UNIVERSITÉ DE NEUCHÂTEL
FACULTÉ DES LETTRES ET SCIENCES HUMAINES
INSTITUT DE LANGUE ET LITTÉRATURE ANGLAISES
&
UNIVERSITEIT ANTWERPEN
FACULTEIT LETTEREN EN WIJSBEGEERTE
DEPARTEMENT TAALKUNDE

Measurements of Grammaticalization: Developing a quantitative index for the study of grammatical change

Thèse présentée pour l'obtention du titre de docteur ès lettres en anglais auprès de l'Université de Neuchâtel par

Proefschrift voorgelegd tot het behalen van de graad van doctor in de Taalkunde aan de Universiteit Antwerpen te verdedigen door

David Correia Saavedra

Le 7 juin 2019

Membres du jury :
Martin Hilpert, co-directeur de thèse, professeur, Université de Neuchâtel
Peter Petré, co-directeur de thèse, professeur, Universiteit Antwerpen, Belgique
Elena Smirnova, professeure, Université de Neuchâtel
Hendrik de Smet, professeur, KU Leuven, Belgique
Marianne Hundt, professeure, Université de Zürich

This dissertation was funded by the following SNF projects :

“Why is there regularity in grammatical semantic change? Reassessing the explanatory power of asymmetric priming”
(SNF Grant 100015_149176)

“Measurements of Grammaticalization: Developing a quantitative index for the study of grammatical change”
(SNF Grant P1NEP1_171937)

UNIVERSITÉ DE
NEUCHÂTEL

Faculté des Lettres et Sciences Humaines
Espace Louis-Agassiz 1
CH - 2000 Neuchâtel
Tél : +41 (0)32 718 17 00
Fax : +41 (0)32 718 17 71
secretariat.lettres@unine.ch

IMPRIMATUR

La Faculté des lettres et sciences humaines de l'Université de Neuchâtel, sur les rapports de M. Martin Hilpert, co-directeur de thèse, professeur, Université de Neuchâtel ; M. Peter Petré, co-directeur de thèse, professeur, Universiteit Antwerpen, Belgique ; M. Hendrik de Smet, professeur assistant, KU Leuven, Belgique ; Mme Marianne Hundt, professeure, Université de Zürich ; Mme Elena Smirnova, professeure, Université de Neuchâtel, autorise l'impression de la thèse présentée par M. David Correia Saavedra en laissant à l'auteur la responsabilité des opinions énoncées.

Neuchâtel, le 7 juin 2019

Le doyen
Pierre Alain Mariaux

Abstract

There is a broad consensus that grammaticalization is a process that is gradual and largely unidirectional: lexical elements acquire grammatical functions, and grammatical elements can undergo further grammaticalization (Hopper and Traugott 2003). This consensus is based on substantial cross-linguistic evidence. While most existing studies present qualitative evidence, this dissertation discusses a quantitative approach that tries to measure degrees of grammaticalization using corpus-based variables, thereby complementing existing qualitative work. The proposed measurement is calculated on the basis of several parameters that are known to play a role in grammaticalization (Lehmann 2002, Hopper 1991), such as token frequency, phonological length, collocational diversity, colligate diversity, and dispersion. These variables are used in a binary logistic regression model which can assign a score to a given linguistic item that reflects its degree of grammaticalization.

Grammaticalization can be conceived in synchrony and in diachrony. The synchronic view of grammaticalization is concerned with the fact that some items are more grammaticalized than others, which is commonly referred to as gradience. The diachronic view is concerned with the development of grammatical elements over time, whereby elements become increasingly more grammatical through small incremental steps, which is also known as gradualness. This dissertation proposes studies that deal with each of these views.

In order to quantify the gradience of grammaticalization, data from the British National Corpus is used. 264 lexical and 264 grammatical elements are selected in order to train a binary logistic regression model. This model can rank these items and determine which ones are more lexical or more grammatical. The results indicate that the model makes successful predictions overall. In addition, generalizations regarding grammaticalization can be supported, such as the relevance of key variables (e.g. token frequency, diversity to the left of a given item) or the ranking of morphosyntactic categories as a whole (e.g. adverbs are on average in between the lexical and grammatical categories).

