* (1.12)
1.2 – 3.1	*go from* (1.42); *and about* (1.55); *and during* (1.74); *use [Verb] for* (2.18); *teaching in* (2.23); *guilty of* (2.29); *but with* (2.38); *house in* (3.04)
3.1 – max.	*money for* (6.86); *more of* (8.17); *and in* (8.81); *seems like* (9.57); *all of* (10.33); *one of* (13.49); *lot of* (14.31)

Table 4.29: Word-pairs at exemplary G scores

Wherever the predictor 'hesitation type is chosen, it separates the filled pauses from unfilled pauses and discourse markers (with one exception, though; see Split 14 in Figure 4.2). The resulting terminal nodes show that *uh*, *um* and clusters thereof have a significantly greater propensity to occur within the prepositional phrase than unfilled pauses and discourse markers.

Crucially, the predictors most frequently selected and those ranked highest according to their variable importance scores are those relating to the word-pair which brackets the prepositional phrase boundary. Wherever leaves are particularly homogenous, i.e. where around 80% of hesitations are placed at the same position (e.g. Node 9 in Figure 4.4) or where one position is (almost) never chosen (e.g. Node 7 in Figure 4.10), these tend to be created by means of splits based on predictors which describe attractions holding across the prepositional phrase boundary. This indicates that prepositional phrase boundaries are locations where hesitations are particularly good indicators of chunking.

Analyses of the contents of the terminal nodes in the trees indicate that the hedges *kind of* and *sort of,* quantifying expressions, such as *one of, all of* or *lots of*, and a small number of further *of*-collocates, e.g. *out of,* are especially unlikely to be disfluent. Data-points containing these expressions are assigned to separate nodes due to these expressions' high frequency, high lexical gravity G or high direct transitional probability. This means that the words in these pairs are prone to co-occur and unlikely to be interrupted by hesitations, which indicates that they should be strongly chunked. These expressions will therefore be analysed in more detail in Section 4.6.

Conversely, terminal nodes in several trees which were characterised by hesitations occurring at the prepositional phrase boundary contained many cases where the prepositional phrase was preceded by coordinating conjunctions or where the prepositional phrase boundary coincided with the hiatus of a repetition or self-correction. These data-points, being strong attractors of hesitations, will be further analysed in Section 4.7.

4.5 Comparison of Predictors

This section draws on the results from the regression analyses in Section 4.4. It takes a meta approach, combining results from all six analyses with the aim of evaluating how well different measures of association fared at predicting hesitation placement. This information will allow for conclusions later on about the way in which the mind keeps track of co-occurrence patterns in speech and how chunking can best be modelled.

CART trees and random forests do not provide the odds ratios, separate significance values or other measures of effect strength or variance resolution that we are familiar with from more commonly used methods of regression. They each offer their own, particular option, though. In CART trees, how frequently and early (i.e. high up) predictors are chosen can serve as an indicator of their predictive power. The earlier a predictor is chosen, the more it will generally contribute to impurity reduction, because it relates to all or a great proportion of the data. A predictor chosen further down in the tree may bring great impurity reduction for the two nodes it relates to, but these may be so small that the overall effect is a lot weaker. Thus, a predictor chosen early represents a factor of importance to all or most data-points, while a predictor chosen late is only of relevance to a small sub-set of the data. However, as pointed out in other sections, marginally outperformed predictors may never appear in a tree, leading to their performance being underestimated. Therefore, exploiting the means of random forests is a better option.

As we have seen in Figures 4.3, 4.5, 4.7, 4.9, 4.11 and 4.13, random forests provide so-called variable importance scores. The higher the score, the more the predictor contributed towards the prediction accuracy of the model (cf. Strobl, Malley and Tutz 2009b:335). Due to the fact that they quantify predictors' contribution to variance resolution, variable importance scores provide a much more accurate means of predictor comparison than any judgements based on single trees (see Section 3.3.3.2 for more information about the logic of the score and the method of comparison applied here).

Figure 4.14 shows mean variable importance scores and the respective standard errors for all types of predictors. Means were calculated across all phrase types and all types of transitions. In the plot, predictors are shown in order of increasing mean variable importance. On average, backwards transitional probability had the least effect on model performance, followed by word frequencies, bigram (i.e. pair) frequencies, the mutual information score, direct transitional probability, lexical gravity G and the type of hesitation, but differences in mean predictive power are minor. Backwards transitional probability performs significantly worse than the mutual information score ($p<.01$, based on a Wilcoxon rank-sum test), but otherwise there are no significant

differences in performance between the various measures of association strength. Importantly, none of the association measures' performance differs significantly from that of bigram frequency (based on separate Wilcoxon rank-sum tests, here preferred to ANOVA because the data is not normally distributed).

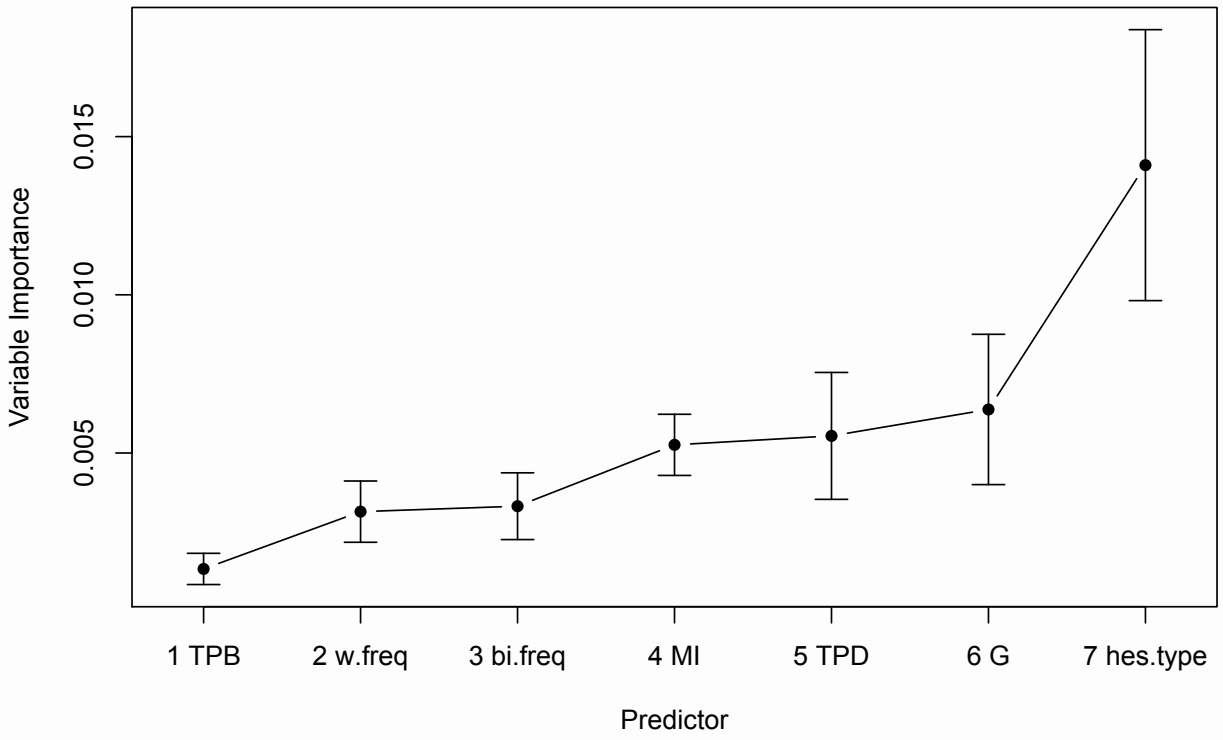


Figure 4.14: Mean variable importance scores by type of predictor. The dot indicates the mean for each group and the error bars show the standard error. TPB = backwards transitional probability; w.freq = word frequency; bi.freq = bigram frequency; MI = Mutual Information score; G = lexical gravity G; hes.type = hesitation type.[36]

There are indications, however, that averaging scores for all transitions, as done in Figure 4.14, may obscure some important effects. The analyses in Section 4.4 above repeatedly indicated that the prepositional phrase boundary plays a particularly important role for hesitation placement. Across all CART trees, measures describing the cohesiveness (or lack thereof) of those word-pairs spanning the prepositional phrase boundary were chosen as splitting criteria noticeably more often than measures relating

[36] The graphs in this section were drawn based on variable importance scores obtained in R Versions 2.15.2 and 3.0.0, while graphs in Section 4.4 were generated based on data obtained in R Version 2.13.1. It turned out that the later versions provide slightly different scores than the old version. Fortunately, due to the large number of trees used, these differences should be negligible (cf. also Strobl, personal communication, 2013).

to any other transition. Additionally, the highest-ranked predictors according to the variable importance scores of random forests always related to the pairs bridging the prepositional phrase boundary. These observations suggest that associations between the words in the pair at the prepositional phrase boundary have a more profound influence on hesitation placement than relations holding within the prepositional phrase and that they should therefore be analysed as separate groups.

Figure 4.15 shows variable importance scores relating to 'X + Preposition' pairs separately from all other scores. Effects of word frequencies and hesitation type are no longer shown. The grey dots in the graph visualise that associations between words within the phrase have minimal influence on hesitation placement in prepositional phrases. All predictor's mean performance in these contexts is rated lower than 0.005.

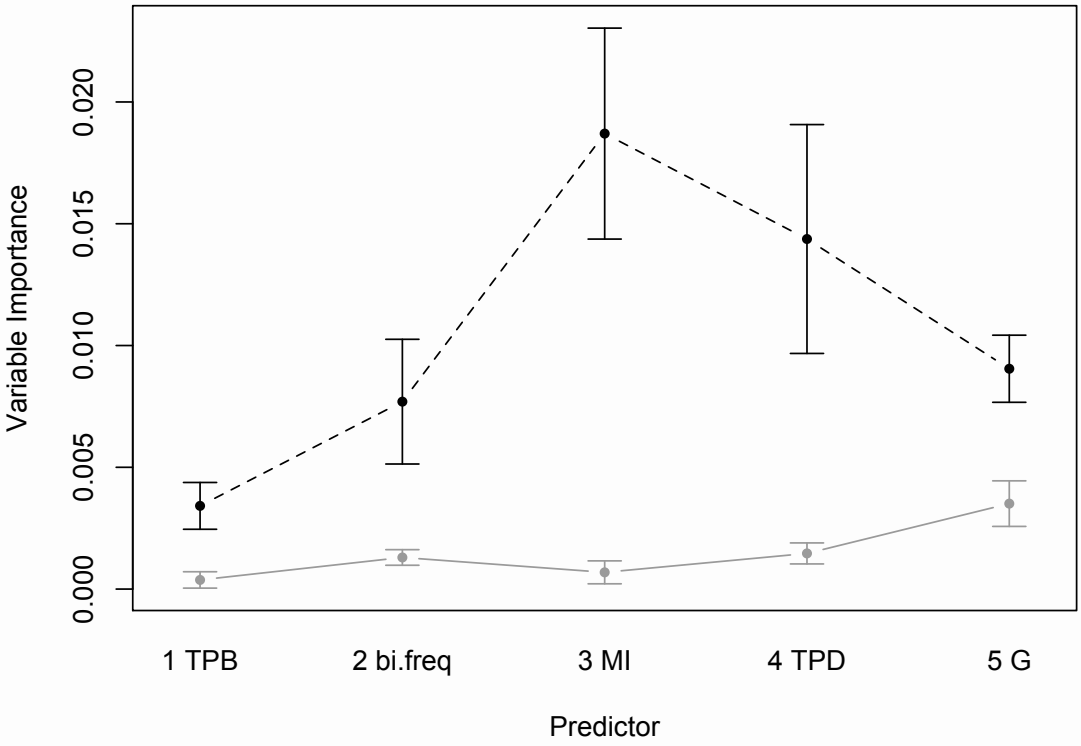


Figure 4.15: Separate mean variable importance scores for prepositional phrase boundary pairs and mid-prepositional phrase pairs. Points connected by the black dashed line show values for those pairs bracketing the prepositional phrase boundary, while points connected by the solid grey line show values for all transitions within the phrase. The dot indicates the mean for each group and the error bars show the standard error. TPB = backwards transitional probability; w.freq = word frequency; bi.freq = bigram frequency; MI = Mutual Information score; G = lexical gravity G; hes.type = hesitation type.

The black dots indicate how much the predictors contributed to the overall performance of the models wherever they described associations holding between the words left and right of the prepositional phrase boundary. Table 4.30 shows that all predictors perform significantly or even highly significantly better when relating to the prepositional phrase boundary than when relating to other word-pairs. These marked differences in performance suggest that the predictor ranking in Figure 4.14 results mainly from predictors' performance when relating to the 'X + Preposition' pair.

Importantly, Figure 4.15 also reveals that absolute co-occurrence frequency and measures of the relative chance of co-occurrence are not equally predictive of hesitation placement. In the case of the the 'X + Preposition' pair, there is an indication that the mutual information score outperforms co-occurrence frequency ($p<.1$, based on a Wilcoxon rank-sum test).

Predictor	**Freq.**	**TPD**	**TPB**	**MI**	**G**
Significance level	$p<.01$	$p<.001$	$p<.01$	$p<.001$	$p<.05$

Table 4.30: A comparison of predictors' performance at the phrase boundary vs. within the phrase. Given are p-values for the difference in predictors' variable importance scores when applied to the word-pair cutting across the phrase boundary and to all other transitions. Results are based on separate Wilcoxon rank-sum tests.

These findings indicate that the relations between those words framing the prepositional phrase boundary – and potentially generally the boundary of any larger constituent – are of particular importance for the placement of hesitations. As a result, the average random forest is far more likely to predict whether a hesitation will be placed at the prepositional phrase boundary or not than whether it will be placed at any particular location within the prepositional phrase.

4.6 Chunking across the Prepositional Phrase Boundary

So far, each phrase type was analysed separately and the focus was on whether the absolute co-occurrence frequency of words or their relative chance to co-occur was reflected in the placement of hesitations. In the CART trees employed for these analyses, terminal nodes containing few or no hesitations before the preposition were mainly comprised of structures in which the preposition is *of* and the 'X + Preposition' pair as a whole is either a quantifying expression, such as *all of*, or a hedge like *kind of*. Therefore, all of these cases, as well as some further *of*-collocates which tend to be placed in the same nodes by the CART trees are extracted from the data sets and examined more closely.

4.6.1 Quantifier + *of*

The group of quantifiers considered here contains determiners such as *all*, *some*, *much*, *many* and *each*, as exemplified by (109) to (111)[37].

(109) you have to save all of your *uh* vacation time (sw3523.A.s55)
(110) and I think also some of the *uh* car companies (sw4114.B.s47)
(111) how to isolate each of the different *[pause]* muscle (sw2789.B.s222)

Also included are cases where a quantity noun like *lot* or *amount* combines with *of* (cf. Quirk et al. 1985:264). (112) to (114) show some examples from the data.

(112) I haven't found a lot of *[pause] uh* fiction books (sw2792.B.s46)
(113) we have a tremendous amount of *um* sunny days (sw3148.B.s101)
(114) you uh put about two tablespoons of *uh [pause]* water (sw2608.A.s71)

Finally, numbers followed by *of* are also included, as these are yet another way of quantifying the following noun phrase. (115) provides an example.

(115) but I noticed in one of the sales *[pause]* catalogs (sw2299.A.s19)

According to these criteria, all phrases introduced by 'Quantifier + *of*' were extracted from the datasets. These amounted to 289 cases. Table 4.31 shows the complete set of 'Quantifier + *of*' combinations in the newly-formed subset. *A lot of* and *one of* are by far the quantifying expressions most frequently found in the data, followed by *some of* and *all of*.

[37] Examples are shown here with more left context than considered in the analysis.

(a) lot of – 92	any of – 5	(a) bit of – 5	hours of – 1	one of – 53
lots of – 8	each of – 2	(a) couple of – 9	months of – 1	two of – 1
many of – 3	both of – 1	(a) bunch of – 4	year of – 1	thousands of – 1
much of – 8	some of – 24	part of – 2	years of – 1	trillions of – 1
more of – 10	several of – 1	amount of – 5	ounce of – 1	(a) fifth of – 1
most of – 8	various of – 1	amounts of – 2	tablespoons of – 1	
all of – 17	half of – 1	(a) number of – 3	pounds of – 1	
(the) whole of – 1	less of – 2	percent of – 6		
		percentage of – 2		
		(a) portion of – 1		
		(the) majority of – 1		
		(a) gamut of – 1		

Table 4.31: Quantifier + of expressions in the dataset. Numbers indicate the frequency of occurrence in the dataset[38].

Across all 'Quantifier + *of*' data-points, hesitations are placed before the preposition in only 21 cases, corresponding to a rate of 7.26%. Compared to a 47.17% chance of a hesitation being placed before the preposition across all other data points, this rate is highly significantly reduced ($p<.001$; based on a 2x2 chi-square test). This low interruption rate can be interpreted as a first indication that 'Quantifier + *of*' sequences are far more strongly chunked than the average 'X + Preposition' sequence.

It follows from the chunking definition given in Chapter 1 that the words forming strong chunks should co-occur more frequently than weakly associated words or that their relative chance of co-occurrence should be higher than that of other pairs. Whether this is the case can be judged in two ways: we can investigate whether there is an attribute such as high frequency, or high transitional probability common to all these data-points and we can analyse the performance of the *ctree* models.

If *ctree* models, which are based on information about absolute and relative chances of co-occurrence of the words in a sequence, are able to predict that hesitations are highly unlikely to occur before the preposition in phrases where the preposition is *of* and where it is preceded by a quantifier, this can be interpreted as evidence that the lack of hesitations in these contexts results from the structures' high frequency or strong attractions between the quantifier and *of*. In fact, the trees only predict that hesitations occur after the quantifier in 46 cases, in 13 of these correctly so. Half of the cases (n=17) where the models wrongly predict that hesitations will interrupt the sequence are found in Node 8 in the CART tree for 'Preposition Determiner Adjective Noun' (Figure 4.12) – a node which is generally very noisy. Furthermore, the *ctree* models correctly predict where the hesitation is placed in 68.85% of phrases introduced by 'Quantifier +

[38] Frequencies as given here should not be confused with overall frequencies in the corpus, which are much higher.

of', which constitutes an above-average success rate. This means that 'Quantifier + *of*' structures repel hesitations and the *ctree* models pick this up well.

In addition to this, based on the splitting criteria chosen by the individual trees, I investigated whether there is an attribute like high frequency, or high transitional probability common to all 'Quantifier + *of*' pairs, i.e. which describes their similar behaviour. Such a single characteristic could, however, not be found; rather, all 'Quantifier + *of*' pairs are characterised by a specific combination of attributes. Except for two types, they all have a positive mutual information score and the highest possible direct transitional probability for any given mutual information score in an 'X +Preposition' word-pair, which is illustrated by Figure 4.16. Notably, it is not co-occurrence frequency which characterises this group as a whole.

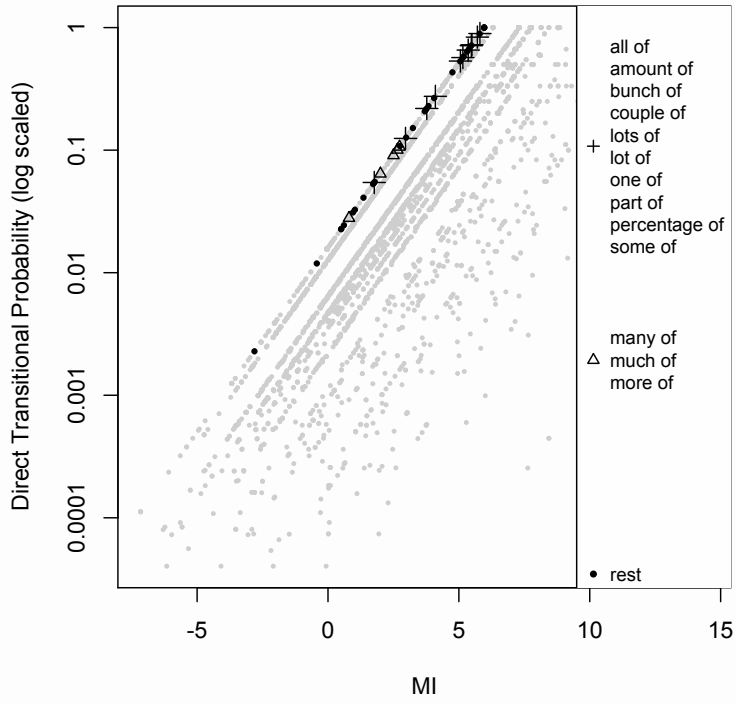


Figure 4.16: Mutual information scores and direct transitional probability of 'Quantifier+of' pairs.[39]

In summary, 'Quantifier + *of*' sequences are extremely unlikely to be interrupted by hesitations. Both the CART-tree models and the random forests recognise this and the CART-tree models rarely predict hesitations to occur in these contexts. The trees use a number of different predictors, like co-occurrence frequency, the mutual information score and direct transitional probabilities to form groups which contain many

[39] Due to the fact that they did not receive the same POS tag throughout the corpus, some quantifiers appear more than once.

'Quantifier + *of*' combinations. Thus each token of a 'Quantifier + *of*' combination must at least show one 'chunky' characteristic, i.e. it was assigned to a terminal node based on co-occurrence frequency, transitional probability etc. For instance, tokens of *one of* were assigned to Node 9 in the model in Figure 4.12 due to the high frequency of *one of* and the fact that, in this dataset, this sequence is never interrupted by hesitations (i.e. there are no hesitations at Position 1 in Node 9). In the analysis shown in Figure 4.10, on the other hand, the lack of hesitations before the preposition was correlated with a high direct transitional probability (Split 1) and a high lexical gravity G (Split 5). When investigating the entire group of 'Quantifier + *of*' combinations, no single characteristic was found which applies to all members of the group. It is particularly interesting that not all members of the group are highly frequent, yet, in the majority of cases, 'Quantifier + *of*' combinations are uttered as a single, uninterrupted unit. This and the additional finding that all of these expressions have very high mutual information scores and the highest possible direct transitional probability for any given mutual information score in an 'X+Preposition' word-pair suggest that this ratio of scores may be the best predictor of chunking strength in this context.

Furthermore, 'Quantifier + *of*' combinations provide evidence that chunking across the phrase boundary is possible and, in fact, common. Results confirm previous findings that 'Quantifier + *of*' combinations are treated very much like one cohesive unit (Clark and Wasow 1998; Vogel Sosa and MacFarlane 2002; Bybee 2007b) and that in 'Quantifier *of* Y' sequences the quantifier is not the headword being followed by a postmodifying prepositional phrase, but that in this case Y is the headword which is modified by 'Quantifier + *of*' (Sinclair 1991:85-6).

4.6.2 Further *of* Collocates

There are a number of further expressions which the CART trees often group together with the quantifying expressions and which are hardly interrupted by hesitations. These are the hedges *kind(s) of, type(s) of, sort(s) of* and *form(s) of* and the collocates *out of* and *(in) terms of*.

Hedges are "deintensifier[s]" (Lakoff 1973:471), used to "downtone the assertiveness of a segment of discourse" (Carter and McCarthy 2006:223). *Sort of* and *kind of* are among the most frequently described hedges. Here, *type of* and *form of* and the plural forms *sorts of, kinds of, types of* and *forms of* were also included, as they serve the same functions in the data. Results show that hedges are mostly preceded by *some*, *all* or *any* as in (116) to (119).

(116) to have *uh* some sort of *uh* solar power (sw4796.B.s43)

(117) congress is able to attach all kinds of *[pause] uh* funny amendments (sw4333.B.s45)

(118) she started some type of *um [pause]* national *[pause]* organization (sw2065.B.2173)

(119) prohibit them from committing any kind of *[pause]* prohibitive act (sw2051.B.s91)

However, not all cases express vagueness. In questions like (120), they take on the opposite function, i.e. are used to ask for more specific information.

(120) what kinds of *uh* sweat shirts (sw3349.A.s6)

Sinclair (1991:89-90) offers a more general terminology, encompassing these different semantic functions. In his classification, hedges fall under the category of "supporting nouns", where a semantically reduced N1, "which [is] rarely used alone" (Sinclair 1990 as quoted in Sinclair 1991:89) "offer[s] some kind of support to N2" (Sinclair 1991:89).

	No. of Tokens in the Data-Set	**Interrupted by Hesitation**	**Predicted Correctly by *ctree* Models**
kind of	38	0	33
kinds of	10	0	10
type of	12	0	10
types of	2	0	2
sort of	8	0	4
sorts of	6	0	5
form of	5	1	3
forms of	2	1	2
out of	20	3	13
terms of	18	0	14

Table 4.32 Interruption of hedges as well as 'out of' and 'terms of' by hesitations.

Table 4.32 reveals that all hedges as well as *out of* and *terms of*[40] are strong repellents of hesitations. Within the complete set of 121 tokens, only five (4.13%) are interrupted by hesitations. This is a highly significantly reduced rate ($p<.001$; based on a 2x2 chi-square test) compared to a 44.62% chance for a hesitation to be placed before the preposition across all other data points. At a misclassification rate of 17.24%, the *ctree* models perform above average on these structures.

Figures 4.17 and 4.18 show that this set of collocations is also characterised by the pattern of positive mutual information scores and the highest possible direct transitional probability for any given mutual information score. This pattern and the excellent

[40] The transition from *in* to *terms* is not part of the analysis.

performance of the *ctree* models suggest that we are, in fact, dealing with further evidence of chunking in violation of traditional phrase boundaries.

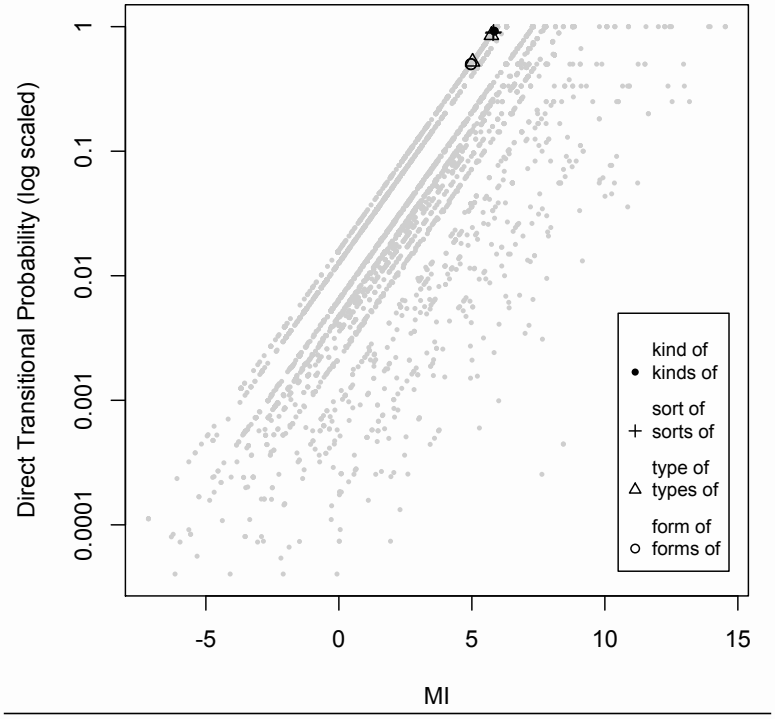


Figure 4.17: Mutual information scores and direct transitional probability of hedges.

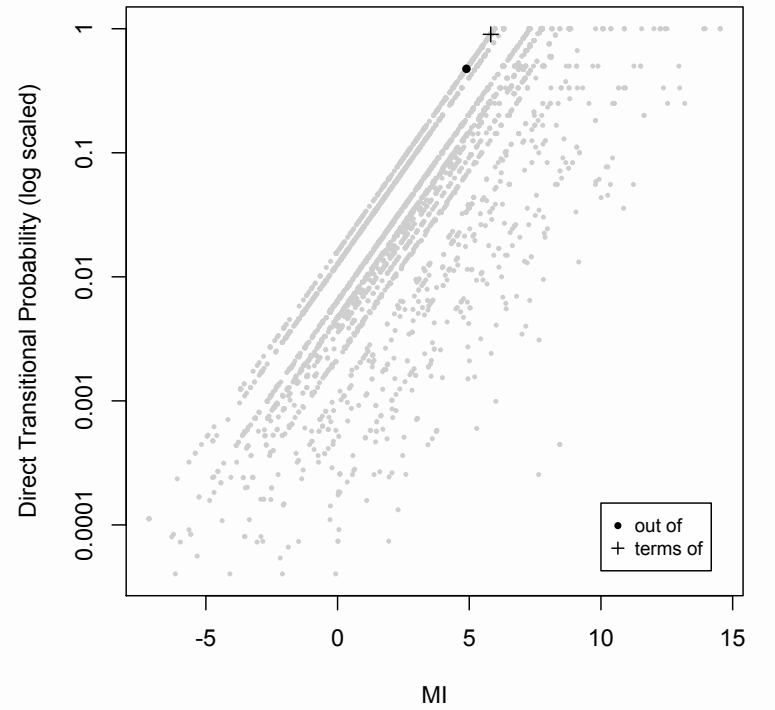


Figure 4.18: Mutual information scores and direct transitional probability of 'out of' and 'terms of'.

4.7 Hesitation-Attracting Pairs

To investigate whether there is further support of the finding that a ratio of probabilistic measures of association reflects chunking strength and is more apt at doing so than co-occurrence frequency, the opposite relations were also tested. This was achieved through investigation of whether word-pairs that are very frequently interrupted by hesitations are characterized by low direct transitional probabilities and low mutual information scores. To do this, I analyzed data points in clusters containing almost exclusively hesitations before the preposition.

4.7.1 Coordinating Conjunction & Preposition

In many terminal nodes in the CART trees, where the predominance of hesitations before the preposition suggests that the words to both sides of the boundary do not form a chunk, the word preceding the prepositional phrase tends to be a coordinating conjunction. Therefore, all instances of hesitations occurring in a phrase which is preceded by one of the coordinating conjunctions *and*, *but* or *or* are extracted. There are a total of 183 such data-points in the combined sets of prepositional phrases; (121) to (123) are three exemplary cases.

(121) but *uh* at work (2139.A.s123)

(122) and *uh* in business technology (sw3450.B.s9)

(123) or *[pause] you know* at the wrong time (sw4325.A.s102)

And is by far the most frequent of the three: 121 hesitations occur in phrases introduced by *and*, 51 in phrases introduced by *but* and only eleven in phrases introduced by *or*.

Of 183 hesitations which occur in a phrase introduced by a coordinating conjunction, 148 (80.87%) are placed immediately after the conjunction, which is significantly more than the 41.6% chance of a hesitation being placed before the preposition across all other data points ($p<.001$; based on a 2x2 chi-square test). The models also perform very well at predicting these hesitations, making the right judgement in 149 cases (81.42%), which significantly exceeds the general performance ($p<.001$).

All trees assign data-points introduced by a coordinating conjunction to one or two particular leaves only. In most cases, however, these leaves are very large and coordinating conjunctions only amount to about ten per cent of the tokens in these leaves.

Crucially, when compared to phrases introduced by hedges and quantifiers, phrases introduced by a coordinating conjunction not only display the opposite hesitation pattern but are also characterised by a complementary pattern of mutual information scores and direct transitional probabilities; mutual information scores are low, even negative in most cases, and direct transitional probabilities are also very low (see Figure 4.19).[41]

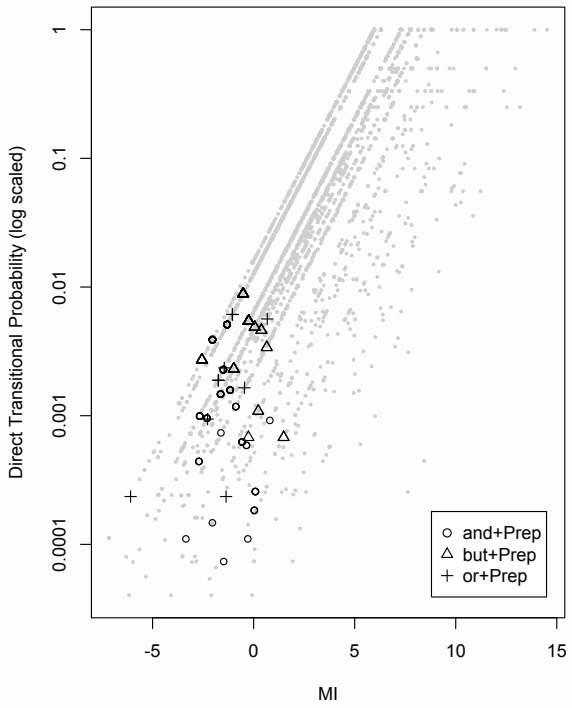


Figure 4.19: Mutual information score and direct transitional probability of 'Coordinating conjunction + Preposition' word-pairs.

The presented evidence indicates that coordinating conjunctions do not tend to form chunks with prepositions to their right. Despite the fact that some of the most frequent word-pairs in the corpus contain a coordinating conjunction (e.g. *and I*: 2,566; *but I*: 1,821; *and it*: 1,605), specific 'Coordinating conjunction + Preposition' sequences are neither frequent, nor are coordinating conjunctions and prepositions strongly attracted , as evident from Figure 4.19.

The fact that the position after the conjunction is such a strong attractor of hesitations could result from the lack of attraction, i.e. hesitations are placed where attractions are low or lowest. Thus we might be dealing with a 'standard' case where

[41] The fact that subordinating conjunctions are not treated in this chapter is not meant to indicate that they do not show the same effect; they are merely too infrequent as a group to warrant analysis. *Because* is the most frequent with only 19 instances in the data-set (17 of which are followed by a hesitation, so the pattern is the same).

words to both sides of the prepositional phrase boundary do not form a unit (or strong chunk) in the sense that they are not predominantly used together (after all, *and* is followed by 2,765 different types in the corpus, *or* by 895 and *but* by 484). On the other hand, there is evidence that coordinating conjunctions are so frequently followed by hesitations that the hesitations themselves may have become part of a chunk, particularly in the cases of *and uh* and *but uh*. This hypothesis will be explored in more detail in Chapter 5.

4.7.2 Repetitions & Self-Corrections

Filled and unfilled pauses as well as discourse markers are often shown to be placed at the hiatus of disfluent repetitions and self-corrections (cf. Clark 1996; Clark and Wasow 1998; Heeman and Allen 1999). The prepositional phrase dataset allows for this hypothesis to be tested as the context preceding the prepositional phrases was not controlled for repetitions and self-corrections. As a consequence, in 639 cases, the 'X + Preposition' pair frames the hiatus of a repetition or self-correction, such as in (124).

(124) very authentic reproductions <u>of the</u> *[pause]* <u>of the</u> actual stuff (sw2177.B.s77)

Such sequences as *the of* in the above example are highly interesting from the point of view of this study. They may occur very frequently in spoken language, yet we intuitively do not expect the words to form a chunk, i.e. to be mentally represented as a unit. The fact that in 443 (69.33%) of the 639 phrases which start with the replacement part of a repair sequence, the hesitation is placed at the hiatus of the repair, i.e. at the prepositional phrase boundary, confirms the impression that we are not dealing with mentally cohesive pairs.

For the extraction procedure and the coding, a repetition was defined as any exact repetition of one or more previous words (cf. Maclay and Osgood 1959:24; Clark 1996:264-5). There was no need to specifically exclude reduplication serving semantic purposes (e.g. *very very tall guy*) as semantic reduplication is not possible in the context of prepositional-phrase repetitions. Importantly, truncated words or phrases which were then restarted, such as in (129), were considered self-corrections, not repetitions, because we do not know the original intentions of the speaker. The person uttering (125) may originally have started to produce *glitter* and interrupted himself because he realised that the intended expression required the word *glitz*.

(125) I I would settle <u>for the gli-</u> *[pause]* <u>for the glitz</u> (sw4796.B.s86)

Self-corrections were restricted to same-turn self-initiated actions[42] (cf. Schegloff, Jefferson and Sacks 1977:367). They encompass substitutions, such as (126), deletions, and additions, such as (127) (cf. Clark 1996:264)

(126) who works for the aut- *[pause]* for the Audi dealer (sw2965.A.s153)

(127) a tendency to find out *uh* the *uh [pause]* about the different areas (sw2024.B.s7)

Only sequences which can clearly be identified as repairs are counted. Cases such as (128), where the 'repetition' or 'correction' serves to clarify are not counted.

(128) I do live in the better *[pause] well* in the best part (sw4346.A.s16)

Both repetitions and self-corrections will sometimes collectively be referred to as 'repair'. The term 'hiatus' refers to the point between the "original delivery" and the repair, be it in the form of a correction or a repetition (Clark 1996:258).

The number of hesitations placed before the preposition when this position constitutes the hiatus of a repair sequence highly significantly exceeds the rate of hesitations placed in this position across all other data-points ($p<.001$; based on a 2x2 chi-square test). Furthermore, the *ctree* models predict hesitation placement correctly in 431 of the 639 data-points, which significantly exceeds the *ctree* performance on all other data-points ($p<.01$; based on a 2x2 chi^2-test).

So far, disfluent combinations of words show one of the typical characteristics of 'unchunky' pairs; they are frequently interrupted by hesitations. Based on the *ctree* models' above average performance at predicting hesitations in these contexts, I expect them also to show the typical pattern of low mutual information scores and low direct transitional probabilities which has turned out to be characteristic of hesitation-attracting pairs at the prepositional phrase boundary.

Figure 4.20 shows that the majority of disfluent 'X + Preposition' pairs show the expected low scores. Yet there are far more exceptions here than in the previous subsets.; some of the self-corrections, in particular, reach very high scores on both scales.

[42] It is theoretically possible, though, that the interlocutor initiated the correction by uttering a short feedback signal like *huh?*. Due to the fact that in Switchboard NXT the two interlocutor's speech is documented in two separate transcripts, this possibility could not be eliminated. Yet it is unlikely because pauses of more than one second in length were excluded and thus the exchange between the speakers would have to be extremely fast. Importantly, even if other-initiated, we would not expect a speaker to interrupt a chunk in a repair sequence.

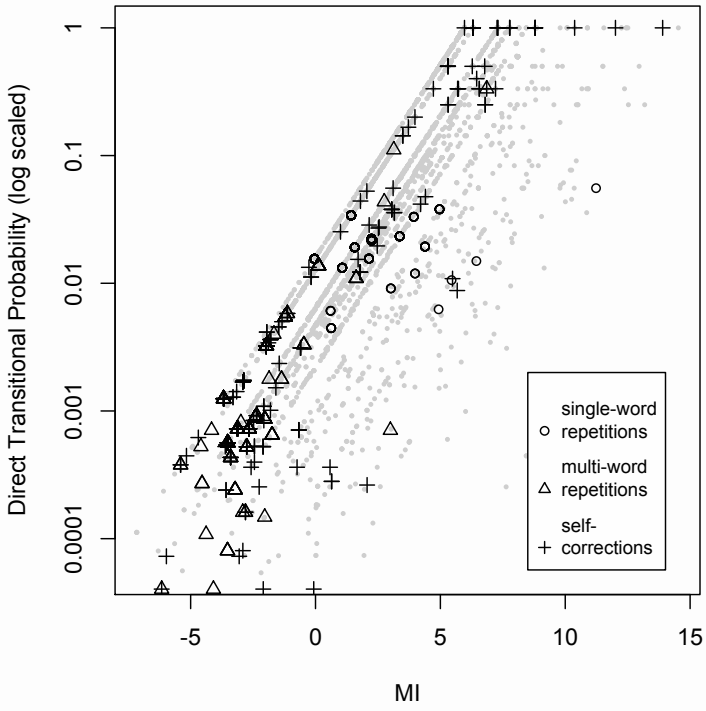


Figure 4.20: Mutual information score and direct transitional probability of disfluent repetitions and self-corrections.

The high direct transitional probabilities are due to the fact that self-corrections often contain truncated words which only appear once in the corpus in exactly this form, such as *aut-* in (130). Transitional probability thus reflects that *aut-* is followed by *for* in 100% of cases it appears in the corpus, even though 100% means only a single case. The mutual information score is similarly skewed in the case of hapax legomena.

The case of single-word repetitions is even more complex. They appear in a single cloud in the middle of the diagram in Figure 4.20, receiving mutual information scores between zero and five and reaching comparatively high transitional probabilities. Sequences of repeated prepositions are furthermore often highly frequent. As prepositions appear at the beginnings of phrases, they make ideal candidates for time-buying devices and are thus frequently repeated. (129) lists the 14 repair bigrams which appear more than 25 times each in the corpus (in order of decreasing frequency). The majority are single-word repetitions. The most frequent, *in in*, appears 336 times in the corpus, corresponding to rank 172 in the bigram frequency list.

(129) in in, of of, for for, on on, the in, with with, that in, to to, as as, at at, from from, than than, a in, the of

Theoretically, the high frequency and relatively strong attractions between the words in these pairs should be indicative of rather strong chunkiness. While combinations like *to to* cannot appear in fluent sentences in English, it would still be possible for them to

have developed into hesitation devices themselves, i.e. that the speaker mentally already has *to to* available as a pre-packaged, longer, time-buying version of the preposition. The large number of hesitations intervening between single-word repetitions, however, undermines this argument.

	No. of Tokens in the Data-Set	**Hesitation at Hiatus**	**Predicted Correctly by *ctree* Models**
Single-word Repetitions	374	247 (66.04%)	243 (64.97%)
Multi-word Repetitions	121	88 (72.73%)	88 (72.73%)
Self-corrections	144	108 (75%)	100 (69.44%)

Table 4.33: Hesitation placement and model performance in repetitions and self-corrections

Finally, repair types are not equally distributed across phrase types. Multi-word repetitions are absent in all phrases which do not contain a determiner. So, if speakers repeat the beginning of a prepositional phrase, they only repeat material up to the first content word in the phrase. If there is only one word preceding the first content word multi-word repetitions are impossible.

4.8 Summary & Discussion

This chapter investigated how far speakers' choice of where to hesitate in the context of prepositional phrases is influenced by chunking within the phrase and across the phrase boundary. These six phrase types were selected for analysis:

- Preposition Noun
- Preposition Determiner Noun
- Preposition Noun Noun
- Preposition Determiner Noun Noun
- Preposition Adjective Noun and
- Preposition Determiner Adjective Noun.

The placement of hesitations in the data was extremely varied, with hesitations occurring in all kinds of transitions with no single preferred position. Only the position before the second content word in longer noun phrases was dispreferred. Overall, the pattern proposed by earlier findings was confirmed (Maclay and Osgood 1959; Goldman-Eisler 1968; Clark and Clark 1977; Shriberg 1994; Biber et al. 1999; Bortfeld et al. 2001): hesitations occur predominantly at the phrase boundary and before the first content word in the phrase.

In the chapter, Lounsbury's (1954) claim that hesitations are placed where direct transitional probability is lowest was put to the test. It was found that hesitations were significantly more likely to be placed where direct transitional probability was lowest than at any other transition in the phrase. The result was the same for the following other measures of association:

- co-occurrence frequency
- backwards transitional probability (only in the case of three out of six phrase types)
- the mutual information score
- lexical gravity G

Subsequent analyses, based on Classification and Regression Trees (CART Trees) and random forests provided the following results.

Evidence of Chunking – The principal hypothesis that there is frequency-induced chunking, i.e. that mental connections between words are strengthened the more likely

the words are to be used together, is supported here. Speakers should prefer to utter a strong chunk in a single stretch rather than interrupting it to hesitate. Therefore, within the context of this work, significant negative correlations between the attraction in word-pairs and the presence of hesitations is considered evidence of chunking. This evidence was provided by the regression models for the phrase types 'Preposition Noun', 'Preposition Determiner Noun' and 'Preposition Determiner Adjective Noun'. These showed that pairs of words which are prone to co-occur are interrupted significantly less frequently than 'un-attracted' word-pairs.

The non-significant performance of the frequency-based models for the remaining three structures should not be interpreted as counter-evidence. This is primarily because all CART trees *do* split the data into subgroups, which means that they actually find significant effects. These are merely not strong enough to warrant generalised claims. All evidence for chunking across the prepositional phrase boundary (see below) is also found in these datasets. Furthermore, particularly among the 'Noun Noun' sequences, there are many references to towns, schools and employers. To the person who comes from *Warren*, *Michigan*, went to *Lincoln High School* and now works for *General Motors*, all of these names are highly familiar, oft-used items, while among the average cross-section of American adults, which is represented in the corpus, these are rarely (if ever) used. So what is surely chunked in the idiolect of the Warren speaker cannot be expected to be chunked in the lexicon of the average American speaker, as approximated by the corpus.

Chunking across the prepositional phrase boundary – Results show that chunking across the prepositional phrase boundary is possible and, in fact, common. *Of* and its left collocates, in particular, often form cohesive, uninterrupted units, like *all of, one of, out of, kind of* and *terms of*. This indicates that sequentiality is not always basic to (currently accepted) constituent structure (cf. Bybee 2007b:314) and confirms hypotheses and earlier results by Sinclair (1991), Altenberg (1998), Clark and Wasow (1998), Pullum and Huddleston (2002), Vogel Soza and MacFarlane (2002), Bybee (2007b) and Beckner and Bybee (2009).

Hesitations are particularly good indicators of chunking across the prepositional phrase boundary. Effects were stronger for word-pairs crossing the phrase boundary than for those within the phrase, which does not indicate that strong chunks do not form within the phrase. This is evident from the fact that the average frequency, MI score, etc. of pairs bridging the phrase boundary does not significantly exceed that of pairs within the phrase (see Table 4.5 and Figures C.1 to C.5 in the Appendix). Therefore, we can conclude that chunking at the phrase boundary simply has a much stronger effect on hesitation placement than chunking within the phrase – at least at the bigram level.

Role of the embedded noun phrase – Results do not indicate that nouns are generally more strongly attached to preceding determiners, adjectives or nouns – because these are part of the same phrase – than to prepositions because these are part of the superordinate prepositional phrase (cf. Bybee 2007b). This can be deduced from the distribution of hesitations within the phrases and from the mean strengths of the relations between the words. Hesitations are not a very good indicator of chunking within the prepositional phrase, though, as evidenced by the fact that relations between the noun and its antecedents did not generally have a profound influence on hesitation placement.

Relative chance of co-occurrence versus absolute co-occurrence frequency – Overall, probabilistic measures of association do not perform better at predicting hesitation placement than absolute co-occurrence frequency. The selected probabilistic measures performed on par with frequency. An increase in information contained in a measure did not lead to a significant improvement in performance. Occam's razor would suggest we stick to the simplest measure of chunking strength, i.e. frequency of co-occurrence.

However, analyses of relations holding across the prepositional phrase boundary indicate that in this context the mutual information score may outperform co-occurrence frequency ($p<.1$). Lexical gravity G, the most complex among the measures, still performs no better than frequency, though.

Structures which were unlikely to be interrupted by hesitations, such as quantifying expressions followed by *of*, displayed a characteristic pattern of high direct transitional probability and high mutual information score. Co-occurrence frequency, in turn, did not characterise the group as a whole. Frequently interrupted structures, such as coordinating conjunctions followed by prepositions, in turn, displayed the opposite pattern. Thus we can conclude that – at least at the prepositional phrase boundary – the mutual information score and direct transitional probability reflect chunking at both ends of the spectrum; they characterise 'chunky' sequences as well as very 'unchunky' pairs and do so better than absolute co-occurrence frequency. These results indicate that the mind not only keeps track of absolute co-occurrence frequencies but also of the relative chance of co-occurrence of two words.

Type vs. token frequency effects – Results strongly suggests that 'Quantifier+*of*' structures are chunked. The question arises of whether these are chunked as a construction on what Bybee (2010; see Section 2.2.1.1 above) calls the "partially mixed" level or whether they are chunked as individual MWUs. The overall homogeneity of the group in terms of lack of hesitations as well as its characteristic mutual information/direct transitional probability ratio allows for tentative conclusions that the abstract category 'Quantifier + *of*' may also be cognitively represented.

Methodological issues – All analyses in this chapter rely on a combination of statistical analyses which are relatively new to linguistics and thus still in need of evaluation. In this case, CART trees and random forests were the only method available which could deal with the conditions of the study. Analyses were extremely computationally intensive and had to be conducted on more powerful hardware than offered by current home computers, but in return did not crash or run into problems when faced with multinomial outcomes and correlated predictors. Individual CART trees, particularly, were more versatile than anticipated. The clusters created by these trees facilitated pattern recognition as these meant that the data was prearranged for the investigator. CART trees further turned out to be far more reliable statistical tools than anticipated. It has been cautioned that single trees may provide unreliable results (cf. Strobl, Malley and Tutz 2009b). For this reason, all results were backed up by random forests, which rely on thousands of trees. Forest results generally indicated stronger effects than the individual trees did, but this appeared to be an effect of over-fitting to the data because the out-of-bag sample (see Section 3.3.3.2) always confirmed CART results.

Hesitations also proved better indicators of chunking than cautioned by studies of repairs (cf. Fox and Jasperson 1995; Kapatsinski 2005). These studies find that speakers are so inclined to recycle back to the nearest constituent boundary that they care little for the associations between the words in the surrounding context. This tendency has not been observed for pauses and may apply rather to clause than to phrase boundaries.

Repetitions and self-corrections point to the limits of this type of analysis. Hesitations are strongly preferred at the hiatus of disfluencies, yet some repetitions of phrase-initial function words (such as *in in* and *of of*) are so frequent that frequency-based models must assume that these are rather highly attracted word-pairs. The models therefore do not anticipate hesitations in these contexts. A larger 'window' of analysis, i.e. providing models with information about four- or five-word strings instead of just two-word strings, would solve this problem yet would require data from extremely large corpora.

5 Hesitation Placement in Sentence-Initial Structures

The previous chapter showed that hesitation placement can serve as an indicator of frequency-based chunking. The present chapter applies the same methodology to sentence-initial contexts, in order to analyse whether frequency-based chunking is also observable in and across other types of phrases. In several sets of sentence-initial subject-verb sequences of increasing complexity, it will be tested whether a speaker's choice of hesitation placement is influenced by chunking, i.e. by the attraction between words. Of particular interest is not only the relation between the subject and the verb, such as in (130), and attractions within the verb phrase, but also hesitation placement in structures where the subject is preceded by sentence-initial elements (SEs), such as conjunctions and adverbs (see (131)).

(130) a. *well uh* we live (sw3317.B.s3)
b. we *uh [pause]* saw (sw3342.A.s91)

(131) a. *uh* and where you are (sw3313.B.s11)
b. and *uh [pause]* actually it was (sw2107.A.s151)
c. and then they *[pause] um you know* go (sw2065.B.s222)

Throughout this work, hesitation placement is analysed as an indicator of the processing and mental representation of multi-word units. It is expected that the more frequently words are used together or the stronger the attraction between them, the more likely they are to be produced 'en block', i.e. uninterrupted by hesitations.

Results show that the sentence boundary is a far stronger attractor of hesitations than lower-order constituent boundaries. Particularly where there is no coordinating conjunction, sentence adverb or discourse marker at the beginning of the sentence, speakers opt to hesitate before starting to articulate the sentence in the overwhelming majority of cases. Nevertheless, the data provides evidence of chunking. First of all, frequent coordinating conjunctions, particularly *and* and *but*, have merged with following pause fillers. Sequences like *and uh* and *but uh* are chunked and serve as one longer time-buying device. Furthermore, whether speakers hesitate in the verb phrase or between the subject and the verb depends on the frequency and cohesiveness of the subject-verb combination and the word-pairs in the verb phrase. Thus there is chunking in the verb phrase and between the subject and the verb. Negative constructions in particular are prone to being chunked.

5.1 Background & Previous Research

Sentence beginnings up to and including the verb phrase are interesting objects for a study of this kind for a number of reasons. Of particular interest in the present analysis are the relation between the subject and the following finite verb, the relationship between the finite and non-finite verbs, and the competition between structural and frequency-based factors in general. Focal questions are how much variation in hesitation placement there is in the proximity of a major syntactic boundary and whether this variation is influenced by the frequencies of word-pairs in the sentence and the relations holding between the words in the pairs. There will be particular emphasis on chunks forming between SEs and hesitations. The remainder of this section provides some brief information about these aspects of interest.

Relations within the verb phrase – Bybee and Torres Cacoullos (2009) show that frequency of use has an influence on the grammaticalisation of Spanish progressive constructions, such as *estás hablando* (you are speaking). Increased usage of originally lexical verbs with locative meanings in these constructions has lead to their grammaticalisation into auxiliaries (see Bybee 2010:148). The authors show that over time, as grammaticalisation has progressed, the likelihood of the sequence being interrupted by intervening elements has diminished and that some of the most frequent types have developed into MWUs which can be accessed as single units and are thus even less likely to be interrupted (see Bybee 2010:149).

Turning to English, Bybee (2010:151-64) claims that many negated verb phrases with *can't*, such as *can't remember* or *can't think*, are MWUs. This claim is based on the observation that in these cases the negative construction is more frequent than its positive counterpart and semantically does not really represent a negation of the positive statement (Bybee 2010:152-3). Even the larger constructions in which negative and positive *can*-constructions are predominantly used differ. *Can think of*, for instance, is predominantly used in relative clauses, while *can't think of* is not (Bybee 2010:154-5).

Based on these findings, I expect that many English verb constructions are also chunked and thus unlikely to be interrupted by hesitations.

Relation between the subject and the verb – Bybee argues that some traditionally-assumed constituent boundaries should be reconsidered (Bybee 2010:137-8). Based on the fact that English auxiliaries may contract with the preceding subject, but not with the following verb, and thus 'prefer' to form chunks across constituent boundaries, Bybee argues that constituency is not as rigid as previously assumed and should instead be viewed as a gradient concept (Bybee 2010:136-8).

Chunked clitics, however, may only be the most obvious and strongest case of cross-boundary chunking, as, in their case, phonetic reduction is so strong (and conventionalised) that it is represented in writing; other subject-verb combinations may well be strongly attracted, too, and somewhere on the cline of being chunked.

There are some results from studies of hesitation placement which can be taken as indications that subject-verb attractions may indeed be stronger than relations across other phrase boundaries. While it has been shown that speakers generally prefer to place hesitations at constituent boundaries (cf. e.g. Maclay and Osgood 1959:33; Clark and Clark 1977:267-8; Swerts 1998:489; Biber et al. 1999:1054, 1060; see also Section 2.3.2), Cook (1971:138) finds significantly fewer filled pauses than expected before verbs and auxiliaries (cf. also Maclay and Osgood 1959:31). Yet there are no studies which test whether this is an effect of usage frequency.

The sentence as the central unit in speech planning – At the beginning of a sentence, the speaker needs to lay the ground plan for its syntactic structure (cf. Auer 2009:4 who refers to the unit of the "turn" or "turn component"). Results from psycholinguistic experiments suggest that speakers generally defer this task until they have finished producing the previous sentence. Power (1986:378), for example, shows that in a secondary tracking task, error rates increase towards the end of a clause if it is immediately followed by another which is part of the same sentence. He concludes that in consecutive clauses, some of the planning for the second clause is already done while articulating the first. Importantly, however, if the first and second clauses belong to different sentences, error rates decrease towards the end of the first clause, indicating that the new sentence is only planned after articulating the first.

Power's results lead to the conclusions that (a) the sentence is the central unit of speech planning and (b) the processing load at the beginning of the sentence is higher than at locations within the sentence. This hypothesis is borne out by the finding that hesitation and discourse marker rates before the first word in a sentence exceed those at other locations (cf. for example Schiffrin 1987:328; Holmes 1988:331; Shriberg 1996:12; Biber et al. 1999:1086). In this respect, the conditions under which speakers deviate from the default pattern and place hesitation within the sentence are of particular interest.

Sentence-initial hesitation chunks – Apart from being units of planning, sentences are also units in discourse. This means that sentence beginnings often coincide with the start of a new turn and the end of most sentences mark possible turn-completion points (cf. De Ruiter, Mitterer and Enfield 2006:531). Thus, if a speaker wants to keep the turn, it is imperative that he or she start articulating the first words of the sentence and not pause

for long (cf. Sacks, Schegloff and Jefferson 1974:718-9). Consequently, for the speaker, starting a sentence means accomplishing a cognitively demanding task under immense time pressure.

There are at least two common strategies to alleviate the pressure. The first, here referred to as the 'start-first-hesitate-later' strategy, is to start producing the initial word(s) of the sentence to keep the turn and then to hesitate within the sentence (cf. Clark 1996:269; Clark and Wasow 1998:208). The second, the 'dummy-first' strategy, also involves starting to speak as soon as possible in order to keep the turn; only in this case speakers start with a time-buying device *before* uttering the first words of the sentence. The time-buying function can be fulfilled by pause fillers and discourse markers, but potentially also by "appositional beginnings", such as "*well, but, and, so*" (Sacks, Schegloff and Jefferson 1974:719). These "redundant linking words" (Holmes 1988:329)

> satisfy the constraints of beginning. But they do that without revealing much about the constructional features of the sentence thus begun, i.e. without requiring that the speaker have a plan at hand as a condition for starting. (Sacks, Schegloff and Jefferson 1974:719)

Results from a range of studies indicate that speakers often produce SEs, such as *and*, *but*, *you see* or *because* – either as part of the start-first-hesitate-later or the dummy-first strategy – and hesitate after them, i.e. before producing the subject of the sentence. This results in the repeated combined use of SEs and discourse markers or pause fillers.

- Jurafsky et al. (1998:2) find that in a subset of Switchboard 22.6% of *and* and 19% of *that* are followed by hesitations (i.e. filled and unfilled pauses as well as repetitions), compared to only 11.7% of *the,* 11% of *I* and 3.5% of *you* being followed by disfluencies.
- Holmes (1988:337-8) reports that in spontaneous speech, connectives are prone to being followed by hesitations, the highest rates being reached by coordinating conjunctions (on average, 4.1% of coordinating conjunctions are followed by filled pauses and 16.3% unfilled pauses), complementisers introducing complements (e.g. *that, to*; these are, on average followed by filled pauses in 2.4% of cases and by unfilled pauses in 12.2% of cases) and conjunctions introducing adverbials (e.g. *when, because, if*; on average followed by filled pauses in 2.8% of cases and unfilled pauses in 10.8% of cases).
- Altenberg (1998:112) shows that in the London-Lund corpus, *and, but, that* and *because* are frequently followed by the discourse markers *you know*, *I mean* and *you see*.

I argue that some of these combinations, particularly *and uh* and *but uh*, occur so frequently that they result in an as of yet under-researched kind of chunk: a fixed combination of a word and a hesitation. Such chunks are then used as longer time-buying devices in the dummy-first strategy (cf. Altenberg 1998:113), which further increases their usage frequency and which presumably is also partly responsible for the high frequencies reported in the aforementioned studies.

Evidence from Clark and Wasow (1998) supports the chunking hypothesis. In Switchboard and the London-Lund corpus, the authors find that there is often no intervening time between a word and the filler following it. The authors argue that in order to utter words in such quick succession, they must have been planned together. Many fillers are even cliticised onto the previous word, thus forming a single phonological word, such as "I.muh" or "to.wuh" (Clark and Wasow 1998:229). Clark and Fox Tree (2002:101) emphasise that this is particularly common for fillers following conjunctions, resulting in "an.duh", "bu.tuh"and "so.wuh".

The present study examines whether SEs and hesitations merely happen to be frequently used together or whether they actually form chunks and if they do so, whether this is an effect of usage frequency.

5.2 Data & Predictors

The type of analysis conducted throughout this work requires that only structurally similar sequences be compared. Therefore, not all hesitations occurring in sentence-initial contexts were considered for analysis. Instead, only hesitations within a limited set of structures were selected. The following sections describe the selected structures and provide information concerning the technical aspects of the extraction procedure.

5.2.1 Selection of Contexts

Starting out with a subject and a simple finite verb phrase, eight different types of sentence-initial contexts were selected for analysis. Tables 5.1 shows all included sequences.

Structure	**Example**	**n**
Subject Verb(finite)	*well [pause] um* I seem (sw2247.B.s57)	2,576
Subject Verb(finite) Verb(non-finite)	it was *uh* working (sw3663.A.100)	612
Subject Verb(finite) *not* Verb(non-finite)	*uh* I don't agree (sw2502.A.s154)	340
SE Subject Verb(finite)	and *uh [pause]* I thought (sw2010.A.s9)	1,660
SE Subject Verb(finite) Verb(non-finite)	and *uh you know* you can eat (sw3473.A.s76)	429
SE Subject Verb(finite) *not* Verb(non-finite)	but *um [pause]* I don't know (sw2241.A.s44)	225
SE SE Subject Verb(finite)	and *uh* fortunately we agreed (sw2005.B.s104)	367
SE SE Subject Verb(finite) Verb(non-finite)	*well* ironically enough I'm sitting (sw2433.A.s1)	108
	Total	6,317

Table 5.1: Structure-types included in the analysis, examples and number of data-points (i.e. hesitations) per type.

The simplest sequence, 'Subject Verb(finite)' is extended in a stepwise pattern:

- The verb phrase may be complex (i.e. consist of a finite and a non-finite form each) or negated (i.e. have the structure 'Verb(finite) *not* Verb(non-finite').

- The subject may be preceded by a maximum of two typical sentence-initial elements (SEs), such as adverbs and coordinating and subordinating conjunctions. This narrow focus means that some typical longer sentence-initial constructions (e.g. *as far as, so right now, but then again*) are excluded. Random samples however showed that in sentences containing three and four words before the verb, the percentage of disfluent combinations (e.g. *and and then*) disproportionately increased.

The maximum structure containing all permitted elements would be 'SE SE Subject Verb(finite) *not* Verb(non-finite)' which will, however, not be analysed, because it is too rare in the corpus and thus only very few hesitations occur in this context.

Save for one exception (see Section 5.2.2.4 below) there are no restrictions on how sentences continue after these initial sequences. In a sentence like (132), for instance, the search heuristics only pick up *has* as part of the verb phrase because of the intervening adverb. Therefore, the sequence is categorised as 'Subject Verb(finite)' and only the placement of the *uh* will be analysed.

(132) he *uh* has thoroughly *um* read the book

Crucially, the first word in the structure has to be the first word in the sentence. This ensures that the actual sentence boundary is part of the window of analysis. It is therefore possible to describe interactions between frequency effects and the effect of the sentence boundary as a major attractor of hesitations.

5.2.2 Retrieval Procedure & Definitions

Data was exclusively drawn from the Switchboard NXT corpus. Hesitations and their respective contexts were extracted from the corpus with the help of the software *R* (R Development Core Team 2009). All necessary scripts were developed by me.

Table 5.2 lists the parts of speech sequences which were permitted by the automatic corpus search. Not all sequences theoretically permitted will occur in fluent sentences in English. The following sections explain how groups were defined and which further steps were taken to ensure that repair sequences were excluded.

While the dataset was partially hand-annotated, and false hits were manually excluded, the search procedure remains a heuristic due to the large number of data-points. Tagging and parsing of the corpus were not checked throughout because random samples show that both are very accurate considering we are dealing with disfluent stretches of spoken language.

Sentence-initial Elements		Subject	Finite Verb	not	Non-finite Verb
coordinating conjunction	coordinating conjunction	existential *there* personal pronoun	modal 's, form of BE verb base form verb past tense verb past participle verb, 3rd sing.	not n't	modal verb base form verb past tense verb past participle verb gerund/ present participle
subordinating conjunction	subordinating conjunction				
adverb	adverb				
comparative adv	comparative adv				
superlative adv	superlative adv				
interjection	interjection				
discourse marker	discourse marker				
wh-adverb	wh-adverb				

Table 5.2: Parts of speech permitted by the automatic search heuristics

5.2.2.1 Sentences

In the context of spoken language, the concept of a sentence merits some explanation. The definition applied here follows corpus annotation and is detailed in Section 3.1.1.4.

5.2.2.2 Sentence-initial Elements (SEs)

In order to obtain a set of frequent structure types and to avoid retrieving embedded sentences, only conjunctions, adverbs, discourse markers (except those classified as hesitations; see Section 3.1.2.3) and interjections were permitted before the subject (see Table 5.2).

In the study of hesitation placement in prepositional phrases, it was confirmed that if an utterance contains a repair sequence, any further hesitations are likely to be placed at the hiatus of the repair. Results furthermore indicated that this behaviour is not necessarily an effect caused by lack of statistical attractions in the repair sequence. To avoid the noise thus caused in the data, sequences of SEs containing truncations (e.g. *bu-* instead of *but*), repetitions (e.g. *and and*) or self-corrections (e.g. *with if* or *after when*) were not permitted. If one of the target hesitations occurred in a sentence containing one of the aforementioned disfluencies, the data-point was deleted, save in one specific case. If a hesitation occurred in a sentence starting with two SEs, the first being *and* or *but*, the data-point was not deleted even if the combination of SEs was seemingly ungrammatical. Standard (written language) grammar would consider *and so*, *and because* or *and if* grammatically correct, while deeming *and but* incorrect. In spoken language, however, the initial *and* may actually serve the same function in all cases – presumably to buy time. Therefore all SE combinations starting with *and* or *but*, except repetitions of these, were permitted.

5.2.2.3 Subjects

As shown in Table 5.2, only personal pronouns and existential *there* were permitted in subject position. This ensures that the subject is consistently one word long. Pronouns had to be marked as the subject on the syntax level of annotation in order to be considered for analysis. Existential *there* was permitted because it is a one-word-element which can stand in subject position and in most respects act as the grammatical subject of the clause, which additionally contains a notional subject (cf. Quirk et al. 1985:1403-5).

Subjects consisting of a single noun were rejected as these would have been prone to consist of proper names, which, as shown in the prepositional phrase study, tend to be highly frequent in the speech of individual speakers, but infrequent in a general corpus which makes their chunking behaviour hard to predict.

5.2.2.4 Verb Phrases

The verb phrase had to be finite, meaning that the first or only verb had to be finite; if a second verb followed, it had to be a non-finite form. Permitted part-of-speech tags were chosen accordingly, e.g. no gerunds as the first verb and no third person singular forms as the second (see Table 5.1). Yet, parts of speech alone do not fully distinguish between finite and non-finite verbs. Therefore, the output of the automatic extraction procedure was manually checked and non-permissible sequences deleted.

Again, disfluencies other than the target hesitations were not allowed and data-points containing repetitions, self-corrections or truncated verbs were deleted. Where the non-finite form was disfluent, the structure was analysed up to the finite verb. Clear cases of embedded clauses were also deleted. These were mostly cases of reported speech, as in (133), or such cases as (134).

(133) and when they say *you know [pause]* buy [one get one free it's hard to resist] (sw3142.A.s98)

(134) and now I guess *[pause] you know* being [in my forties I just kind of mellowed out a little bit] (sw2331.A.s242)

The retrieval heuristics interpreted sequences where *not/n't* was followed by an adverb, as in (135) a. and (136) a., as simple, non-negated 'Subject Verb(finite)' sequences (i.e. as (135) b. and (136) b.). This happened because 'Subject Verb(finite) *not*' is not included as a data-set in the present study, the heuristics therefore shortened the sequence to a smaller, permitted sequence.

(135) a. *you know I mean* it wasn't even funny (sw2436.B.s58)

b. *you know I mean* it was

(136) a. *you know* we don't really do any gardening (sw3728.B.s158)

b. *you know* we do

This was undesirable because of the fact that *ca* and *wo* do not exist as words without the clitic. Therefore, all data-points where *not* is cliticised and not directly followed by a non-finite verb were excluded.

Throughout all analyses, finite verbs will be coded as V(fin) and non-finite verbs as V(inf).

5.2.2.5 Hesitations

The hesitations considered in the present analysis are filled pauses (i.e. *uh* and *um*) and the discourse markers *well*, *you know*, *I mean* and *like*. Importantly, unfilled pauses are not covered here. In contrast to the previous analyses of chunking in prepositional phrases, which only considered sentence-medial phrases, this chapter exclusively deals with sentence-initial contexts, which poses a problem for the analysis of unfilled pauses. If the beginning of a sentence also constitutes the start of a new turn, then a sentence-initial pause will simultaneously be turn-initial, meaning that it cannot be attributed to one of the speakers; it might equally well have resulted from speaker A pausing after his turn than from speaker B stalling before starting his turn. Including only unfilled pauses which occur sentence-medially or before non-turn-initial sentences would result in severe imbalances in the design. Consequently, unfilled pauses are not considered in this study. Where they occur together with pause fillers or discourse markers, they are transcribed in examples; otherwise they are ignored.

Where two or more hesitations occur in a sequence not interrupted by other words (e.g. *uh [pause] well*), these are referred to as a 'hesitation cluster' and treated as one data-point. Due to the large variety of hesitation combinations in clusters, no distinction is made between different cluster types. Instead, individually occurring discourse markers and all clusters containing at least one discourse marker are covered by the predictor *dm*, while individually occurring pause fillers and all other clusters are subsumed under the category *u* (for more information see Section 3.1.3).

Hesitations occurring within or before one of the structures defined above are analysed, while those after the final word in the structure are not taken into consideration. If there are two hesitations in different positions in the same structure, these are treated as two different data-points.

5.2.3 Distribution of Hesitations

In total, the dataset consists of 6,317 hesitations, made up of 4,236 individually-occurring hesitations and 2,181 hesitation clusters. Table 5.3 and Figure 5.1 show the distribution of hesitations. They illustrate that placement in sentence-initial contexts displays characteristic patterns. Most notably, where no SE precedes the subject, almost all hesitations are placed at the sentence boundary. Further patterns will be discussed in the light of previously suggested factors in Section 5.3.

Structure	**before SE 1**	**before SE 2**	**before Subj**	**before V(fin)**	**before *not***	**before V(inf)**	**Total**
Subject Verb(fin)			2,535	41			2,576
Subject Verb(fin) Verb(inf)			564	7		41	612
Subject Verb(fin) *not* Verb(inf)			330	3	0	7	340
SE Subject Verb(fin)		449	1,171	40			1,660
SE Subject Verb(fin) Verb(inf)		116	268	8		37	429
SE Subject Verb(fin) *not* Verb(inf)		41	178	2	1	3	225
SE SE Subject Verb(fin)	74	216	64	13			367
SE SE Subject Verb(fin) Verb(inf)	24	57	20	0		7	108
						Total	6,317

Table 5.3: Distribution of hesitations across the pre-verbal sentence-initial contexts

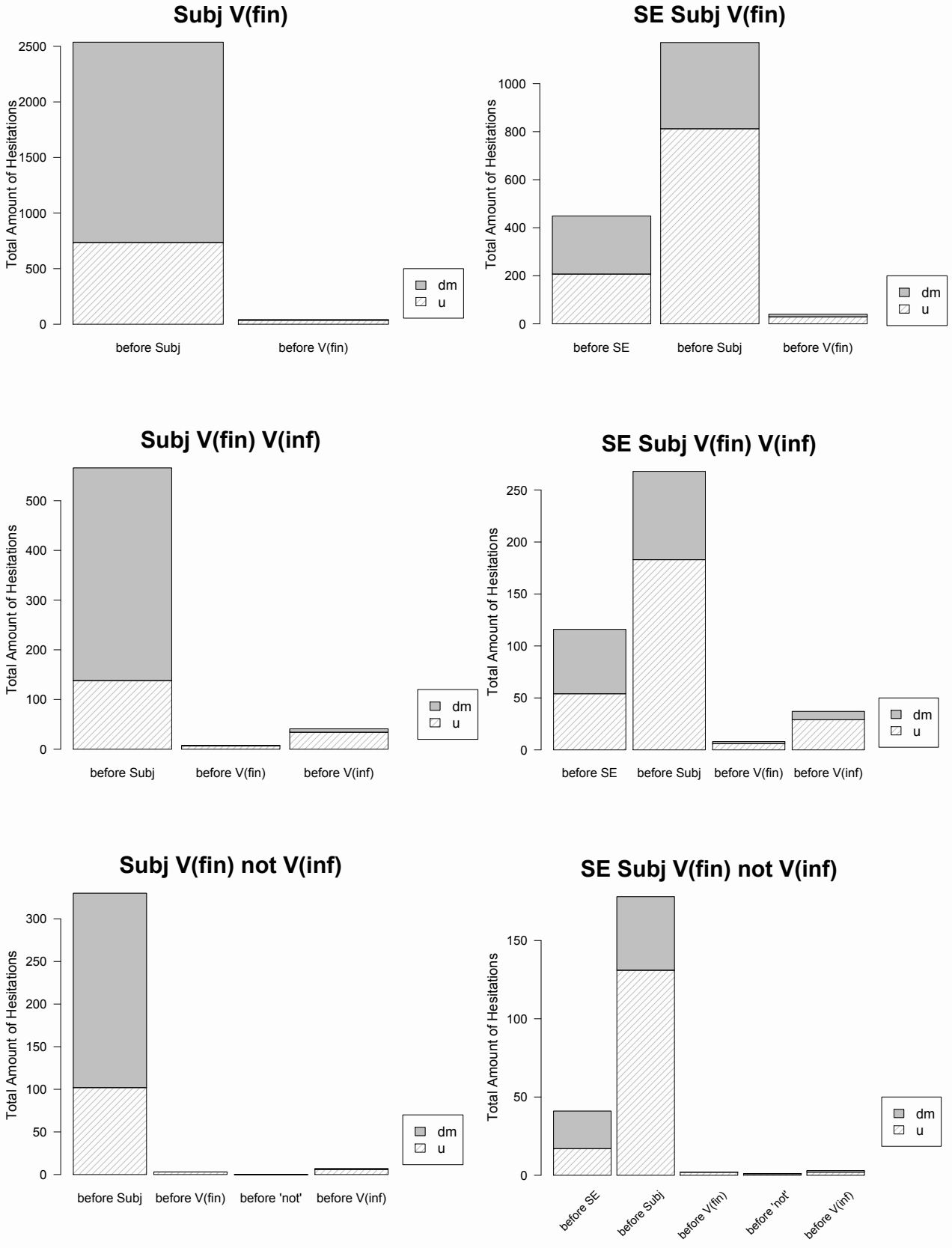


Figure 5.1: Distribution of discourse-markers (dm) and filled pauses (u) across sentence-initial structures (Part 1).

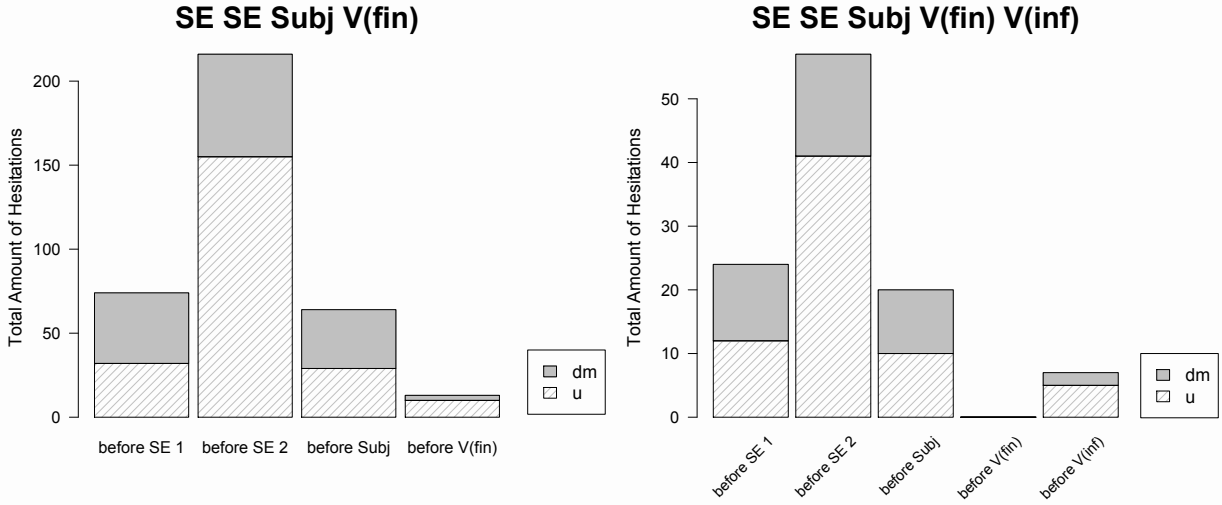


Figure 5.1: Distribution of discourse-markers (dm) and filled pauses (u) across sentence-initial structures (Part 2).

5.2.4 Predictors

This study draws on the following frequency-based measures of association. For the full description of their calculation see Section 3.3.1 and for an abridged description see Section 4.2.4.

- Bigram frequency (*bi.freq.NXT*)
- Direct transitional probability (*TPD.bi.NXT*)
- Backwards transitional probability (*TPB.bi.NXT*)
- Mutual information score (*MI.NXT*)
- Lexical gravity G (*G.NXT*)

Additional predictors are:

- Word frequency (*w.freq.NXT*)
- Hesitation type (*hes.type*)

Predictors were calculated for each word and transition in a given structure. However, it is important to note that sentence-initial contexts pose certain restrictions on the calculation of measures of association. The beginning of a sentence often also constitutes the beginning of a turn, and was thus clearly not planned with the previous sentence which was uttered by another speaker. Even within one speaker's turn, cross-sentence word pairs form neither semantic nor formal units and are therefore highly unlikely to form chunks (cf. Beckner and Bybee 2009:30-1; see also Power 1986:378).

Therefore, word-pairs cutting across a sentence boundary were not included in the calculation of bigram frequencies, transitional probabilities etc. Consequently, no measures of association can be provided for the first word in a sentence and its preceding context.

Furthermore, some finite verbs and negations are cliticised onto the preceding word. Obviously, in these cases, no hesitations intervene. Yet, whether a word is an enclitic is not included in the analysis as a separate factor because any strongly associated chunk will show some degree of phonetic reduction and merging even if not indicated in the spelling (cf. Bybee 2010:37-45). Consequently, if chunking is a frequency effect and predictable from co-occurrence frequencies and other measures of association, then these predictors should suffice to predict that hesitations do not occur before a clitic. Nevertheless, the effect of both verbal and *not* clitics on hesitation placement was tested in a pilot study, which considered these two in addition to the predictors listed above. As expected, information about cliticisation does not provide the model with significant additional information and is consequently disregarded.

	n		**Freq.**	**TPD**	**TPB**	**MI**	**G**
SE + SE	**475**	mean	249.14	0.04	0.1	1.68	6.63
		sd	387.98	0.1	0.16	2.38	5.17
SE + Subject	**2,789**	mean	1,150.12	0.11	0.06	1.8	10.82
		sd	1,174.68	0.09	0.05	0.93	4.73
Subject + Verb(finite)	**6,317**	mean	1,691.95	0.08	0.42	3.61	8.66
		sd	1,982.12	0.11	0.31	0.95	4.43
Verb(finite) + *not*	**565**	mean	3,039.16	0.49	0.34	5.1	8.58
		sd	2,032.03	0.25	0.22	1.19	3.34
***not* + Verb(non-fin.)**	**565**	mean	689.28	0.08	0.46	4.75	7.72
		sd	656.55	0.07	0.32	1.54	4.88
Verb(finite) + Verb(non-fin.)	**1,149**	mean	103.03	0.04	0.18	4.85	3.47
		sd	161.91	0.07	0.2	1.99	4.46

Table 5.4 Mean values and standard deviation (sd) of frequency, direct transitional probability (TPD), backwards transitional probability (TPB), mutual information score (MI) and lexical gravity G (G) of all transition types in the dataset.

5.2.5 Frequency-based Characteristics of all Transitions

Associations in a word-pair depend greatly on the parts-of-speech of the words in the pair. This means that nouns, for example, form different kinds of relations to their left and right context than conjunctions or verbs. This section briefly characterises each type of transition in the dataset. Table 5.4 shows an overview of mean values and Figures E. 1 to E.5, found in the Appendix, provide a visual comparison.

5.2.5.1 Sentence-initial Element & Sentence-initial Element

Due to the fact that SEs from many different parts of speech are permitted in the dataset, this is the most diverse group linguistically. This diversity is evident in the comparatively high type/token ratio of 0.34 (160 types vs. 475 tokens; on average only 2.96 tokens to the type) and the low mean bigram frequency. There are three specific word pairs which stand out:

- At 59 occurrences, *and so* is the type with the greatest currency in this group (frequency in the corpus: 769).
- *And then* is the member of the group with both the highest frequency in the corpus (1,360; 33 occurrences in the data-set) and the highest lexical gravity G of the group (14.93). Even taking into consideration that lexical gravity G strongly correlates with log frequency, this a very high G score, indicating strong cohesion.
- *Ironically enough* is the only member of the group with a direct transitional probability of 100%, which indicates that *ironically* is always followed by *enough*. At 11.53 it also carries the highest mutual information score in the group. Both of these results reflect that *ironically enough* constitutes a single constituent.

Generally, this group contains a lot of pairs beginning with a coordinating conjunction. (137) shows that when ordered according to corpus frequency, eight out of the ten most frequent pairs in this group are combinations containing a coordinating conjunction:

(137)

and then (1,360)	and when (159)
and so (769)	and now (127)
right now (334)	but then (124)
and if (243)	and but (116)
and just (199)	pretty much (114)[43]

[43] Number in brackets is the bigram frequency in the corpus.

According to the mutual information score, on the other hand, the highest-ranked pairs are mostly two-word sentence adverbs, as listed in (138); the numbers given in brackets are mutual information scores.

(138)

ironically enough (11.53)	around here (6.95)
deep down (10)	pretty much (6.73)
even though (8.31)	up here (6.49)
right now (7.75)	until recently (6.4)

Pairs containing coordinating conjunctions tend to receive a lower mutual information score; so much so that when the types in this group are ordered according to the mutual information score, only 32.5% of the types in the first quartile contain a coordinating conjunction, while 70% in the fourth quartile do. This is further confirmation that the mutual information score reflects which pairs function as constituents and which do not.

5.2.5.2 Sentence-initial Element & Subject

This group is characterised by an extremely low type/token-ratio of 0.03. The 2,789 tokens in the group are only made up of 93 types, corresponding to 29.99 tokens to the type. This is partly due to the fact that only personal pronouns are permitted in subject position. It is also due to a small set of recurrent conjunctions and adverbs: *and* (n=1,047), *but* (n=571), *so* (n=223), *if* (n=147) and *when* (n=104) jointly occur in 75% of the bigrams in this group. In terms of chunking, this group is interesting because it simultaneously has the highest mean lexical gravity G, the lowest backwards transitional probability in the dataset and a very low mean mutual information score. This means that some measures rate these bigrams as very cohesive, while other measures indicate the opposite.

5.2.5.3 Subject & Finite Verb

This is the only type of transition which features in all clause types. It shows the typical properties of a function-word-content-word sequence – low direct and high backwards transitional probabilities – resulting from the fact that a small group of closed-class items is followed by a large group of open-class items.

Almost half the tokens in this group (3,027 out of 6,317) start with *I*, owing to the medium and task. Switchboard participants were encouraged to talk about their personal opinion on a given topic, which leads to frequent use of the first person

pronoun. The most frequent member of the group, however, is *it's* (n=580; frequency in the corpus = 6,717), followed by *I think* (n=387), *I do* (n=381) and *I'm* (n=339).

5.2.5.4 Finite Verb & *not*

348 out of the 565 data-points in this group are *don't*. This is visible in the boxplots (see Appendix E) where the third quantile is collapsed and coincides with the median and there are no top whiskers or outliers. Overall, the group has a very low type/token ratio of 0.05 (on average 26 tokens per type). This is not only due to the large proportion of *don't*, but also to the fact that only auxiliaries can function as the operator in negated finite verb phrases with *not*. Of course, the direct transitional probability of *ca n't* and *wo n't* is 100%, due to the fact that the forms *ca* and *wo* only exist in this combination.

5.2.5.5 *not* & Non-finite Verb

In this group, the verbs occurring with *not* are mostly from the large open class of full verbs. *n't know* is the token with the highest corpus frequency (1,498), the highest direct transitional probability (0.16) and the highest lexical gravity G (12.08). Combinations of *not* and a lower-frequency verb reach the highest backwards transitional probabilities and mutual information scores, indicating that these verbs are statistically likely to be used in negative constructions without intervening adverbs. (139) lists those pairs which reach a backwards transitional probability of 100%.

(139) n't job n't encompass not persecuted n't foul

5.2.5.6 Finite Verb & Non-finite Verb

This final set of transitions is characterised by extreme variability, which is evident in the high type/token ratio of 0.46. On average, there are 2.16 tokens to the type (1,149 tokens; 532 types). The bigrams in this group are generally infrequent, as indicated by the low mean bigram frequency showing in Figure E.1 in the Appendix. The pair with the highest corpus frequency in the set is *'ve got* (frequency in the corpus = 555); *kept dismissing* carries the highest mutual information score (13.15), owing to the fact that *dismissing* only occurs once in the corpus and that *kept* is not very frequent either (frequency in the corpus = 86). The future marker *'re going* receives the highest lexical gravity G (11.7). Due to the high degree of variation, the set is dominated by hapax legomena, an exemplary selection of which is listed in (140). In many cases, the low frequency is surprising as the first as well as the second element are frequent

individually; in these cases, hapax status must result from the small size of the Switchboard NXT corpus.

(140)

am calling	like driving	should expect
can guess	‘ll weed	tried reading
did compete	‘m hiring	was mating
enjoy walking	might try	were shooting
have screened	‘re burning	would crawl

5.3 Previously Suggested Factors

This section provides a brief overview of factors which have been previously suggested as determinants of hesitation placement. Most studies cited are based on a different set of hesitations (mostly filled and unfilled pauses). They are referred to here in order to show that their findings can be generalised to describe the behaviour of fillers and discourse markers in the present dataset (for more information on the cited studies and other proponents of these claims see Section 2.3.2). I will show, though, that no single factor explains the given variation and nor do all structural factors taken together. Rather, some of them offer competing explanations for placement in a certain position or make contradictory predictions.

The sentence boundary attracts hesitations (cf. Cook 1971; Clark and Clark 1977; Shriberg 1994; Biber et al. 1999). – Overall, hesitations are more likely to be placed at the sentence boundary than within the sentence. Of 6,317 hesitations, 4,133 are placed at the sentence boundary, corresponding to 65.4% placed at the boundary versus 34.6% within the structures. This difference is highly significant ($p<.001$ based on a chi-square test) regardless of whether we expect a 50/50 distribution or whether we take into consideration that there are actually 22 transitions within the structures and only eight at sentence boundaries. However, speaker behaviour differs considerably from structure to structure: while in simple 'Subject Verb(finite)' structures 98.4% of hesitations are placed before the sentence, only 18.2% in 'SE Subject Verb(finite) *not* Verb(non-finite)' structures are.

The noun phrase boundary attracts hesitations (cf. Maclay and Osgood 1959; Clark and Clark 1977; Goldman-Eisler 1968; Bortfeld et al. 2001). – The vast majority of hesitations (81.2%) is placed before the subject. Particularly if there are no SEs and thus the pre-subject position coincides with the sentence boundary there is almost no variation in placement. In these cases, 97.2% of hesitations are placed before the subject. This suggests that the factors 'sentence boundary' and 'noun phrase boundary' interact; if both boundaries coincide, speakers virtually always hesitate before the sentence.

The transition after the first SE attracts hesitations (cf. Cook 1971:138; Holmes 1988; Altenberg 1998; Jurafsky et al. 1998). – The second most frequently used position for hesitating is after the first SE, no matter what follows. 67.77% of hesitations which occur in a context containing at least one SE are placed after the first (or only) SE. In sentences with a single SE, hesitation placement after the first SE necessarily coincides

with placement before the subject. Thus there is some overlap between these two positions. (For an in-depth analysis of this phenomenon see Section 5.6.)

The verb phrase boundary repels hesitations (cf. Cook 1971; Maclay and Osgood 1959:31). – The verb phrase boundary is strongly dispreferred: across all 6,317 data-points, only 1.8% of hesitations are placed before the finite verb. While this phenomenon has been described for full NP subjects elsewhere, we can assume that the effect is stronger for 'Pronoun + Verb' combinations.

Hesitations within the verb phrase are dispreferred (cf. Cook 1971:138). – Out of 1,714 hesitations occurring in structures containing a complex verb phrase – and thus offering the possibility to place hesitations within the verb phrase – only 96 (5.6%) are actually placed inside it, i.e. preceding *not* or the non-finite verb. Still, there is a placement pattern in complex verb phrases. In negated ones, there are significantly fewer hesitations before the non-finite verb and significantly more hesitations before the subject than in non-negated ones (both significant at the $p<.001$ level, based on a chi-square test). This means that in negated verb phrases, hesitations tend to be shifted to the position before the subject, so that the subject, finite verb, *not* and non-finite verb are produced as one unit.

Sentence-initial contexts attract discourse markers (cf. Schiffrin 1987:328; Fraser 1990:389; Biber et al. 1999:1086; Saz Rubio 2007:76). – Discourse markers constitute 55.22% of data-points – a stark contrast to the prepositional phrase dataset, where only 13.15% of data-points were discourse markers.

Some of these structural factors offer different explanations for placement in the same position. For example, 'noun phrase boundary' and 'after the first SE' both predict (141), or make competing predictions, e.g. 'noun phrase boundary' predicts (142) while 'after the first SE' predicts (143).

(141) SE *uh* Subject Verb(finite)
(142) SE SE *uh* Subject Verb(finite)
(143) SE *uh* SE Subject Verb(finite)

Only if factors are ranked can they predict the observed pattern (e.g. 'after the first SE' must rank higher than 'noun phrase boundary' because we find more 'SE *uh* SE Subject Verb(finite)' than 'SE SE *uh* Subject Verb(finite)'). However, this kind of ranking as well as some of the factors themselves (e.g. 'speakers tend to place hesitations after the first SE') are structurally unmotivated and therefore have no explanatory power. This

leads to the conclusion that structural factors provide fairly accurate descriptions about where hesitations are placed, but offer no explanation as to why they are placed where they are. In short, they can predict placement but they cannot explain variation.

Probabilistic arguments, on the other hand, should be more apt to explain variation. While structural arguments rely on generalisations along the lines of 'noun-phrase boundaries (generally) attract hesitations', a probabilistic analysis takes into account the characteristics of each individual transition. Hence a claim along probabilistic lines of argumentation is, for example, 'hesitations are placed where the transitional probability is lowest'. Unfortunately, this claim, first put forward by Lounsbury (1954), which was addressed in some detail in Section 4.3, cannot be tested here. Applying this hypothesis to data requires that we know the transitional probabilities of all transitions in the given context. In the case of sentence-initial structures, however, no transitional probabilities can be given for the hesitation position before the first word in the sentence as there is, per se, no word before the first word in a sentence. We could opt for claiming that this means that the transitional probability at this position is zero. Yet this would be an oversimplification, because sentence adverbs and coordinating conjunctions in particular are only used at the beginning of a sentence and are therefore expectable in this location.

5.4 Analyses by Structure

This section provides analyses of chunking in the eight selected sentence-initial structures. Analyses are based on the predictors listed in Section 5.2.4 above and described in more detail in Section 3.3.1. They deal with hesitation placement and ultimately chunking in sentence-initial contexts. The regression methods employed for this purpose are CART trees and random forests which test whether there are correlations between the dependent and independent variables. In this case, they test whether the placement of hesitations can be predicted from the associations between the words in the surrounding context. To this purpose, the statistical algorithm divides the data into non-overlapping subgroups. The resulting groups will be homogenous in at least two respects. For example, sentences containing a highly frequent subject-verb combination will only be separated from the rest if the hesitation behaviour in the resulting high- and low-frequency groups differs significantly. Both 'daughter' groups will thus be more homogenous than the 'parent' group with respect to both their hesitation behaviour and the frequency-range of subject-verb combinations. The final (or 'terminal') sub-groups resulting from this splitting procedure can be further analysed both quantitatively and qualitatively.

The method is explained in detail in Section 3.3.3, which also explains how the performance of these models is statistically evaluated. Furthermore, the analysis of hesitation placement in 'Preposition Noun' phrases in Section 4.4.1 serves as a model analysis where every step taken is commented. In this chapter, only the immediately relevant information about the method will be briefly repeated.

Section 5.4.9 provides a summary of results. An overall quantitative evaluation of the prediction strength of the selected measures of association follows in Section 5.5. The remaining sections in this chapter provide analyses building on the findings from Section 5.4.

Table 5.5 explains the labelling of hesitation positions, words and transitions. Across all clause types, transitions are always labelled the same, meaning that, for example, Word 3 is always the subject, despite the fact that it may not necessarily be the third word in a particular type of sentence. In this way, parallels and discrepancies in hesitation placement in different structures are easily discernible.

Element		SE		SE		Subject		Verb(fin)		*not*		Verb(inf)
Hesitation Position	1		2		3		4		5		6	
Word		word 1		word 2		word 3		word 4		word 5		word 6
Bigram/Transition		--------- Bigram 1 ----------				---------- Bigram 3 ----------				---------- Bigram 5 ----------		
				---------- Bigram 2 ----------				---------- Bigram 4 ----------				
								---------- Bigram 6 ---------- excluding *not*; hesitation position in-between = 6				

Table 5.5: Legend to data labelling

5.4.1 Subject Verb(finite)

At 2,576 data-points, this is by far the largest dataset. While 'Subject Verb(finite)' is a common way to begin a sentence, sentences only consisting of a personal pronoun and a simple verb are rare – only 166 sentences in this dataset do not continue after the verb. These are mostly (self-) interrupted utterances like (144) and (145).

(144) *I mean uh* it's (sw2005.A.s187)

(145) *well* she's (sw3099.A.s203)

There is extremely little variation in hesitation placement in this context (see Figure 5.1). This lack of variation owes to the fact that the first word in the sentence is the subject; hence a hesitation placed before the subject is simultaneously placed at the sentence boundary and so the two preferred positions for hesitation placement coincide (see Section 5.3). Consequently, 98.41% of hesitations in this set are placed at the sentence boundary. Only five discourse markers and 36 filled pauses are placed before the verb. This means that only a marginal 1.59% of hesitation behaviour deviates from the norm.