The gradualness of grammaticalization is investigated using a selection of twenty elements that have grammatical and lexical counterparts in English (e.g. *keep*). The Corpus of Historical American English (1810s-2000s) is used to retrieve the relevant data. The aim is to check how the different variables and the grammaticalization scores develop over time. The main theoretical value of this approach is that it can offer an empirically operationalized way of measuring unidirectionality in grammaticalization, as opposed to a more qualitative observation based on individual case studies.

Keywords: grammaticalization, gradience, gradualness, binary logistic regression, quantitative linguistics, corpus linguistics, grammatical change.

Résumé

Il existe un large consensus sur le fait que la grammaticalisation est un processus graduel et largement unidirectionnel : les éléments lexicaux acquièrent des fonctions grammaticales et les éléments grammaticaux peuvent ensuite se grammaticaliser d'avantage (Hopper et Traugott 2003). Ce consensus est fondé sur des observations consistantes entre plusieurs langues. Alors que la plupart des études existantes présentent des données qualitatives, cette thèse propose une approche quantitative qui tente de mesurer les degrés de grammaticalisation à l'aide de variables basées sur des corpus, complémentant ainsi le travail qualitatif existant. La mesure proposée est calculée sur la base de plusieurs paramètres connus pour jouer un rôle dans la grammaticalisation (Lehmann 2002, Hopper 1991), tels que la fréquence (token), la longueur phonologique, la diversité des collocats, la diversité des colligats et la dispersion. Ces variables sont utilisées dans un modèle de régression logistique binaire qui peut attribuer un score à un élément linguistique donné qui reflète son degré de grammaticalisation.

La grammaticalisation peut être conçue en synchronie et en diachronie. La vision synchronique de la grammaticalisation concerne le fait que certains éléments sont plus grammaticalisés que d'autres, ce que l'on appelle communément la gradience. La vue diachronique concerne le développement d'éléments grammaticaux au fil du temps, les éléments devenant de plus en plus grammaticaux par petits pas incrémentiels, ce qu'on appelle aussi la gradualité. Cette thèse propose des études qui traitent de chacun de ces points de vue.

Pour quantifier la gradience de la grammaticalisation, des données du British National Corpus sont utilisées. 264 éléments lexicaux et 264 éléments grammaticaux sont sélectionnés pour entraîner un modèle de régression logistique binaire. Ce modèle peut classer ces éléments et déterminer ceux qui sont plus lexicaux ou plus grammaticaux. Les résultats indiquent que le modèle fait des prédictions réussies dans l'ensemble. De plus, des généralisations concernant la grammaticalisation peuvent être soutenues, comme la pertinence des variables clés (p.ex. la fréquence token et la diversité à gauche d'un élément donné) ou le classement des catégories morphosyntaxiques dans leur ensemble (p.ex. les adverbes se situent en moyenne entre les catégories lexicales et grammaticales).

La gradualité de la grammaticalisation est étudiée à l'aide d'une sélection de vingt éléments qui ont des équivalents grammaticaux et lexicaux en anglais (p.ex. *keep*). Le Corpus of Historical American English (1810s-2000s) est utilisé pour récupérer les données pertinentes. L'objectif est de vérifier l'évolution dans le temps des différentes variables et des scores de grammaticalisation. La principale valeur théorique de cette approche est qu'elle peut

offrir une façon empiriquement opérationnelle de mesurer l'unidirectionnalité en grammaticalisation, par opposition à une observation plus qualitative basée sur des études de cas individuels.

Mots-clés: grammaticalisation, gradience, gradualité, régression logistique binaire, linguistique quantitative, linguistique des corpus, changement grammatical.

Samenvattig

Er bestaat een brede consensus dat grammaticalisatie als proces geleidelijk en grotendeels in één richting verloopt: lexicale elementen verwerven grammaticale functies, en grammaticale elementen kunnen verdere grammaticalisatie ondergaan (Hopper en Traugott 2003). Deze consensus is gebaseerd op substantieel cross-linguïstisch bewijs. Waar echter de meeste bestaande studies kwalitatief bewijs leveren, werkt dit proefschrift een kwantitatieve benadering uit die de graad van grammaticalisatie tracht te meten aan de hand van corpusgebaseerde variabelen, complementair aan bestaand kwalitatief werk. De voorgestelde maat wordt berekend op basis van verscheidene parameters waarvan bekend is dat ze een rol spelen in grammaticalisatie (Lehmann 2002, Hopper 1991), zoals tokenfrequentie, fonologische lengte, collocationele diversiteit, colligationele diversiteit, en spreidingsgraad. Deze variabelen worden gebruikt in een binair logistisch regressiemodel dat een score kan toekennen aan een specifiek taalitem, die de grammaticalisatiegraad ervan weerspiegelt.