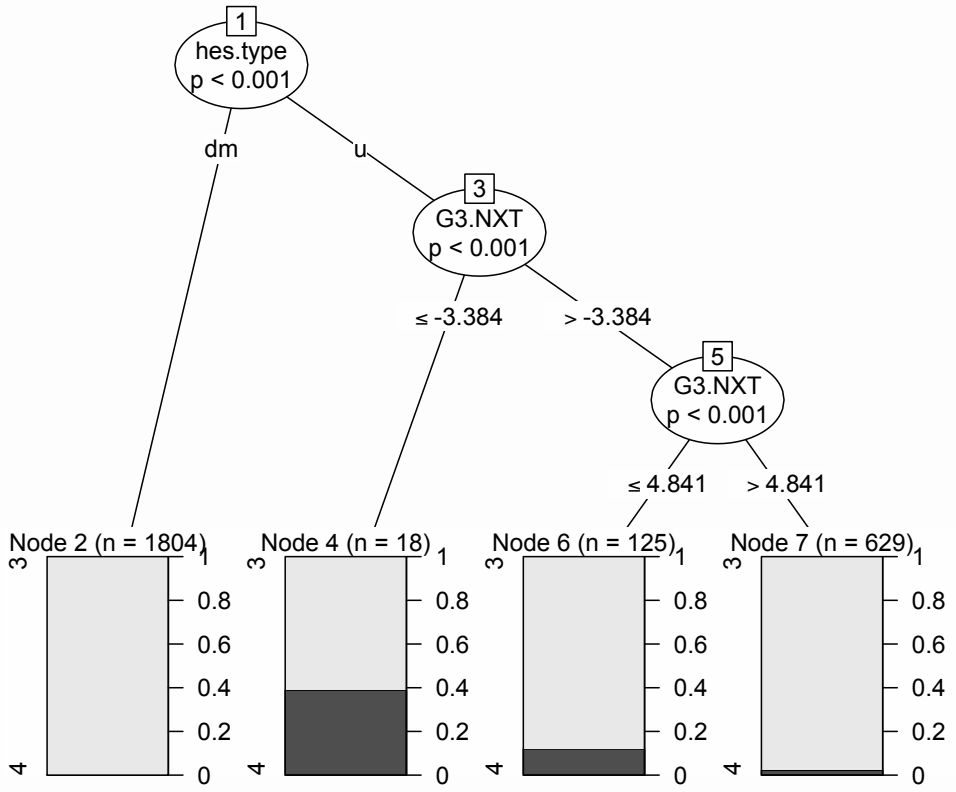
Crucially, *ctree* does actually grow a tree for the data. Trees are only grown if the algorithm finds significant effects for at least one of the predictors. Figure 5.2 shows the resulting tree. The first split in the tree is based on hesitation type, separating the discourse markers from the pause fillers *uh* and *um*. None of the frequency-based predictors has an effect on the placement of discourse markers, which are almost exclusively placed at the phrase boundary.

The three terminal nodes for pause fillers, however, show a gradual effect of lexical gravity G: hesitation placement before the verb is generally unlikely, yet the stronger the attraction between the subject and the verb, the less likely the subject-verb sequence is interrupted by filled pauses. Where subject and verb strongly repel (Node 4), 38.89% of filled pauses are placed before the verb, while only 2.23% of hesitations are placed before the verb where subject and verb strongly attract (Node 7). (146) to (149) show four cases of hesitations placed before the verb; the former two are taken from Node 4, the latter from Node 7. These examples show that the less attracted combinations generally contain low-frequency verbs while the verb in strongly attracted pairs is a frequent form.

(146) they *[pause] uh [pause]* auctioned [some tools and things like that] (sw2837.A.s98)

(147) *you know* I *uh* camped [in the boy scouts] (sw3750.B.s13)

(148) I *uh [pause]* get [a lot of my news driving home from work ...] (sw4345.B.s1)



List of Abbreviations			
		Word Frequencies	**Bigram Measures**
w.freq	Word Frequency	w3 Subject	bi3 Subject + Verb(fin)
bi.freq	Bigram Frequency	w4 Finite Verb	
TPD	Direct Transitional Probability		
TPB	Backwards Transitional Probability		
MI	Mutual Information Score		
G	Lexical Gravity G		

Figure 5.2: Ctree results for the structure 'Subject Verb(finite)'. Labels at terminal node bar graphs (here: 3, light, and 4, dark) indicate hesitation position before the corresponding words (w3=Subject; w4=Finite Verb).

(149) *well* I *uh* am [an assistant teacher and in business technology] (sw3450.B.s9)

Overall, the performance of the *ctree* model is poor, as it fails to find a condition under which placement before the verb is the preferred option. Thus all terminal nodes make the same prediction: hesitations are placed before the pronoun (see Table 5.6).

		Model Predictions		
	Hesitation Position	**pre Subj (3)**	**pre V(fin) (4)**	**Total**
Actual Distribution	**pre Subj (3)**	2,535	0	2,535
	pre V(fin) (4)	41	0	41
	Total	2,576	0	2,576

Table 5.6: Performance of ctree model for 'Subject Verb(finite)'. Corresponds to cforest performance and cforest out-of-bag predictions (ntree=2,000, mtry=5, seed=813).

ctree performance is evaluated by comparing the number of correct predictions to the prediction accuracy of a baseline classifier which generalises from the most frequent behaviour to all data-points. The most frequent behaviour in this case being placement before the subject, the baseline model predicts that all hesitations are placed before the subject, resulting in a misclassification rate of 1.59% – the same as the *ctree*'s.

Even a random forest growing 2,000 trees based on different subsets of data-points and predictors does not find conditions under which hesitations are predominantly placed before the verb and the variable importance scores generated by *cforest* emphasise how marginal the frequency effect actually is. Figure 5.3 illustrates that even the strongest effect only receives a score of 0.00015, which is so low that all effects must be considered unstable and non-significant[44]. The fact that bigram frequency is actually ranked highest, while lexical gravity G is rated as noise (i.e. it receives a negative score) must not be interpreted because variable importance scores of insignificant predictors will fluctuate around zero (cf. Strobl, Malley and Tutz 2009b:343).

In summary, in sentence-initial 'Subject Verb(finite)' sequences where the hesitation 'slot' before the subject coincides with the sentence boundary, there is almost no variation in discourse marker placement (they are almost exclusively placed at the boundary) and little variation in the placement of filled pauses. A *ctree* analysis suggests that the more attracted the subject and the verb, the less likely it is for filled pauses to be

[44] Strobl et al. (2009b:336) advise against interpreting the absolute values of variable importance scores, as these depend on the characteristics of the forest, i.e. the number of data-points, predictors, trees etc. In this case, however, numbers are so low and vary so much depending on the random seed that such evaluations are warranted.

placed before the verb. Re-evaluation of effect strength with the help of a *cforest*, however, reveals that, overall, any frequency effects are non-significant and unreliable.

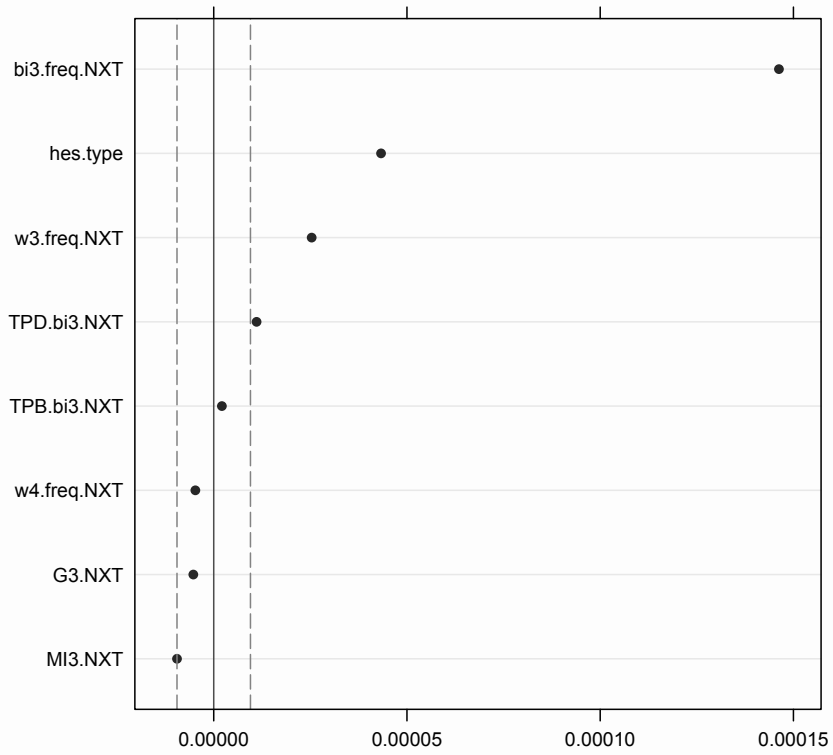


Figure 5.3: Variable importance of predictors for 'Subject Verb(finite)' (mtry=5, ntree=2,000, seed=813, OOB=false, results from R version 2.15.2/3.0.0).

5.4.2 Subject Verb(finite) Verb(non-finite)

Hesitation behaviour in the context of sentence-initial 'Subject Verb(finite) Verb(non-finite)' structures is very much in line with the kind of behaviour detailed in the previous section. As the subject is the first word in the sentence, hesitations are predominantly placed at the sentence boundary. Out of 612 hesitations, only a mere seven discourse markers and 41 pause fillers (7.8% of hesitations) are placed within the structure, leaving very little scope for the frequency-based predictors.

Interestingly, when placing hesitations within the sentence, speakers interrupt the verb phrase rather than place hesitations at the verb-phrase boundary. This is also reflected in the result of the *ctree* model (Figure 5.4). The only frequency-based predictor selected by the model is the attraction (i.e. lexical gravity G) between the verbs (Splits 2 and 5 in Figure 5.4), indicating that the more attracted the two parts of the verb phrase, the less likely the phrase is to be interrupted by hesitations. While different cut points are chosen (i.e. 1.746 in Split 2 and -1.73 in Split 5), this effect holds for both discourse markers and pause fillers, though it is stronger for pause fillers. (150) to (153) show exemplary members of each of the four terminal nodes.

(150) you should *uh* pursue [that I think ...] (sw2121.A.s216)

(151) *um* I've been [out of Texas about ten years] (sw2938.B.s67)

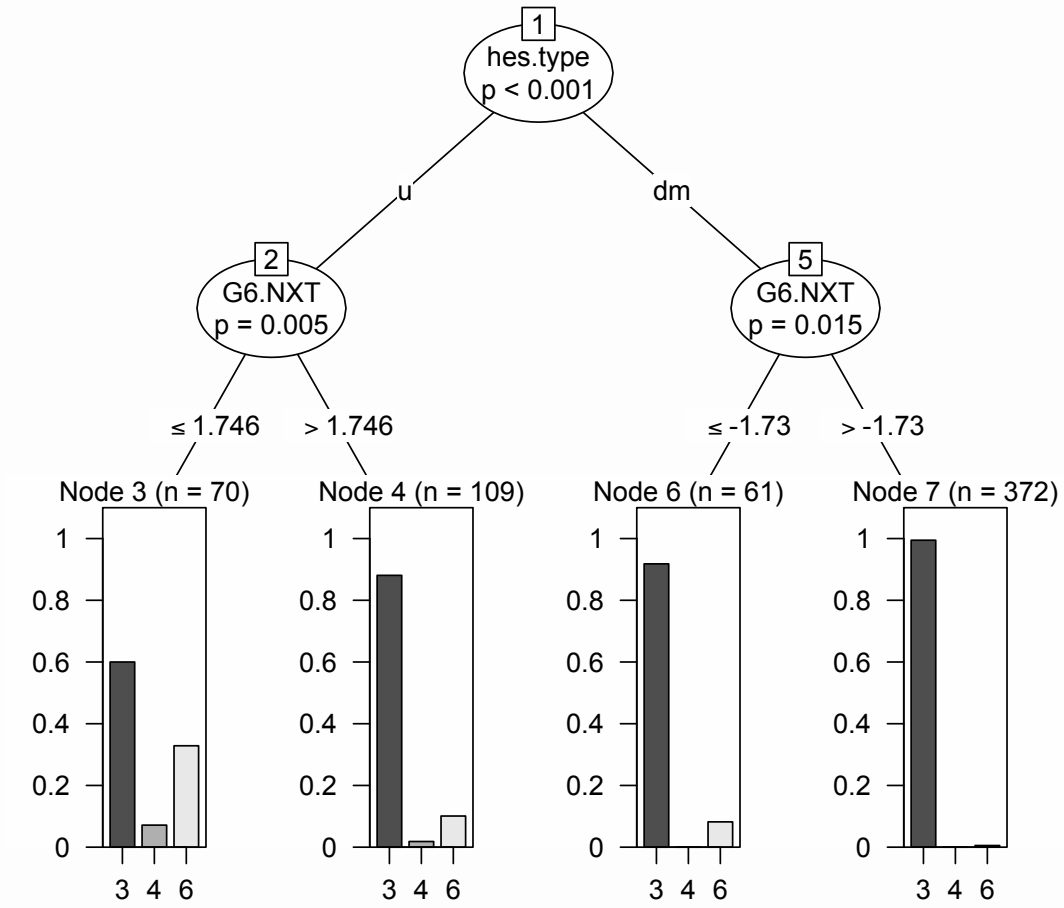
(152) *I mean* I was astounded (sw2441.B.s103)

(153) *well* I'm trying [to get back in shape for softball this spring] (sw2229.A.s25)

Typically, structures in Nodes 3 and 6 – corresponding to (150) and (152) – have infrequent non-finite verbs, while those in Nodes 4 and 7 contain enclitic finite and frequent non-finite verbs – see (151) and (153). Nevertheless, *ctree* again fails to find conditions under which hesitation placement at the phrase boundary is not the preferred option (see Table 5.7). Consequently, model performance does not exceed the performance of the baseline model.

		Model Predictions			
	Hesitation Position	**pre Subj (3)**	**pre V(fin) (4)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre Subj (3)**	564	0	0	564
	pre V(fin) (4)	7	0	0	7
	pre V(inf) (6)	41	0	0	41
	Total	612	0	0	612

Table 5.7: Performance of ctree model for 'Subject Verb(finite) Verb(non-finite)'



List of Abbreviations

		Word Frequencies	**Bigram Measures**
w.freq	Word Frequency	w3 Subject	bi3 Subject + Verb(fin)
bi.freq	Bigram Frequency	w4 Finite Verb	bi6 Verb(fin) + Verb(inf)
TPD	Direct Transitional Probability	w6 Non-finite Verb	
TPB	Backwards Transitional Probability		
MI	Mutual Information Score		
G	Lexical Gravity G		

Figure 5.4: Ctree results for the structure 'Subject Verb(finite) Verb(non-finite)'. Labels at terminal node bar graphs (here: 3, 4 and 6) indicate hesitation position before the corresponding words (w3=Subject; w4=Finite Verb; w6=Non-finite Verb).

Cforest predictions hardly differ. In only two cases does *cforest* accurately predict hesitation placement before the non-finite verb, which is not a significant improvement compared to a baseline model (based on a chi-square-test).[45] The variable importance scores (see Figure 5.5) further emphasise the small effect size. At a variable importance score of 0.009, even the highest ranked effect – that of hesitation type – is not strong. Most of the other predictors have non-significant effects, as indicated by their scores which vary around zero.

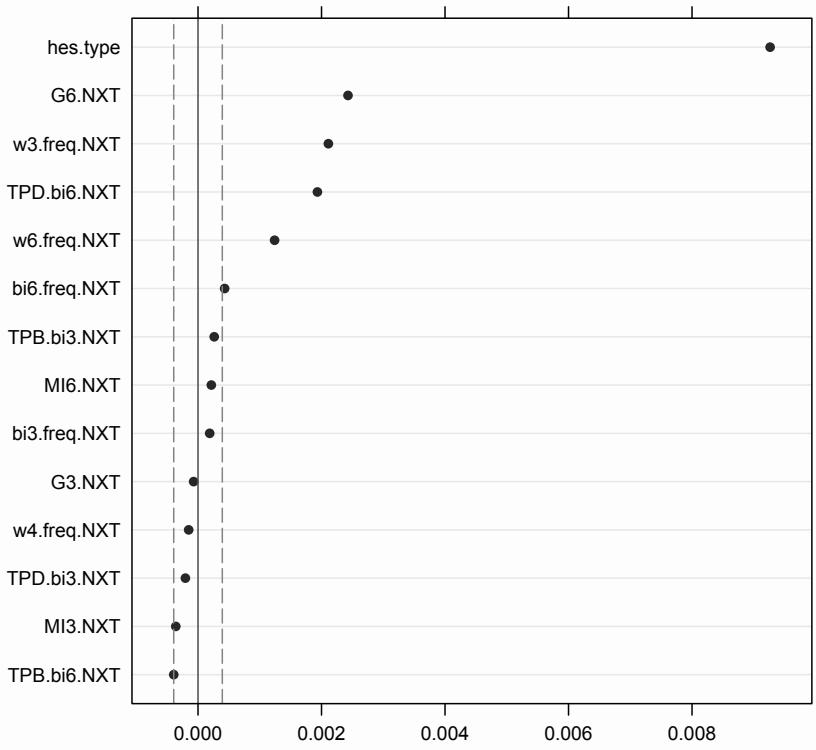


Figure 5.5: Variable importance of predictors for 'Subject Verb(finite) Verb(non-finite)' (mtry=5, ntree=2,000, seed=95, OOB=false, results from R version 2.15.2/3.0.0)

[45] For tables with exact *cforest* and out-of-bag predictions see Appendix G.

5.4.3 Subject Verb(finite) *not* Verb(non-finite)

The 'Subject Verb(finite) *not* Verb(non-finite)' dataset is another case in point of the influence of structural factors on hesitation placement. Of a total of 340 hesitations only ten (2.9%) deviate from the norm, i.e. are not placed at the sentence boundary (see Table 5.8). Due to the fact that this pattern is observable in all structures where the sentence boundary coincides with the noun phrase boundary, i.e. where hesitation placement before the sentence coincides with placement before the subject, this pattern is likely to be caused by this factor.

This dataset is characterised by frequent multi-word expressions which make up the entire structure. 106 data-points are *I don't know*, a further eight are variants with *did* or non-enclitic *do not*, seven data-points are *I don't/can't/do not/cannot remember* and 15 are *I don't think*. Due to their high frequency, these multi-word units might be less likely to be interrupted than other sequences. Semantically they themselves work as markers of processing or retrieval problems, as the speaker explicitly mentions that he does not remember something.

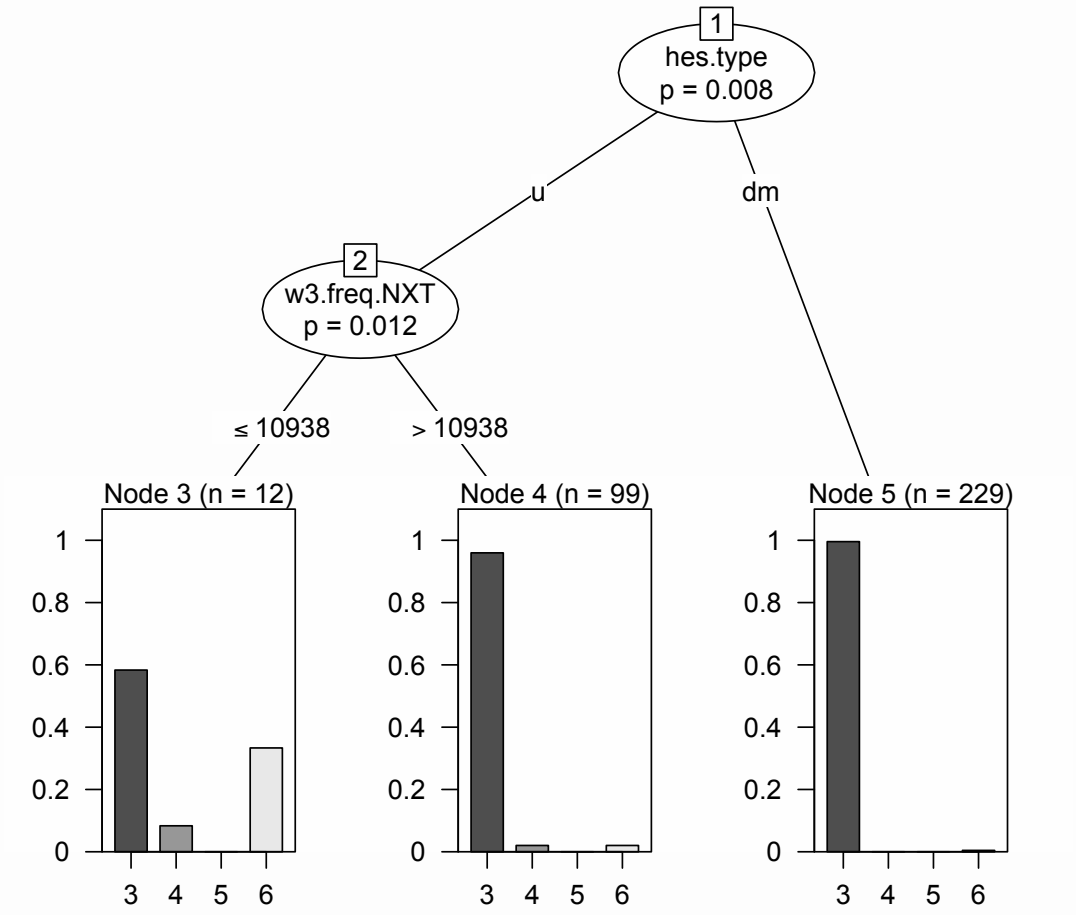
A *ctree* model fitted to the data grows a tree (see Figure 5.6), despite the fact that the placing of only ten data-points deviates from the norm and can thus be explained by the model. As a result, the tree unsurprisingly fails to find conditions where hesitation placement at a location other than the phrase boundary is preferred. Consequently, overall model performance does not exceed baseline performance.

		Model Predictions				
	Hesitation Position	**pre Subj (3)**	**pre V(fin) (4)**	**pre *not* (5)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre Subj (3)**	330	0	0	0	330
	pre V(fin) (4)	3	0	0	0	3
	pre *not* (5)	0	0	0	0	0
	pre V(inf) (6)	7	0	0	0	7
	Total	340	0	0	0	340

Table 5.8: Performance of ctree model for 'Subject Verb(finite) not Verb(non-finite)'. Corresponds to cforest performance and cforest out-of-bag predictions (ntree=2,000, mtry=5, seed=783).

Splits in the tree (Figure 5.6) will not be discussed here, because random forest results clearly refute that any of the measures of attraction are predictive of hesitation placement. Figure 5.7 shows that variable importance measures for all predictors are extremely low and effects are non-significant and unstable.

In conclusion, frequency of words and word-pairs, and attraction between words have no influence on hesitation placement in this context. Hesitations are consistently placed at the sentence boundary, which coincides with placement before the subject.



List of Abbreviations

		Word Frequencies	**Bigram Measures**
w.freq	Word Frequency	w3 Subject	bi3 Subject + Verb(fin)
bi.freq	Bigram Frequency	w4 Finite Verb	bi4 Verb(fin) + *not*
TPD	Direct Transitional Probability	w5 *not*	bi5 *not* + Verb(inf)
TPB	Backwards Transitional Probability	w6 Non-finite Verb	
MI	Mutual Information Score		
G	Lexical Gravity G		

Figure 5.6: Ctree results for the structure 'Subject Verb(finite) not Verb(non-finite)'. Labels at terminal node bar graphs (here: 3, 4, 5 and 6) indicate hesitation position before the corresponding words (w3=Subject; w4=Finite Verb; w5=not; w6=Non-finite Verb).

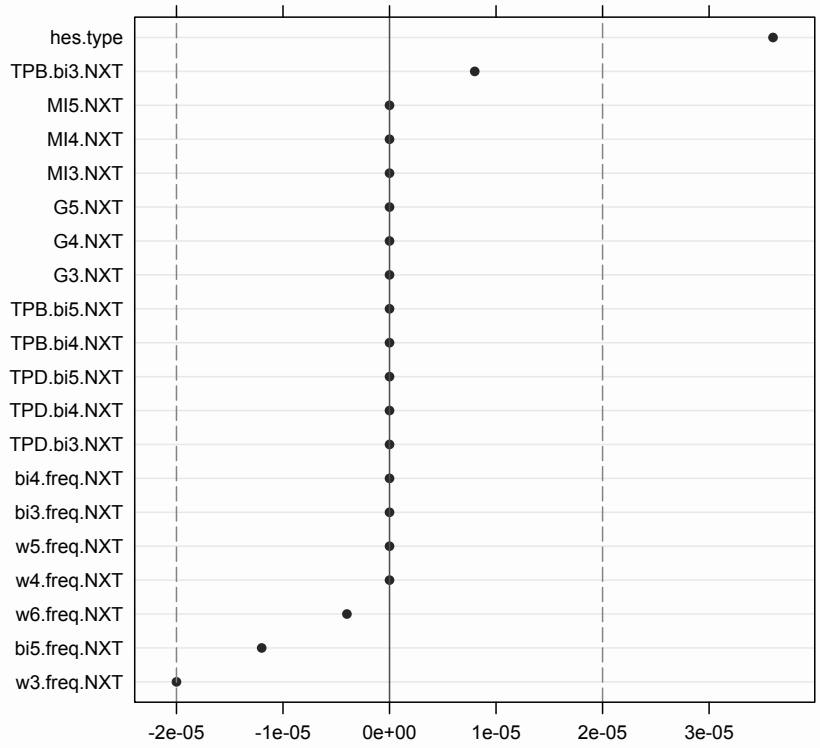


Figure 5.7: Variable importance of predictors for 'Subject Verb(finite) not Verb(non-finite)' (mtry=5, ntree=2,000, seed=783, OOB=false, results from R version 2.15.2/3.0.0).

5.4.4 SE Subject Verb(finite)

'SE Subject V(finite)' and the following datasets differ from the previous ones in that the subject is not the first word in the sentence, leading to competition between hesitation placement at the sentence boundary and placement before the subject. This competition results in far more variation than in the 'non-competition contexts' (see Section 5.4.1 to Section 5.4.3). Thus there is much more variation in this and the following datasets than in the previous ones.

In this dataset, 449 hesitations are placed at the sentence boundary, 1,171 before the subject and 40 before the verb. Hence, hesitating before the subject is statistically the unmarked case, with 29.46% of data-points deviating from this pattern.

A *ctree* fitted to the data is extremely successful at predicting speakers' behaviour (see Table 5.9). It produces the correct outcome in 1,390 cases, corresponding to a misclassification rate of 16.27%. By contrast, the baseline model only predicts 1,171 outcomes correctly (misclassification rate: 29.46%). The difference in performance is highly significant ($p<.001$; residuals: 6.4 and -9.9).

		Model Predictions			
	Hesitation Position	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**Total**
Actual Distribution	**pre SE (2)**	300	149	0	449
	pre Subj (3)	81	1,090	0	1,171
	pre V(fin) (4)	9	31	0	40
	Total	390	1,270	0	1,660

Table 5.9: Performance of ctree model for 'SE Subject Verb(finite)'

Figure 5.8 shows the resulting tree, which is highly interesting because the deciding factor for hesitation placement is the frequency of the SE (see Split 1). If the SE is highly frequent, hesitations are placed after it, while they are predominantly placed before it if the SE is less frequent. This effect could be interpreted as counter-evidence to a theory of chunking because it suggests that hesitation placement is not predominantly determined by the association strength between words, but by the frequency of individual words. Additionally, the effects of Split 9 are apparently incompatible with the concept of cognitively represented chunks; the stronger the subject is attracted to the verb, the more often hesitations are placed before the subject. A chunking theory would generally expect the opposite – the stronger the attraction between words, the higher the chunking strength and the *less* likely are hesitations between them.

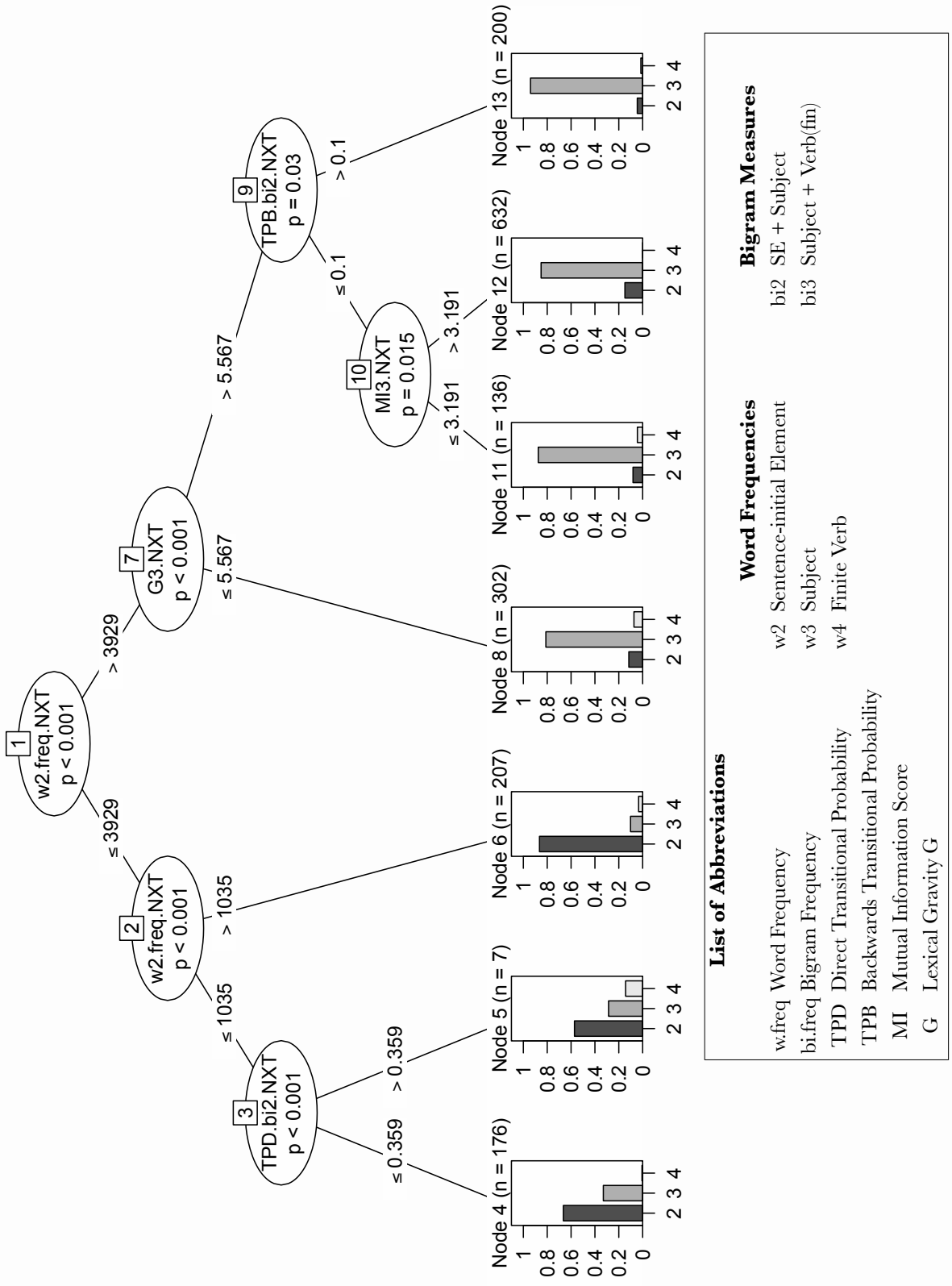


Figure 5.8: Ctree results for the structure 'SE Subject Verb(finite)'. Labels at terminal node bar graphs (here: 2, 3 and 4) indicate hesitation position before the corresponding words (w2=Sentence-initial element; w3=Subject; w4=Finite Verb).

Crucially, however, there are indications that these effects can be interpreted as evidence of chunking. Here, certain SEs appear to be stored together with the following hesitation. This can be deduced from the first split which, based on the frequency of the SE, separates a small group of SEs from the rest to form the right part of the tree. Only *and*, *but*, *or*, *so*, *oh, uh-huh* and *that* reach a corpus frequency of 3,929 or above. *Or*, *uh-huh* and *that* are actually rare in the present dataset (combined n = 12), so they can be ignored. There are indications that the remaining set of semantically unspecific elements is used in fixed combinations with certain hesitations. In this dataset, we find, for example, only 30 instances of *uh and*, but 462 cases of *and uh*; there are no instance of *uh so*, but 61 cases of *so uh*. Node 13 indicates that this effect may be strongest for *and uh*, as this node, consisting entirely of sentences beginning in *and,* is the most homogeneous in terms of hesitation behaviour. Thus 'odd' Split 9 is explained: it separates a set of *and*-sentences from the other data-points.

If certain SEs do indeed form chunks with hesitations, then these chunks should also be evident in the trees of the following structures. A more detailed quantitative analysis of this phenomenon is postponed until Section 5.6.

Finally, a *cforest* of 2,000 trees is fitted to the data. At 1,404 correct predictions (misclassification rate: 15.42%), the forest confirms that the model performs highly significantly above baseline ($p<.001$; residuals: 6.81 and -10.54).

Cforest furthermore offers the possibility to test whether results can be generalised. It accomplishes this by applying its predictions to previously unconsidered data. This is possible because each tree in the forest is based on a random subset of data-points, so that there is always a second subset, the out-of-bag data, which the algorithm did not consider when growing the tree. This data can be used to check whether predictions hold for unseen data and are therefore reliable (for more information, see Section 3.3.3.2). The out-of-bag misclassification rate is 15.78%, thus even conservative out-of-bag predictions confirm that effects are highly significant ($p<.001$, based on a chi-square test; residuals: 6.63, -10.27).[46]

Importantly, *cforest* variable importance scores, shown in Figure 5.9, confirm that SE frequency (w2.freq) is by far the most important predictor in this model. Interestingly though, any predictor describing association strength between the subject and the verb (i.e. relating to Bigram 3) is rated as non-significant, while the relation between the SE and the subject (i.e. Bigram 2) is shown to have some influence on hesitation placement. Furthermore, while the previous models always made a distinction between discourse markers and pause fillers, in this case hesitation type is rated very low.

---

[46] For an overview of exact *cforest* and out-of-bag predictions see Tables I.1 and I.2 in Appendix I.

Overall, these are very strong models with a high predictive accuracy. Both *ctree* and *cforest*, however, rely heavily on a single strong predictor: the frequency of the SE. Therefore, model results can only be interpreted as evidence in favour of multi-word chunking if the following models continue to provide consistent evidence that SEs and hesitations can form chunks. (SE-hesitation chunks will also be investigated in Section 5.6.)

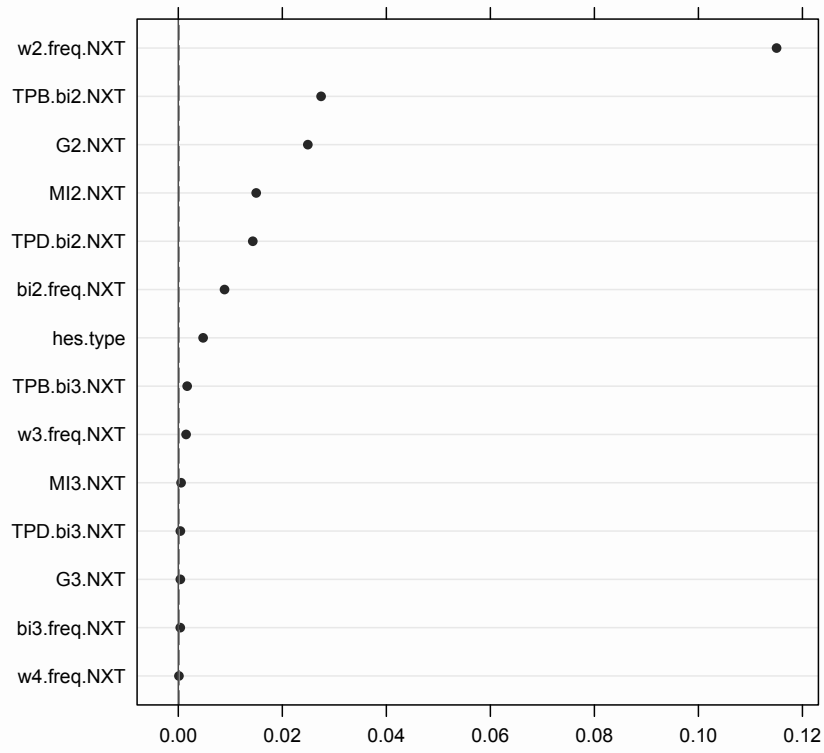


Figure 5.9: Variable importance of predictors for 'SE Subject V(fin)' (mtry=5, ntree=2,000, seed=777, OOB=false, results from R version 2.15.2/3.0.0).

5.4.5 SE Subject Verb(finite) Verb(non-finite)

Hesitation placement in this structure conforms to the pattern displayed in the previously analysed dataset. Of 429 hesitations, 268 are placed before the subject and 116 are placed at the sentence boundary. Only 45 are placed within the verb phrase.

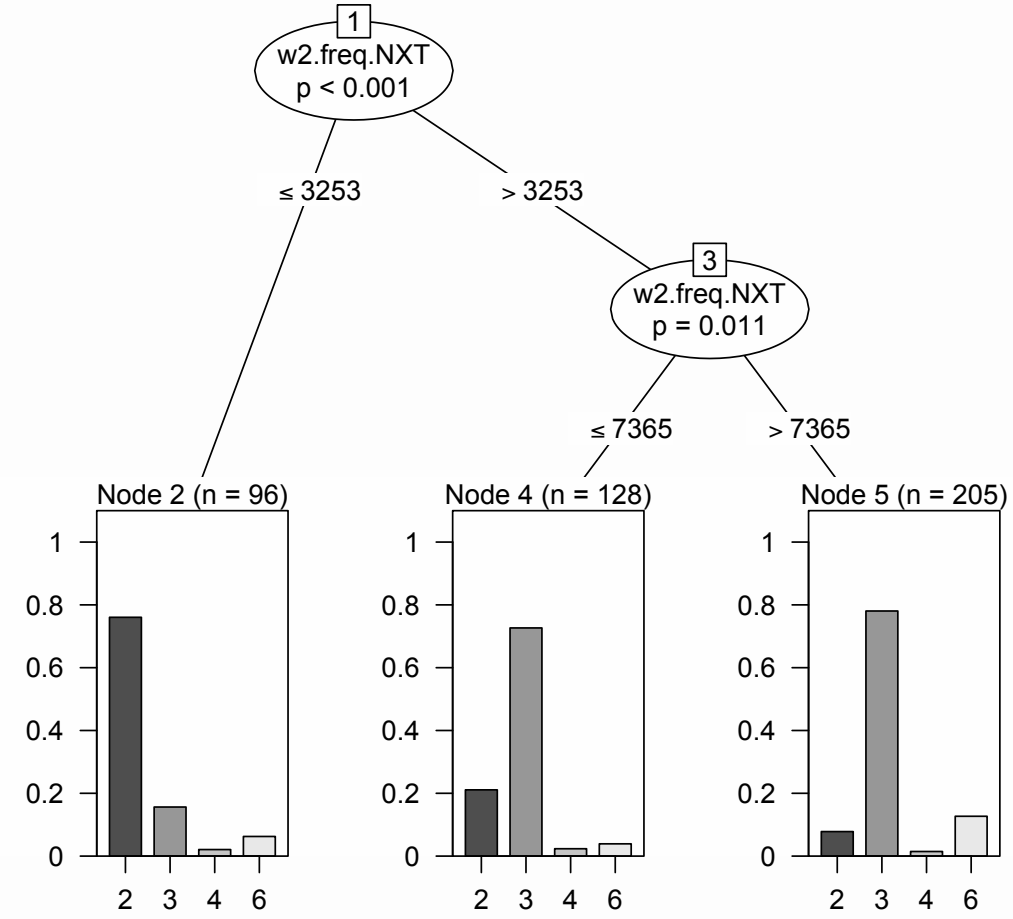
A *ctree* model fitted to the data predicts 326 outcomes correctly, corresponding to a misclassification rate of 24.01%. The corresponding baseline model, in turn, produces 268 correct predictions (misclassification rate: 37.53%). The *ctree* result highly significantly exceeds this baseline ($p<.001$; based on a chi-square test; residuals: 3.54 and -4.57). The summary of model predictions in Table 5.10 shows that the model finds conditions where placement before the sentence or before the subject is preferred, yet fails to determine circumstances under which hesitations are moved into the verb phrase.

		Model Predictions				
	Hesitation Position	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre SE (2)**	73	43	0	0	116
	pre Subj (3)	15	253	0	0	268
	pre V(fin) (4)	2	6	0	0	8
	pre V(inf) (6)	6	31	0	0	37
	Total	96	333	0	0	429

Table 5.10: Performance of ctree model for 'SE Subject Verb(finite) Verb(non-finite)'

An analysis of the tree (see Figure 5.10) reveals two gradual frequency effects. In the terminal leaves, the proportion of hesitations placed at the sentence boundary decreases from left to right, while the proportion of hesitations before the subject increases. This shift in proportions can be analysed as a gradual effect because both decisions in the tree are based on the same predictor, i.e. the frequency of the SE. The directionality of the effect is the same as in the model for 'Subject Verb(finite) Verb(non-finite)', namely the more frequent the SE, the greater the chance of hesitations following rather than preceding it.

The fact that splits are only based on SE frequency means that each terminal node contains data-points beginning with a different set of SEs. A particular type of SE will be assigned to one terminal node only. Node 2, despite being the smallest, contains the greatest variety of SEs because only highly frequent SEs are permitted in the other nodes. (154) lists the range of SEs in Node 2 and (155) shows a typical data-point from this node.



List of Abbreviations

		Word Frequencies		**Bigram Measures**	
w.freq	Word Frequency	w2	Sentence-initial Element	bi2	SE + Subject
bi.freq	Bigram Frequency	w3	Subject	bi3	Subject + Verb(fin)
TPD	Direct Transitional Probability	w4	Finite Verb	bi6	Verb(fin) + Verb(inf)
TPB	Backwards Transitional Probability	w6	Non-finite Verb		
MI	Mutual Information Score				
G	Lexical Gravity G				

Figure 5.10: Ctree results for the structure 'SE Subject Verb(finite) Verb(non-finite)'. Labels at terminal node bar graphs (here: 2, 3, 4 and 6) indicate hesitation position before the corresponding words (w2=Sentence-initial Element; w3=Subject; w4=Finite Verb; w6=Non-finite Verb).

(154)

if_IN (3,253)	also_RB (538)	somewhere_RB (111)
because_IN (2,706)	why_WRB (401)	cause_IN (80)
when_WRB (2,521)	usually_RB (391)	actually_UH (78)
then_RB (2,236)	sometimes_RB (368)	once_IN (77)
there_RB (2,017)	since_IN (295)	unfortunately_RB (72)
now_RB (1,732)	once_RB (259)	anyway_UH (69)
as_IN (1,690)	anyway_RB (238)	plus_CC (68)
how_WRB (1,615)	whether_IN (219)	originally_RB (60)
where_RB (1,452)	because_RB (145)	hopefully_RB (47)
like_IN (1,323)	recently_RB (142)	occasionally_RB (46)
maybe_RB (700)	while_IN (139)	otherwise_RB (42)
actually_RB (616)	unless_IN (120)	plus_IN (1)

(155) *um* as I was sitting [there ...] (sw3215.A.s3)

(155) is a typical example of a data-point in Node 2 in the sense that the hesitation is placed at the sentence boundary (i.e. in Position 2). In contrast, in Nodes 4 and 5 the preferred location to hesitate is before the subject (i.e. in Position 3). (156) and (157) show exemplary data-points from Nodes 4 and 5 respectively. The difference between the two nodes is that Node 4 contains the SEs *but, oh, or* and *so,* while Node 5 exclusively contains *and*.

(156) but *um* I've been [real pleased] (sw2566.B.s280)

(157) and *uh* I'll be [honest with you] (sw3573.B.s10)

The correlation between SE frequency and hesitation behaviour is exactly the same as in the previous dataset. In the case of lower-frequency SEs, hesitation placement before the SE is preferred, while in the case of highly-frequent *and*, *but*, *oh*, *or* and *so*, hesitations are predominantly placed after the SE; this effect is most pervasive for *and*. This can be interpreted as additional evidence that a small group of SEs can be employed as hesitation devices. Combinations of one of these SEs and filled pauses and discourse markers have merged into longer time-buying devices, which are used sentence-initially. This effect is strongest for the most frequent member of the group (i.e. *and*).

In this case, a *cforest* does little more than confirm *ctree* results. At misclassification rates of 23.31% and 24.24% (for exact predictions see Tables J.1 and J.2 in the Appendix), both forest and out-of-bag predictions are almost the same as the *ctree* result, thus confirming the highly significant result. Of particular interest are the variable importance scores shown in Figure 5.11. These exactly repeat findings from the 'SE Subject Verb(finite)' dataset, namely that the frequency of the SE is the decisive factor, that measures relating to the cohesion between the SE and the subject (i.e. Bigram 2)

have some minor influence, and that all other factors have virtually no influence on hesitation placement.

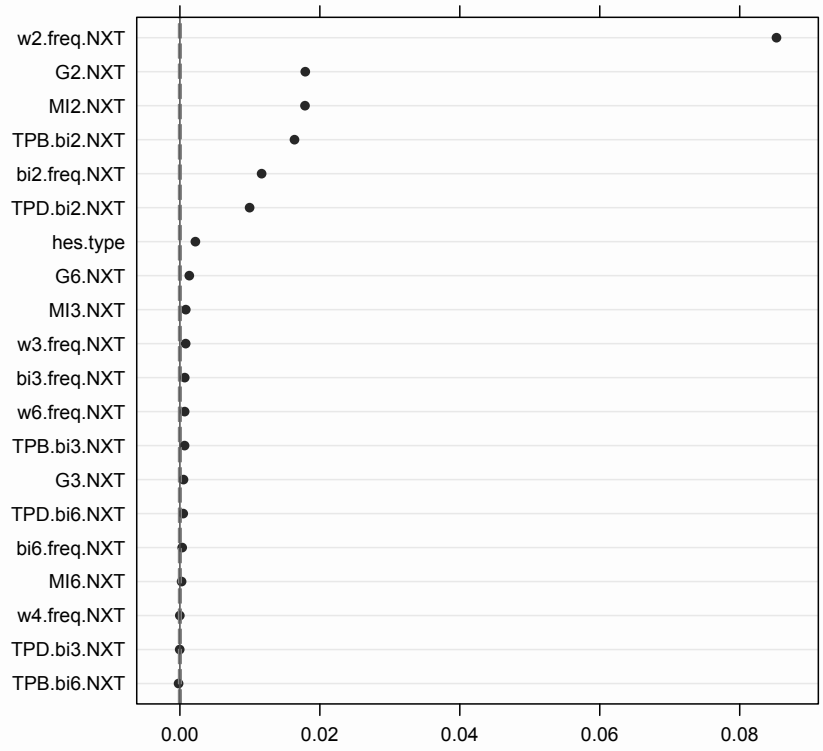


Figure 5.11: Variable importance of predictors for 'SE Subject Verb(finite) Verb(non-finite)' (mtry=5, ntree=2,000, seed=923, OOB=false, results from R version 2.15.2/3.0.0).

5.4.6 SE Subject Verb(finite) *not* Verb(non-finite)

The 'SE Subject Verb(finite) *not* Verb(non-finite)' dataset is dominated by two longer multi-word units which together make up more than half the data-points. Of 225 data-points, 99 are SE + *I don't know* and 15 SE + *I don't think*. This could be the cause for there being a mere six hesitations placed within the verb phrase. The lack of hesitations in the verb phrase is, however, a pattern common to both datasets containing negated verb phrases (see Section 5.4.3).

A *ctree* fitted to the data (see Figure 5.12) only classifies 183 data-points correctly (see Table 5.11). This corresponds to a misclassification rate of 18.67%, a result which does not significantly exceed the performance of the baseline model (based on a chi-square test), which classifies 178 data-points correctly.

		Model Predictions					
	Hesitation Position	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**pre *not* (5)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre SE (2)**	15	26	0	0	0	41
	pre Subj (3)	10	168	0	0	0	178
	pre V(fin) (4)	0	2	0	0	0	2
	pre *not* (5)	0	1	0	0	0	1
	pre V(inf) (6)	0	3	0	0	0	3
	Total	25	200	0	0	0	225

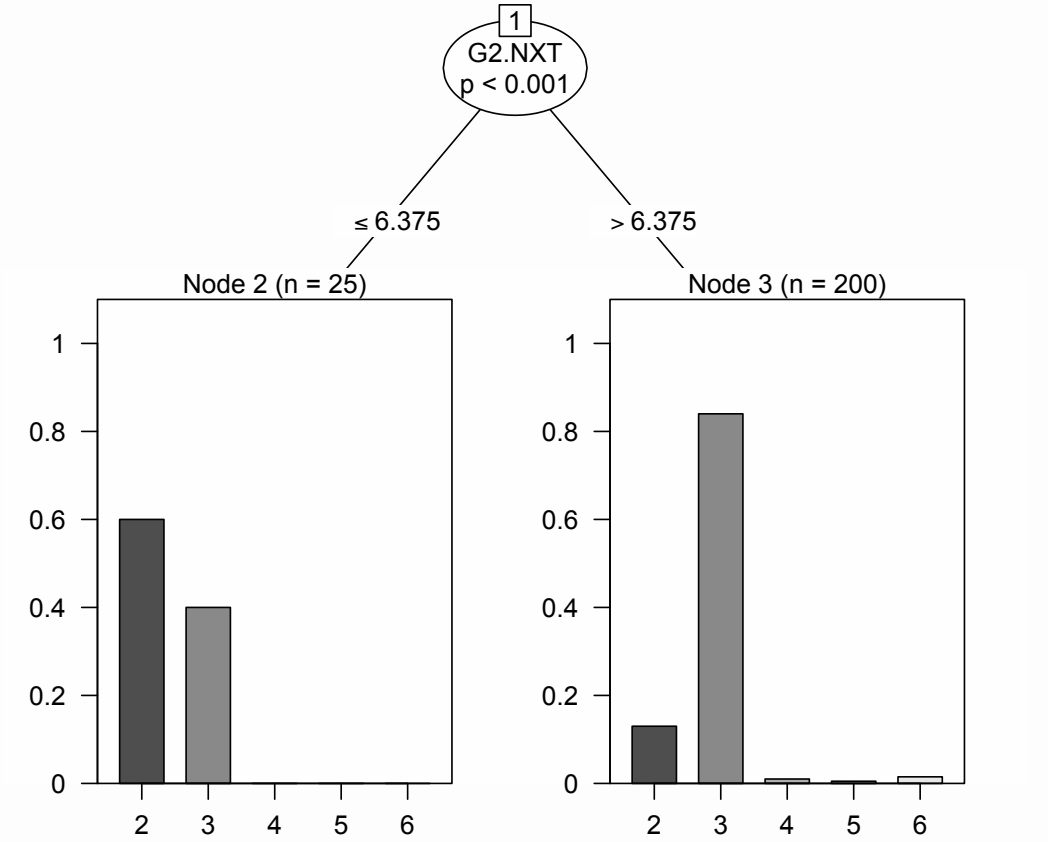
Table 5.11: Performance of ctree model for 'SE Subject Verb(finite) not Verb(non-finite)'

The tree creates only two terminal leaves. The only split appears to contravene chunking principles, as it indicates that the stronger the cohesion between the SE and the subject (i.e. the higher lexical gravity G for Bigram 2), the more likely this pair is to be interrupted by hesitations. An analysis of the data-points in each leaf reveals, however, that the split merely leads to a separation of *so*, *but*, *and, if*, *because, or*, *then* and *oh* from the other SEs.

These high-frequency SEs are assigned to Node 3, where hesitations are predominantly placed after the SE. *and* and *but* jointly make up more than 75% of data-points in this node (n(*and*)=85; n(*but*)=91).

All other SEs, like *personally*, *gosh* or *actually*, as well as a total of six tokens of *or*, *so* and *because* are assigned to Node 2, where hesitations are predominantly placed before the SE. In five of the six higher-frequency tokens in this node, the hesitation follows the SE, as is typical in the case of frequent SEs.

Thus the pattern observed in the previous datasets is repeated: infrequent SEs are preceded by the hesitation, as in (158), an example from Node 2, while frequent SEs are followed by it, as in (159), taken from Node 3.



List of Abbreviations				
		Word Frequencies		**Bigram Measures**
w.freq	Word Frequency	w2	Sentence-initial Element	bi2 SE + Subject
bi.freq	Bigram Frequency	w3	Subject	bi3 Subject + Verb(fin)
TPD	Direct Transitional Probability	w4	Finite Verb	bi4 Verb(fin) + *not*
TPB	Backwards Transitional Probability	w5	*not*	bi5 *not* + Verb(inf)
MI	Mutual Information Score	w6	Non-finite Verb	
G	Lexical Gravity G			

Figure 5.12: Ctree results for the structure 'SE Subject Verb(finite) not Verb(non-finite)'. Labels at terminal node bar graphs (here: 2, 3, 4, 5 and 6) indicate hesitation position before the corresponding words (w2=Sentence-initial Element; w3=Subject; w4=Finite Verb; w5=Non-finite Verb).

(158) *well [pause]* luckily it hasn't gotten [that bad here] (sw4320.B.s10)

(159) and *uh [pause] uh* I can't understand [why anyone would abandon ...] (sw2719.A.s10)

Both the *cforest* and its out-of-bag predictions fail to perform significantly above baseline. Despite the non-significant performance, *cforest* variable importance scores (see Figure 5.13) confirm the pattern observable in previous datasets. If a factor has some predictive power, it is always related to either the 'SE Subject' word pair or just the SE. While 'hesitation type' receives a positive score, its effect is very weak.

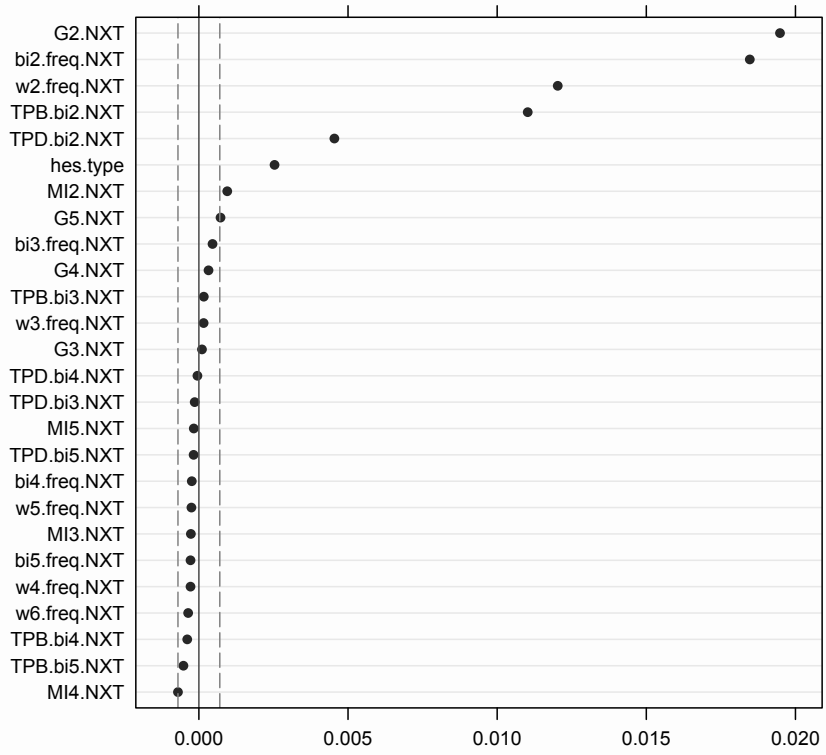


Figure 5.13: Variable importance of predictors for 'SE Subject Verb(finite) not Verb(non-finite)' (mtry=5, ntree=2,000, seed=1,321, OOB=false, results from R version 2.15.2/3.0.0).

5.4.7 SE SE Subject Verb(finite)

This and the following datasets consist of sentence-initial pre-verbal structures introduced by two SEs. There is a characteristic hesitation placement pattern in such structures. Placement before the second SE is the preferred option, followed by placement at the sentence boundary or before the subject. Hesitations are rarely moved into the verb phrase (see also Figure 5.1).

Ctree grows a comparatively complex tree on the 367 data-points in this set (see Figure 5.14). At a misclassification rate of 32.7% (247 correct predictions), model performance significantly exceeds the baseline rate of 41.14% (216 correct predictions; $p<.01$; residuals: 2.11 and -2.52). Like most *ctree* models in this study, the one for 'SE SE Subject Verb(finite)' fails to find conditions under which hesitation placement in the verb phrase is the preferred option (see Table 5.12).

		Model Predictions				
	Hesitation Position	**pre SE (1)**	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**Total**
Actual Distribution	**pre SE (1)**	44	30	0	0	74
	pre SE (2)	13	203	0	0	216
	pre Subj (3)	3	61	0	0	64
	pre V(fin) (4)	2	11	0	0	13
	Total	62	305	0	0	367

Table 5.12: Performance of ctree model for 'SE SE Subject Verb(finite)'

The first split in the tree shows that strong attractions between the SEs mean that the 'SE SE' sequence is not interrupted and hesitations are rather placed at the sentence boundary. Strongly attracted 'SE SE' sequences, found in the two small leaves 12 and 13, are generally cases where both SEs together form an adverbial such as *even though, around here, deep down* and – most frequently – *right now*.

In each of the terminal leaves 6, 7, 9 and 10, at least 75% of first SEs are *and* or *but*. Notably, these are leaves where hesitation placement after the first SE (i.e. in Position 2) is preferred. The only nodes which contain neither of the two conjunctions are numbers 4, 12 and 13 – all conditions where hesitations are preferentially placed at the sentence boundary. Yet, in contrast to several previous models, the split separating the frequent SEs from the infrequent ones is not the topmost one (it is Split 3) and other factors take effect.

As already mentioned, highly attracted SEs, according to the mutual information score, are not separated by hesitations (Split 1). Furthermore, if the second SE is likely to be preceded by the first, i.e. if backwards transitional probability is high, there are fewer hesitations separating the two than if backwards transitional probability is low (Split 2).

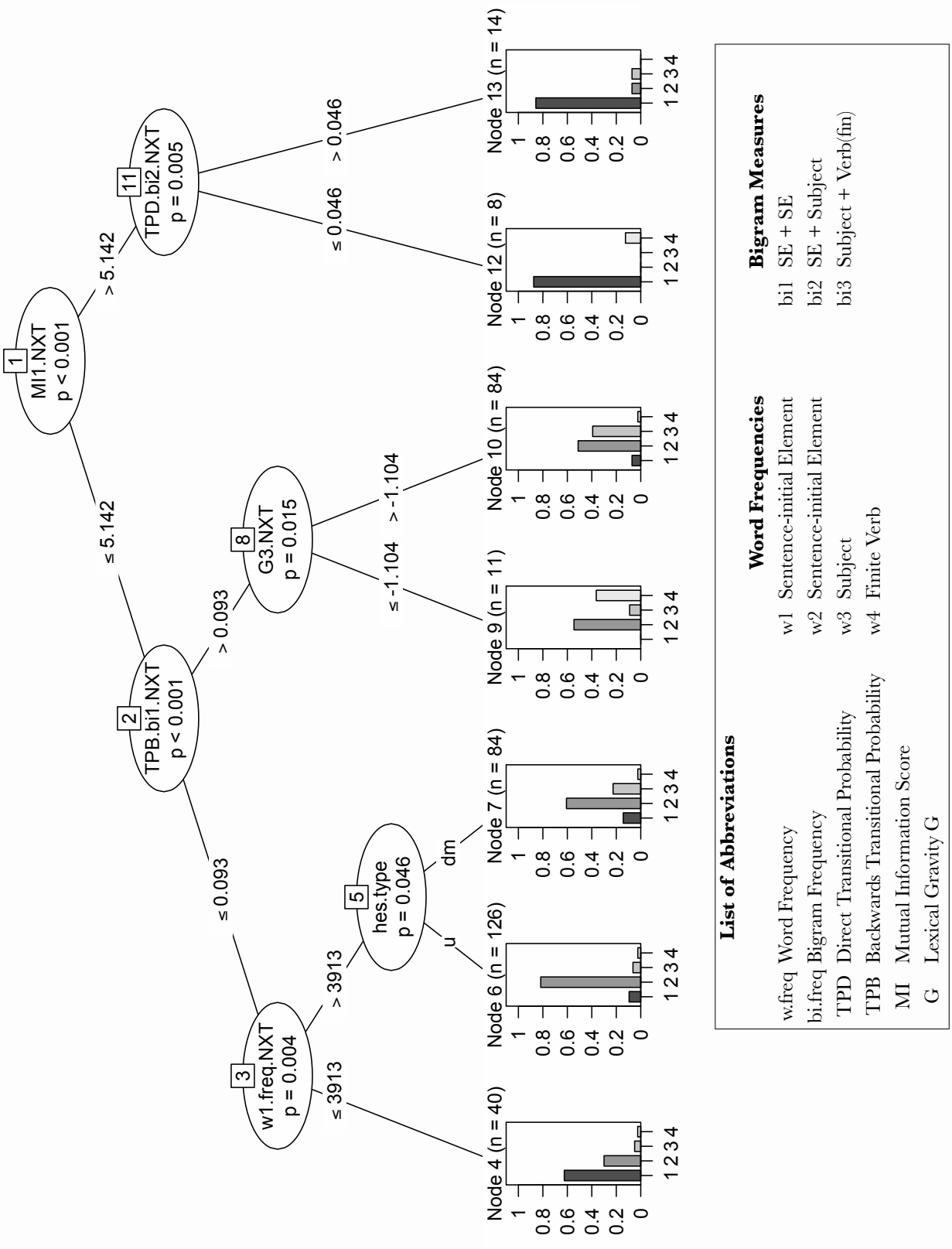


Figure 5.14: Ctree results for the structure 'SE SE Subject Verb(finite)'. Labels at terminal node bar graphs (here: 1, 2, 3 and 4) indicate hesitation position before the corresponding words (w1= Sentence-initial Element1; w2=Sentence-initial Element2; w3=Subject, w4=Finite Verb).

Such high backwards transitional probabilities are displayed particularly by *and so* and *and then,* which together occur 71 times in Nodes 9 and 10.

There is even an effect of the relation between the subject and the verb, seen in Split 8. The more strongly the subject and the verb attract according to lexical gravity G, the fewer hesitations we find before the verb, i.e. between the two. The following examples illustrate the difference between strong subject-verb attraction (illustrated by (160), taken from Node 10) and weak attractions between the subject and the verb (illustrated by (161), taken from Node 9).

(160) and *uh* so I guess [you just have to take the two problems ...] (sw3798.B.s132)

(161) and finally they *uh* carried [her out into the courtyard] (sw3038.B.s32)

A *cforest* of 2,000 trees predicts 264 outcomes correctly, corresponding to a misclassification rate of 28.07%. It thus shows that effects are very highly significant ($p<.001$; residuals: 3.27 and -3.91). This significance level is confirmed by the more conservative out-of-bag predictions ($p<.001$; residuals: 2.38 and -2.85), which still reach 251 correct predictions (misclassification rate: 31.6%).

Finally, *cforest* variable importance scores, shown in Figure 5.15, emphasise that effects in structures with two SEs do not differ from those in structures with one SE. It is the frequency of the first SE (here Word 1) and the relation between this SE and the following word which mostly determine hesitation placement. Furthermore, 'hesitation type' has some influence and relations between the subject and the verb are rated irrelevant. This apparent contradiction between *ctree* and *cforest* results concerning the role of subject-verb relations results from the fact that *cforest* grows 2,000 different trees. Re-evaluated across 2,000 trees, the effect of subject-verb relations does not reach significance – presumably because hesitations are rarely placed before the verb.

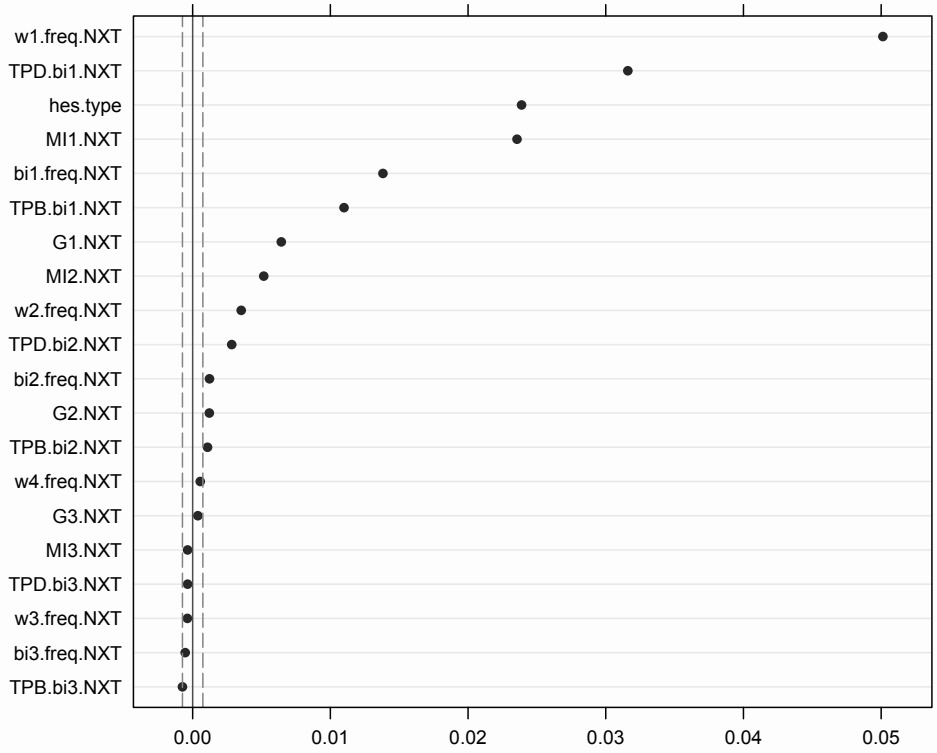


Figure 5.15: Variable importance of predictors for 'SE SE Subject V(fin)' (mtry=5, ntree=2,000, seed=604, OOB=false, results from R version 2.15.2/3.0.0).

5.4.8 SE SE Subject Verb(finite) Verb(non-finite)

The last structure to be analysed in this study is 'SE SE Subject Verb(finite) Verb(non-finite)'. At only 108 data-points it is the smallest dataset. Nevertheless, *ctree* finds effects in the data, therefore growing a tree (see Figure 5.16); it is in fact the only tree able to predict three different outcomes (see Table 5.13). It determines conditions for preferential placement before the first SE, the second SE and the subject.

		Model Predictions					
	Hesitation Position	**pre SE (1)**	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre SE (1)**	17	7	0	0	0	24
	pre SE (2)	2	53	2	0	0	57
	pre Subj (3)	1	12	7	0	0	20
	pre V(fin) (4)	0	0	0	0	0	0
	pre V(inf) (5)	2	3	2	0	0	7
	Total	22	75	11	0	0	108

Table 5.13: Performance of ctree model for 'SE SE Subject Verb(finite) Verb(non-finite)'

Overall, the model predicts 77 out of 108 data-points correctly, corresponding to a misclassification rate of 28.7%, which highly significantly exceeds the baseline performance of 57 correct predictions ($p<.001$; residuals: 2.65 and -2.8).

Placement before the first SE – and thus at the sentence boundary – is preferred if the SE is not highly frequent (see Node 2). The cut-off point in the 3,000s (Split 1) is in accordance with similar splits in the trees in previous sections. Infrequent SEs are assigned to Node 2. Semantically, both SEs in this node often form a unit, like a sentence adverb; (162) and (163) show some examples.

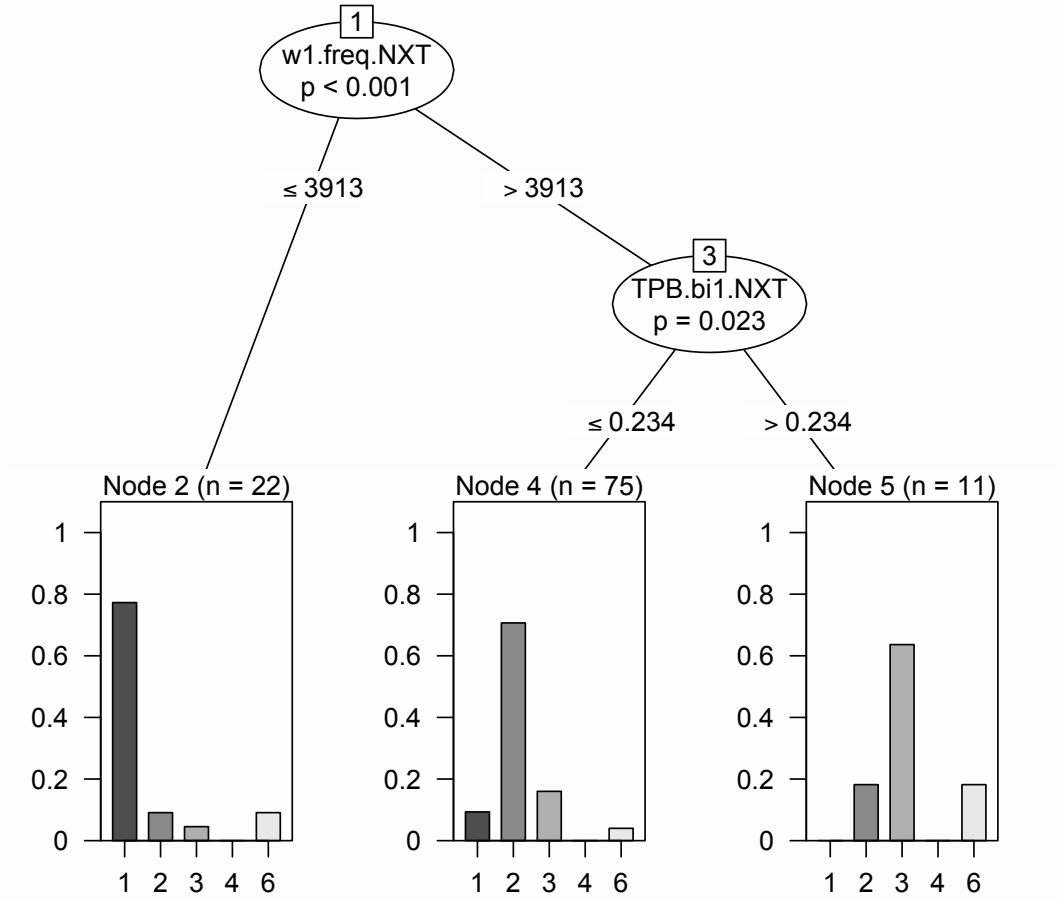
(162) *well [pause] uh* until recently I was taking [the Wall Street Journal] (sw3569.A.s3)

(163) *uh* not if it's done [fairly] (sw2314.B.s25)

Nodes 4 and 5 are dominated by *and* and *but* as first SEs. In Node 4, 80% of data-points begin with one of the two coordinating conjunctions (n(*and*) = *46*; n(*but*) = 14). Other SEs in the node are *so*, *oh*, *just* and *or*. Consequently, the usual pattern for highly frequent SEs emerges: hesitations tend to be placed after them, as is the case in (164).

(164) but *[pause] you know* when we're having [guests ...] (sw2124.A.s79)

In small Node 5, however, the association between the two SEs is stronger, indicated by a higher backwards transitional probability. Eight out of the eleven data-points in this



List of Abbreviations

		Word Frequencies		**Bigram Measures**	
w.freq	Word Frequency	w1	Sentence-initial Element	bi1	SE + SE
bi.freq	Bigram Frequency	w2	Sentence-initial Element	bi2	SE + Subject
TPD	Direct Transitional Probability	w3	Subject	bi3	Subject + Verb(fin)
TPB	Backwards Transitional	w4	Finite Verb	bi6	Verb(fin) + Verb(inf)
MI	Mutual Information Score	w6	Non-finite Verb		
G	Lexical Gravity G				

Figure 5.16: Ctree results for the structure 'SE SE Subject Verb(finite) Verb(non-finite)'. Labels at terminal node bar graphs (here: 1, 2, 3, 4 and 6) indicate hesitation position before the corresponding words (w1= Sentence-initial Element1; w2=Sentence-initial Element2; w3=Subject, w4=Finite Verb; w6=Non-finite Verb).

node begin with *and then*. Thus speakers prefer not to interrupt the SE sequence and move the hesitation further into the sentence, as in (165).

(165) and then *you know [pause]* you can have [cookouts and stuff like that] (sw2485.B.s189)

A *cforest* produces the same number of correct predictions as the individual tree (see Table M.1 in Appendix M). Performance on out-of-bag data-points is poor, however. Only 65 data-points are predicted correctly, which does not significantly exceed baseline performance. This discrepancy between *cforest* and out-of-bag performances may result from the small number of data-points.

Finally, variable importance scores again show that the single most effective predictor of hesitation placement is the frequency of the first SE (see Figure 5.17). Apart from measures relating to the coherence in the SE SE word pair, all other frequencies and measures of association are irrelevant for predicting hesitation placement.

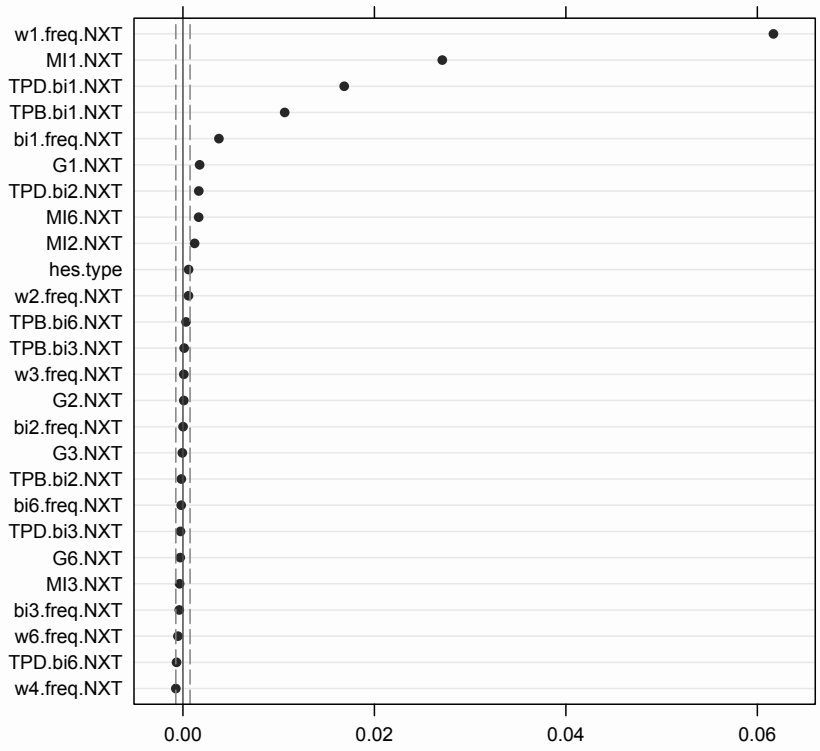


Figure 5.17: Variable importance of predictors for 'SE SE Subject Verb(finite) Verb(non-finite)' (mtry=5, ntree=2,000, seed=902, OOB=false, results from R version 2.15.2/3.0.0).

5.4.9 Summary

The previous sections detailed the analyses of 6,317 hesitations placed in or immediately preceding one of eight different pre-verbal sent-initial structures. The regression methods employed for analysis were Classification and Regression Trees (CART trees) and random forests (see Section 3.3.3), which were mainly provided with predictors reflecting statistical attractions between all word-pairs in the analysed sequences. Analyses were conducted separately for each type of sentence beginning.

For four of the eight selected structures, CART trees and random forests performed significantly better than simple models which overgeneralise from the most frequent outcome to all data-points. These structures are

- SE Subject Verb(finite)
- SE Subject Verb(finite) Verb(non-finite)
- SE SE Subject Verb(finite)
- SE SE Subject Verb(finite) Verb(non-finite)

Tables 5.14 to 5.16 provide an overview of analyses. *ctree* results (Table 5.14) are generally consistent with *cforest* predictions (Table 5.15), and both are confirmed by the most conservative results based on out-of-bag data (Table 5.16). Only in the case of 'SE SE Subject Verb(finite) Verb(non-finite)' do out-of-bag results not confirm the highly significant effects found by *ctree* and *cforest* models, which might be due to the small size of the dataset (n=108).

Structures where analyses do not yield significant results share certain characteristics. The three structures where the subject is not preceded by an SE, and where therefore the noun phrase boundary coincides with the sentence boundary, show very little variation; here, hesitations are almost exclusively placed at the sentence boundary. Placement patterns in structures containing negated verb phrases are not explained by the trees. 'Subject Verb(finite) *not* Verb(non-finite)' sequences are characterised by a distinct lack of hesitations, which are instead placed after the first SE or before the sentence. Whether this is due to characteristics of negated verb phrases will be investigated in Section 5.7.

Structure	**MCR**	**Sig. level**	**Residuals**	
Subject Verb(finite)	1.59%	non-sig.	-	-
Subject Verb(finite) Verb(non-finite)	7.8%	non-sig.	-	-
Subject Verb(finite) *not* Verb(non-finite)	2.9%	non-sig.	-	-
SE Subject Verb(finite)	16.27%	p<.001	6.4	-9.9
SE Subject Verb(finite) Verb(non-finite)	24.01%	p<.001	3.54	-4.57
SE Subject Verb(finite) *not* Verb(non-finite)	18.67%	non-sig.	-	-
SE SE Subject Verb(finite)	32.7%	p<.01	2.11	-2.52
SE SE Subject Verb(finite) Verb(non-finite)	28.7%	p<.001	2.65	-2.8

Table 5.14: Performance of the CART trees. Given are misclassification rates (MCR), p-values based on chi-square tests and the residuals of the chi-square tests.

Structure	**MCR**	**Sig. level**	**Residuals**	
Subject Verb(finite)	1.59%	non-sig.	-	-
Subject Verb(finite) Verb(non-finite)	8.17%	non-sig.	-	-
Subject Verb(finite) *not* Verb(non-finite)	2.9%	non-sig.	-	-
SE Subject Verb(finite)	15.42%	p<.001	6.81	-10.54
SE Subject Verb(finite) Verb(non-finite)	23.31%	p<.001	3.73	-4.81
SE Subject Verb(finite) *not* Verb(non-finite)	16%	p<.1	0.82	-1.6
SE SE Subject Verb(finite)	28.07%	p<.001	3.27	-3.91
SE SE Subject Verb(finite) Verb(non-finite)	28.7%	p<.001	2.65	-2.8

Table 5.15: Performance of random forests. Given are misclassification rates (MCR), p-values based on chi-square tests and the residuals of the chi-square tests.

Structure	**MCR**	**Sig. level**	**Residuals**	
Subject Verb(finite)	1.59%	non-sig.	-	-
Subject Verb(finite) Verb(non-finite)	7.8%	non-sig.	-	-
Subject Verb(finite) *not* Verb(non-finite)	2.9%	non-sig.	-	-
SE Subject Verb(finite)	15.78%	p<.001	6.63	-10.27
SE Subject Verb(finite) Verb(non-finite)	24.24%	p<.001	3.48	-4.49
SE Subject Verb(finite) *not* Verb(non-finite)	21.33%	non-sig.	-	-
SE SE Subject Verb(finite)	31.6%	p<.001	2.38	-2.85
SE SE Subject Verb(finite) Verb(non-finite)	39.81%	non-sig.	-	-

Table 5.16: Performance of the random forests' out-of-bag sets. Given are misclassification rates (MCR), p-values based on chi-square tests and the residuals of the chi-square tests.

Across all trees, word frequencies and lexical gravity G are the most popular splitting criteria: both are selected six times, while bigram frequency is never chosen. An analysis of *cforest* variable importance scores following in Section 5.5 will reveal whether this means that G actually outperforms bigram frequency or whether this results from the known correlation between lexical gravity G and bigram frequency.

Finally, selected predictors most frequently relate to the first SE or the bigram containing the first SE. Thus, the frequency of the first SE functions as the splitting criterion six times, while other word frequencies are only selected once, and the bigram containing the first SE is selected six times while all other transitions combined are selected eight times. Four of the splits based on the frequency of the first SE are made at frequencies between 3,253 and 3,929. The latter results could be indicators that there are SE-hesitation chunks. This hypothesis will be further investigated in Section 5.6.

5.5 Comparison of Predictors

The performance of the various predictors can be compared by evaluating how often and how high up in the trees different types of predictors are chosen[47]. The previous sections' description of recurrent *ctree* decisions revealed that lexical gravity G is the predominant bigram-related predictor in the present dataset, while bigram frequency is never selected as a splitting criterion. However, because of the fact that CART splits will underrepresent marginally outperformed and collinear predictors, *cforest* variable importance scores offer a more reliable and comprehensive method of evaluation.

For a comparison of variable importance, all scores resulting from the analyses of hesitation placement in sentence-initial contexts were pooled by predictor type. For instance, all scores relating to word frequencies were grouped. No distinction was made according to type of structure or transition[48].

Figure 5.18 shows the results. From bigram frequency (bi.freq) through to 'hesitation type', differences are minor and not statistically significant (based on Wilcoxon rank-sum tests). The only predictor which stands out is word frequency (w.freq). This finding is in line with *ctree* decisions. In the trees, word frequencies were chosen as splitting criteria in seven cases and consequently were among the most popular predictors. However, Wilcoxon rank-sum tests show that the performance of word frequency does not significantly exceed that of any of the other predictors.

Additionally, it appears that not all words and transitions in the structures have a uniformly strong influence on hesitation placement. The splits made in the *ctrees* and the distribution of the *cforest* variable importance scores suggest that the frequency of the first SE and its associations with the following word play a more prominent role in determining where hesitations are placed than any of the other predictors. For an evaluation of whether this perceived effect is indeed statistically significant, the variable importance scores pertaining to the frequency of the first SE as well as to those predictors describing the frequency and cohesion of the pairs '1stSE 2ndSE' and '1stSE Subject' are separated from those pertaining to all other types of transitions, namely

[47] For an explanation of why it matters how early (or high up) predictors are chosen in a CART tree, see Section 4.5.

[48] For an argumentation on why comparison of variable importance scores across different studies is warranted in this case see Section 3.3.2.2.

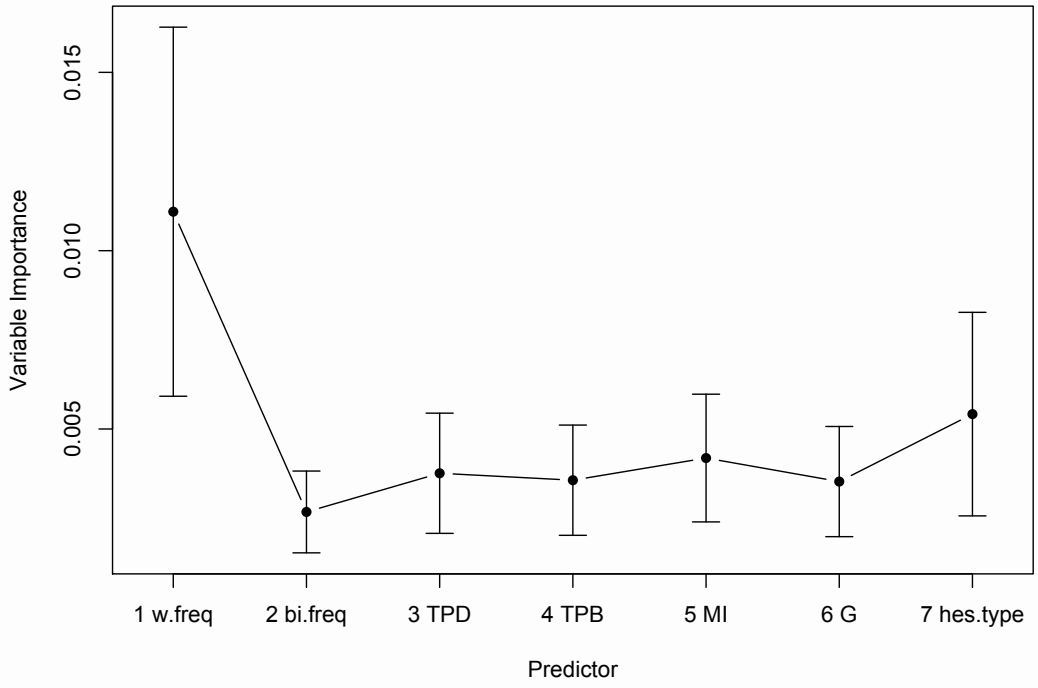


Figure 5.18: Variable importance measures by type of predictor (1= word frequencies, 2= bigram frequencies, 3= direct transitional probabilities, 4= backwards transitional probabilities, 5= mutual information scores, 6= lexical gravity G, 7= hesitation type). The dot indicates the mean for each group and the error bars show the standard error.

- 2ndSE Subject
- Subject Verb(finite)
- Verb(finite) *not*
- *not* Verb(non-finite)
- Verb(finite) Verb(non-finite)

This means that scores as shown in Figure 5.18 are split into two groups: those pertaining to the first SE and those relating to any other word or a transition which does not include the first SE.

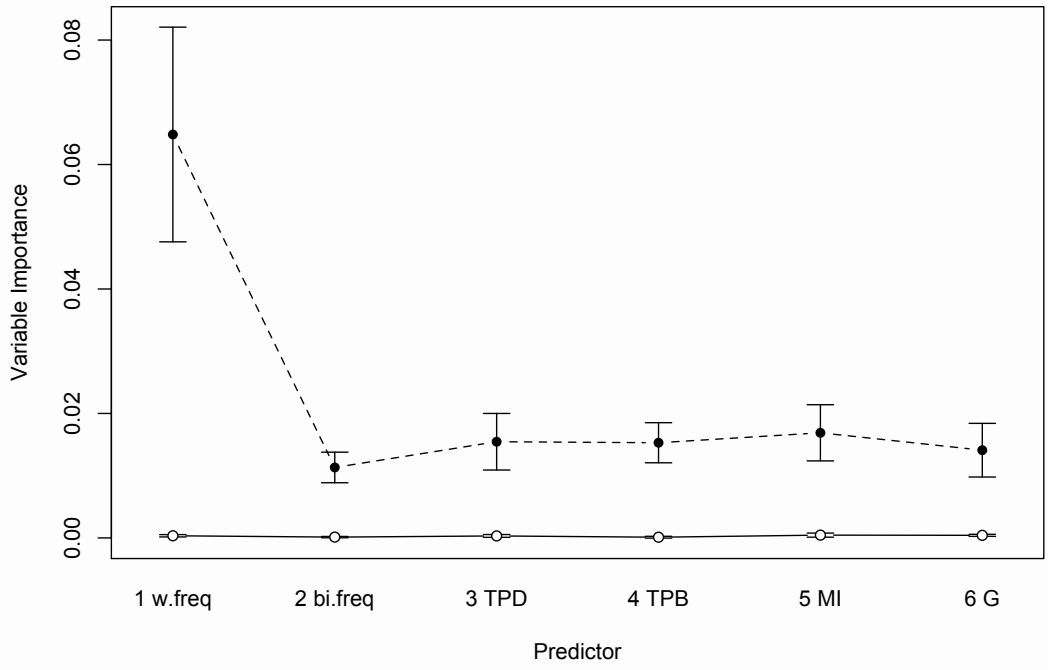


Figure 5.19: Variable importance scores by type of predictor (1= word frequencies, 2= bigram frequencies, 3= direct transitional probabilities, 4= backwards transitional probabilities, 5= mutual information scores, 6= lexical gravity G) and transition. Points connected by the dashed line show values for the first SE (in the case of word frequencies) or the bigram containing the first SE (for all other predictors), while points connected by the solid line show values for all other words and transitions. The dot indicates the mean for each group and the error bars show the standard error.

Predictor	**w.freq**	**bi.freq**	**TPD**	**TPB**	**MI**	**G**
Significance Level	p<.001	p<.001	p<.001	p<.001	p<.01	p<.01

Table 5.17: Difference in predictors' variable importance scores when applied to the bigram containing the first SE (or, in the case of word frequencies, the first SE) and to all other transitions. Results are based on separate Wilcoxon rank-sum tests.

Predictor	**bi.freq**	**TPD**	**TPB**	**MI**	**G**
Significance Level	p<.05	p<.1	p<.05	p<.1	p<.1

Table 5.18: Difference in performance compared to word frequency of the first SE. Results are based on separate Wilcoxon rank-sum tests.

The difference in performance between these two groups can be seen in Figure 5.19. The dashed line at the top shows the mean variable importance of the frequency of the first SE as well as of predictors describing the associations between the first SE and the word following it. The solid line at the bottom indicates mean variable importance scores of predictors pertaining to any other transition. The graph emphasises that relations between the subject and the verb as well as within the verb phrase have almost no influence on hesitation placement. As predictors, these relations (represented by the solid line) consistently receive scores close to zero. Section 5.7 will deal with verb-phrase relations and their lack of effect on hesitation placement in more detail. For each type of predictor, the difference between the two groups' scores is highly significant. Table 5.17 shows the exact levels of significance. For example, it provides evidence that statistically the mutual information score of the first bigram has a significantly greater influence on where a hesitation will be placed than the mutual information score of any of the other transitions.

Table 5.18 shows that when comparing the predictors represented by the dashed line, i.e. those relating to the first SE and its associations, there is a tendency for word frequency to outperform measures of association. This suggests that it is not really the relation between the first SE and the following word which has an effect on hesitation placement, but possibly the relation between the first SE and the following hesitation. Section 5.6 will address this phenomenon in more detail.

5.6 Sentence-Initial 'Dummy Chunks'

One of the aims of the analysis of sentence-initial structures is to determine whether recurrent combinations of an SE and a hesitation can be mentally chunked and if so whether this is an effect of their frequent combined use. This section combines data from all structures which contained at least one SE and investigates this point in more detail.

As already introduced in some detail in Section 5.1, Holmes (1988) and Jurafsky et. al (1998) as well as Altenberg (1998) describe close links between certain SEs and hesitations. Their studies show that conjunctions and sentence adverbs, particularly semantically unspecific ones – such as *and* or *but* – are frequently followed by filled pauses and other hesitations. Some SEs (i.e. *and, that*) are even more likely to be followed by pause fillers than other parts of speech which are also prone to appearing sentence-initially, such as determiners and personal pronouns (Jurafsky et al. 1998:2).

In light of Clark (1996:269), Clark and Wasow (1998) and Sacks et al. (1974:719), this finding can be interpreted as an indication that speakers make use of 'dummy' SEs. If speakers are pressed to start speaking before they have finished planning their sentence, they use hesitations and "redundant linking words" (Holmes 1988:329), i.e. certain SEs, to buy time (cf. Altenberg 1998:113).

Results from Clark and Wasow (1998) as well as Clark and Fox Tree (2002) already provide strong evidence that chunking between filled pauses and words is indeed possible. The former study finds that in Switchboard and the London-Lund corpus fillers are often uttered after a word without an intervening pause, sometimes resulting in resyllabification of the final consonant (Clark and Wasow 1998:229). The latter study points out that conjunctions are particularly prone to undergoing this process, leading to such phonological words as "an.duh" or "bu.tuh" (Clark and Fox Tree 2002:101).

Taken together, the finding that some SEs can be used as time-buying devices and the finding that some words and hesitations form phonological chunks suggest that such chunks should be highly likely to occur at sentence beginnings, where speakers use 'dummy' SEs but also hesitations. Indeed, Altenberg finds that specific conjunctions, discourse markers and adverbs frequently occur in sequence – so often, in fact, that they satisfy his criterion for classification as "recurrent word combination[s]" (threshold: 20 instances per million words; Altenberg 1998:101-2). He states that combinations such as *and you know* or *because I mean* are "routinised sentence or clause openers", which he calls "frames" (Altenberg 1998:112-3). He notes that

> [t]o what extent each choice [of element in the frame] restricts the choice of the next item [in the frame] is unclear, but the relative frequency of the

> various combinations suggests the existence of certain pragmatic restrictions. (Altenberg 1998:113)

Still, Altenberg remains doubtful that frames are phraseological units or even collocations. My approach moves away from pragmatic or semantic motivations and instead asks whether there is frequency-based evidence that such frames are cognitively represented chunks.

5.6.1 Definition

Chunking at this mixed word/hesitation level cannot be modelled like chunking between subsequent words. So far, I have argued that chunking is brought about through high co-occurrence frequency or strong statistical attractions between subsequent words. The absence of hesitations has been interpreted as an indicator that the chunking process thus postulated is indeed happening. This line of argumentation first of all requires knowledge about the frequency of the pair and associations between the words in it, which is not available in this case, as all hesitations were removed prior to the calculation of bigram-related measures. Hence I have no record of how often *and* and *uh* or *but* and *I mean* occur together in the Switchboard NXT corpus. Secondly, the definition takes the absence of hesitations as an indicator for chunking, which is not possible here, as the hesitation is analysed as part of the potential chunk. Fortunately however, the literature only mentions 'dummy chunks' containing vocalised hesitations, i.e. filled pauses and discourse markers. Thus the absence of unfilled pauses within the purported chunks can still be interpreted as an indicator of chunking status. In addition to this definition, the following analysis rests on two assumptions.

SE frequency determines chunking. – I expect that the more frequent an SE, the easier it is to retrieve and consequently the more likely it is to be used as a dummy element. Elements which take some time to retrieve cannot be used as time-buying devices because they do not come to mind fast enough to serve this purpose. Therefore, I expect more frequent SEs to be more likely to form dummy chunks than low-frequency SEs.

Importantly, this prediction concerning which SEs may be used as time-buying devices relies on frequency alone and does not refer to semantic criteria, which previous studies have generally used as arguments for why SEs become dummy elements. Of course, ideally, a time-buying device requires no knowledge of the structure or content of the following sentence (cf. Sacks, Schegloff and Jefferson 1974:719), so semantics must play a certain role in the development of time-buying devices. However, exclusively semantic motivations for frame status are at risk of resulting in logical fallacies if interpreted as predictions. The fact that SEs in time-buying function are semantically

redundant (Holmes 1988:328 even talks about "meaningless stereotyped expressions"), does not automatically imply that any semantically empty SE will be used as a time-buying device.

Chunked sequences should be in a specific order. – Two types of sequence are possible: the hesitation may precede the SE (e.g. *well then*) or it may follow the SE (e.g. *so uh*). I expect the hesitation to follow the SE in chunked sequences because if highly-frequent, easily-retrievable SEs enter into chunks, SE retrieval should be fast, not requiring a hesitation. Hesitation may follow after the SE to buy time to plan the rest of the sentence. This is in line with findings from previous studies which emphasise that conjunctions and adverbs tend to be followed by hesitations (cf. Holmes 1988; Jurafsky et. al 1998; but see Altenberg (1998:113), who lists frequent combinations in which the discourse marker *well* precedes SEs, e.g. *well of course, well you know*). It is furthermore confirmed by the discovery that resyllabification only occurs in 'SE hesitation' sequences, never in 'hesitation SE' combinations (e.g. an.duh but *uh.wand; Clark and Wasow 1998:229; Clark and Fox Tree 2002:101). Finally, all SE-based splits in Section 5.4 confirm this hypothesis; the more frequent an SE, the more likely the hesitation is to follow it.