Grammaticalisatie kan sychroon of diachroon worden opgevat. Het synchrone perspectief betreft het gegeven dat sommige items meer gegrammaticaliseerd zijn dan andere, een gegeven dat algemeen bekend staat als de grammaticalisatiegradiënt. Het diachrone perspectief heeft betrekking op de ontwikkeling van grammaticale elementen in de loop der tijd, waarbij elementen in steeds grammaticaler worden door kleine incrementele stapjes, een fenomeen dat bekend staat als de geleidelijkheid van grammaticalisatie. Dit proefschrift presenteert studies die elk van deze perspectieven omvatten.

Om de grammaticalisatiegradiënt te kunnen kwantificeren, wordt gebruik gemaakt van data uit het British National Corpus. 264 lexicale en 264 grammaticale elementen werden geselecteerd om een binair logistisch model te trainen. Dit model kan deze items rangschikken en bepalen welke eerder lexicaal zijn en welke eerder grammaticaal. De resultaten geven aan dat het model over het geheel genomen succesvolle voorspellingen maakt. Bovendien bieden ze steun aan generalisaties met betrekking tot grammaticalisatie, zoals de relevantie van sleutelvariabelen (bv. tokenfrequentie, diversiteit links van een item) of de rangorde van morfosyntactische categorieën in haar geheel (bv. bijwoorden situeren zich gemiddeld genomen tussen de lexicale en grammaticale categorieën).

De geleidelijkheid van grammaticalisatie wordt onderzocht aan de hand van een selectie van twintig elementen met grammaticale en lexicale tegenhangers in het Engels (bv. *keep*). Het Corpus of Historical American English (1810s-2000s) werd gebruikt om de relevante data op te halen. Het doel is om na te gaan hoe de verschillende variabelen en

grammaticalisatiescores zich ontwikkelen over de tijd. De belangrijkste theoretische waarde van deze benadering is dat ze een empirisch geoperationaliseerde manier kan bieden om unidirectionaliteit in grammaticalisatie te meten, tegenover een meer kwalitatieve observatie gebaseerd op individuele case studies.

Sleutelbegrippen: grammaticalisatie, gradiënt, geleidelijkheid, binair logistisch regressiemodel, kwantitatieve taalkunde, corpuslinguïstiek, grammaticale verandering.

Acknowledgements

Writing this dissertation was an arduous process, but I was fortunate to be surrounded by the right people. First and foremost, I would like to thank my main supervisor, Martin Hilpert, for his constant support since we started working together in 2014. I would also like to thank my co-supervisor, Peter Petré, whom I met during my FNS Doc.Mobility stay at the University of Antwerp in 2017. Your combined feedback was extremely helpful; I always felt that you respected my ideas and that you gave me a lot of freedom, which is one of my main draws to the academic world.

There are many other people from this academic world that I need to thank as well, including those who introduced me to it. I am greatly indebted to Dorota Smyk-Bhattacharjee who supervised my MA thesis and who introduced me to Martin Hilpert, which encouraged me to pursue an academic career. I also need to thank Aris Xanthos, who co-supervised my MA thesis, for encouraging me to pursue an academic career, and for helping me again years later with my FNS Doc.Mobility grant application.

I am also grateful to the English department of the University of Neuchâtel who was welcoming and quite supportive throughout my PhD. I enjoyed our parties and outings, which were always a nice break from work. I would also like to thank Samuel Bourgeois and Susanne Flach for their support, especially during my colloque de thèse.

During my time as an FNS doctoral student in Neuchâtel, I was also extremely lucky to travel around and present my work at various conferences. This resulted in an exciting collaboration with linguists from the University of Mainz. I am greatly thankful to Linlin Sun, Walter Bisang and Andrej Malchukov for their interest in my work and for their help in applying some of the ideas in this dissertation to Chinese data.

My mobility stay at the University of Antwerp was certainly the most memorable part of my PhD. I would like to thank my office mates Lynn Anthonissen, Sara Budts, Will Standing, Emma-Louise Silva and Odile Strik for the friendly and stimulating environment at our office, and, of course, for our lunches at the Komida. The ambitious projects you work on were inspiring and certainly motivated me to get on with my own.

In addition, I am also grateful to all my friends who are outside of academia and who always helped me put things into perspective. Finally, the last sentence goes to my parents, who have been supportive since the very beginning. Maman, papa, je suis extrêmement chanceux de vous avoir dans ma vie et sans votre soutien cette thèse serait sans doute médiocre, voire inexistante.

Table of contents

1. Introduction 1
2. Theoretical background 15
2.1 Defining grammaticalization 15
2.2 Grammaticalization features 19
2.2.1 Lehmann's parameters 19
2.2.2 Hopper's principles 24
2.3 The cline-based view 25
2.4 The unidirectionality hypothesis 28
2.5 Degrammaticalization 31
2.6 Grammaticalization across languages 34
2.7 The construction-based view of grammaticalization 36
2.8 Summary 41
3. Grammaticalization parameters quantified 43
3.1 Token frequency 44
3.2 Letter counts 47
3.3 Collocate diversity 49
3.4 Colligate diversity 53
3.5 Dispersion 55
3.6 Corpus-based works on grammaticalization 58
4. Data and methods 65
4.1 Statistical methods 65
4.1.1 Linear regression 65
4.1.2 Binary logistic regression 69
4.1.3 Mixed effects modelling 77
4.1.4 Principal component analysis 77
4.2 Corpora 82
4.2.1 The British National Corpus (BNC) 82
4.2.2 The Corpus of Historical American English (COHA) 83
5. Quantifying grammaticalization in synchrony 87

5.1 Introduction 87
5.2 Data 88
5.3 Retrieving the relevant measures from the British National Corpus 90
5.4 Version 1: Written BNC (no punctuation) 94
5.4.1 Principal component analysis 94
5.4.2 Binary logistic regression 98
5.4.3 Discussion 105
5.4.4 Including mixed effects 108
5.5 Version 2: Written BNC (with punctuation) 109
5.5.1 Principal component analysis 110
5.5.2 Binary logistic regression 112
5.5.3 Discussion 113
5.6 Version 3: Spoken BNC (no punctuation) 114
5.6.1 Principal component analysis 114
5.6.2 Binary logistic regression 116
5.6.3 Discussion 121
5.7 Re-calibrating the study (modal verbs) 122
5.8 Overview 127
5.9 Conclusions 128
6. Diachronic study 131
6.1 Introduction 131
6.2 Data 133
6.2.1 Selection of cases 133
6.2.2 Distinguishing lexical and grammatical counterparts 136
6.2.3 Keep 140
6.2.4 Quit 143
6.2.5 Have 144
6.2.6 Go 146
6.2.7 Die 147
6.2.8 Help 148
6.2.9 Use 149

6.2.10 Long 150
6.2.11 Well 152
6.2.12 Soon 152
6.2.13 Regard 153
6.2.14 Respect 154
6.2.15 Spite 155
6.2.16 Terms 157
6.2.17 View 157
6.2.18 Event 158
6.2.19 Matter 159
6.2.20 Means 160
6.2.21 Order 161
6.2.22 Back 162
6.3 General results 164
6.3.1 Individual variables 164
6.3.2 Grammaticalization score over time 170
6.3.3 Overview 174
6.4 Individual results 174
6.4.1 Individual variables 175
6.4.2 Grammaticalization score over time 189
6.4.3 Overview 197
6.5 Theoretical relevance of the findings and methodological concerns 198
7. Conclusions 201
7.1 Main findings 201
7.2 Generalizations regarding grammaticalization 205
7.3 Current and future challenges 207
7.4 Concluding remarks 214
References 215
Monographs and articles 215
Corpora 232
Dictionaries and databases 232