5.6.2 Data

Data was exclusively drawn from the dataset of sentence-initial structures. Out of all 6,317 hesitations, those were included which immediately preceded or followed the first or only SE in a structure. Hence *because I mean* in (166) was included, while *otherwise uh* in (167) was not.

(166) because *I mean* when you rent [a video] (sw2435.A.s328)

(167) but otherwise *uh* they have [to be in one of their paper bags] (sw2249.A.s151)

This decision was made based on the variable importance scores from random forests. The frequency of the first SE ('w1.freq') is ranked as a much better predictor of hesitation placement than the frequency of the second SE ('w2.freq'), the former receiving a score 14 to 107 times higher than the latter.[49] This can be interpreted as an indication that chunking between the second SE and a hesitation is unlikely (presumably because the second word in a sentence is much less likely to be a place-holding device than the first word). These choices resulted in a new dataset of 2,594 data-points.

[49] See also Figure 5.19 where the mean predictive value of the frequency of the first SE is indicated by the first dot on the dashed line while the predictive value of the frequency of the second SE is included in the mean indicated by the first dot on the solid line.

In a second step, only the SE and the hesitation were retained and the rest of the structure was discarded. In the case of hesitation clusters, such as *well uh [pause]* in (168), only the hesitation closest to the SE was retained. If an unfilled pause intervened between a filled pause or discourse marker and the SE, both the closest vocalised hesitation and the unfilled pause were retained. In (168), *uh* and the pause were retained.

(168) *well uh [pause]* maybe I am [into some thing occasionally...] (sw2024.A.s59)

Based on the presence or absence of an unfilled pause and on the order of the SE and the vocalised hesitation, four sequences are possible. Table 5.19 lists all options, which have been termed types A, B, C and D, and the number of data-points which fall into each category. The sequence *uh [pause] maybe* in (171), for instance, was coded as type A. According to the definition of a chunk and the additional assumptions outlined above, only types B and D can be interpreted as chunks, with frequency-based chunking resulting in type D.

Type	**Sequence**					**n**
A	hesitation	[pause]	SE			149
B	hesitation		SE			555
C			SE	[pause]	hesitation	298
D			SE		hesitation	1,592

Table 5.19: Types of sequence and number of data-points per sequence

Thus data-points were coded for the following set of parameters[50]:

SE type – e.g. *and*. In total, there are 102 types of SEs in the data-set.

SE frequency – Corpus frequency of the SE. Frequencies range from 27,202 (*and*) to 1 (*ironically*).

Sequence – One of the four types listed in Table 5.19, which subsume information about the order in which the hesitations and the SE appear and whether or not a pause intervenes.[51]

[50] All selection and coding procedures were conducted automatically with the help of R scripts.

[51] In this case, the problems with sentence-initial pauses described in Section 5.2.2.5 do not apply. Due to the fact that we are exclusively dealing with unfilled pauses occurring in clusters with other hesitations, all pauses could be reliably extracted.

5.6.3 Analysis & Results

This section investigates whether there are frequency effects in the ordering of sentence-initial SEs and hesitations. It specifically tests whether the ordering assumption made above is tenable. Are highly frequent SEs more likely to precede the hesitation than lower frequency SEs? In other words, does the chance of outcome type D increase with SE frequency?

In order to investigate whether higher-frequency SEs behave differently from lower-frequency SEs, they need to be treated as separate groups. Analyses predicting the likelihood of every SE type separately, based on its frequency, are not possible as 38 SEs only occur once in the data-set and 27 only two or three times. This means that for more than half the types in the data-set, token frequencies are too low for this type of analysis. Instead, the data is split into frequency bins. Instead of relying on subjectively created bins, a CART tree (run in R) based on the sole predictor 'SE frequency' (SE.freq) was used to establish how (and whether at all) the data can be grouped according to frequency.

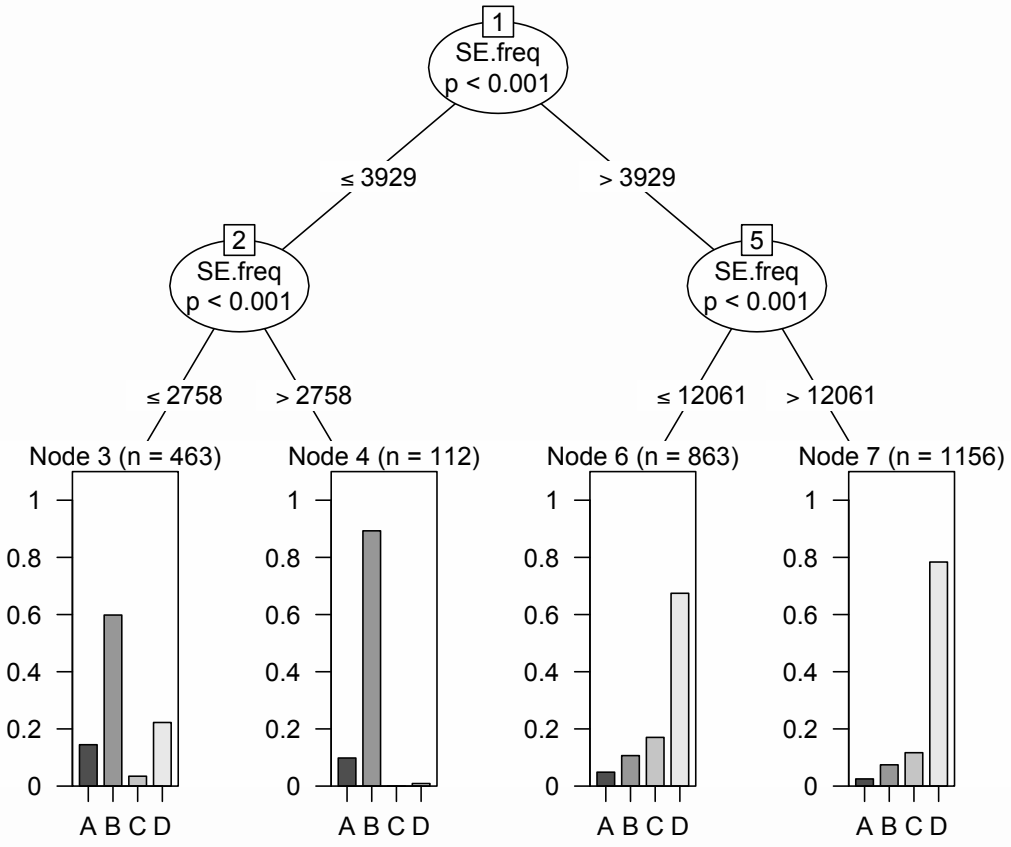


Figure 5.20: Influence of SE frequency (SE.freq) on the ordering of SE hesitation sequences. For a legend to outcome types A, B, C and D see Table 5.19.

Figure 5.20 shows the resulting *ctree*, which confirms the effects found in the analyses of the individual structures (Section 5.4): SEs with a corpus frequency above 3,929 behave differently from the lower-frequency ones (indicated by Split 1 in Figure 5.20). In the two high-frequency leaves (Nodes 6 and 7), the preferred sequence is D, 'SE hesitation', while lower-frequency SEs tend to be preceded by the hesitation; i.e. pattern B, 'hesitation SE', dominates. All four terminal leaves display very homogenous behaviour: between 59.83 and 89.29% of data-points in each leaf follow the same sequencing pattern.

Interestingly though, these effects are not gradual. The chance of outcome D does not continuously increase with SE frequency, but instead abruptly skyrockets at a threshold of 3,929. Hence it appears that high and low frequency SEs display a fundamentally different behaviour. In order to gain a better understanding of the nature of this behaviour, the following sections will separately describe each of the four terminal nodes in order of decreasing SE frequency with a particular focus on the SE types they encompass.

Node	**Total n**	**Content**		
		SEs	**n**	**Frequency in the Corpus**
Node 7	1,156	and	1,156	27,202
Node 6	863	yeah	1	12,061
		in	1	9,872
		uh-huh	1	7,637
		but	605	7,365
		so	167	6,352
		just	4	5,737
		oh	61	4,901
		or	23	4,248
Node 4	112	that	3	3,929
		not	5	3,913
		if	90	3,253
		like	14	3,123
Node 3	463	89 types, e.g.		
		when	54	2,521
		then	14	2,236
		see	15	222
		anyway	1	69

Table 5.20: Additional information about terminal nodes in Figure 5.20.

Node 7 – This node is entirely comprised of tokens of *and*. At a frequency of 27,202, *and* is the only SE whose frequency exceeds the splitting criterion of 12,061 (see Split 5 in the tree). Hence the behaviour displayed is not the behaviour of highly frequent SEs in general, but merely that of the most highly frequent one. Here, in 78.37% of cases, the outcome is '*and* hesitation', highly significantly exceeding the combined frequency of A to C ($p<.001$, based on a chi-square test).

Figure 5.21 shows that *and uh* dominates this group. The frequency of *and uh* very highly significantly exceeds the combined frequency of its competitors *uh [pause] and*, *uh and* and *and [pause] uh*.[52] Furthermore, *and um* and *and I mean*, while far less frequent overall, also significantly exceed the frequency of their competitors.

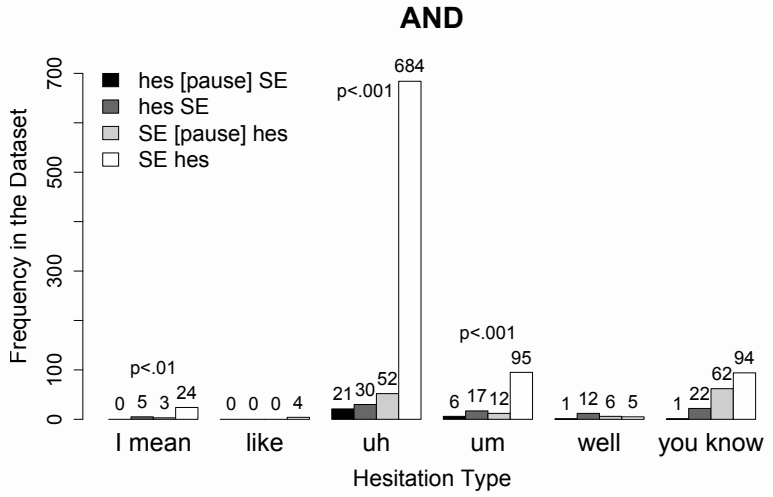


Figure 5.21: Ordering and choice of hesitations occurring in combination with 'and'

Node 6 – Node 6 is comprised of SEs with a frequency between 3,930 and 12,061 and thus contains the SEs *yeah*, *in*, *uh-huh*, *but*, *so*, *just*, *oh* and *or* (see Table 5.20). *But* alone makes up over 70% of data-points in this node, while *yeah*, *in*, *uh-huh* and *just* are rarely used in combination with hesitations and are therefore rare in the data. Despite the fact that the hesitation pattern in this node is slightly less homogenous than in Node 7, 67.44% of data-points still display sequence D. Thus the frequency of 'SE hesitation'

[52] This and the following significance levels result from chi-square tests which were calculated comparing the combined frequency of outcomes A to C and that of D to an expected 50/50 distribution. This simplified method of comparison was chosen because comparison of outcomes A, B, C and D in a 2x2 table was not possible throughout due to empty cells. Assuming a 75/25 distribution would have accurately reflected random variation, yet the present option was chosen because it sets a higher benchmark and thus ensures that only clearly chunked combinations are discussed as such. Crucially, this method of evaluation also corresponds to the way the performance of the tree is evaluated.

still highly significantly exceeds the combined frequency of A to C (p<.001, based on a chi-square test).

Figures 5.22 and 5.23 exemplarily show the combination patterns of *but* and *oh*. The chart for *but* strongly resembles that for *and*. In both cases, *but uh/and uh* is the predominantly chosen combination and *but um/and um* are also significantly more frequent than their competitors. In the case of *oh*, we see a different pattern. Here, *oh well* is the predominant combination. The collocation *oh well* is in fact so well established as a conversational device that it has long received a separate sub-entry in the OED ("well, adv.", OED).

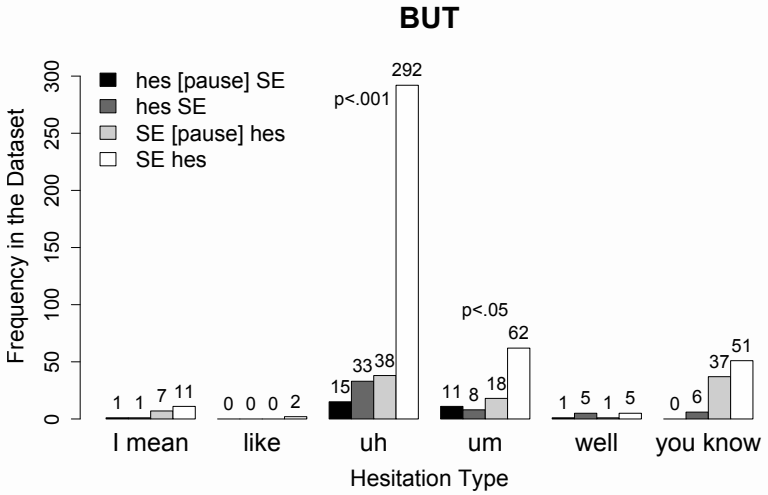


Figure 5.22: Ordering and choice of hesitations occurring in combination with 'but'

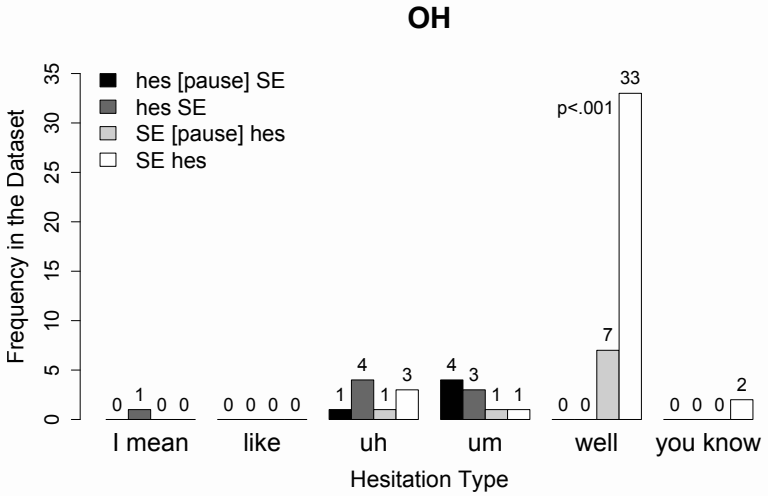


Figure 5.23: Ordering and choice of hesitations occurring in combination with 'oh'

Node 4 – Node 4 contains SEs with a frequency between 2,759 and 3,929, a band so narrow it only applies to *that, not, if* and *like*. Hesitation behaviour in this node is more homogenous than in any of the other nodes. 89.29% of data-points display pattern B, 'hesitation SE', which highly significantly exceeds the combined frequency of A, C and D ($p<.001$, based on a chi-square test). Thus there is a very abrupt reversal of behaviour between Nodes 4 and 6. The proportion of outcome D drops from 67.44% in Node 6 to a single instance in Node 4 (0.89%). As *that* and *not* are rare in the data-set (see Table 5.20), this is mostly due to the behaviour of *if* and *like*. All 14 *like* data-points display pattern B and 79 out of 90 *if* data-points do. Figure 5.24 shows that there is no pattern which is as dominant as in the previous examples, but there are several patterns which can be considered chunked, namely *I mean if, well if* and *you know if*.

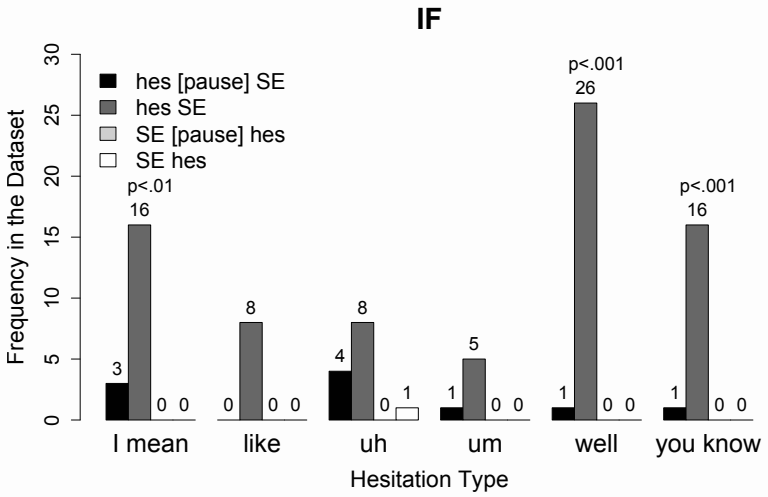


Figure 5.24: Ordering and choice of hesitations occurring in combination with 'if'

Node 3 – This node contains SEs with a corpus frequency between one and 2,758. It is the least homogenous of the four terminal nodes. At 59.83%, outcome B, 'hesitation SE', is the most popular (highly statistically exceeding the frequency of A, C and D combined, based on a chi-square test), followed by D (22.25%). The node not only comprises the lowest-frequency SEs, but the greatest range of SEs (89 different types; see Table 5.20), thus some of the variety in behaviour might result from diversity.

Figures 5.25 and 5.26 show results for *when* and *then*. These indicate that we still find combinations in this node which are likely to be chunked. As expected, in these chunks the discourse marker precedes the SE. *Well then* is a particularly interesting case as it is lexicalised. The OED describes it as "introducing a conclusion or further statement, or implying that one can naturally be drawn or made" ("well, adv.", OED). Interestingly, in

the present data-set we find cases of both the lexicalised, see (169), and the compositional meaning, see (170).

(169) *well* then I have [a friend at school that has a boyfriend that's a lawyer] (sw2220.A.s247)

(170) *uh well* then you must know [a lot more about this than I do] (sw2749.A.s5)

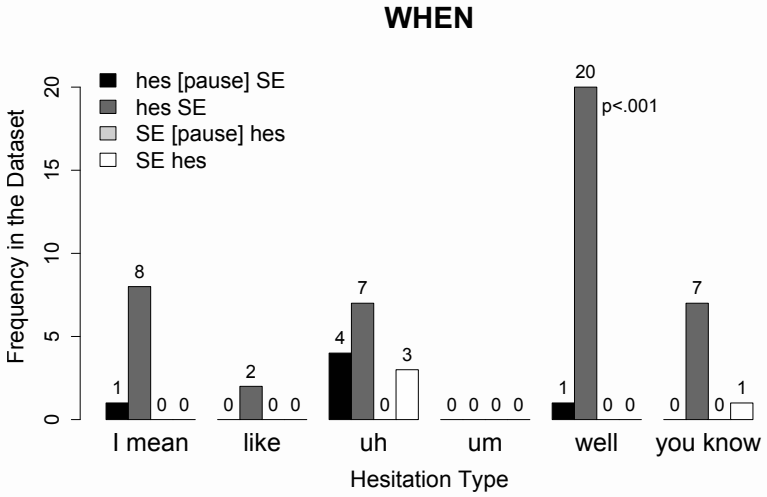


Figure 5.25: Ordering and choice of hesitations occurring in combination with 'when'

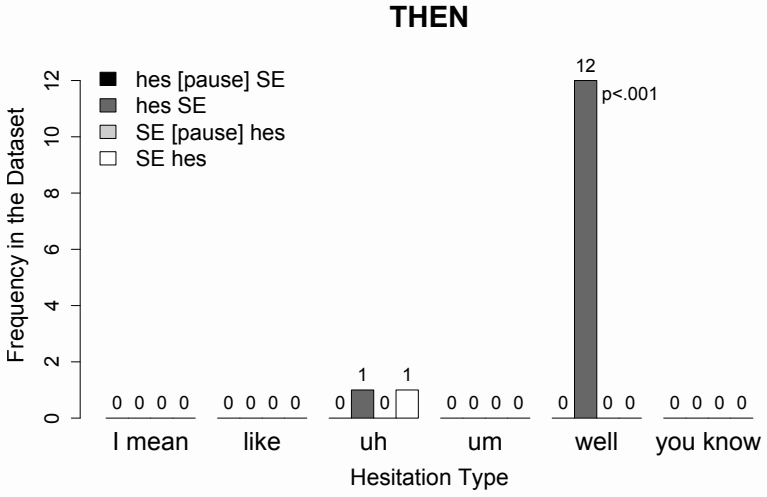


Figure 5.26: Ordering and choice of hesitations occurring in combination with 'then'

These results can be summarised and interpreted as follows:

- There are highly significant frequency effects in the data. The *ctree* model shows that whether speakers hesitate before or after the first SE in a sentence can be successfully predicted from the usage frequency of the SE.
- The direction of the effect confirms the hypothesis. Highly frequent SEs are more easily retrievable than their low-frequency counterparts, the former coming to mind so fast that no time-buying device is necessary. Hence, hesitations triggered by the planning of the rest of the sentence follow them (see Nodes 6 and 7). Lower-frequency SEs, on the other hand, take longer to retrieve, hence the speaker prefers to hesitate before them (see Nodes 3 and 4).
- CART results suggest that the difference between high-frequency-induced and low-frequency-induced behaviour is not one of degree, but categorical. Behavioural trends appear not to change continuously along a scale from high-frequency to low-frequency SEs, but instead seem to be reversed at a threshold of an SE frequency of 3,929. This observation needs to be interpreted with caution, though, as logarithmic developments and Zipf distributions may be disguised in categorical divisions of data as performed by CART trees.
- The proportion of intervening pauses declines with increasing SE frequency. From Node 6 to Node 7, the proportion of 'SE [pause] hesitation' sequences decreases from 20.16% to 12.97% (i.e. the change in the ratio of C to D outcomes is significant at the $p<.001$ level, based on a chi-square test).
- For most types of SEs, there is only one predominant combination which provides further evidence that these SEs are not randomly followed by any kind of hesitation but instead have formed chunks with one specific hesitation.

All of these results provide evidence in favour of the hypothesis that chunking does not only take place between words, but can also affect combinations of words and vocalised hesitations. The more frequently SE-hesitation combinations are used, the more likely they are to become entrenched until they can be retrieved and uttered as a single unit in a fixed order without intervening unfilled pauses.

The low degree of noise in the leaves, particularly in the higher-frequency ones, and the fact that the sequence of terminal leaves shows a non-reversing, non-fluctuating pattern suggest that the frequency of the SE is, in fact, the driving force behind these effects. If the formation of particular combinations were mainly due to semantics, we would find either less homogenous groups or more groups consisting of a single type only (such as Node 7). Furthermore, we would expect a very low-frequency group to

emerge which behaves like the two high-frequency leaves 6 and 7, because the latter contain relatively semantically unspecific SEs. This criterion can also be applied to many interjections at the lower end of the SE frequency scale, such as *hey*, *gee*, *ooh* and *man* or such conjunctions and adverbials as *plus* or *anyhow* which are also infrequent.

Contrary to the hypothesis, however, there are indications that hesitations and SEs can also undergo chunking if the SE follows the hesitation.

- From Node 3 to Node 4 – i.e. with increasing SE frequency – the preference for 'hesitation SE' gets stronger. Node 4 is the least noisy of all terminal leaves.
- Again, the proportion of intervening pauses declines with increasing SE frequency. From Node 3 to Node 4, the proportion of interrupted 'hesitation SE' sequences decreases from 19.48% to 9.9% (i.e. the change in the ratio of A to B outcomes is significant at the $p<.05$ level, based on a chi-square test).
- In Node 4, where there is a clear preference for hesitating before uttering the SE, the reverse order almost never occurs. Thus the order 'hesitation SE' appears to be chunked, blocking 'SE hesitation'.

I argue that this is also a case of frequency-induced chunking. In the context of the above *ctree* model, SEs in Node 4 appear infrequent in comparison to the two higher-frequency nodes, yet they still have a corpus frequency in the 3,000s, making them highly frequent words. If these SEs were preceded by hesitations because they are difficult to retrieve, we would expect far more unfilled pauses after the vocalised hesitation. Finally, the observation that some of these combinations have taken on lexicalised meanings shows that these 'hesitation SE' combinations are chunked and can be used as pre-assembled utterance launchers or time-buying devices.

5.7 Chunking & Hesitation Placement in the Verb Phrase

Firstly, the question of whether verb phrase components are particularly likely to form chunks and secondly of how likely chunks are to form between subjects and the verbs following them are two aspects of particular interest in this analysis. Bybee and colleagues (Bybee 2010; Bybee and Torres Cacoullos 2009), in particular, have drawn attention to chunking in negated verb-phrases, providing conclusive evidence that constructions like *can't think of* have become independent from their positive counterparts (Bybee 2010:154-5). Concerning subject-verb chunking, i.e. chunking across the verb phrase boundary, Bybee (2010:136-8) draws attention to auxiliaries which, as clitics, can form extremely strong chunks with the preceding subject, evident from the strong phonetic reduction. Figures from studies of hesitation placement (cf. Maclay and Osgood 1959:31; Cook 1971:138) seem to support this hypothesis. In these studies, the verb phrase boundary proved a far weaker attractor of hesitations than other constituent boundaries.

The overview of my data, provided in Table 5.3 and Figure 5.1 showed that here the vast majority of hesitations is neither placed in the verb phrase nor at its boundary. Of 6,317 hesitations, a mere 114 (1.8%) are placed at the verb phrase boundary, i.e. preceding the finite verb. Of those 1,714 hesitations which occur in sentences with a complex verb phrase, only 96 (5.6%) are placed within the verb phrase. (Speakers cannot place hesitations in simple single-word verb phrases because these obviously contain no hesitation-relevant transition.)

Thus, overall, speakers display a strong tendency to hesitate closer to the beginning of the sentence and avoid the verb phrase to do so. In respect to the question of whether this is an effect of frequency-induced chunking between the subject and the verb or the components of the verb phrase, the analyses conducted in Section 5.4 were not very conclusive. Models hardly, if ever, chose verb relations as predictors of hesitation placement, which could be interpreted in three ways:

1. *Statistical explanation*: The proportion of hesitations placed in the verb phrase is so small that the algorithm simply 'does not bother' with it. Any predictor explaining why hesitations are sometimes placed in the verb phrase explains the behaviour of such few data-points that its influence is deemed insignificant.

2. *Processing explanation*: In sentence-initial contexts, the attractions between verb phrase components do not have much of an influence on hesitation placement – the tendency to place hesitation at the sentence boundary is just too strong. Furthermore, many hesitations occur in dummy combinations which can only be used at the absolute beginning of the sentence.

3. *Chunking explanation*: Hesitations hardly occur in the verb phrase or before the verb because subject-verb combinations and verb phrase components are mostly chunked. The little variation we find merely points to a few exceptions.

4. *Zero-hypothesis explanation*: Frequency of use does not influence language processing in the verb phrase. Chunking does not explain the existing variation.

Of these explanations, we already know from the analyses conducted in Section 5.6 that number two plays a role; 'dummy chunks' like *and uh* and *but uh* are common phenomena. Explanation number one is certain to also influence results, considering that the percentage of hesitations placed in and at the boundary of the verb phrase ranges from only 1.6% (for 'Subject Verb(finite)') to 10.5% (for 'SE Subject Verb(finite)').

Explanations three and four have not been addressed so far. They are contradictory and represent the extreme ends on a scale between no chunking and constant chunking in the verb phrase.

To address explanation three, I will compare the absolute co-occurrence frequency and probabilistic chance of co-occurrence of the three verb phrase pairs

- Verb(finite) *not*
- *not* Verb(non-finite)
- Verb(finite) Verb(non-finite)

to the two pairs which do not contain a verb, namely

- SE SE
- SE Subject

Any information used for this analysis is visualised in Table 5.4, which was discussed in Section 5.2.5, and in the graphs in Appendix E. As the table and graphs show, verb phrase pairs are not generally more frequent than non-verb phrase pairs, neither do they generally score higher on the probabilistic scales. Wilcoxon rank sum tests (here preferred to t-tests because the data is not normally distributed), in fact, show that the frequency of the non-verb phrase pairs highly significantly exceeds that of the verb-phrase transitions. The same is true for direct and backwards transitional probability and lexical gravity G. Only the mutual information score rates the verb phrase pairs as a group more cohesive than the non-verb phrase pairs (based on a Wilcoxon rank sum test). This finding refutes categorical explanation three; if verb-phrase pairs are not measurably more cohesive than non-verb phrase pairs, the absence of hesitations in the

verb phrase cannot be explained by the argument that verb phrases are generally chunked.

Nevertheless, further comparisons of this sort reveal a number of interesting facts about negated verb phrases. These stand out as particularly cohesive types of verb phrases. The frequency, direct transitional probability and mutual information score of 'Verb(finite) *not*' pairs highly significantly exceeds that of the two non-verb phrase bigram types (based on Wilcoxon rank sum tests). Furthermore, '*not* Verb(finite)' pairs are rated highly significantly more cohesive than 'Verb(finite) Verb(non-finite)' pairs on all scales except the mutual information score (based on Wilcoxon rank sum tests). These two facts taken together offer explanations for the hesitation placement pattern in verb phrases, observable in Table 5.3:

- Except for a single case, hesitations are not placed before *not*. This is due to the high average cohesiveness of 'Verb(finite) *not*' pairs, which in turn is due to the high scores achieved by *don't* and the extremely high rate of contracted forms in general.
- In negated verb phrases, far fewer hesitations are placed before the non-finite verb than in non-negated verb phrases, which could be explained by the fact that '*not* Verb(non-finite)' is on average more cohesive and frequent than 'Verb(finite) Verb(non-finite)'.

These findings about the strong cohesiveness in negated verb phrases are in line with Bybee (2010) and Bybee and Torres Cacoullos (2009), who draw attention to the chunkiness of negated verb phrases.

I also compared the 'Subject Verb(finite)' bigram, which bridges the verb phrase boundary, to the two pre-verbal transitions. The frequency and cohesiveness of this group of transitions significantly exceeds that of the pre-verbal transitions on all scales except lexical gravity G (based on Wilcoxon rank sum tests). This can be interpreted as an indication that 'Subject Verb(finite)' pairs are more likely to be chunked than 'SE SE' and 'SE Subject' pairs and that hesitations are consequently less likely to be placed before the finite verb than at positions earlier in the sentence.

Results presented so far show that, at the bigram level, verb phrase components are not always chunked. Some transitions, in fact, appear less chunky than pre-verbal sentence-initial types of pairs. There are, however, indications that at least some negative constructions and subject verb combinations are chunked. The arguments discussed suffer from one major shortcoming, though: they assume that the group as a whole will reflect the behaviour of individual data-points. In positions where the group as a whole was comparatively frequent or showed strong attractions between words and

contained few hesitations, the absence of hesitations was interpreted as a result of high frequency or strong attractions. Such lines of argumentation are vulnerable to criticism because they do not rest on analyses of individual data-points. Only if we know that those pairs in the group which are interrupted by hesitations are actually the ones of low-frequency, low-transitional probability etc. do we have more profound evidence of chunking.

Consequently, in a second step of the analysis, verb phrase transitions were split into a hesitant and a fluent group each. If highly frequent or measurably attracted pairs are chunked, then the groups of fluent pairs should be more frequent and more attracted than the disfluent groups.

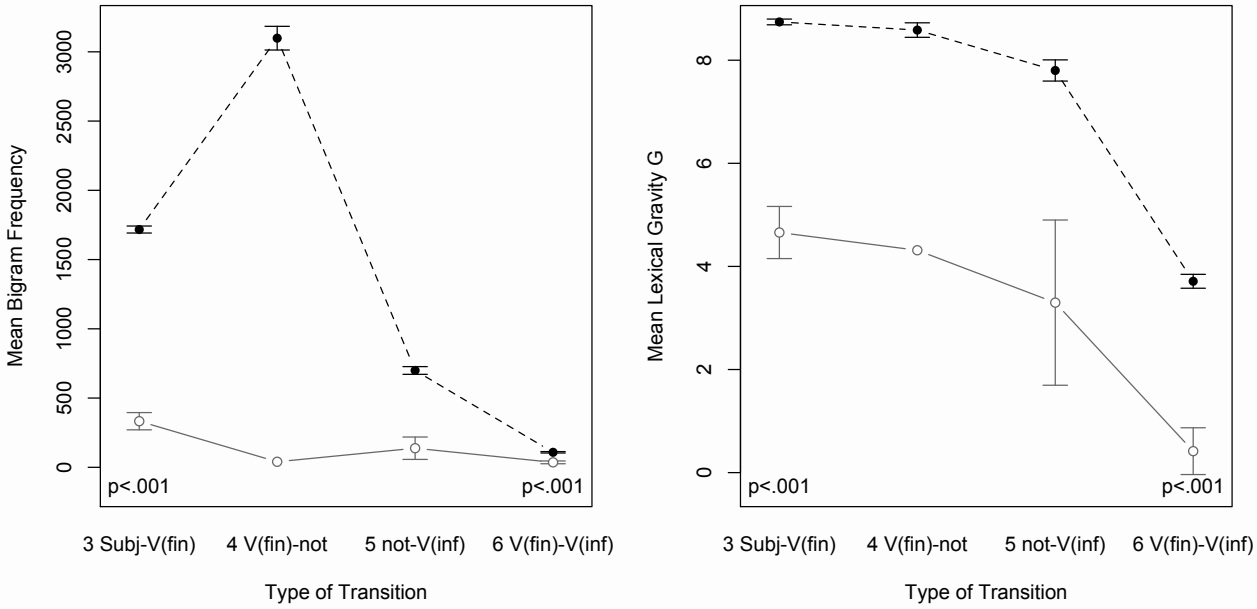


Figure 5.27: Comparison of the frequency and lexical gravity G of fluent (dashed black line) and hesitant (solid grey line) verb-phrase transitions.

Figure 5.27 exemplarily shows the results for bigram frequency (first diagram) and lexical gravity G (second diagram). The graphs show that fluent pairs (shown in black) are highly significantly more frequent and receive a highly significantly higher G score than the interrupted pairs (shown in grey). The given significance levels are based on Wilcoxon rank sum tests which were preferred to t-tests because the data is not normally distributed. Note that no significance tests were conducted for 'Verb(finite) *not*' and '*not* Verb(non-finite)' because only one 'Verb(finite) *not*' pair (out of 565) and ten '*not* Verb(non-finite)' pairs (out of 565; see Table 5.4) are hesitant. Results according to the

other measures of association are similar, except backwards transitional probability rates the disfluent cases of 'Verb(finite) Verb(non-finite)' as marginally more cohesive than the fluent cases ($p>.1$, see Appendix N).[53] In summary, the presented analyses constitute evidence refuting explanation number four. There clearly is chunking in the verb phrase and across its boundary.

[53] Interestingly, the pattern observed for hesitant and fluent prepositional phrase pairs does not emerge. In the case of 'X Preposition' pairs discussed in Section 4.6, very fluent groups of bigram types were characterised by a positive mutual information score and received the highest direct transitional probability for this score. Disfluent pairs, in turn, were characterised by a low mutual information score and low direct transitional probabilities (see, for example, Figures 4.17 and 4.19). Figure O.1 in Appendix O shows that this pattern is not observable here.

5.8 Summary & Discussion

In this chapter, I analysed the placement of all 6,317 hesitations which were uttered in the corpus in the context of the following pre-verbal sentence-initial structures, where 'SE' stands for 'sentence-initial element' and encompasses coordinating and subordinating conjunctions, adverbs, discourse markers (other than those classified as hesitations for the purpose of this study) and interjections.

- Subject Verb(finite)
- Subject Verb(finite) Verb(non-finite)
- Subject Verb(finite) *not* Verb(non-finite)
- SE Subject Verb(finite)
- SE Subject Verb(finite) Verb(non-finite)
- SE Subject Verb(finite) *not* Verb(non-finite)
- SE SE Subject Verb(finite)
- SE SE Subject Verb(finite) Verb(non-finite)

Hesitations here serve as an indicator of the chunkiness of the word-pairs in the sentence. It is expected that speakers will be less likely to interrupt chunky pairs than non-chunky ones. In order to see whether the chunkiness of a sequence results from its usage frequency or the likelihood of the two words to co-occur, regression analyses are used to predict hesitation placement based on absolute co-occurrence frequencies and the following measures of relative co-occurrence frequency.

- direct transitional probability
- backwards transitional probability
- the mutual information score
- lexical gravity G

These analyses lead to the following results.

Evidence of frequency-induced chunking – Overall, models only predict significantly more hesitations correctly than simple, 'predictorless' models in four out of eight cases. So for the four structures

- SE Subject Verb(finite)
- SE Subject Verb(finite) Verb(non-finite)

- SE SE Subject Verb(finite)
- SE SE Subject Verb(finite) Verb(non-finite)

we know that chunking has a significant effect on the placement of hesitations. Those datasets where overall model performance is not rated as significant share a crucial characteristic. Hesitation placement in them is extremely homogenous; in the overwhelming majority of cases, hesitations are placed at the sentence boundary or after the first SE. In those structures where the subject is the first word in the sentence, only between 1.6% and 7.8% of hesitations are not placed at the sentence boundary. In 'SE Subject Verb(finite) *not* Verb(non-finite)' sentences, the majority of hesitations are placed after the SE. Here, 20.9% of hesitations are placed elsewhere, but this is still significantly less variation than in the four sets where models show significant effects. In these four sets, between 29.5% and 47.2% of hesitations are not placed at the most popular location. Thus I conclude that models do not produce significant effects in four sub-sets of the data because the level of variation in these sets is simply too small to generate significant effects. The following sections summarise factors which lead to this lack of variation.

Sentence-initial hesitation chunks – The data provides conclusive evidence that speakers utilise chunked combinations of sentence-initial elements (SEs; in this case predominantly conjunctions) and hesitations as longer time-buying and turn-keeping devices as well as lexicalised sentence launchers. Which SEs enter into such chunks and the order of the SE and the hesitation in the chunk depends strongly on the frequency of the SE. The more frequent the SE, the more likely it is to be part of a 'dummy' chunk and to be the first element in the chunk. This means that highly frequent *and* and *but* form *and uh* and *but uh*, while less frequent *if* and *then* form *well if* and *well then*. Truly infrequent SEs do not appear to enter into dummy chunks.

The ordering of the elements results from highly frequent SEs being more easily retrievable than lower-frequency ones. Frequent SEs are so easily retrievable that no time-buying device is necessary and hesitations triggered by the planning of the rest of the sentence follow them, thus 'SE hesitation' chunks form. Lower-frequency SEs, on the other hand, may take time to retrieve, hence speakers hesitate before them and 'hesitation SE' chunks emerge.

Although there was some evidence of such dummy chunks in the prepositional phrase data-sets, they appear to be a predominantly sentence-initial phenomenon, possibly resulting from the need for longer time-buying devices before having planned substantial parts of the sentence, i.e. at a point when the speaker does not yet have material available which he could repeat to buy time.

The sentence as the central unit in speech planning – The previously described findings are in line with Power (1986) who claims that the sentence is the central unit in speech planning. His results indicate that mid-sentence, speakers start planning a clause while uttering the preceding one. If the clauses are part of different sentences, however, speakers do not start planning the second before finishing the first. This leads to the conclusion that the planning load at the beginning of the sentence is particularly high.

In the present data, hesitations are predominantly utilised before the start of the sentence. A far smaller proportion is uttered within the sentence than within the previously-analysed prepositional phrases. This becomes particularly evident if we keep in mind that a large proportion of hesitations which appear to occur after the first word in the sentence are actually part of a dummy chunk and thus form one longer hesitation together with the first word (which in such cases is mostly semantically-unspecific *and* or *but*).

Relations between the subject and the verb – Bybee (2007a, 2010) claims that constituency is a result of frequent co-occurrence, while, on the other hand, pointing out that some traditionally defined constituent boundaries should be reconsidered. She particularly draws attention to enclitics like '*ll* and '*m*, which form a unit with the preceding subject. So, chunking of elements to both sides of the verb phrase boundary seems possible and, in fact, common. Data from Cook (1971) as well as Maclay and Osgood (1959) indicates that the rate of hesitations placed at verb phrase boundaries is, indeed, lower than that at other types of phrase boundaries.

Results obtained in the present study are in line with these findings. Hesitation rates at the verb phrase boundary stand in stark contrast to those at the prepositional phrase boundary. While the prepositional phrase boundary is a popular location to hesitate, only 114 out of 6,317 hesitations (1.8%) are placed at the verb phrase boundary. Analyses showed that 'Subject Verb(finite)' combinations are, on average, chunkier, i.e. more frequent and likely to co-occur, than the combinations which precede them in the sentence and that their fluency not only results from the tendency to hesitate at the sentence boundary, but also from their 'chunkiness'.

Relations within the verb phrase – Interestingly, on average, verb phrase transitions are no more frequent or cohesive than the pre-verb-phrase transitions in this set, yet few hesitations are moved into the verb phrase. Thus particularly 'Verb(finite) Verb(non-finite)' combinations are unlikely to be strongly associated according to the selected measures of association. They are even rated less attracted than pre-verbal sentence-initial types of pairs. Negated constructions, particularly 'Verb(finite) *not*' pairs, however seem chunked. They are comparatively frequent, measurably attracted and only a in a

single case a hesitation intervenes. Analyses show that hesitations are only moved into the verb phrase when the associations between the words in the phrase are weaker than in other verb phrases.

6 Discussion & Conclusion

This study investigates how multi-word sequences are mentally represented and how this representation is shaped by different usage-based factors. I approach these issues by means of an analysis of where speakers hesitate in speech. I hypothesise that sequences of words should be more strongly represented in the mind the more likely the words in them are to occur together. As a consequence of the representations of common sequences such as *I don't know* and *I'm trying* being stronger than the representations of rare combinations such as *I don't recall* or *I am attempting*, speakers should be more likely to utter the former without interruptions than the latter.

Underlying these assumptions is Bybee's Linear Fusion Hypothesis which states that "items that are used together fuse together" (Bybee 2007b:316). Bybee postulates that sequential links develop between items that are used together frequently, so that the items in the sequence prime and automate each other. This chunking process is supposed to start with the first encounter of a sequence. Bybee furthermore argues that the mind is organised as a network of exemplars wherein every new token encountered is stored as a new exemplar which is strengthened through repeated use (Bybee 2006:716). Initially, representation is still weak, so that it is easier for speakers to access sequences by their parts; however, through repeated use the representation of the sequence is strengthened and it becomes more easily accessible as a whole (cf. Bybee 2010:36; Bybee 2006:716-7).

While Bybee (2010:97) argues that the degree of chunkiness of a sequences is determined by the amount of times the words in the sequences have occurred together, other studies, such as for example Wiechmann (2008), use measures of the relative chance of the words to co-occur as a determinant of chunking strength. The present study evaluates these different approaches by empirically testing whether speakers' hesitation behaviour can better be modelled by absolute co-occurrence frequency or by relative measures of association such as transitional probabilities, the mutual information score and lexical gravity G. Speakers should be less likely to interrupt a strong chunk in order to hesitate than to interrupt a weakly-chunked sequence. In this way, we can explain that in the following examples highly-frequent *we've got* is uttered as an uninterrupted unit, while the much rarer sequence *we've enjoyed* is interrupted.

(171) *you know* we've got (sw2331.A.s133)

(172) we've *uh [pause]* enjoyed (sw2316.A.s154)

The placement of hesitations is analysed with the help of Classification and Regression Trees (CART Trees) and random forests. These non-parametric methods of regression select predictors which help them to separate the data into ever smaller and more homogenous subgroups. It is also an important aspect of this work to evaluate whether these new statistical methods can profit linguistic analyses.

For my studies, I employ the Switchboard NXT corpus, which consists of transcripts of telephone conversations conducted in American English. I model chunking on the so-called bigram level, meaning that only two-word strings of surface-level word-forms are taken into consideration. Bigram frequencies and measures of association are calculated based on the Switchboard NXT corpus. Furthermore, more than 11,000 filled and unfilled pauses as well as discourse markers occurring in the context of prepositional phrases and sentence-initial structures are extracted from the corpus.

In a first study, hesitation placement in prepositional phrases is analysed. Six types of prepositional phrases of different complexity are selected for analysis, ranging from "Preposition Noun" to "Preposition Determiner Adjective Noun". In the data, hesitations are most commonly placed at the prepositional phrase boundary and before the first content word in the phrase, but they also occur at all other transitions. Analyses show significant effects in three out of six phrase types. Results thus confirm that the more likely two words are to co-occur, the less likely speakers are to place a hesitation between them. Importantly, CART trees select not only frequency of co-occurrence as a predictor, but also measures of association. In fact, more splits are made based on measures of association than based on absolute frequency of co-occurrence and some trees never use absolute frequency as a predictor. Random forests furthermore score the usefulness of predictors in a model. These scores show that there is no significant difference in performance between the measures of association and co-occurrence frequency.

Results indicate that chunking across the prepositional phrase boundary is very common and that across all phrase types analyses are best at predicting chunking in this position. Importantly, chunks consisting of the words to both sides of the prepositional phrase boundary are characterised as a group by a specific ratio of high mutual information scores and high direct transitional probabilities. The variable importance scores awarded by random forests show that in these cases the mutual information score marginally outperforms absolute co-occurrence frequency.

In the three phrase types where effects did not reach significance, we find evidence for a phenomenon which so far has received little attention: For the most part, the 'chunking inventories' of all speakers of a language community overlap, resulting from speakers having to form sentences by the grammatical rules of the language and using a

common set of lexicalised social formulae. Additionally, however, every speaker forms idiolectal chunks. The ones found in the present database relate mostly to speakers' personal interests and their environment. In particular, we see that names of hometowns (e.g. *Boise, Idaho*) as well as schools and employers (e.g. *Richardson Symphony*) are mentioned so frequently and fluently that they must be considered chunked for the individual speakers although the corpus frequencies do not reflect this.

A separate study tests whether speakers predominantly place hesitations between the least associated words in a phrase. Results reveal that according to co-occurrence frequency and all other measures of association (except backwards transitional probability) speakers are most likely to place hesitations where words have the smallest chance to co-occur. Thus, this analysis confirms that direct transitional probability, the mutual information score and lexical gravity G perform on par with absolute co-occurrence frequency.

A second set of studies then investigates hesitation placement in sentence-initial contexts. Eight different sequences are selected for analysis. These differ in terms of the complexity of the verb phrase and the number of sentence-initial elements permitted before the subject (such as adverbs, interjections and coordinating conjunctions). This means that contexts range from 'Subject Verb(finite)' to 'Sentence-Initial Element Sentence-Initial Element Subject Verb(finite) *not* Verb(non-finite)'. The distribution of hesitations in this dataset differs strongly from that in the prepositional phrase dataset in so far as preferences for placement in certain positions are much stronger. Speakers predominantly hesitate at or very near the beginning of a sentence. Few hesitations are placed after the subject or within the verb phrase.

This dataset furthermore reveals that hesitations themselves can become part of chunks. In particular, highly frequent coordinating conjunctions, such as *and* and *but*, often merge with following fillers. In this way, longer time-buying devices like *and uh* and *but uh* emerge.

Model performance on these data-sets only reaches significance in four cases. This is mostly due to the fact that in all three structures where no sentence-initial elements precede the subject, hesitations are almost exclusively placed at the sentence boundary. Some of the little variation there is in these contexts can be explained by means of the frequency-derived predictors. Overall, the variable importance scores obtained from the random forests of this set of analyses confirm that measures of association perform on par with co-occurrence frequency.