Softwares and programming languages 232

Appendices 233

Appendix A – Additional tables (chapter 5) 233

A1. Lexical and grammatical elements database 233

A2. Top 100 and bottom 100 items using the written BNC, punctuation included 237

Appendix B – Random samples for classification data (chapter 6) 239

B1-A. Keep 239

B1-B. Keep 243

B2. Quit 248

B3. Have 251

B4. Go 257

B5. Die 260

B6. Help 262

B7. Use 265

B8. Long 266

B9. Well 268

B10. Soon 270

B11. Regard 272

B12. Respect 274

B13-A. Spite 276

B13-B. Spite 277

B14. Terms 280

B15. View 282

B16. Event 284

B17. Matter 285

B18. Means 287

B19. Order 289

B20-A. Back 291

B20-B. Back 294

Appendix C – Additional figures (chapter 6) 297

C1. Collocate diversity (R1) 297

C2. Collocate diversity (L1) 299

C3. Colligate diversity (L1) 301

Appendix D – Scripts 303

D1. BNC written script (section 5.4) 303

D2. Selection of lexical and grammatical counterparts (section 6.2) 306

D2-A. Auxiliary example (keep) 306

D2-B. Complex preposition example (as well as) 308

D3. COHA script example (sections 6.3 and 6.4) (keep) 309

1. Introduction

This dissertation deals with the phenomenon of grammaticalization. The aim that is pursued in this study is to make grammaticalization quantitatively measurable. In line with this aim, the chapters of this dissertation will develop a corpus-based approach to the quantification of grammaticalization that takes well-known aspects of the phenomenon into account and operationalizes them, based on synchronic and diachronic corpus data from English, in ways that allow other researchers to implement and test the approach on the basis of other linguistic data.

The term grammaticalization presupposes the notion of grammar, which therefore should be the starting point of the discussion. In this dissertation, grammar is viewed as a set of principles that governs how human beings form and understand words and sentences (Huddleston and Pullum 2002). Trying to describe and understand what these principles are and how they work is one of the most important endeavours of contemporary linguistic research. Understanding how language is produced and understood in greater details can provide insightful information about how the human mind works, but it can also have concrete applications such as enhancing how effectively people learn languages or creating more advanced artificial intelligence for language processing and translation tasks. One of the central design features of human languages is that they have elements that fulfil grammatical functions. Grammatical elements, such as markers of tense, definiteness, case, or evidentiality, to name just a few, are used to indicate relationships between the other elements involved in a linguistic utterance. Grammaticalization is the linguistic field of study that aims at deepening our understanding of where they came from and how they develop. A common definition is that grammaticalization is "the change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions and, once grammaticalized, continue to develop new grammatical functions" (Hopper and Traugott 2003: 18). Grammatical elements rarely occur by themselves and they usually require other elements in order to have meaning. They also tend to be phonologically short and semantically abstract. Because of this, grammatical elements can be relatively obscure to speakers and machines alike.

Existing research on grammaticalization has unearthed a rich catalogue of insights on how grammatical elements come into being and develop with regard to their structure in meaning. A lot is known about the qualitative aspects of grammaticalization. Researchers that

are familiar with the relevant literature will be able to identify grammaticalized elements even in languages that they are only superficially familiar with.

However, in the current literature, there is no consensus on how grammaticalization could be quantitatively measured. This is problematic because grammaticalization is widely recognized as a gradual phenomenon (Traugott and Trousdale 2010). Given a continuum from lexical elements to fully grammaticalized elements, where exactly should a specific grammaticalizing element be placed at a given point in time? Is it possible to compare two grammaticalizing elements? Let us consider a concrete example. Nouns for human body parts are lexical items that have gained multiple grammatical functions in many languages (Heine and Kuteva 2002: 46-50). An example from English is the word *back.* In its grammaticalized use, one of its functions is that of a prepositional adverb when used in combination with certain verbs (e.g. *come back*, *go back*). This development involves further smaller steps, where *back* came to be used in more abstract ways that do not necessarily involve a spatial component (e.g. *write back, talk back*), which translates to a broadening of the verbs with which *back* can appear. These more abstract uses indicate further grammaticalization, which leads to the question whether this phenomenon can be quantified in a systematic way. Given two grammatical markers, is there a principled way of stating which one of the two is grammaticalized to a stronger degree? The focus of this dissertation is an attempt to develop a method for measuring such degrees. The first main research question, then, is:

RQ1: Can the gradualness of grammaticalization be quantified?