In summary, through the analysis of the placement of hesitations in various prepositional-phrase and sentence-initial contexts, this work provides empirical evidence of chunking. On the most basic level, it shows that speakers tend to produce sequences

of words fluently in which the words are likely to co-occur based on both absolute and relative frequency of co-occurrence. When they need to hesitate, they utter such chunks uninterruptedly and prefer to place the hesitation elsewhere in the surrounding context. On a more abstract level, it demonstrates that the mind appears to track not only absolute co-occurrence frequency but also relative chances of co-occurrence.

These results allow for a number of conclusions, pertaining to both methodological questions and to questions related to linguistic model building. In the following, these questions will be addressed, starting with methodological issues and then moving on to concerns in model building.

What are the advantages of CART trees and random forests?

On a methodological level, the analyses attest that Classification and Regression Trees (CART trees) and random forests, which are just becoming known in linguistics, can greatly benefit the field. Both single trees as well as random forests operate based on a simple mechanism. They repeatedly split the data, resulting in ever smaller, yet more homogenous, subgroups. With each split, two subgroups are obtained which differ from each other in terms of two aspects: hesitation placement and the splitting criterion. This type of procedure allows for conclusions concerning under which conditions speakers make the same decisions. Table 6.1 illustrates the possible outcomes of a split based on frequency of co-occurrence.

	- Frequent	**+ Frequent**
- Hesitations	A	B
+ Hesitations	C	D

Table 6.1: Possible outcome types.

In a very basic model, which simply investigates the presence or absence of a hesitation in between two words, splitting the data according to the usage frequency of all pairs in the dataset means that one group is created which contains the higher-frequency pairs ('+ Frequent') and one group which contains the lower-frequency pairs ('- Frequent'). As the split will only be made if this separation also leads to two groups which differ in terms of their hesitation behaviour, we also find that one group will contain more hesitations ('+ Hesitations') than the other ('- Hesitations'). The pairs in the '+ Hesitations' group will thus be more likely to be interrupted than those in the '- Hesitations' group. Table 6.1 shows that this leads to four possible combinations of features in a group.

In the context of this analysis, outcome B is the most important because it confirms my usage-based chunking hypothesis: Where we find that the higher-frequency pairs are less likely to be interrupted by hesitations we can argue that the frequently used pairs are 'chunkier' than their low-frequency counterparts.

If one resultant group has attribute combination B, the other group must necessarily display outcome C, which indicates that with a reduction of usage frequency comes a reduction of fluency. The combined outcome of B and C shows that more practised sequences are more likely to be pronounced fluently. From this, we can conclude that in frequent sequences, the words evoke each other so that once the first word has been retrieved, the second follows automatically and no further pauses are needed. We can further conclude that if the speaker needs a time-buying device due to other planning difficulties in the sentence, frequent pairs, just like single words, are unlikely to be interrupted and any hesitations are rather placed elsewhere in the vicinity.

The other possible combination of outcome types is A and D. Result A, as such, does not present counter-evidence to a theory of chunking. It merely reveals the obvious, namely that even infrequent sequences may be uttered fluently. As the majority of two-word pairs in an average conversation are uttered fluently, this is a default situation. As the result of a split in a CART tree, however, A will always be paired with D. The combination of A and D constitutes counter-evidence of the chunking theory. If oft-used pairs are more likely to be disfluent than more rarely used pairs, we cannot argue that the former are more cohesive or more easily retrievable in combination.

Across all prepositional phrase datasets, not a single split leads to result A-D. Instead, all splits provide results B and C. In the dataset of sentence-initial structures, three splits emerge which create outcomes A and D (Split 9 in Figure 5.8, Split 1 in Figure 5.12 and Split 11 in Figure 5.14). All of these, however, result from chunks forming between a sentence-initial element and a hesitation, which are difficult for the models to handle due to the setup of the study.

A final, so far undiscussed, option of CART trees is simply not to split the data. Where the frequency of the pairs has no effect on their chance of being interrupted, no split is made. All models discussed in this study create at least one split.

Crucially, the analyses conducted in this study show that absolute co-occurrence frequency and measures of associations between words have the same influence on processing, i.e. according to all criteria we find that the higher the score awarded to a pair of words (e.g. its frequency or its transitional probability) the less likely speakers are to interrupt the pair to hesitate.

Overall, results show that CART trees and random forests are useful tools for such types of analyses. Not only can they handle the fact that speakers mostly have more than

two options available where they can place a hesitation and that some of the predictors are necessarily correlated, CART trees also objectively group the data into homogenous subgroups which facilitate comparisons across several parallel analyses. Random forests complement the individual trees by evaluating whether results can be generalised. They do this by means of testing the predicted effects on unseen data. Results of these 'out-of-bag' tests furthermore show that a single CART tree generally already provides a reliable result.

What is the best measure of chunking strength?

Earlier, I postulated that measures of the relative chance of co-occurrence should be more apt at predicting chunking than absolute co-occurrence frequency. The need to compare different means of measuring chunking strength arises from Bybee's (2010) argument that absolute co-occurrence frequency, as approximated by the bigram frequency in a corpus, should be the most important determinant of chunking. If this were indeed the case, we would find that the predictive value of bigram frequency would be much higher than that of measures which evaluate the relative chance of two words to co-occur. Yet objective measurements obtained with the help of random forests reveal that this is not the case. Neither bigram frequency nor lexical gravity G, a measure of the relative chance of co-occurrence which strongly correlates with pair frequency, outperforms transitional probabilities and the mutual information score.

I furthermore hypothesised that predictors should rank in a specific order. The more complex the calculation of a probabilistic measure of association, the better a predictor of chunking it should be, because the better is its understanding of the distributions of words in English. Thus co-occurrence frequency should be outperformed by transitional probabilities, which, in turn, should be outperformed by the mutual information score and finally lexical gravity G.

This hypothesis is falsified. More complex measures neither outperform bigram frequency, nor can the performance of all predictors be ranked by their complexity. In certain contexts, individual measures or combinations of measures stand out as particularly successful. In the case of chunking across the prepositional phrase boundary, for example, there is an indication that the mutual information score is more predictive of hesitation placement and consequently of chunking than absolute co-occurrence frequency. There is also an indication that an interaction of the mutual information score and direct transitional probability delimits particularly hesitant from particularly fluent types of combinations. Such advantages of one predictor over all others appear to be highly context-specific, though. I conclude that none of the tested measures can be advocated as the single best predictor of chunking, as differences in

performance are minor. Modelling improves however if, as done here, measures of the relative chance of co-occurrence are combined with absolute co-occurrence frequency.

Is chunking a gradual process or are there threshold levels?

Overall, results tie in with Bybee's (2010) hypothesis that chunking is a gradual process, already starting at low frequencies. Though, in the present case, CART tree splits near the lower end of a scale were rather made for lexical gravity G and transitional probabilities. Thus, we could say that chunking starts at very low chances of co-occurrence.

Furthermore, the data shows no indication of confirming the opposite hypothesis, namely that chunking is an abrupt mechanism, in the sense that only high-frequency sequences, or words that are highly likely to co-occur, are stored holistically and no effects are expectable for low-frequency sequences or words which are not very likely to co-occur. Figure 6.1 illustrates the shape of the effect curve we would expect to find if chunking were an abrupt mechanism. We would expect no chunking effects below the threshold and full 'chunkedness' afterwards. In CART trees, all splits should thus be made at the exact same level or within a very narrow range. It follows that only one predictor may be successful because data-points which lie to the left and the right of a threshold on one scale do not necessarily do so on another.

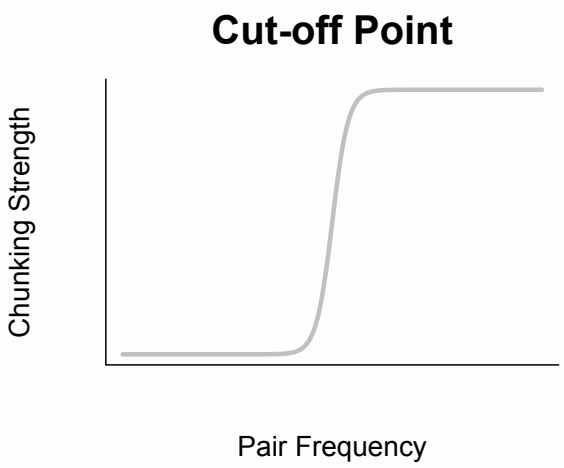


Figure 6.1: Correlation between frequency of co-occurrence and chunking strength in a threshold model.

These assumptions are not borne out by the data. First of all, there was no one predictor which stood out as the sole determinant of chunking. Secondly, on most scales, splits do not fall within a single, narrow range. Only splits in the prepositional phrase data-set made according to the mutual information score fall in a range between scores of 1.8 and 3.5. (In the data-set of sentence-initial structures, MI is only once chosen as a

predictor, for a split at a value of 5.1.) Figure C.4 in the Appendix however shows that 'X + Preposition' pairs, for which the majority of these splits were made, mostly fall in a range between scores zero and five and therefore splits within this range are highly expectable. A small difference in the splitting point within this range has a large effect on the resulting grouping of the data. The CART trees thus do not provide evidence that chunking is an abrupt mechanism.

How do phrase structure and frequency effects interact?

There is another aspect of Bybee's work which can be addressed by means of this data: the role of phrases and constituents in processing. Bybee (2007b) holds that constituents as postulated by phrase structure grammar are not cognitively real. She argues that constituency is a gradual concept derived from patterns of combined usage and furthermore that some traditionally-assumed constituent boundaries should be reconsidered. Earlier studies of hesitation placement, on the other hand, have frequently shown that hesitations often fall at phrase boundaries (cf. e.g. Maclay and Osgood 1959; Goldman-Eisler 1968; Biber et al. 1999) and thus suggest that these are units of encoding.

What both approaches have in common is that they assume that hesitations are placed at the boundaries of units of encoding. They merely differ in their conceptualisation of what these units are: syntactic or frequency-derived. Thus we need to evaluate which of the two factors primarily influences hesitations placement – and should therefore be considered primary. The present study was designed to allow for this by restricting the analysed sequences syntactically, thus keeping syntactic form stable.

In the data-sets, sentence boundaries and prepositional phrase boundaries proved to be strong attractors of hesitations. Speakers tend to place hesitations at prepositional phrase boundaries unless the boundary is obscured by a strong chunk stretching across it, such as, for example, *one of* or *all of*. This might initially suggest that the model should include phrase boundaries as a separate factor influencing hesitation placement.

However, the large number of chunks which violate the phrase boundary cannot be explained if we assume that the prepositional phrase in general is a unit of processing. Furthermore, there are very few hesitations placed at the verb phrase boundary and we find many cases where the finite verb is cliticised onto the subject. These findings rather suggest that hesitations fall at the boundaries of frequency-derived units, which often, but not always, coincide with traditionally-assumed phrases.

The present study can only offer tentative conclusions, though, because at the bigram level neither absolute co-occurrence frequency nor relative measures of association always predict the strong unit-like behaviour we see in some types of word pairs, particularly in the verb phrase and among subject-verb pairs. Further studies operating

with a larger window of analysis, i.e. longer n-grams, would have to confirm the suspicion that where speakers place hesitations at phrase boundaries, these boundaries, in fact, coincide with the boundaries of usage-based units.

Which kind of model best explains the effects found in the present study?

There are two basic types of models which can be employed to explain frequency effects like chunking: ones that postulate holistic storage and ones that do not. In the latter type of model only the parts (in this case words) need to be stored as units and anything larger than them is 'stored' in the form of connections between the parts (and possibly other processing units), so that a large network emerges. Frequency effects occur because combined use strengthens these connections. This concept stems from connectionist or parallel-distributed processing frameworks such as McClelland and Rumelhart's (1981). It is also found in simple recurrent networks, such as Elman (1990) and corresponds to what was termed a 'distributed account' in Kapatsinski and Radicke's (2009) study detailed in Section 2.2.2.

In a connectionist network either co-occurrence frequency, the probabilistic measures of co-occurrence or an interaction of several of these reflect the intensity of the connections between words and thus chunking strengths. This leads to the prediction that the chunkier the sequence, the less likely it is for hesitations to intervene – which is indeed the kind of effect we find.

An exemplar model, on the other hand, belongs to the former group and assumes that tokens of all grain-sizes are stored in memory. If a new token is identical to an already stored exemplar, it strengthens the representation of this exemplar. If the new exemplar differs from previous experiences, it is stored as a new exemplar and located close to similar exemplars. Importantly, this also holds for sequences longer than a word and for more abstract constructions, meaning that chunks are stored holistically from the first time they are encountered. Initially, when the whole has not yet been used very often, it is easier for the speaker to access the parts, while with increasing frequency of combined use it becomes easier to access the whole. The model by Bybee (2006; 2010), which was at the focus of this study, is an exemplar model; furthermore, Langacker's (2000) Dynamic Usage-based Model shares central characteristics with exemplar models.

In an exemplar model, co-occurrence frequency and/or the probability of co-occurrence reflect the strength of the representation of the chunk. Like connectionist models, exemplar models predict that the chunkier the sequence, the less likely it is for hesitations to intervene.

We might argue, however, that these two types of models do not make the same predictions throughout: The exemplar model predicts that when speakers access the

whole, they produce it as a unit and therefore without intervening hesitations. When the speaker retrieves the sequence by means of accessing the parts, hesitations might intervene. Yet as the parts of a chunk must, by necessity, be at least as frequent as the chunk, it follows that there should always be competition between the parts and the whole.

Many sequences which behaved very chunky in the present dataset contain at least one highly frequent element. Therefore, in the case of sequences like *much of, any of, out of, of course* or *and uh*, an exemplar model additionally has to explain why we do not find signs of competition between the parts and the whole, i.e. why we do not find an increase in hesitations in sequences containing a highly frequent word and why word-frequencies were generally poor predictors of hesitation placement. Thus the model needs an additional explanatory factor to account for the lack of competition from the parts. This factor could be semantics, i.e. when the meaning of the chunk is no longer fully compositional the connections to its components begin to weaken. In Langacker's (2000) aforementioned model, for instance, the level of activation of a node in the network – and thus ultimately the degree of competition between nodes – is not only determined by entrenchment (i.e. their overall frequency), but, among else, also by the degree of semantic fit between the target and the structure. Langacker holds that longer structures are never fully compositional, so that the whole always offers a better semantic fit than the parts.

Yet a semantic filter of this sort not only renders a theory considerably more complex, the degree of semantic fit is also hard to determine objectively and the filter is consequently not easily incorporated in a statistical model. As the data showed that pairs scoring a high MI are (parts of) semantic units, we could interpret the MI score – and possibly other types of probabilistic measures of co-occurrence – as reflecting semantic unity though certainly not the degree of semantic fit between the target and a structure.

Many existing models are, in fact, more complex than the ones sketched so far. Some incorporate both the possibility of holistic storage and connection strengthening. So-called 'localist' connectionist approaches, such as the one advocated by Kapatsinski and Radicke (2009; see Section 2.2.2), for example, assume that a network of connections between words emerges, but that additionally sequences longer than a word can be represented by a single node in the network.

Bybee also describes both holistically stored exemplars and strengthening of 'sequential links' between words (cf. e.g. Bybee 2007b). Importantly, Bybee assumes that connections between the whole and its parts may weaken (Bybee 2010:52) – a process which may, for example, be caused by shifts in the meaning of the whole and which may explain the absence of signs for competition between a chunk and its parts in the

present data. Another explanation which can be deduced from the assumptions of the model would be that through entrenchment additional sequential links develop between the parts because the mind has registered that when the first word of the chunk has been retrieved, the second one is likely to follow. As a consequence, repeated access to the parts also leads to fewer hesitations in the string, so that at a surface level we can no longer distinguish between access to the whole and access to the parts. Most importantly, however, in this framework, chunking is only a small gear-wheel in a large clockwork of interacting effects. Which gear eventually propels storage, processing and development of a specific structure, cannot easily be statistically deduced from frequencies alone; semantics also play a crucial role.

In conclusion, exemplar models which are based on the idea that every new token encountered is stored as a new entry in the lexicon can account better for the mapping of new meanings onto existing strings, which is necessary to explain grammaticalisation and lexicalisation. Due to their need to incorporate semantic material, such models, however, need more complex assumptions in order to describe the effects of chunking on a processing phenomenon like hesitation placement.

Consequently, a model which conceptualises chunking as entrenchment in the sense of strengthened connections between the nodes of the constituent parts can explain the different choices in hesitation placement in the present data with fewer assumptions than a model which conceptualises chunks as holistically stored exemplars. Thus I conclude that hesitation placement is best modelled by means of a connectionist model which operates based on absolute and relative co-occurrence frequencies.

In summary, this study corroborates the claim that corpora can be used for psycholinguistic model building and contributes knowledge towards how frequency effects can be implemented in linguistic models. It provides evidence that the mind keeps track of such usage-based factors as absolute and relative co-occurrence frequencies. A model which operates based on these factors can explain hesitation placement significantly better than models which do not take these factors into account. Nevertheless, not all of the speakers' choices could be explained. For one thing, we will never be able to predict *if* a speaker needs to hesitate at all, but models can be improved concerning their predictions about *where* a speaker will place hesitations.

First of all, the present study only modelled chunks on a two-word level. Of course, this is only one of many levels at which the mind keeps track of usage-based factors. As explained above, a model which also incorporates the associations holding between words in longer sequences will presumably be able to explain even more of the speakers' choices. Such models easily become extremely computationally intensive though, as the relations in larger sequences are more complex to model than those in bigrams.

Furthermore, they require very large corpora for the extraction of longer sequences and the calculation of the relations holding within them.

Secondly, the present study took the precaution to use only spoken data for the calculation of frequencies, transitional probabilities and the like. This choice was made because exemplar models assume that any available information about words and larger strings is stored in the exemplar and that this may include information about register and medium. Therefore, medium and frequency may interact in processing, so that in a conversation spoken chunks may get more activated than written ones. Results of large-scale psycholinguistic studies have since shown that disregarding the medium in favour of using large amounts of data does not appear to lead to noisier results – the opposite may even be the case, namely that large amounts of data can lead to more stable results. Thus it would be interesting to see in how far the use of large-scale mixed-medium corpora would provide more stable or noisier results in a study like the present one.

Thirdly, new measures of association have since been developed. Such measures as, for example, Delta P (cf. Gries 2013) might more adequately reflect factors leading to chunking. Furthermore, frequency of use not only influences the mental representation of units, it also affects processing. Future models with the aim to fully account for the placement of hesitations would therefore additionally have to take recency effects such as priming into consideration.

Finally, it would be highly interesting both from a psycholinguistic and a sociolinguistic point of view to compare the chunking inventories of different speakers in order to see how they differ and which linguistic and extra-linguistic factors determine which mental representations are strengthened.

Appendices

Appendix A: Switchboard NXT Terminals Layer – Additional Information

Treebank3 Transcript (terminals)	
msstateID	alignment with the MS-State Transcript
nite:id	sentence number and word number in the sentence
nite:start; nite:end	start and end times of words
orth	orthographic transcription
pos	see Table A.2
punc	location of a full stop, question mark or exclamation mark
sil	element; mostly brackets filled pauses, repetitions and self-corrections
trace	trace of moved syntactic elements
word	marks all information concerning each orthographic word

Table A.1: Details terminals layer of Switchboard NXT (based on Calhoun et al. 2010:394 and Switchboard in NXT – Data Summary)

POS Tagset / Lexical Categories		
Value	**Penn Treebank**	**Switchboard NXT**
Adjective	JJ	JJ
Adjective, comparative	JJR	JJR
Adjective, superlative	JJS	JJS
Adverb	RB	RB
Adverb, comparative	RBR	RBR
Adverb, superlative	RBS	RBS
Cardinal number	CD	CD
Coordinating conjunction	CC	CC
Determiner	DT	DT
Existential *there*	EX	EX
Foreign word	FW	FW
Discourse marker		
Interjection	UH	UH
Phonetic editing signal		
Modal	MD	MD
Noun, singular or mass	NN	NN
Noun, plural	NNS	NNS
Particle	RP	RP
Predeterminer	PDT	PDT
Proper noun, singular	NNP	NNP
Proper noun, plural	NNPS	NNPS
Possessive ending	POS	POS
Personal pronoun	PRP	PRP
Possessive pronoun	PP$	PRP$
Preposition	IN	IN
Subordinating conjunction		
‘*s* as a form of *BE*		BES
to	TO	TO

POS Tagset / Lexical Categories		
Value	**Penn Treebank**	**Switchboard NXT**
Verb, base form	VB	VB
Verb, past tense	VBD	VBD
Verb, gerund / present participle	VBG	VBG
Verb, past participle	VBN	VBN
Verb, non-3rd ps. sing. present	VBP	VBP
Verb, 3rd ps. sing.present	VBZ	VBZ
***wh*-adverb**	WRB	WRB
***wh*-determiner**	WDT	WDT
***wh*-pronoun**	WP	WP
Possessive *wh*-pronoun	WP$	
Partial word, POS unclear		XX

Table A.2: Part-of-speech values relating to spoken language (cf. Calhoun et al. 2010:394; Marcus, Marcinkiewicz and Santorini 1993:317)

Appendix B: Switchboard NXT Terminals Layer – Additional Information

Syntax				
cat	see Table B.2			
nite:id	sentence number			
nite:start; nite:end	start and end times of sentences			
nt	non-terminal element			
subcat	ADV	Adverbial (other than ADVP or PP)	PRP	Purpose or reason
	DIR	Direction	PRP,TPC	Topicalised purpose or reason
	IMP	Imperative	PUT	Locative complement of *put*
	LOC	Locative	SBJ	Surface subject
	LOC,PRD	Locative predicate	SBJ,UNF	Unfinished surface subject
	MNR	Manner	SEZ	Reported speech
	NOM	Nominal (on relatives and gerunds)	TMP	Temporal
	NOM,TPC	Topicalised Nominal	TMP,UNF	Unfinished Temporal
	PRD	Predicate (other than VP)	TPC	Topicalised
	PRD,PRP	Purpose or reason predicate	UNF	Unfinished
	PRD,UNF	Unfinished Predicate		
wc	word count of the phrase			

Table B.1: Details syntax annotation (based on Calhoun et al. 2010:395 and Switchboard in NXT – Data Summary)

Syntactic Tagset			
Value	**Fidditch**	**Penn Treebank**	**Switchboard NXT**
Adjective phrase	ADJP	ADJP	ADJP
Adverb phrase	ADVP	ADVP	ADVP
Auxiliary phrase	AUX		
Clause introduced by a (possibly empty) subordinating conjunction	SBAR	SBAR	SBAR
Comp node of sbar	COMP		
Conjunction phrase	CONJP		CONJP
Declarative sentence with subject-auxiliary inversion		SINV	
Direct question introduced by *wh*-word or *wh*-phrase	SBARQ	SBARQ	SBARQ
Fragment		FRAG	FRAG
Interjection			INTJ
Interruption point in disfluency			IP
***It*-cleft or "true" cleft**		S-CLF	
Nbar	NBAR		
Not a constituent			NAC
Noun phrase	NP	NP	NP
Genitive noun phrase	NPS		
Parenthetical			PRN
Particle, for words tagged RP			PRT
Prepositional phrase	PP	PP	PP
Reduced relative clause		RRC	
Quantifier phrase	QP		QP
Reparandum in disfluency			EDITED RM
Restart after disfluency			RS
Simple declarative clause	S	S	S
Speech error			TYPO

Syntactic Tagset			
Value	**Fidditch**	**Penn Treebank**	**Switchboard NXT**
Subconstituent of SBARQ excluding *wh*-word or *wh*-phrase	SQ	SQ	SQ
Unlike coordinated phrase			UCP
Verb phrase	VP	VP	VP
***wh*-adjective phrase**	WHADJP	WHADJP	
***wh*-adverb phrase**	WHADVP	WHADVP	WHADVP
***wh*-genitive noun phrase**	WHNPS	WHNP	WHNP
***wh*-noun phrase**	WHNP		
***wh*-prepositional phrase**	WHPP	WHPP	
***wh*-quantifier phrase**	WHQP		
Constituent of unknown or uncertain category	UNK	X	X
Null element – "Understood" subject of infinitive or imperative		*	
Null element – Zero variant of *that* in subordinate clauses		0	
Null element – Trace – marks position where moved *wh*-constituent is interpreted		T	
Null element – Marks position where preposition is interpreted in pied-piping contexts		NIL	

Table B.2: Syntactic values (cf. Bies et al. 1995; Calhoun et al. 2010:395; Hindle 1994:132;. Marcus, Marcinkiewicz and Santorini 1993:321)[54]

[54] While Hindle (1994) does not list the tagset for null elements, they must have also been annotated by Fidditch.

Appendix C: Characteristics of Prepositional Phrase Transitions

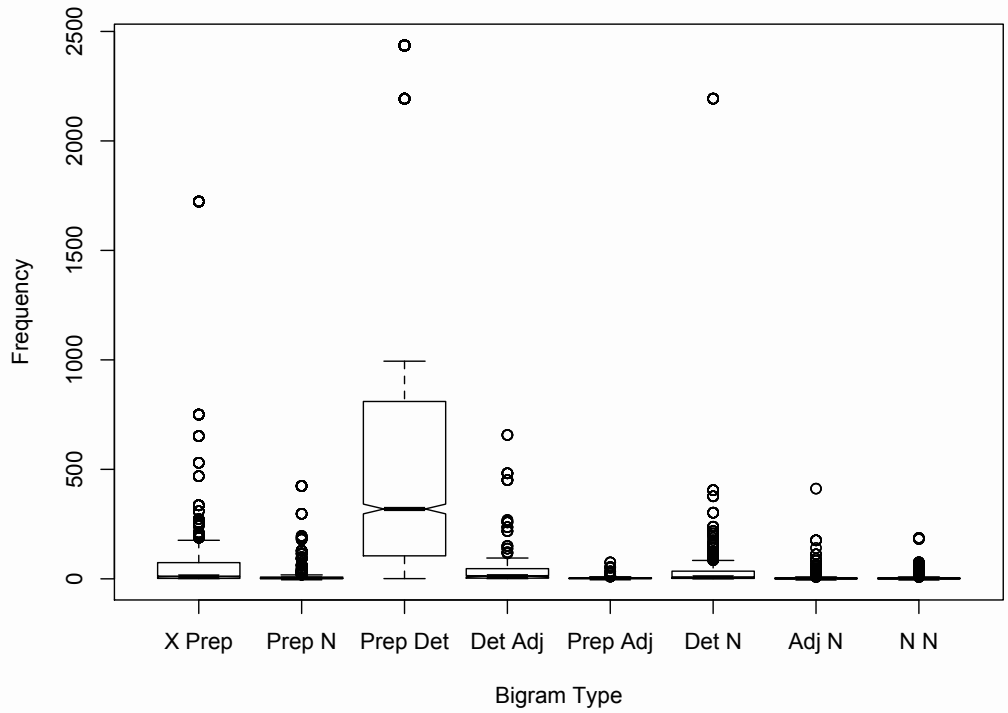


Figure C.1: Frequencies by transition type

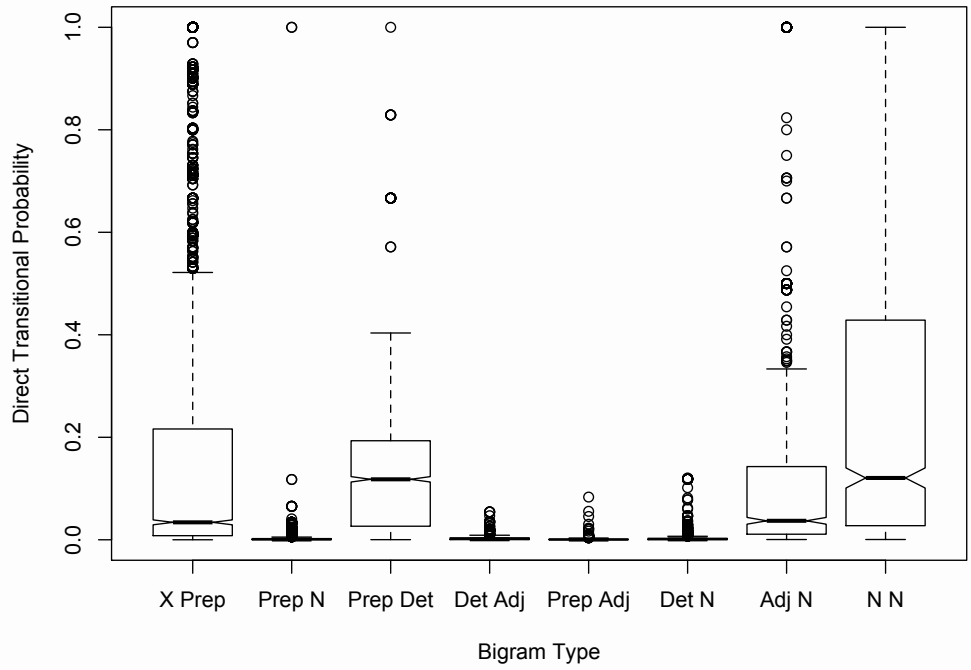


Figure C.2: Direct transitional probabilities by transition type

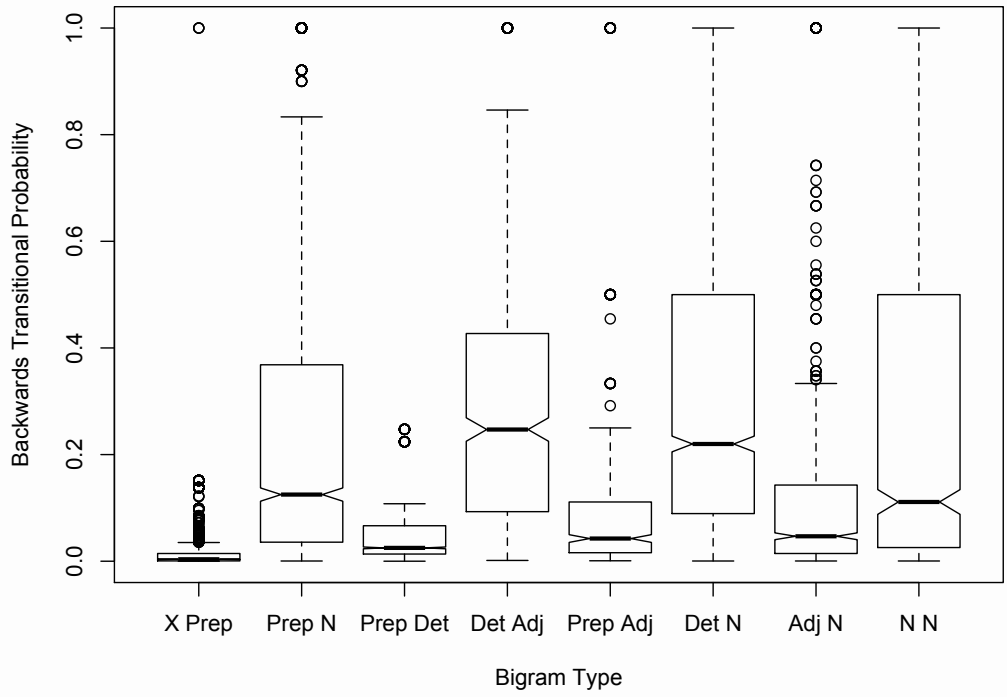


Figure C.3: Backwards transitional probabilities by transition type

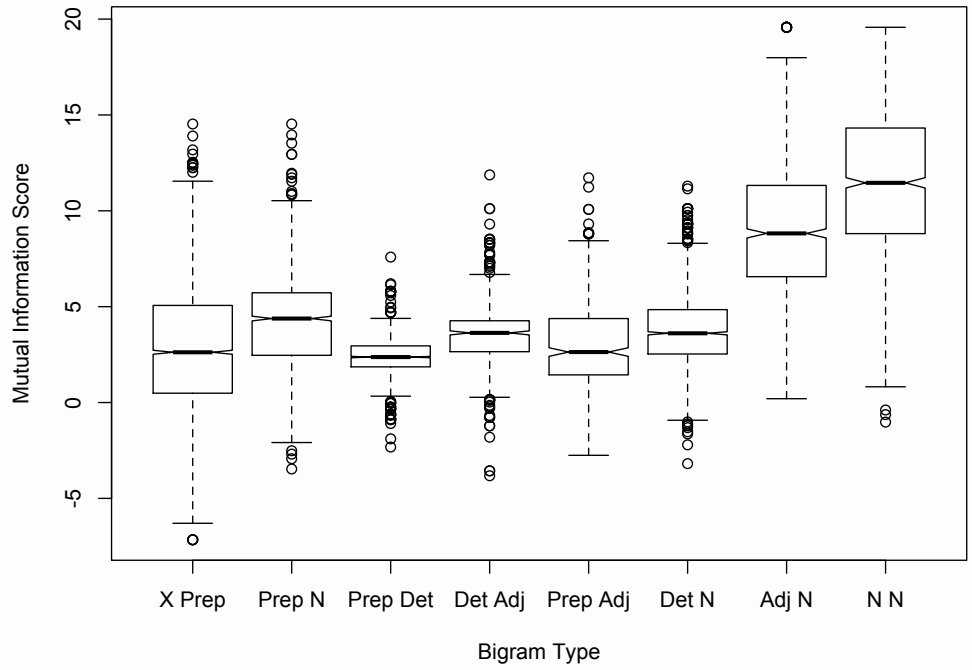


Figure C.4: Mutual Information scores by transition type

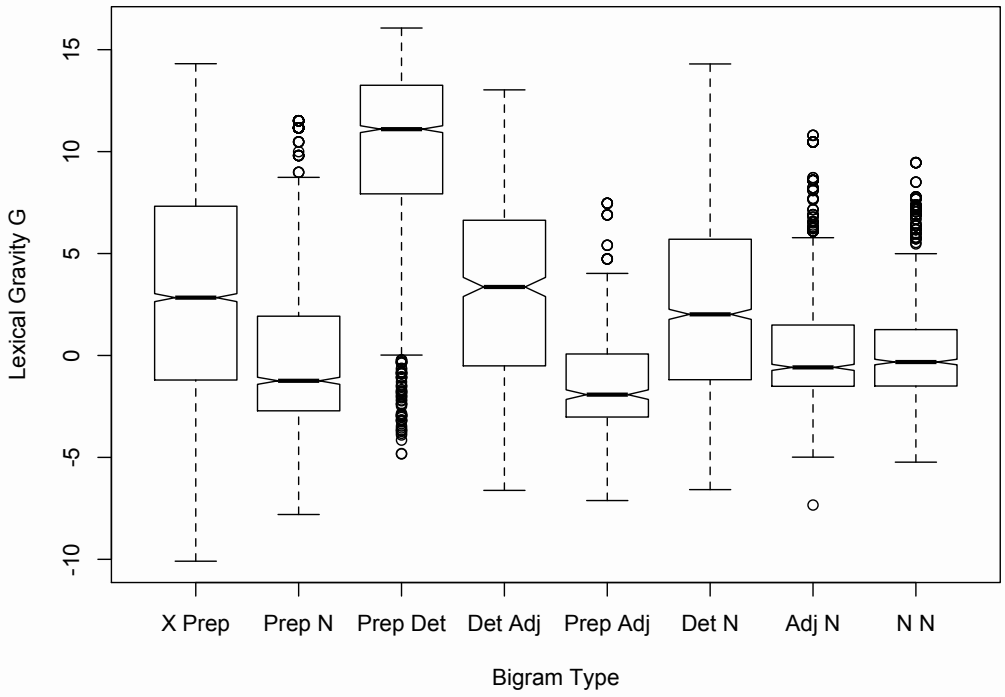


Figure C.5: Lexical gravity G by transition type

Appendix D: Estimation of Best Forest Size for Prepositional Phrase Structures

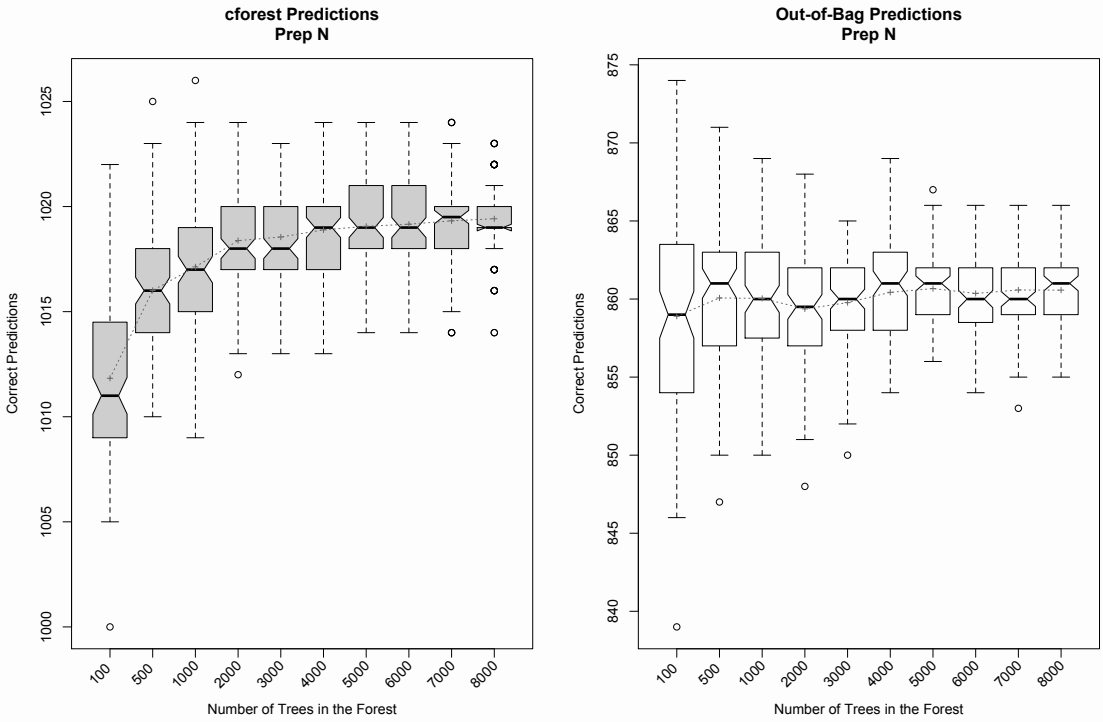


Figure D.1: Correct predictions for 'Preposition Noun' at different forest sizes

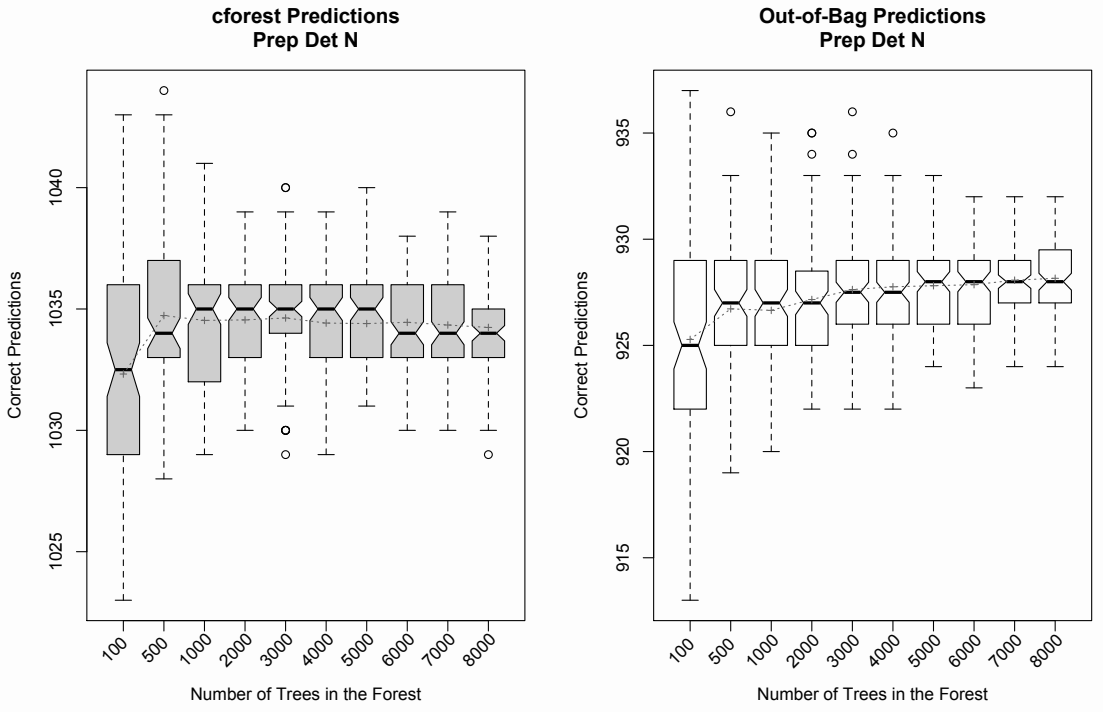


Figure D.2: Correct predictions for 'Preposition Determiner Noun' at different forest sizes

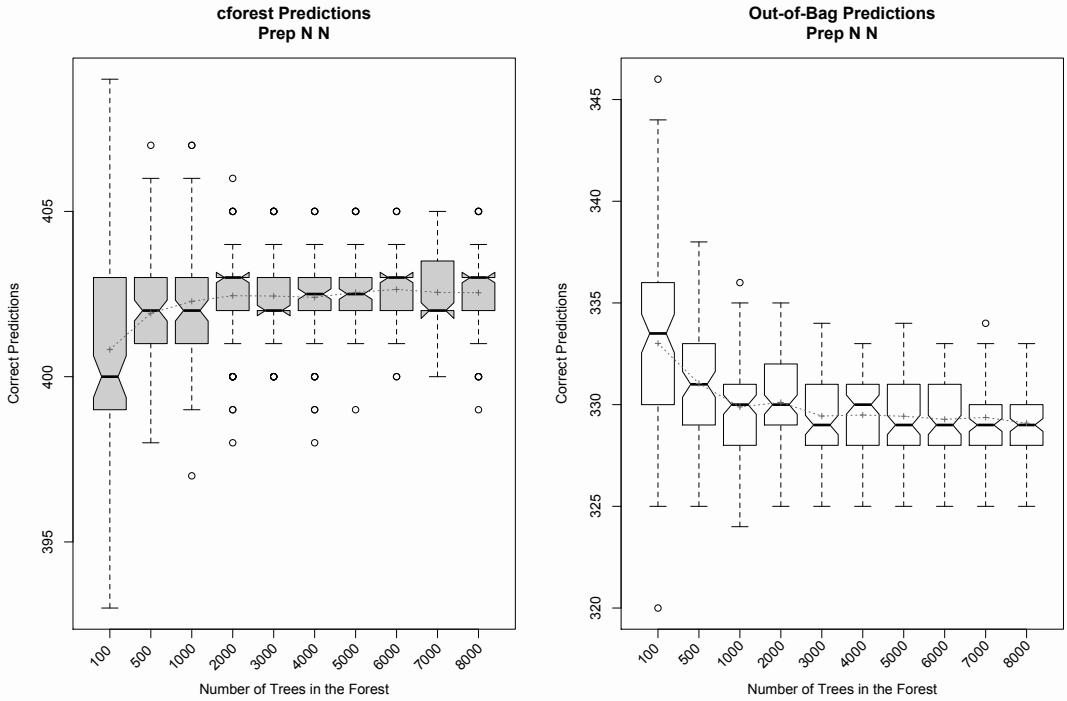


Figure D.3: Correct predictions for 'Preposition Noun Noun' at different forest sizes

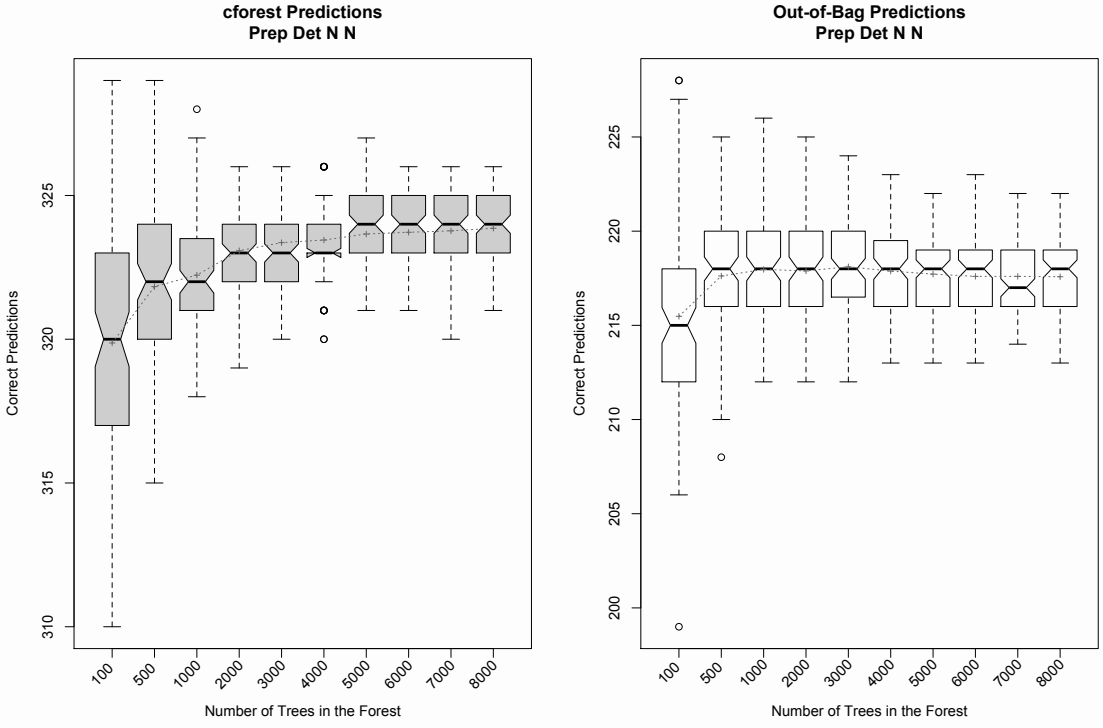


Figure D.4: Correct predictions for 'Preposition Determiner Noun Noun' at different forest sizes

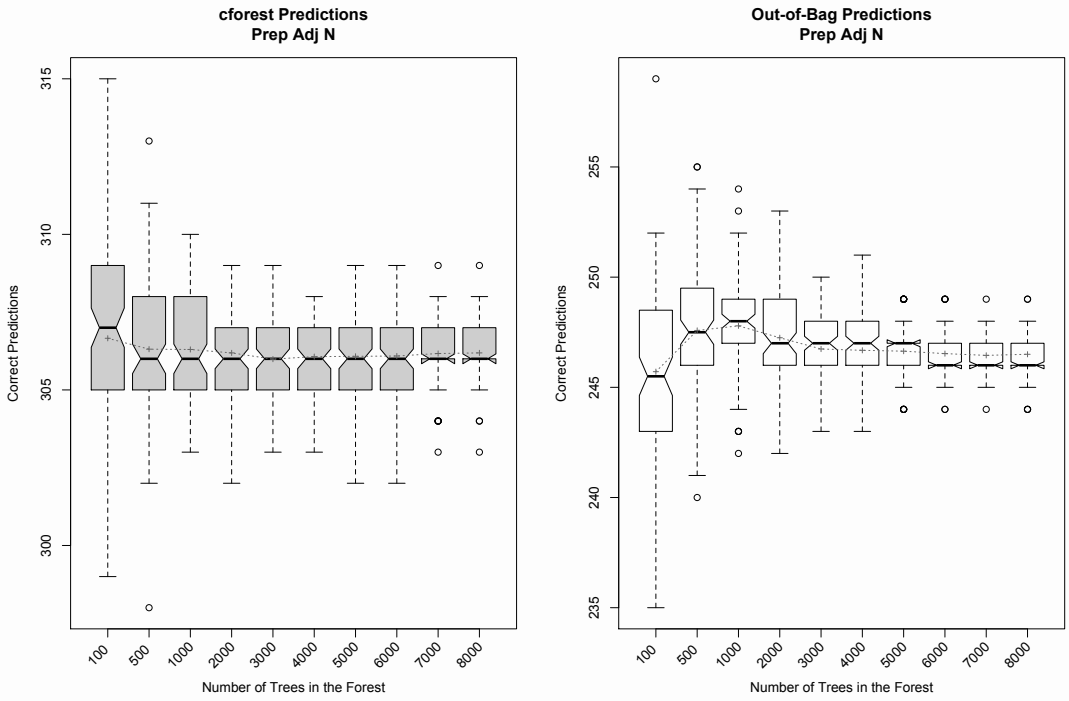


Figure D.5: Correct predictions for 'Preposition Adjective Noun' at different forest sizes

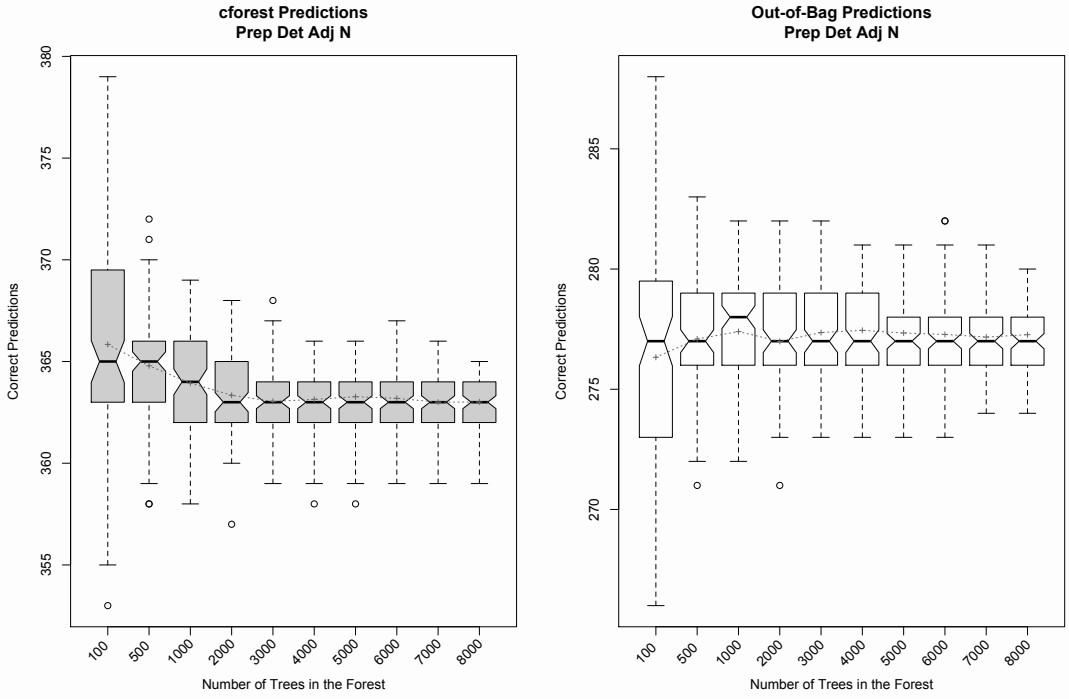


Figure D.6: Correct predictions for 'Preposition Determiner Adjective Noun' at different forest sizes

Appendix E: Characteristics of Pre-Verbal Transitions

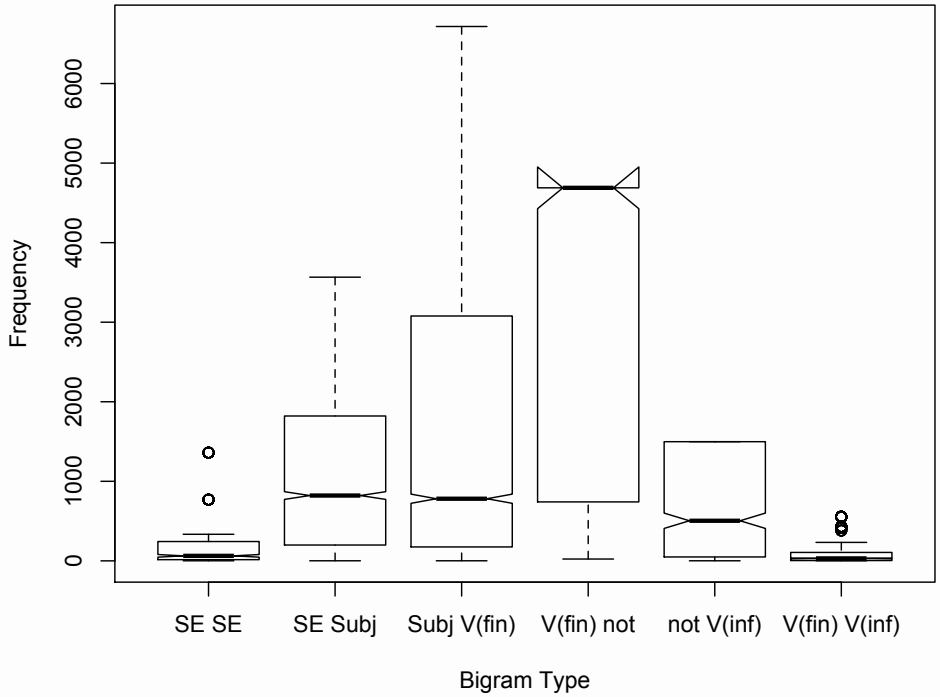


Figure E.1: Frequencies by transition type

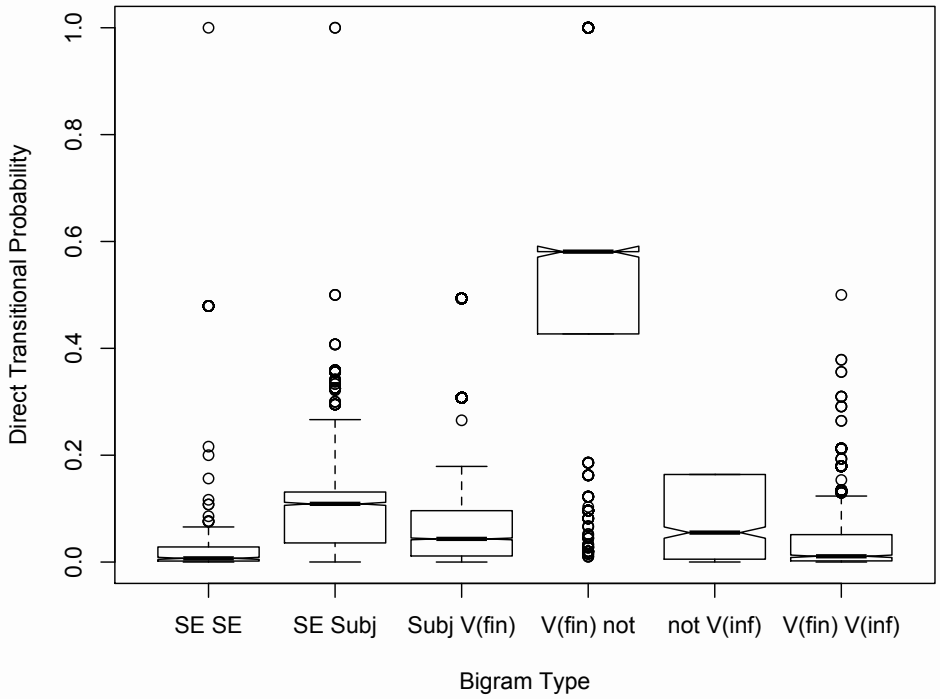


Figure E.2: Direct transitional probabilities by transition type

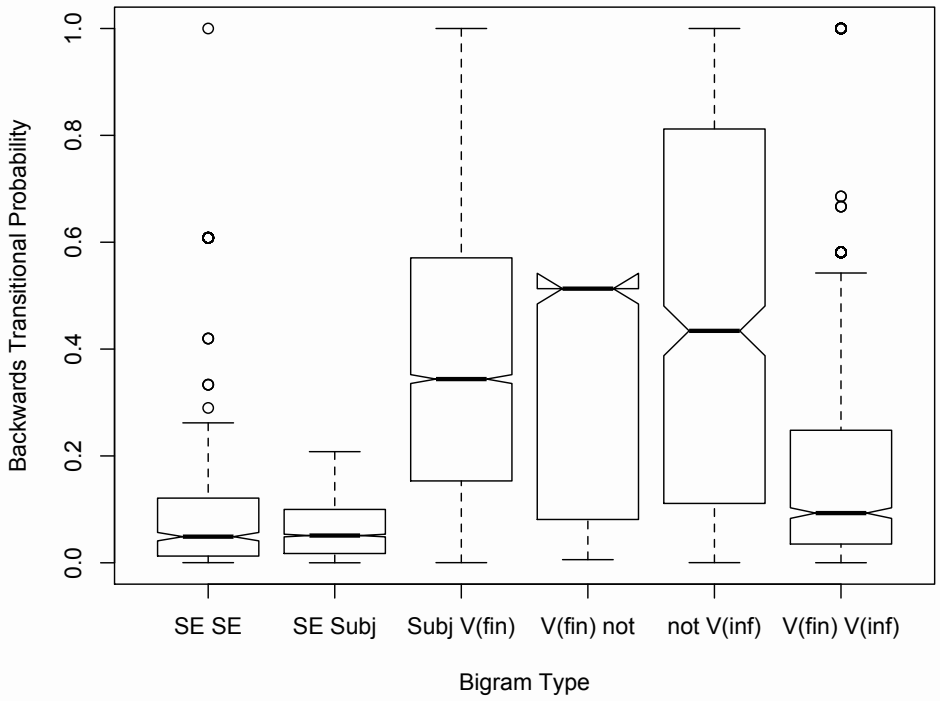


Figure E.3: Backwards transitional probabilities by transition type

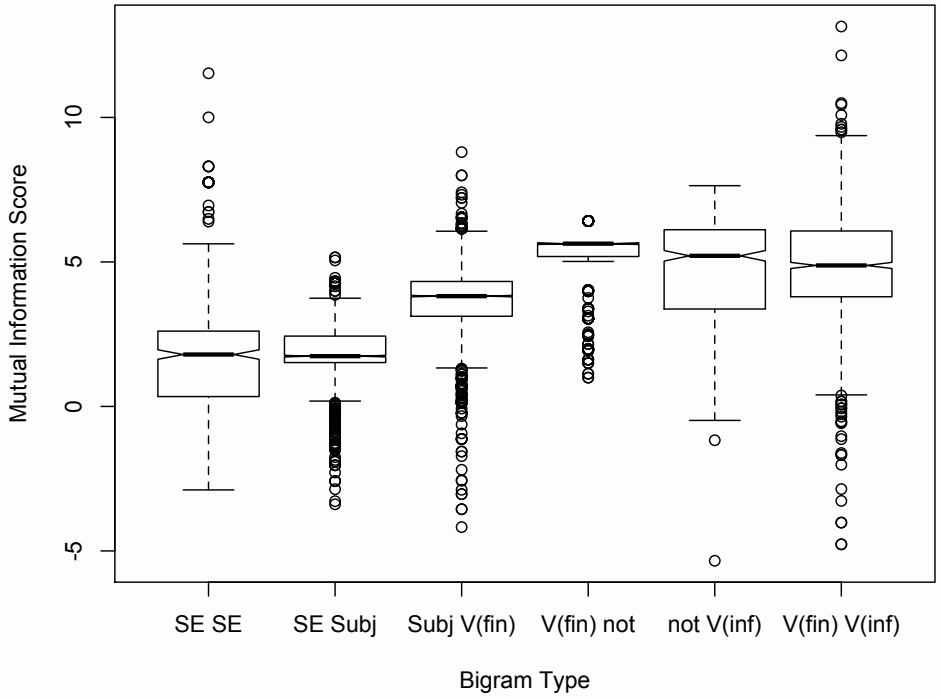


Figure E.4: Mutual information score by transition type

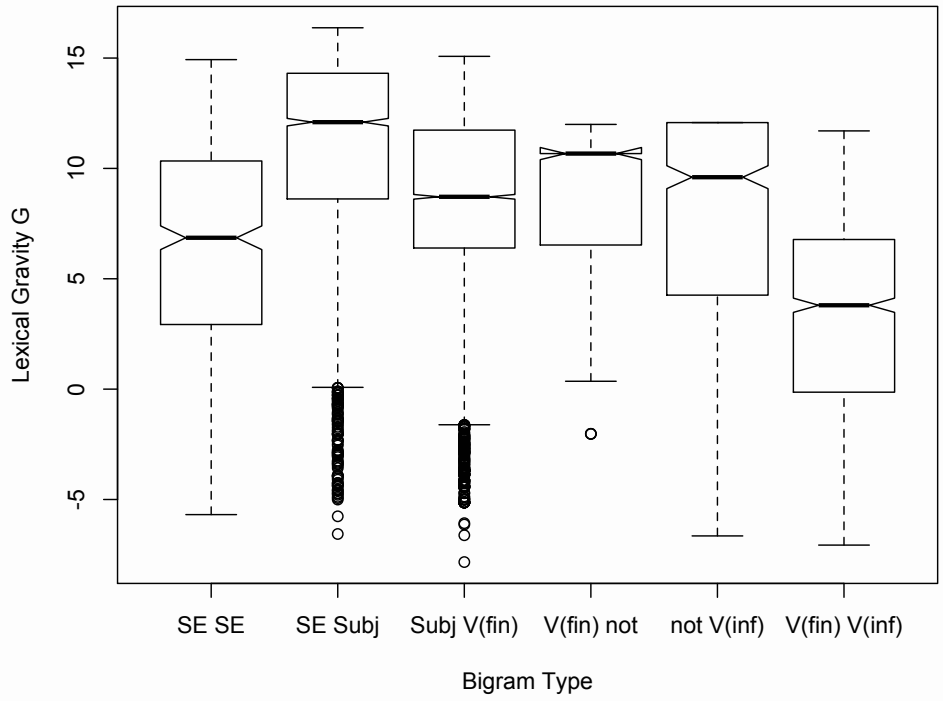


Figure E.5: Lexical gravity G by transition type

Appendix F: Additional Results for 'Subject Verb(finite)'

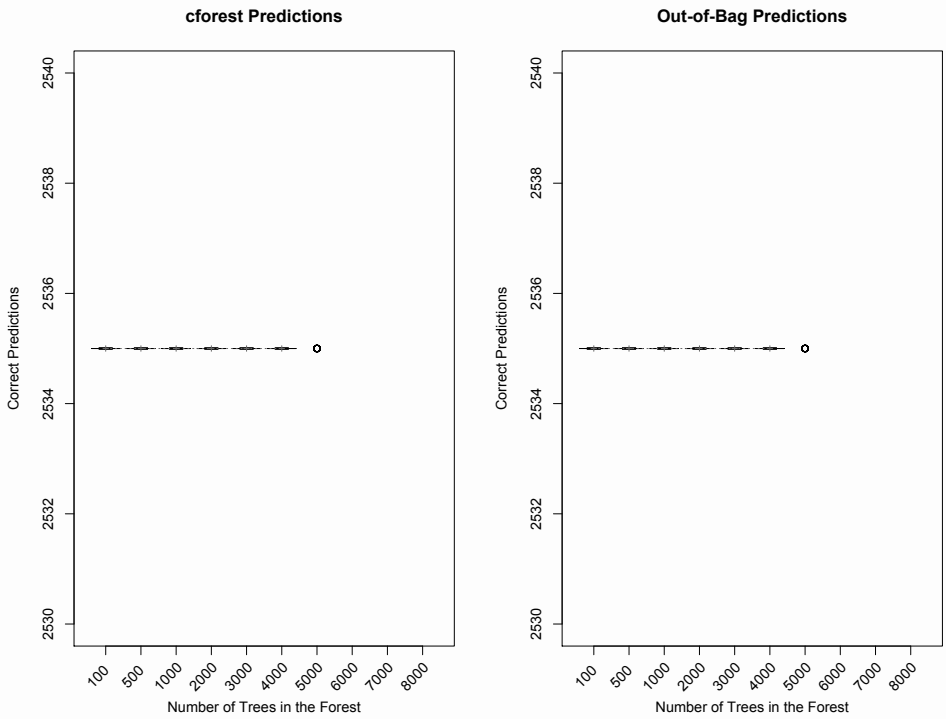


Figure F.1: Correct predictions for 'Subject Verb(finite)' at different forest sizes (mtry=5), based on 100 forests per forest size.[55]

[55] Results for forest sizes above 4,000 trees are missing, because these were too computationally intensive and crashed even if run on bwGRiD hardware.

Appendix G: Additional Results for 'Subject Verb(finite) Verb(non-finite)'

		Model Predictions			
	Hesitation Position	**pre Subj (3)**	**pre V(fin) (4)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre Subj (3)**	564	0	0	564
	pre V(fin) (4)	6	0	1	7
	pre V(inf) (6)	39	0	2	41
	Total	609	0	3	612

Table G.1: Performance of cforest model for 'Subject Verb(finite) Verb(non-finite)' (ntree=2,000, mtry=5, seed=95).

		Model Predictions			
	Hesitation Position	**pre Subj (3)**	**pre V(fin) (4)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre Subj (3)**	564	0	0	564
	pre V(fin) (4)	6	0	1	7
	pre V(inf) (6)	41	0	0	41
	Total	611	0	1	612

Table G.2: Performance of cforest model on out-of-bag data-points of 'Subject Verb(finite) Verb(non-finite)' (ntree=2,000, mtry=5, seed=95).

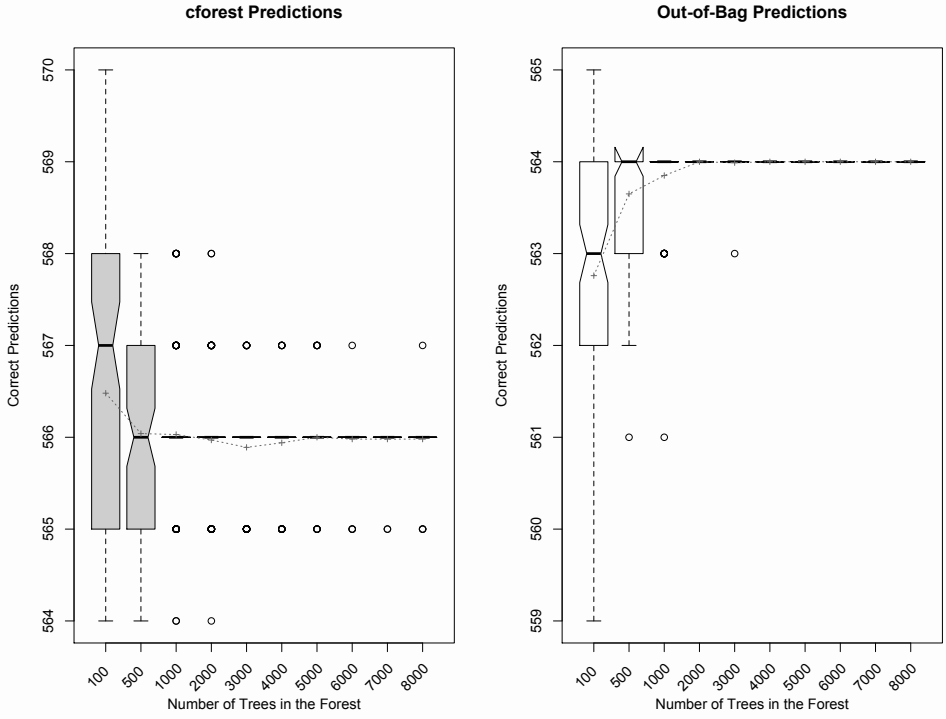


Figure G.1: Correct predictions for 'Subject Verb(finite) Verb(non-finite)' at different forest sizes (mtry=5), based on 100 forests per forest size. The middle of the box is the median, dotted line and crosses indicate the mean.

Appendix H: Additional Results for 'Subject Verb(finite) *not* Verb(non-finite)'

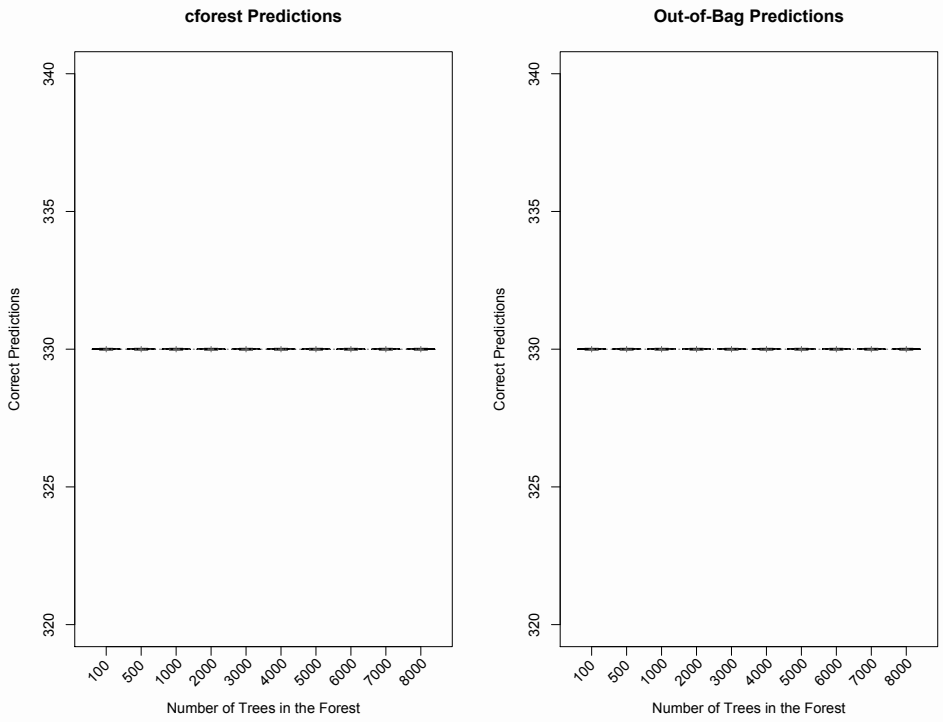


Figure H.1: Correct predictions for 'Subject Verb(finite) not Verb(non-finite)' at different forest sizes (mtry=5), based on 100 forests per forest size.

Appendix I: Additional Results for 'SE Subject Verb(finite)'

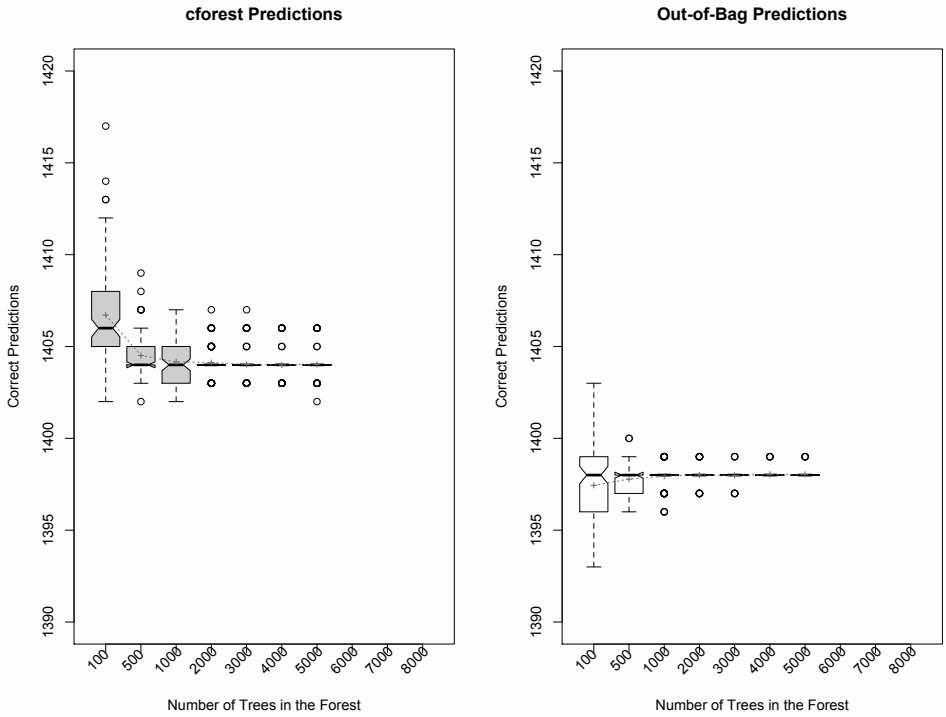


Figure I.1: Correct predictions for 'SE Subject Verb(finite)' at different forest sizes (mtry=5), based on 100 forests per forest size.[56] *The middle of the box is the median, dotted line and crosses indicate the mean.*

		Model Predictions			
	Hesitation Position	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**Total**
Actual Distribution	**pre SE (2)**	299	150	0	449
	pre Subj (3)	67	1,104	0	1,171
	pre V(fin) (4)	9	31	0	40
	Total	375	1,285	0	1,660

Table I.1: Performance of cforest model for 'SE Subject Verb(finite)' (ntree=2,000, mtry=5, seed=777).

[56] Results for forest sizes above 5,000 trees are missing, because these were too computationally intensive and crashed even if run on bwGRiD hardware.

		Model Predictions			
	Hesitation Position	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**Total**
Actual Distribution	**pre SE (2)**	299	150	0	449
	pre Subj (3)	72	1,099	0	1,171
	pre V(fin) (4)	9	31	0	40
	Total	380	1,280	0	1,660

Table I.2: Performance of cforest model on out-of-bag data-points of 'SE Subject Verb(finite)' (ntree=2,000, mtry=5, seed=777).

Appendix J: Additional Results for 'SE Subject Verb(finite) Verb(non-finite)'

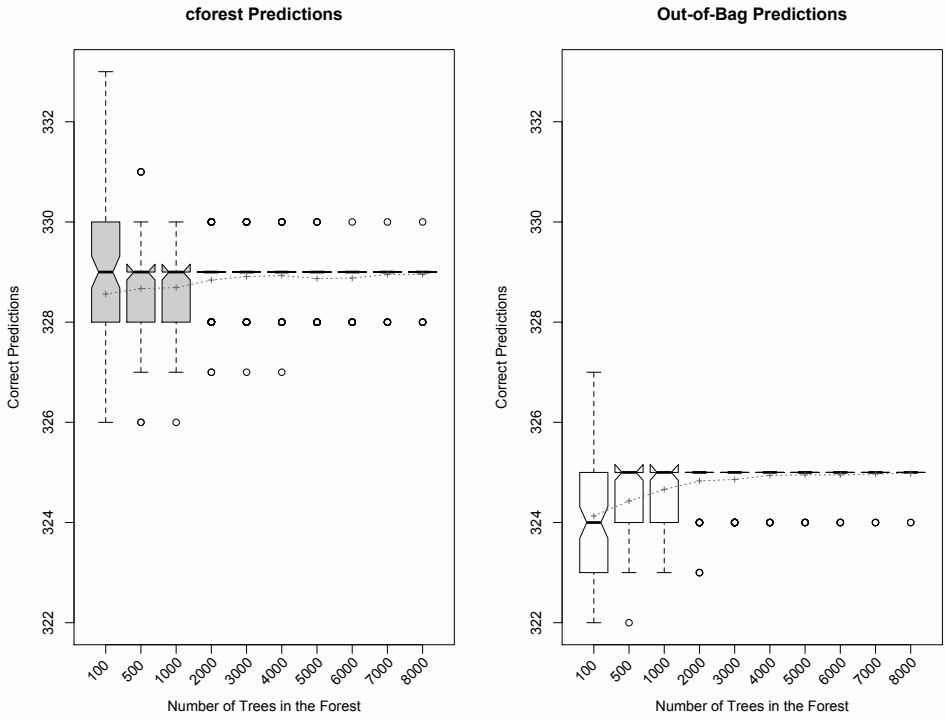


Figure J.1: Correct predictions for 'SE Subject Verb(finite) Verb(non-finite)' at different forest sizes (mtry=5), based on 100 forests per forest size. The middle of the box is the median, dotted line and crosses indicate the mean.

		Model Predictions				
	Hesitation Position	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre SE (2)**	75	41	0	0	116
	pre Subj (3)	14	254	0	0	268
	pre V(fin) (4)	2	6	0	0	8
	pre V(inf) (6)	6	31	0	0	37
	Total	97	332	0	0	429

Table J.1: Performance of cforest model for 'SE Subject Verb(finite) Verb(non-finite)' (ntree=2,000, mtry=5, seed=923).

		Model Predictions				
	Hesitation Position	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre SE (2)**	73	43	0	0	116
	pre Subj (3)	16	252	0	0	268
	pre V(fin) (4)	2	6	0	0	8
	pre V(inf) (6)	6	31	0	0	37
	Total	97	332	0	0	429

Table J.2: Performance of cforest model on out-of-bag data-points of 'SE Subject Verb(finite) Verb(non-finite)' (ntree=2,000, mtry=5, seed=923).

Appendix K: Additional Results for 'SE Subject Verb(finite) *not* Verb(non-finite)'

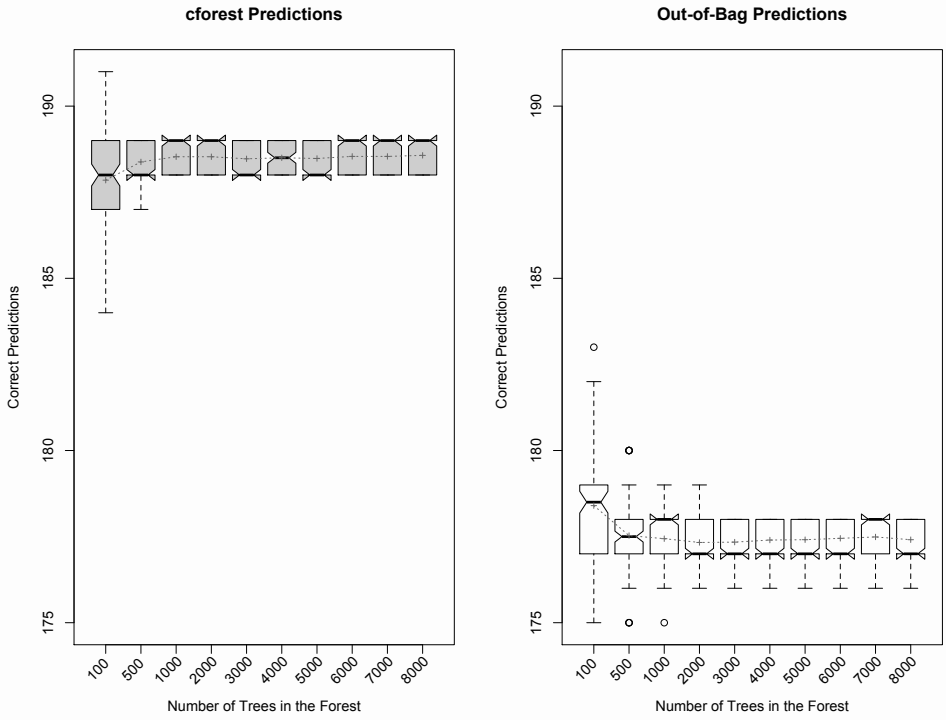


Figure K.1: Correct predictions for 'SE Subject Verb(finite) not Verb(non-finite)' at different forest sizes (mtry=5), based on 100 forests per forest size. The middle of the box is the median, dotted line and crosses indicate the mean.

		Model Predictions					
	Hesitation Position	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**pre *not* (5)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre SE (2)**	12	29	0	0	0	41
	pre Subj (3)	1	177	0	0	0	178
	pre V(fin) (4)	0	2	0	0	0	2
	pre *not* (5)	0	1	0	0	0	1
	pre V(inf) (6)	0	3	0	0	0	3
	Total	13	212	0	0	0	225

Table K.1: Performance of cforest model for 'SE Subject Verb(finite) not Verb(non-finite)' (ntree=2,000, mtry=5, seed=1,321).

		Model Predictions					
	Hesitation Position	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**pre *not* (5)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre SE (2)**	3	38	0	0	0	41
	pre Subj (3)	4	174	0	0	0	178
	pre V(fin) (4)	0	2	0	0	0	2
	pre *not* (5)	0	1	0	0	0	1
	pre V(inf) (6)	0	3	0	0	0	3
	Total	7	218	0	0	0	225

Table K.2: Performance of cforest model on out-of-bag data-points of 'SE Subject Verb(finite) not Verb(non-finite)' (ntree=2,000, mtry=5, seed=1,321).

Appendix L: Additional Results for 'SE SE Subject Verb(finite)'

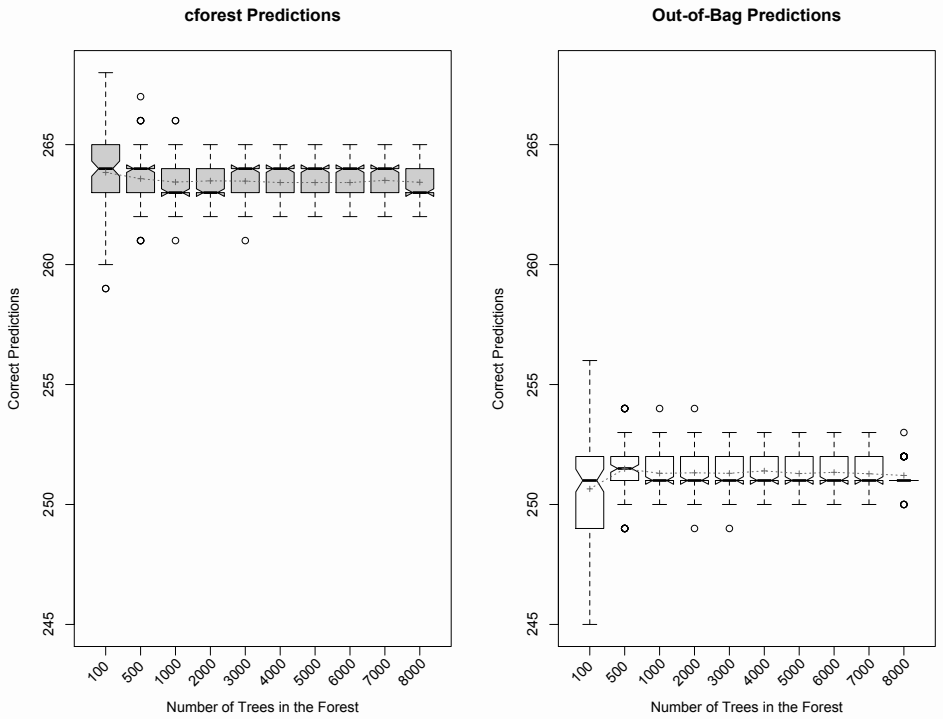


Figure L.1: Correct predictions for 'SE SE Subject Verb(finite)' at different forest sizes (mtry=5), based on 100 forests per forest size. The middle of the box is the median, dotted line and crosses indicate the mean.

		Model Predictions				
	Hesitation Position	**pre SE (1)**	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**Total**
Actual Distribution	**pre SE (1)**	42	30	2	0	74
	pre SE (2)	8	206	2	0	216
	pre Subj (3)	2	46	16	0	64
	pre V(fin) (4)	1	11	1	0	13
	Total	53	293	21	0	367

Table L.1: Performance of cforest model for 'SE SE Subject Verb(finite)' (ntree=2,000, mtry=5, seed=604).

		Model Predictions				
	Hesitation Position	**pre SE (1)**	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**Total**
Actual Distribution	**pre SE (1)**	34	38	2	0	74
	pre SE (2)	10	204	2	0	216
	pre Subj (3)	3	48	13	0	64
	pre V(fin) (4)	1	11	1	0	13
	Total	48	301	18	0	367

Table L.2: Performance of cforest model on out-of-bag data-points of 'SE SE Subject Verb(finite)' (ntree=2,000, mtry=5, seed=604).

Appendix M: Additional Results for 'SE SE Subject Verb(finite) Verb(non-finite)'

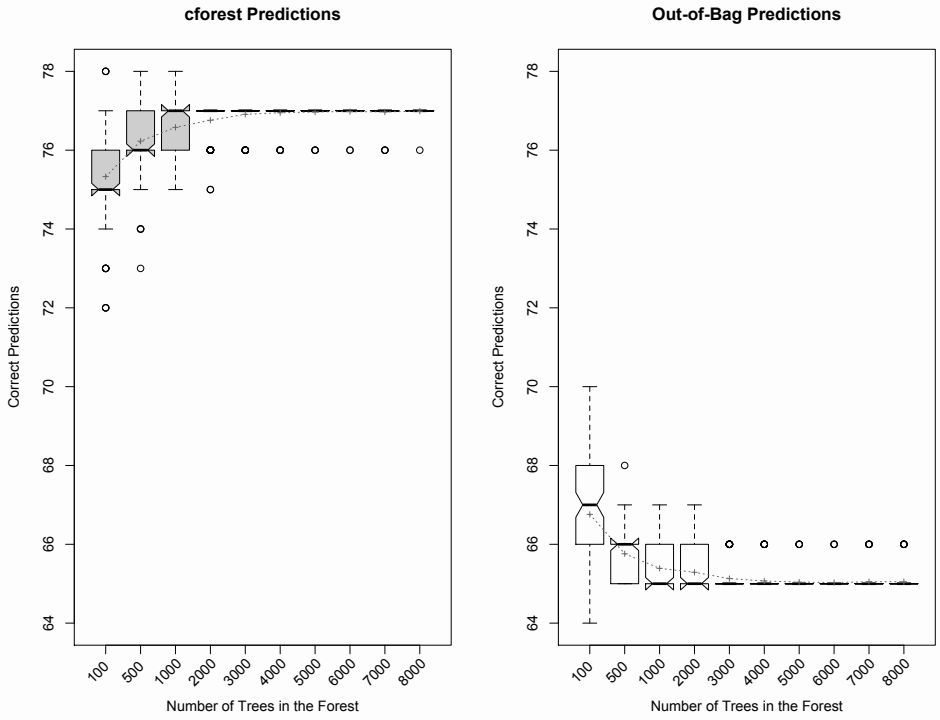


Figure M.1: Correct predictions for 'SE SE Subject Verb(finite) Verb(non-finite)' at different forest sizes (mtry=5), based on 100 forests per forest size. The middle of the box is the median, dotted line and crosses indicate the mean.

		Model Predictions					
	Hesitation Position	**pre SE (1)**	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre SE (1)**	17	7	0	0	0	24
	pre SE (2)	2	55	0	0	0	57
	pre Subj (3)	1	14	5	0	0	20
	pre V(fin) (4)	0	0	0	0	0	0
	pre V(inf) (5)	2	3	2	0	0	7
	Total	22	79	7	0	0	108

Table M.1: Performance of ctree model for 'SE SE Subject Verb(finite) Verb(non-finite)'

		Model Predictions					
	Hesitation Position	**pre SE (1)**	**pre SE (2)**	**pre Subj (3)**	**pre V(fin) (4)**	**pre V(inf) (6)**	**Total**
Actual Distribution	**pre SE (1)**	11	13	0	0	0	24
	pre SE (2)	2	54	1	0	0	57
	pre Subj (3)	1	19	0	0	0	20
	pre V(fin) (4)	0	0	0	0	0	0
	pre V(inf) (5)	2	4	1	0	0	7
	Total	16	90	2	0	0	108

Table M.2: Performance of ctree model for 'SE SE Subject Verb(finite) Verb(non-finite)'

Appendix N: Fluent and Hesitant Verb-Phrase Transitions

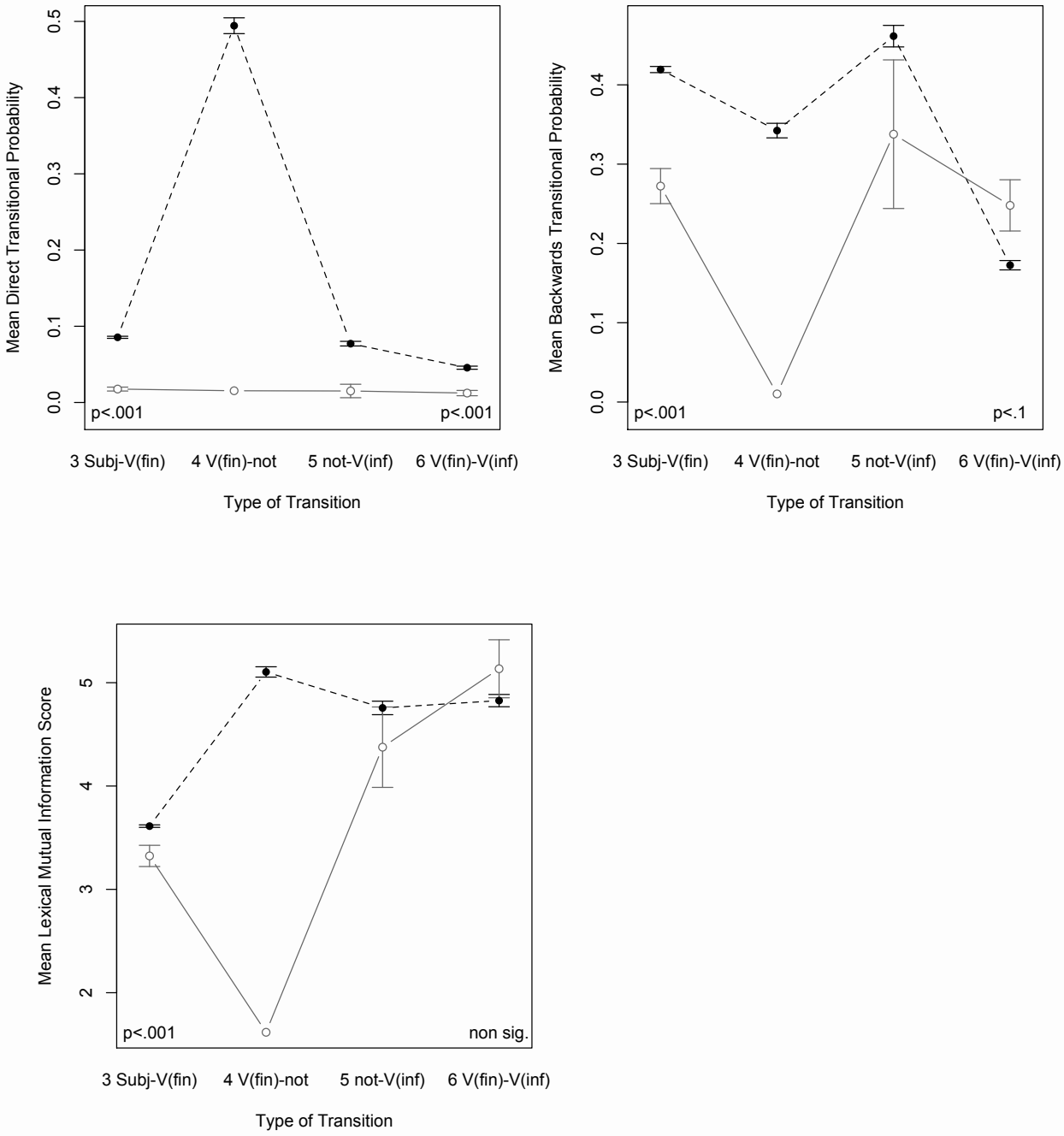


Figure N.1: Comparison of the direct and backwards transitional probability as well as the mutual information score of fluent (dashed black line) and hesitant (solid grey line) verb-phrase transitions.

Appendix O: MI & TPD for Hesitant Verb Phrase Transitions

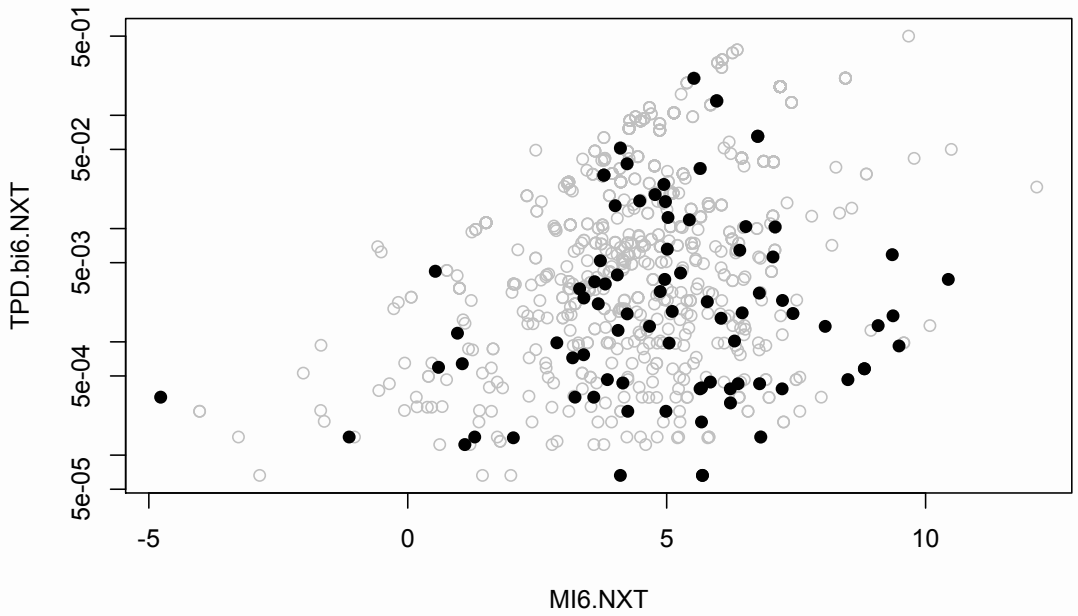


Figure O.1: Mutual information score and direct transitional probability of hesitant 'Verb(finite) Verb(non-finite)' pairs (black) compared to all other 'Verb(finite) Verb(non-finite)' (grey).

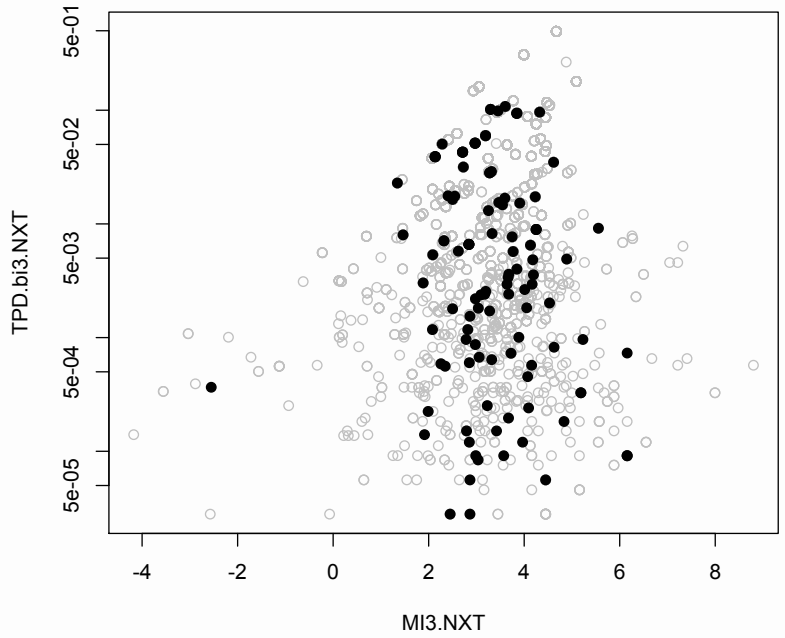


Figure O.2: Mutual information score and direct transitional probability of hesitant 'Subject Verb(finite)' pairs (black) compared to all other 'Subject Verb(finite)' (grey).

Appendix P: R Commands

This section lists exemplary R commands used throughout the present work. Wherever special packages were installed in order to be able to use the command, this is indicated. The R versions used were 2.13.1, 2.15.2 and 3.0.1 (for Mac).

Section 3.1: Opening a corpus consisting of several files

based on Gries (2009a:34)

```
corpus.files <- select.list(dir(scan(nmax=1, what="char")), multiple=T)
setwd("/Users/Ulrike/Documents/Corpora/Switchboard/nxt_switchboard_ann/xml/
 terminals")

whole.corpus <- vector()

for (i in corpus.files) {
 current.corpus.file <- scan(i, what="char", sep="\n", quiet=T)
 cat(i, "\n")
 whole.corpus <- append(whole.corpus, current.corpus.file)
 }
```

Section 3.1: Extraction of the beginning of a sentence

```
parse.file <- file.choose()

for (i in 1:length(syntax.lines)){
 cat(i/length(syntax.lines), "\n")
 current.syntax.pos <- syntax.lines[i]
 result1.parse <- grep("<parse", whole.syntax[current.syntax.pos:1])
 result2.parse <- (current.syntax.pos+1)-result1.parse
 current.parse <- result2.parse[1]
 cat(current.parse, sep="\n", file=parse.file, append=T)
 }
```

Section 3.1: Extraction of the end of a sentence

```
end.parse.file <- file.choose()

for (i in 1:length(syntax.lines)){
 cat(i/length(syntax.lines), "\n")
 current.syntax.pos <- syntax.lines[i]
 result1.end.parse <- grep("</parse",
 whole.syntax[current.syntax.pos:length(whole.syntax)])
 result2.end.parse <- (current.syntax.pos-1)+result1.end.parse
 current.end.parse <- result2.end.parse[1]
 cat(current.end.parse, sep="\n", file=end.parse.file, append=T)
 }
```

Section 3.1: Cluster analysis

```
library(amap)

curv1 <- Dist(tab1, method="correlation", diag=T, upper=T)
clust1 <- hclust(curv1, method="ward")
plot(clust1)
```

Section 3.1: Boxplots

```
quartz()
par(mfrow = c(1,2))
boxplot(length ~ info, data = tab.pall, boxwex = 0.30, at = 1:1 + 0.4,
 notch=T, subset= info == "pause uh", col="#cccccc", ylab="Pause length in
 seconds", main="Length of Pauses Preceding\nFillers and Discourse Markers",
 names)
boxplot(length ~ info, data = tab.pall, add = TRUE, boxwex = 0.30, at = 1:1 +
 0.13, notch=T, subset= info == "pause um", col="#cccccc")
boxplot(length ~ info, data = tab.pall, add = TRUE, boxwex = 0.30, at = 1:1 -
 0.13, notch=T, subset= info == "pause like")
boxplot(length ~ info, data = tab.pall, add = TRUE, boxwex = 0.30, at = 1:1 -
 0.4, notch=T, subset= info == "pause know")
axis(1, at=1:1 +0.4, labels = "pause uh")
axis(1, at=1:1 +0.13, labels = "pause um")
axis(1, at=1:1 -0.4, labels = "pause y.know")
axis(1, at=1:1 -0.13, labels = "pause like")
boxplot(length ~ info, data = tab.allp, boxwex = 0.30, at = 1:1 + 0.4,
 notch=T, subset= info == "uh pause", col="#cccccc", ylab="Pause length in
 seconds", main="Length of Pauses Following\nFillers and Discourse Markers")
boxplot(length ~ info, data = tab.allp, add = TRUE, boxwex = 0.30, at = 1:1 +
 0.13, notch=T, subset= info == "um pause", col="#cccccc")
boxplot(length ~ info, data = tab.allp, add = TRUE, boxwex = 0.30, at = 1:1 -
 0.13, notch=T, subset= info == "like pause")
boxplot(length ~ info, data = tab.allp, add = TRUE, boxwex = 0.30, at = 1:1 -
 0.4, notch=T, subset= info == "know pause")
axis(1, at=1:1 +0.4, labels = "uh pause")
axis(1, at=1:1 +0.13, labels = "um pause")
axis(1, at=1:1 -0.4, labels = "y.know pause")
axis(1, at=1:1 -0.13, labels = "like pause")
```

Section 3.3: Classification and Regression Trees

Growing a tree with standard settings

```
install.packages("party")
library(party)
PDAN.ctree <- ctree(hes.position ~ w0.freq.NXT + w1.freq.NXT + w2.freq.NXT +
 w3.freq.NXT + w4.freq.NXT + bi0.freq.NXT + bi1.freq.NXT + bi2.freq.NXT +
 bi3.freq.NXT + TPD.bi0.NXT + TPD.bi1.NXT + TPD.bi2.NXT + TPD.bi3.NXT +
 TPB.bi0.NXT + TPB.bi1.NXT + TPB.bi2.NXT + TPB.bi3.NXT + G0.NXT + G1.NXT +
 G2.NXT + G3.NXT + MI0.NXT + MI1.NXT + MI2.NXT + MI3.NXT + hes.type,
 data=PDAN.table)
plot(PDAN.ctree)
table(PDAN.table$hes.position, predict(PDAN.ctree))
```

Comparison of results to the baseline model

```
baseline <- c(241,334)
ctree <- c(286,289)
chisq.test(ctree, p = baseline, rescale.p = T)
residuals(chisq.test(ctree, p = baseline, rescale.p = T))
```

Section 3.3: Random Forests

Growing a forest

```
set.seed(1282)
data.controls <- cforest_unbiased(ntree=3000, mtry=5)
PDN.forest <- cforest(hes.position ~ w0.freq.NXT + w1.freq.NXT + w2.freq.NXT +
 w3.freq.NXT + bi0.freq.NXT + bi1.freq.NXT + bi2.freq.NXT + TPD.bi0.NXT +
 TPD.bi1.NXT + TPD.bi2.NXT + TPB.bi0.NXT + TPB.bi1.NXT + TPB.bi2.NXT + G0.NXT
 + G1.NXT + G2.NXT + MI0.NXT + MI1.NXT + MI2.NXT + hes.type, data=PDN.tab,
 controls=data.controls); table(PDN.tab$hes.position, predict(PDN.forest))
```

Plotting variable importance

```
PDAN.varimp <- varimp(PDAN.forest)
dotplot(sort(PDAN.varimp), panel = function(x,y){
     panel.dotplot(x,y, col="darkblue", pch=16)
     panel.abline(v=abs(min(PDAN.varimp)), col="red", lty="longdash", lwd=1)
     panel.abline(v=min(PDAN.varimp), col="red", lty="longdash", lwd=1)
     panel.abline(v=0, col="blue")
}
)
dev.off()
```

Section 3.3: Boxplots for different forest sizes

```
par(mfrow = c(1,2))
boxplot(PV.tab[,1:10], xaxt="n", notch=T, col=c("#cccccc"), ylab="Correct
 Predictions", xlab="Number of Trees in the Forest", main="cforest
 Predictions", ylim=c(2530,2540));  axis(1, at=seq(1, 10, by=1), labels =
 FALSE);text(seq(0.7, 9.7, by=1), par("usr")[3] - 0.35, labels =c("100",
 "500", "1000", "2000", "3000", "4000", "5000", "6000", "7000", "8000"), srt =
 45, pos = 1, xpd = TRUE)
points(mean(PV.tab[,1:10]), pch=3, cex=0.6, col=c("#666666"))
lines(as.data.frame(zoo(cbind(mean(PV.tab[,1:6])))), col="#666666", lty=3)
boxplot(PVoob.tab[,1:10], xaxt="n", notch=T, ylab="Correct Predictions",
 xlab="Number of Trees in the Forest", main="Out-of-Bag Predictions",
 ylim=c(2530,2540)); axis(1, at=seq(1, 10, by=1), labels =
 FALSE);text(seq(0.7, 9.7, by=1), par("usr")[3] - 0.35, labels =c("100",
 "500", "1000", "2000", "3000", "4000", "5000", "6000", "7000", "8000"), srt =
 45, pos = 1, xpd = TRUE)
points(mean(PVoob.tab[,1:10]), pch=3, cex=0.6, col=c("#666666"))
lines(as.data.frame(zoo(cbind(mean(PVoob.tab[,1:6])))), col="#666666", lty=3)
```

Sections 4.2/5.2: Bargraphs

```
setwd("/Users/Ulrike/Documents/Diss/10_PPs/11_Stats_all_in/Bargraphs")

pdf(file="BargraphsPP2.pdf", width=8, height=11)
par(mfrow=c(3,2))
barplot(t(PN2), main="Prep N", ylab="Total Amount of Hesitations", col="grey",
 xlab="Hesitation Placement", density=c(0, 15, 1000), space=0.1, cex.axis=1.1,
 las=1, cex=1.1, cex.lab=1.2, cex.main=2, names=c("before Prep", "before N"))

barplot(t(PDN2), main="Prep Det N", ylab="Total Amount of Hesitations",
 col="grey", xlab="Hesitation Placement", density=c(0, 15, 1000), space=0.1,
 cex.axis=1.1, las=1, cex=1.1, cex.lab=1.2, cex.main=2, names=c("before Prep",
 "before Det", "before N"))

barplot(t(PNN2), main="Prep N N", ylab="Total Amount of Hesitations",
 col="grey", xlab="Hesitation Placement", density=c(0, 15, 1000), space=0.1,
 cex.axis=1.1, las=1, cex=1.1, cex.lab=1.2, cex.main=2, names=c("before Prep",
 "before N1", "before N2"))

barplot(t(PDNN2), main="Prep Det N N", ylab="Total Amount of Hesitations",
 col="grey", xlab="Hesitation Placement", density=c(0, 15, 1000), space=0.1,
 cex.axis=1.1, las=1, cex=1.1, cex.lab=1.2, cex.main=2, names=c("before Prep",
 "before Det", "before N1", "before N2"))

barplot(t(PAN2), main="Prep Adj N", ylab="Total Amount of Hesitations",
 col="grey", xlab="Hesitation Placement", density=c(0, 15, 1000), space=0.1,
 cex.axis=1.1, las=1, cex=1.1, cex.lab=1.2, cex.main=2, names=c("before Prep",
 "before Adj", "before N"))

barplot(t(PDAN2), main="Prep Det Adj N", ylab="Total Amount of Hesitations",
 col="grey", xlab="Hesitation Placement", density=c(0, 15, 1000), space=0.1,
 cex.axis=1.1, las=1, cex=1.1, cex.lab=1.2, cex.main=2, names=c("before Prep",
 "before Det", "before Adj", "before N"))

par(mfrow=c(1,1))
dev.off()
```

Sections 4.2/5.2: Means and standard deviation

mean: `mean(ss.freq, na.rm=T)`

standard deviation: `sd(ss.freq, na.rm=T)`

Sections 4.4/5.4: Wilcoxon rank sum tests

```
vp.freq <- c(vn.freq, VV.freq, nv.freq)
non.vp.freq <- c(SV1.freq, ss.freq)

wilcox.test(vp.freq, non.vp.freq, correct=F)
```

Sections 4.4/5.4: Test for normal distribution

```
plot(density(vp.freq))
shapiro.test(vp.freq)
```

Sections 4.5/5.5: Lineplots

```
lineplot.CI(predictor,x,transition, data = tab.all2, xlab="Predictor",
 ylab="Variable Importance", legend=F)
```

Section 4.6: Plots

```
plot(tab$TPD.bi0.NXT ~ tab$MI0.NXT, log="y", xlab="MI", ylab="Direct
 Transitional Probability (log scaled)", col=c("#cccccc"), pch=20, cex=0.5,
 yaxt="n")
axis(2, at=c(0.0001,0.001,0.01,0.1,1), labels = c("0.0001","0.001",
 "0.01","0.1","1"))
points(tab.out2$TPD.bi0.NXT~tab.out2$MI0.NXT, pch=20)
points(tab.terms2$TPD.bi0.NXT~tab.terms2$MI0.NXT, pch=3)
legend(10, 0.00015, c("out of", "terms of"), pch=c(20,3), bg="white", cex=0.8)
```

```
plot(tab.fluent$TPD.bi0.NXT ~ tab.fluent$MI0.NXT, log="y", xlab="MI",
 ylab="Direct Transitional Probability (log scaled)", col=c("#cccccc"),
 pch=20, cex=0.5, yaxt="n")
axis(2, at=c(0.0001,0.001,0.01,0.1,1), labels = c("0.0001","0.001",
 "0.01","0.1","1"))
points(rep.tab$TPD.bi0.NXT ~ rep.tab$MI0.NXT, pch=1, cex=0.8)
points(mwrep.tab$TPD.bi0.NXT ~ mwrep.tab$MI0.NXT, pch=2)
points(sc.tab$TPD.bi0.NXT ~ sc.tab$MI0.NXT, pch=3)
legend(8.5, 0.002, c("single-word\nrepetitions\n", "multi-word\nrepetitions
 \n", "self-\ncorrections\n"), pch=c(1,2,3), bg="white", cex=0.8)
```

Section 5.7: Bargraphs

```
and.tab <- all2[all2$AdvConj.short=="and_CC",c(7,6)]
and.tab$hes.type <- as.factor(and.tab$hes.type)
and.tab$combi <- as.factor(and.tab$combi)
a <- table(and.tab)
bla <- barplot(t(a), beside=T, main="AND", cex.names=1.7, cex.main=2,
 cex.axis=1.37, col=c("#000000", "#666666", "#cccccc", "#ffffff"),
 ylim=c(0,740), xlab="Hesitation Type", ylab="Frequency in the Dataset",
 cex.lab=1.55)
barplot(t(a), beside=T, main="AND", cex.names=1.7, cex.main=2, cex.axis=1.37,
 col=c("#000000", "#666666", "#cccccc", "#ffffff"), ylim=c(0,740),
 xlab="Hesitation Type", ylab="Frequency in the Dataset", cex.lab=1.55)
legend("topleft", c("hes [pause] SE","hes SE","SE [pause] hes","SE hes"),
 cex=1.5, bty="n", fill=c("#000000", "#666666", "#cccccc", "#ffffff"))
text(x=bla,
 y=c(a[1],a[7],a[13],a[19],a[2],a[8],a[14],a[20],a[3],a[9],a[15],a[21],a[4],a[
 10],a[16],a[22],a[5],a[11],a[17],a[23],a[6],a[12],a[18],a[24]),
 labels=c(a[1],a[7],a[13],a[19],a[2],a[8],a[14],a[20],a[3],a[9],a[15],a[21],a[
```

```
 4],a[10],a[16],a[22],a[5],a[11],a[17],a[23],a[6],a[12],a[18],a[24]), pos=3,
 col="black", cex=1.25)
text(x=c(3,12.5,18), y=c(a[19]+50,a[21]-50,a[22]+50), labels=c("p<.01", "p<.
 001", "p<.001"), pos=3, col="black", cex=1.25)
```

Bibliography

Acton, Eric K. (2011): "On gender differences in the distribution of *um* and *uh*." *University of Pennsylvania Working Papers in Linguistics* 17 (2 Selected Papers from NWAV 39). 1-9.

Altenberg, Bengt (1998): "On the phraseology of spoken English: The evidence of recurrent word-combinations." In Cowie, Anthony P. (Ed.): *Phraseology: Theory, Analyis, and Application.* Oxford: OUP. 101-22.

Arnon, Inbal and Neal Snider (2010): "More than words: Frequency effects for multi-word phrases." *Journal of Memory and Language* 62. 67-82.

Arppe, Antti, Gaëtanelle Gilquin, Dylan Glynn, Martin Hilpert and Arne Zeschel (2010): "Cognitive corpus linguistics: five points of debate on current theory and methodology." *Corpora* 5 (1). 1-27.

Atkinson, Elizabeth J. and Terry M. Therneau (2000): *An Introduction to Recursive Partitioning Using the RPART Routines.* Mayo Foundation.

Auer, Peter (2009): "On-line syntax: Thoughts on the temporality of spoken language." *Language Sciences* 31. 1-13.

Baayen, R. Harald (2008): *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge: CUP.

Barlow, Michael (2000): "Usage, blends, and grammar." In Kemmer, Suzanne and Michael Barlow (Eds.): *Usage-Based Models of Language.* Stanford, CA: CSLI Publications. 315-45.

Beattie, G. and Brian Butterworth (1979): "Contextual probability and word frequency as determinants of pauses in spontaneous speech." *Language and Speech* 22. 201-11.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman and Tom Schoeneman (2009): "Language is a complex adaptive system: position paper." *Language Learning* 59 (Supplement 1). 1-26.

Beckner, Clay and Joan Bybee (2009): "A usage-based account of constituency and reanalysis." *Language Learning* 59 (Supplement). 27-46.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory and Daniel Gildea (2003): "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation." *Journal of the Acoustical Society of America* 113 (2). 1001-24.

Biber, Douglas and Susan Conrad (2003): "Register variation: A corpus approach." In Schiffrin, Deborah, Deborah Tannen and Heidi Hamilton (Eds.): *The Handbook of Discourse Analysis*. Malden, MA: Blackwell. 175-96.

Biber, Douglas, Susan Conrad and Viviana Cortes (2004): "If you look at... : Lexical bundles in university teaching and textbooks." *Applied Linguistics* 25 (3). 371-405.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (1999): *Longman Grammar of Spoken and Written English.* Harlow: Pearson.

Bies, Ann, Mark Ferguson, Karen Katz and Robert MacIntyre (1995): *Bracketing Guidelines for Treebank II Style.* Department of Computer and Information Science, University of Pennsylvania.

Bod, Rens (2010): "Probabilistic linguistics." In Heine, Bernd and Heiko Narrog (Eds.): *The Oxford handbook of Linguistic Analysis.* Oxford: OUP. 633-62.

Boomer, Donald S. (1965): "Hesitation and grammatical encoding." *Language and Speech* 8. 148-58.

Boomer, Donald S. and Allen T. Dittmann (1962): "Hesitation Pauses and Juncture Pauses in Speech." *Language and Speech* 5. 215-20.

Bortfeld, Heather, Silvia D. Leon, Jonathan E. Bloom, Michael F. Schober and Susan E. Brennan (2001): "Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender." *Language and Speech* 44. 123-47.

Bortz, Jürgen (2005): *Statistik für Human- und Sozialwissenschaftler.* (6th ed.). Heidelberg: Springer.

Branigan, Holly P., Martin J. Pickering, Simon P. Liversedge, Andrew J. Stewart and Thomas P. Urbach (1995): "Syntactic priming: Investigating the mental representation of language." *Journal of Psycholinguistic Research* 24 (6). 489-506.

bwGRiD, member of the German D-Grid initiative, funded by the Ministry for Education and Research (Bundesministerium für Bildung und Forschung) and the Ministry for Science, Research and Arts Baden-Wuerttemberg (Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg). http://www.bw-grid.de.

Bybee, Joan (2002): "Phonological evidence for the exemplar storage of multiword sequences." *SSLA* 24 (2). 215-21.

Bybee, Joan (2006): "From usage to grammar: The mind's response to repetition." *Language* 82 (4). 711-33.

Bybee, Joan (2007a): *Frequency of Use and the Organization of Language.* Oxford: OUP.

Bybee, Joan (2007b): "Sequentiality as the basis of constituent structure." In Bybee, Joan (Ed.): *Frequency of Use and the Organisation of Language.* Oxford: OUP. 313-35. (Reprinted from Talmy Givón and Bertram F. Malle (Eds.): *The Evolution of Language out of Pre-Language.* Amsterdam: John Benjamins. 2002. 107-132.)

Bybee, Joan (2007c): "Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change." In Bybee, Joan (Ed.): *Frequency of Use and the Organisation of Language.* Oxford: OUP. 235-64. (Reprinted from *Language Variation and Change* 14. 2002. 261-290.)

Bybee, Joan (2010): *Language, Usage, and Cognition.* Cambridge: CUP.

Bybee, Joan and David Eddington (2006): "A usage-based approach to Spanish verbs of 'becoming'." *Language* 82. 323-55.

Bybee, Joan and Paul Hopper (2001): *Frequency and the Emergence of Linguistic Structure.* Amsterdam: John Benjamins.

Bybee, Joan and James L. McClelland (2005): "Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition." *The Linguistic Review* 22. 381-410.

Bybee, Joan and Joanne Scheibman (2007): "The effect of usage on degrees of constituency. The reduction of *don't* in English." In Bybee, Joan (Ed.): *Frequency of*

Use and the Organisation of Language. Oxford: OUP. 294-312. (Reprinted from *Linguistics* 37(4). 1999. 575-596.).

Bybee, Joan and Rena Torres Cacoullos (2009): "The role of prefabs in grammaticalization: How the particular and the general interact in language change." In Corrigan, Roberta, Edith A. Moravcsik, Hamid Ouali and Kathleen M. Wheatley (Eds.): *Formulaic Language*. Vol. 1: *Distribution and Historical Change*. Amsterdam/Philadelphia: John Benjamins. 187-217.

Calhoun, Sasha, Jean Carletta, Jason Brenier, Neil Mayo, Daniel Jurafsky, Mark Steedman and David Beaver (2010): "The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue." *Language Resources and Evaluation Journal*. 387-419.

Carletta, Jean, Stefan Evert, Jonathan Kilgour, Craig Nicol, Dennis Reidsma, Judy Robertson and Holger Voormann (2009): *Documentation for the NITE XML Toolkit. Revised Version 0.3*. As downloaded from http://groups.inf.ed.ac.uk/nxt/documentation/pdf/documentation.pdf, 26th of October 2010.

Carter, Ronald and Michael McCarthy (2006): *Cambridge Grammar of English.* Cambridge: CUP.

Chambers, John M., William S. Cleveland, Beat Kleiner and Paul A. Turkey (1983): *Graphical Methods for Data Analysis.* Pacific Grove, CA: Wadsworth & Brooks/Cole.

Chomsky, Noam (1965): *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Christenfeld, Nicholas (1994): "Options and ums." *Journal of Language and Social Psychology* 13. 192-9.

Clark, Herbert H. (1996): *Using Language.* Cambridge: CUP.

Clark, Herbert H. (2004): "Pragmatics of language performance." In Horn, Laurence R. and Gregory Ward (Eds.): *The Handbook of Pragmatics.* Malden, MA: Blackwell. 365-82.

Clark, Herbert H. and Jean E. Fox Tree (2002): "Using *uh* and *um* in spontaneous speaking." *Cognition* 84. 73-110.

Clark, Herbert H. and Thomas Wasow (1998): "Repeating words in spontaneous speech." *Cognitive Psychology* 37 (3). 201-42.

Clark, Herbert H. and Eve V. Clark (1977): *Psychology and Language. An Introduction to Psycholinguistics.* New York: Harcourt Brace Jovanovich.

Cook, Mark (1971): "The incidence of filled pauses in relation to part of speech." *Language and Speech* 14 (2). 135-9.

Corley, M. and O. W. Stewart (2008): "Hesitation disfluencies in spontaneous speech: The meaning of *um*." *Language and Linguistics Compass* 2 (8). 589-602.

Croft, William (2001): *Radical Construction Grammar: Syntactic Theory in Typological Perspective.* Oxford: OUP.

Croft, William and David A. Cruse (2004): *Cognitive Linguistics.* Cambridge: CUP.

Crystal, David (1988): "Another look at, well, you know..." *English Today* 4 (1). 47-9.

Daudaravičius, Vidas and Rūta Marcinkevičienė (2004): "Gravity Counts for the boundaries of collocations." *International Journal of Corpus Linguistics* 9 (2). 321-48.

De Ruiter, Jan P., Holger Mitterer and Nick J. Enfield (2006): "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation." *Language* 82 (3). 515-35.

Deese, James (1984): *Thought into Speech: The Psychology of a Language.* Englewood Cliffs, NJ: Prentice Hall.

Deshmukh, Neeraj, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker and Joseph Picone (1998): "Resegmentation of Switchboard." *Proceedings of ICSLP.* Sydney, Australia. 1543-6.

Ellis, Nick C. (2002): "Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition." *SSLA* 24. 143-88.

Ellis, Nick C. (2003): "Constructions, chunking, and connectionism: The emergence of second language structure." In Doughty, Catherine J. and Michael H. Long (Eds.): *Handbook of Second Language Acquisition.* Oxford: Blackwell. 63-104.

Ellis, Nick C. (2008): "Phraseology: The periphery and the heart of language." In Meunier, F. and S. Granger (Eds.): *Phraseology in Lerning and Teaching.* Amsterdam: John Benjamins. 1-13.

Ellis, Nick C., Rita Simpson-Vlach and Carson Maynard (2008): "Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL." *TESOL Quarterly* 24 (3). 375-96.

Elman, Jeffrey L. (1990): "Finding structure in time." *Cognitive Science* 14. 179-211.

Erman, Britt (1987): *Pragmatic Expressions in English: A Study of you know, you see and I mean in Face-to-face Conversation.* Stockholm: Almqvist & Wiksell.

Erman, Britt (2007): "Cognitive processes as evidence of the idiom principle." *International Journal of Corpus Linguistics* 12 (1). 25-53.

Erman, Britt and Beatrice Warren (2000): "The idiom principle and the open choice principle." *Text* 20 (1). 29-62.

Fehringer, Carol and Christina Fry (2007): "Hesitation phenomena in the language production of bilingual speakers: The role of working memory." *Folia Linguistica* 41. 37-72.

Feldstein, Stanley, Marcia S. Brenner and Joseph Jaffe (1963): "The effect of subject sex, verbal interaction and topical focus on speech disruption." *Language and Speech* 6. 229-39.

Field, Andy, Jeremy Miles and Zoë Field (2012): *Discovering Statistics Using R.* London: Sage.

Fillmore, Charles J., Paul Kay and Mary Catherine O'Connor (2003): "Regularity and idiomaticity in grammatical constructions: The case of Let Alone." In Tomasello, Michael (Ed.): *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure.* Mahwah, NJ: Lawrence Erlbaum. 243-70.

Ford, Cecilia E. and Sandra A. Thompson (1996): "Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns." In Ochs, Elinor, Emanuel Schegloff and Sandra Thompson (Eds.): *Interaction and grammar*. Cambridge: Cambridge Univ. Press. 134-84.

Fox, Barbara A., Makoto Hayashi and Robert Jasperson (1996): "Resources and repair: a cross-linguistic study of syntax and repair." In Ochs, Elinor, Emanuel Schegloff and Sandra Thompson (Eds.): *Interaction and Grammar*. Cambridge: CUP. 185-237.

Fox, Barbara A. and Robert Jasperson (1995): "A syntactic exploration of repair in English conversation." In Davis, Philip W. (Ed.): *Alternative Linguistics: Descriptive and Theoretical Modes.* Amsterdam/Philadelphia: John Benjamins. 77-134.

Fox Tree, Jean E. (1995): "Effects of false starts and repetitions on the processing of subsequent words in spontaneous speech." *Journal of Memory and Language* 34 (6). 709-38.

Fraser, Bruce (1990): "An approach to discourse markers." *Journal of Pragmatics* 14. 383-95.

Fraser, Bruce (1999): "What are discourse markers?" *Journal of Pragmatics* 31. 931-52.

Fried, Mirjam and Jan-Ola Östman (2004): "Construction Grammar: a thumbnail sketch." In Fried, Mirjam and Jan-Ola Östman (Eds.): *Construction Grammar in a Cross-language Perspective.* Amsterdam/Philadelphia: John Benjamins. 11-86.

Fung, Loretta and Ronald Carter (2007): "Discourse markers and spoken English: native and learner use in pedagogic settings." *Applied Linguistics* 28 (3). 410-39.

Genuer, Robin, Jean-Michel Poggi and Christine Tuleau (2008): "Random Forests: some methodological insights." *Rapport de Recherche* 6729. 1-32.

Gibbs, Raymond W. (2007): "Idioms and formulaic language." In Geeraerts, Dirk and Hubert Cuyckens (Eds.): *The Oxford Handbook of Cognitive Linguistics.* Oxford: OUP. 697-725.

Gilquin, Gaetanelle (2008): "Hesitation markers among EFL learners: Pragmatic deficiency or difference?" In Romero-Trillo, Jesús (Ed.): *Pragmatics and Corpus Linguistics: A Mutualistic Entente.* Berlin: Mouton de Gruyter. 119-49.

Godfrey, John J. and Edward Holliman (1997): *Switchboard -1 Release 2*. Linguistic Data Consortium. Retrieved July, 15th 2010, from http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S62.

Godfrey, John J., Edward Holliman and Jane McDaniel (1992): "SWITCHBOARD: Telephone speech corpus for research and development." *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1992* 1. I-517-I-20.

Goldberg, Adele (1995): *Constructions. A Construction Grammar Approach to Argument Structure.* Chicago: University of Chcago Press.

Goldman-Eisler, Frieda (1961): "Hesitation and information in speech." In Cherry, Colin (Ed.): *Information Theory*. London: Butterworth. 162-74.

Goldman-Eisler, Frieda (1968): *Psycholinguistics: Experiments in Spontaneous Speech.* New York: Academic Press.

Goldstein, Bejamin A., Eric C. Polley and Farren B.S. Briggs (2011): "Random forests for genetic association studies." *Statistical Applications in Genetics and Molecular Biology* 10 (1). 1-34.

Gregory, Michelle L., William D. Raymond, Alan Bell, Eric Fosler-Lussier and Daniel Jurafsky (1999): "The effects of collocational strength and contextual predictability in lexical production." *CLS* 35. 151-166.

Gries, Stefan Th. (2005): "Syntactic priming: A corpus-based approach." *Journal of Psycholinguistic Research* 34 (4). 365-99.

Gries, Stefan Th. (2008): "Phraseology and linguistic theory. A brief survey." In Granger, Sylviane and Fanny Meunier (Eds.): *Phraseology. An Interdisciplinary Perspective.* Amsterdam/Philadelphia: John Benjamins. 3-25.

Gries, Stefan Th. (2009a): *Quantitative Corpus Linguistics with R: A Practical Introduction.* New York: Routledge.

Gries, Stefan Th. (2009b): *Statistics for Linguistics with R. A Practical Introduction.* Berlin: De Gruyter Mouton.

Gries, Stefan Th. (2010): "Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora." *Proceedings of Corpus Linguistics, University of Liverpool*. 1-14.

Gries, Stefan Th. (2013): "50-something years of work on collocations: what is or should be next …." *International Journal of Corpus Linguistics* 18 (1). 137-65.

Gries, Stefan Th. and Joybrato Mukherjee (2010): "Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes." *International Journal of Corpus Linguistics* 15 (4). 520-48.

Heeman, Peter A. and James F. Allen (1999): "Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue." *Computational Lingusitics* 25 (4). 527-71.

Hieke, Adolf E., Sabine Kowal and Daniel C. O'Connell (1983): "The trouble with 'articulatory' pauses." *Language and Speech* 26. 203-14.

Hilpert, Martin (2013): "Corpus-based approaches to constructional change." In Trousdale, Graeme and Thomas Hoffmann (Eds.): *The Oxford Handbook of Construction Grammar*. Oxford: OUP. 458-77.

Hindle, Donald (1994): "A parser for text corpora." In Atkins, Beryl T. S. and Antonio Zampolli (Eds.): *Computational Approaches to the Lexicon.* Oxford: OUP. 103-51.

Holmes, Virginia M. (1988): "Hesitations and sentence planning." *Language and Cognitive Processes* 3 (4). 323-61.

Hosman, Lawrence A. (1989): "The Evaluative Consequences of Hedges, Hesitations, and Intensifiers.: Powerful and Powerless Speech Styles." *Human Communication Research* 3 (15). 383-406.

Hothorn, Torsten, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro and Mark Van Der Laan (2006): "Survival Ensembles." *Biostatistics* 7 (3). 355-73.

Hothorn, Torsten, Kurt Hornik and Achim Zeileis (2006): "Unbiased recursive partitioning: A conditional inference framework." *Journal of Computational and Graphical Statistics* 15 (3). 651-74.

Howell, Peter and Keith Young (1991): "The use of prosody in highlighting alteration in repairs from unrestricted speech." *Quarterly Journal of Experimental Psychology* 43A (3). 733-58.

Huddleston, Rodney and Geoffrey K. Pullum (2002): *The Cambridge Grammar of the English Language.* Cambridge: CUP.

Jackendoff, Ray (1997): "Twistin' the night away." *Language* 73 (3). 534-59.

Jefferson, Gail (1989): "Preliminary notes on a possible metric which provides for a "standard maximum" silence of approximately one second in conversation." In Roger, Derek and Peter Bull (Eds.): *Conversation.* Clevedon: Multilingual Matters. 166-96.

Jucker, Andreas (1993): "The discourse marker *well*: A relevance-theoretical account." *Journal of Pragmatics* 19. 435-52.

Jurafsky, Daniel, Alan Bell, Eric Fosler-Lussier, Cynthia Girand and William D. Raymond (1998): "Reduction of English function words in Switchboard." *Proceedings of ICSLP-98, Sydney.* 1-4.

Kapatsinski, Vsevolod M. (2005): "Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair." *Berkeley Linguistics Society* 30. 481-92.

Kapatsinski, Vsevolod M. (2010): "Frequency of use leads to automaticity of production: Evidence from repair in conversation." *Language and Speech* 53 (1). 71-105.

Kapatsinski, Vsevolod M. and Joshua Radicke (2009): "Frequency and the emergence of prefabs: Evidence from monitoring. " In Corrigan, Roberta, Edith A. Moravcsik, Hamid Ouali and Kathleen M. Wheatley (Eds.): *Formulaic Language. Vol. 2: Acquisition, Loss, Psychological Reality, Functional Explanations.* Amsterdam: John Benjamins. 499-520.

Kemmer, Suzanne and Michael Barlow (2000): "Introduction: A usage-based conception of language." In Kemmer, Suzanne and Michael Barlow (Eds.): *Usage-Based Models of Language.* Stanford, CA: CSLI Publications. vii-xxviii.

Kowal, Sabine and Daniel C. O'Connell (1993): "Television rhetoric in an age of secondary orality: Psycholinguistic analyses of the speaking performance of Ronald Reagan." *Georgetown Journal of Languages and Linguistics* 1. 174-85.

Lakoff, George (1973): "Hedges: A study in meaning criteria and the logic of fuzzy concepts." *Journal of Philosophical Logic* 2 (4). 458-508.

Langacker, Ronald W. (1987): *The Foundations of Cognitive Grammar. Vol. 1: Theoretical Prerequisites.* Stanford: Stanford University Press.

Langacker, Ronald W. (2000): "A dynamic usage-based model." In Kemmer, Suzanne and Michael Barlow (Eds.): *Usage-Based Models of Language.* Stanford, CA: CSLI Publications. 1-63.

Levelt, Willem J.M. (1983): "Monitoring and self-repair in speech." *Cognition* 14 (1). 41-104.

Levelt, Willem J.M. (1992): "Accessing words in speech production: Stages, processes and representations." In Levelt, Willem J.M. (Ed.): *Lexical Access in Speech Production.* Amsterdam: Blackwell. 1-22.

Levey, Stephen (2006): "The sociolinguistic distribution of discourse marker like in preadolescent speech." *Multilingua* 25. 413-41.

Lounsbury, Floyd G. (1954): "Pausal, juncture and hesitation phenomena." In Osgood, Charles E. and Thomas A. Seboek (Eds.): *Psycholinguistics: A Survey of Theory and Research Problems.* Baltimore.

Lucas, Antoine (2010): *amap: Another Multidimensional Analysis Package.* R package version 0.8-5. <http://CRAN.R-project.org/package=amap>.

Maclay, Howard and Charles E. Osgood (1959): "Hesitation phenomena in spontaneous English speech." *Word* 15. 19-44.

Manning, Christopher D. and Hinrich Schütze (1999): *Foundations of Statistical Natural Language processing.* Cambridge, MA: MIT Press.

Marcus, Mitchell P., Mary Ann Marcinkiewicz and Beatrice Santorini (1993): "Building a large annotated corpus of English: The Penn Treebank." *Computational Linguistics* 19 (2). 313-30.

Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz and Ann Taylor (1999): *Treebank-3*. Linguistic Data Consortium. Retrieved July, 24th 2012, from http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42.

McClelland, James L. and David E. Rumelhart (1981): "An interactive activation model of context effects in letter perception. Part I: An account of basic findings." *Psychological Review* 88. 375-407.

McClelland, James L. and David E. Rumelhart (Eds.) (1986): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Psychological and Biological Models*. (Vol. 2). Cambridge, MA/London: MIT Press/Bradford.

Mel'čuk, Igor (1998): "Collocations and lexical functions." In Cowie, Anthony P. (Ed.): *Phraseology: Theory, Analysis, and Applications.* Oxford: OUP. 23-53.

Meteer, Marie and Ann Taylor (1995): *Dysfluency Annotation Stylebook for the Switchboard Corpus.* Department of Computer and Information Science, University of Pennsylvania.

Moon, Rosamund (1998): "Frequencies and forms of phrasal lexemes in English." In Cowie, Anthony P. (Ed.): *Phraseology: Theory, Analysis, and Application.* Oxford: OUP. 81-100.

Morton, John and John Long (1976): "Effect of word transitional probability on phoneme identification." *Journal of Verbal Learning and Verbal Behaviour* 15. 43-51.

Mukherjee, Joybrato (2007): "Speech is silver, but silence is golden: Some remarks on the function(s) of pauses." *Anglia - Zeitschrift für englische Philologie* 118 (4). 571-84.

Müller, Simone (2005): *Discourse Markers in Native and Non-Native English Discourse.* Amsterdam: John Benjamins.

Newell, Allen (1990): *Unified Theories of Cognition.* Cambridge, MA: MIT Press.

O'Connell, Daniel C. and Sabine Kowal (2004): "The history of research on the filled pause as evidence of the written language bias in linguistics (Linell, 1982)." *Journal of Psycholinguistic Research* 33 (6). 459-74.

O'Connell, Daniel C. and Sabine Kowal (2005): "*Uh* and *Um* revisited: Are they interjections for signalling delay?" *Journal of Psycholinguistic Research* 6 (34). 555-76.

O'Connell, Daniel C., Sabine Kowal and Carie Ageneau (2005): "Interjections in interviews." *Journal of Psycholinguistic Research* 34 (2). 153-71.

Oakes, Michael (1998): *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Pawley, Andrew and Frances Hodgetts Syder (1983): "Two Puzzles for linguistic theory: nativelike selection and nativelike fluency." In Richards, Jack C. and Richard W. Schmidt (Eds.): *Language and Communication.* London/New York: Longman. 191-226.

Pickering, Martin J. and Holly P. Branigan (1999): "Syntactic priming in language production." *Trends in Cognitive Sciences* 3 (4). 136-41.

Pierrehumbert, Janet (2001): "Exemplar dynamics: Word frequency, lenition and contrast." In Bybee, Joan and Paul Hopper (Eds.): *Frequency and the Emergence of Linguistics Structure.* Amsterdam: John Benjamins.

Power, Michael J. (1986): "A technique for measuring processing load during speech production." *Journal of Psycholinguistic Research* 15 (5). 371-82.

Pullum, Geoffrey K. and Rodney Huddleston (2002): "Prepositions and preposition phrases." In Huddleston, Rodney and Goeffrey K. Pullum (Eds.): *The Cambridge Grammar of the English Language.* Cambridge: CUP. 597-661.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1985): *A Comprehensive Grammar of the English Language.* London/New York: Longman.

R Development Core Team (2009): *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Rumelhart, David E. and James L. McClelland (Eds.) (1986): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Foundations.* (Vol. 1). Cambridge, MA/ London: MIT Press/Bradford.

Sacks, Harvey, Emanuel Schegloff and Gail Jefferson (1974): "A simplest systematics for the organisation of turn-taking for conversation." *Language* 50 (4). 696-735.

Santorini, Beatrice (1990): *Part-of-Speech Tagging Guidelines for the Penn Treebank Project. (3rd revision, 2nd printing)*: Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Saz Rubio, Maria Milagros Del (2007): *English Discourse Markers of Reformulation.* Bern: Peter Lang.

Schachter, Stanley, Nicholas Christenfeld, Bernard Ravina and Frances Bilous (1991): "Speech disfluency and the structure of knowledge." *Journal of Personality and Social Psychology* 60 (3). 362-7.

Schegloff, Emanuel, Gail Jefferson and Harvey Sacks (1977): "The preference for self-correction in the organisation of repair in conversation." *Language* 53 (2). 361-82.

Schiffrin, Deborah (1987): *Discourse Markers.* Cambridge: CUP.

Schilperoord, Joost and Arie Verhagen (2006): "Grammar and language production. Where do function words come from?" In Luchjenbroers, June (Ed.): *Cognitive Linguistics Investigations*. Amsterdam/Philadalphia: John Benjamins. 139-68.

Schneider, Ulrike (2014): "CART Trees and Random Forests in Linguistics." In Schulz, Janne C. and Sven Hermann (Eds.): *Hochleistungsrechnen in Baden-Württemberg. Ausgewählte Aktivitäten im bwGRiD.* Karlsruhe: KIT Scientific Publishing. 67-81.

Schourup, Lawrence (1985): *Common Discourse Particles in English Conversation.* New York: Garland.

Schourup, Lawrence (1999): "Tutorial overview: Discourse markers." *Lingua* 107. 227-65.

Shih, Stephanie Sin-yun (2011): *Random Forests for Classification Trees and Categorical Dependent Variables: An Informal Quick Start R Guide.* Stanford University/University of California, Berkley. Retrieved Sept. 28th 2011, from <http://www.stanford.edu/~stephsus/R-randomforest-guide.pdf>.

Shriberg, Elizabeth. (1994). *Preliminaries to a Theory of Speech Disfluencies.* PhD thesis, University of California at Berkeley.

Shriberg, Elizabeth (1996): "Disfluencies in Switchboard." *Proceedings of the International Conference on Spoken Language Processing (Addendum).* Philadelphia, PA. 11-4.

Shriberg, Elizabeth and Andreas Stolcke (1996): "Word predictability after hesitations: A corpus-based study." *Proceedings of the International Conference on Spoken Language Processing (Vol.3)*. 1868-71.

Simpson-Vlach, Rita and Nick C. Ellis (2010): "An academic formulas list: New methods in phraseology research." *Applied Linguistics* 1 (26). 1-26.

Sinclair, John (Ed.) (1990): *Collins COBUILD English Grammar.* London: Collins.

Sinclair, John (1991): *Corpus, Concordance and Collocation.* Oxford: OUP.

Stefanowitsch, Anatol and Stefan Th. Gries (2003): "Collostructions: Investigating the interaction of words and constructions." *International Journal of Corpus Linguistics* 8 (2). 209-43.

Stolcke, Andreas and Elizabeth Shriberg (1996): "Statistical language modeling for speech disfluencies." *Proceedings of ICASSP-96*. 1-4.

Strobl, Carolin, Anne-Laure Boulestreix, Thomas Kneib, Thomas Augustin and Achim Zeileis (2008): "Conditional variable importance for random forests." *BMC Bioinformatics* 9 (307).

Strobl, Carolin, Anne-Laure Boulestreix, Achim Zeileis and Torsten Hothorn (2007): "Bias in random forest variable importance measures: Illustrations, sources and a solution." *BMC Bioinformatics* 8 (25).

Strobl, Carolin, James Malley and Gerhard Tutz (2009a): "An introduction to recursive partitioning: Rationale, application and characteristics of Classification and Regression Trees, bagging and random forests." *University of Munich, Department of Statistics, Technical Report* 55. (Including Supplement). 1-42.

Strobl, Carolin, James Malley and Gerhard Tutz (2009b): "An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests." *Psychological Methods* 14 (4). 323-48.

Swerts, Marc (1998): "Filled pauses as markers of discourse structure." *Journal of Pragmatics* 30 (4). 485-96.

Szmrecsanyi, Benedikt (2006): *Morphosyntactic Persistence in Spoken English. A Corpus Study.* Berlin/New York: Mouton de Gruyter.

Tagliamonte, Sali A. and R. Harald Baayen (2012): "Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice." *Language Variation and Change* 24. 135-78.

Therneau, Terry M., Beth Atkinson and Brian Ripley (2011): *rpart: Recursive Partitioning.* R package version 3.1-50. <http://CRAN.R-project.org/package=rpart>.

Therneau, Terry M. and Elizabeth J. Atkinson (1997): *An Introduction to Recursive Partitioning Using the RPART Routines.* Mayo Foundation.

Tily, Harry, Susanne Gahl, Inbal Arnon, Neal Snider, Anubha Kothari and Joan Bresnan (2009): "Syntactic probabilities affect pronunciation variation in spontaneous speech." *Language and Cognition* 1 (2). 147-65.

Tottie, Gunnel and Filippo Svalduz (2009): "*Er*, *erm*, *uh*, *uhm* – filled pauses in British and American English." Paper presented at ICAME 30, University of Lancaster, May 2009.

Vasilescu, Ioana, Maria Candea and Martine Adda-Decker (2005): "Perceptual salience of language-specific acoustic differences in autonomous fillers across eigth languages." *INTERSPEECH 2005*. 1773-6.

Vogel Sosa, Anna and James MacFarlane (2002): "Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*." *Brain and Language* 83. 227-36.

Wiechmann, Daniel (2008): "On the computation of collostruction strength: Testing measures of association as expressions of lexical bias." *Corpus Linguistics and Linguistic Theory* 4 (2). 253-90.

Wray, Alison (2002): *Formulaic Language and the Lexicon.* Cambridge: CUP.

This book addresses the questions in how far the frequency with which speakers use sequences of words influences how their minds store and process these strings and how such effects can best be modelled.

The studies compiled in the book start out from the usage-based tenet that language use shapes its mental representation. They focus on various types of multi-word strings, such as *I don't know* or *a lot of*, and investigate how their mental representation changes depending on how frequently they are used. Two detailed sets of analyses test whether the more often sequences are used, the more unit-like or 'chunked' they become and furthermore address a number of theoretical and technical questions which mainly arise from details of Bybee's (2010) model of the mind and her understanding of chunking.

The author is the first to present a large-scale corpus analysis which utilises the placement of hesitations in spoken American English – as represented by the Switchboard NXT corpus – together with the innovative statistical procedures of Classification and Regression Trees and random forests. These tools are also used to test whether usage frequency is better modelled as absolute co-occurrence frequency or by means of one (or several) of the many probabilistic measures currently applied in the field (e.g. transitional probabilities, lexical gravity G).

Results not only offer insights about the interrelation between frequency of use, constituent structure and mental storage, but also about the placement of hesitations as well as about strategies for modelling linguistic effects.

Ulrike Schneider studied English, Spanish and Business Studies at the Universities of Giessen (Germany) and Leicester (UK) and received a Diploma degree in Applied Modern Languages and Business from the University of Giessen in 2009. In the same year she joined the research training group "Frequency Effects in Language" at the University of Freiburg (Germany). In 2012 she transferred to the University of Mainz (Germany) where she works as a researcher and lecturer in English Linguistics. This book results from her doctoral dissertation, submitted in February 2014.

ISBN 978-3-928969-57-4

UNI FREIBURG