Grammaticalization is an intensively studied phenomenon and a major field of linguistic research, which means that it inevitably involves many diverging views. There is disagreement on what constitutes an instance of grammaticalization and how it relates to other phenomena of language change, and even whether grammaticalization is a phenomenon in its own right. A logical corollary of these debated issues is that the characteristics of grammaticalization are also not consensually agreed upon. It is likely that disagreement partly stems from the fact that it is relatively difficult to make robust generalizations regarding grammaticalization, because there do not seem to have been attempts at quantifying this phenomenon in a systematic way. This dissertation is an attempt to take a first step in this direction by discussing the extent to which grammaticalization can be measured in terms of quantitative and corpus-based variables. The focus of the dissertation is on English, but the proposed approach is also applicable to other languages. Such a quantitative approach to

grammaticalization enables the analysis of larger datasets in order to support more general claims regarding grammaticalizing elements. It can also be used to highlight which characteristics of grammaticalization are more prototypical and shared by many grammaticalized elements, as opposed to more case-specific features. Furthermore, quantifying the development of grammatical elements is a way to investigate this phenomenon without relying on linguistic intuition.

The quantification of grammaticalization is not a brand new topic. While there have not been attempts to develop a grammaticalization measure in a similar fashion to the present endeavour, several works have taken a quantitative approach on grammatical change (Hilpert and Mair 2015). Many studies have focused on specific grammatical patterns, in particular those that feature competing alternatives (e.g. Mair 2002, Gries and Hilpert 2010, Hilpert 2011). Most of these focus on token frequency or relative frequencies, but there are also examples that focus on type frequency (e.g. Israel 1996).

A notable study that strongly relates to this dissertation is Petré and Van de Velde (2018) who developed a grammaticalization score that relies on counting the number of grammaticalization features on the basis of a list of established features from previous research, such as the ones discussed in Lehmann (2002). The features were selected to measure degrees of grammaticalization of attestations of the *be going to* + *infinitive* construction. The idea is that more features translate to a higher grammaticalization score. However, the approach proposed in this dissertation is more general than this.

The case studies proposed in the following chapters rely on a rather commonly accepted view, namely that languages have contentful elements and functional elements (Hopper and Traugott 2003: 4). Contentful elements provide lexical information and denote entities, objects, actions and qualities. In English, contentful elements are represented in the nominal, verbal, adjectival and adverbial morphosyntactic categories. On the other hand, functional elements provide grammatical information regarding the relationships and links between elements, such as definiteness, possession, causality, spatial locality and referentiality. Traugott and Trousdale (2013: 12) propose the term "procedural" for material with "abstract meaning that signals linguistic relations, perspectives and deictic orientation" (see also Bybee 2002: 111-112 on knowledge of grammar being procedural knowledge). This principally includes articles, determiners, prepositions, conjunctions, auxiliary verbs, pronouns and markers of negation in English. Other languages may have other types of grammatical elements, such as classifiers or postpositions and they may also not include some that are relevant in English (e.g. there are no articles in Chinese). Grammaticalization is the

development from contentful to functional elements, where contentful elements may gain new grammatical functions. In addition, grammaticalization also consists in functional elements gaining further grammatical functions.

Importantly, the distinction between lexical and grammatical elements involves fuzzier boundaries than what has been discussed above. For example, following Langacker (1987), Goldberg and Jackendoff (2004: 532) argue that lexicon and grammar form a "cline". In similar fashion, Lehmann (2002b) discusses cases of adpositions where some of them are "more lexical", whereas others are "more grammatical". Therefore, it should be kept in mind that some of the morphosyntactic categories that are considered as lexical can sometimes be "more grammatical", and vice-versa (see also Taylor 2003, Traugott and Trousdale 2013: 73, Ramat 2011).

The transition from one category to another (van Goethem et al. 2018) is often considered as gradual (Traugott and Trousdale 2010). Gradualness is apparent along various dimensions. Lexical material does not suddenly become grammatical (and grammatical material does not suddenly become even more grammatical), but that this process happens through small incremental steps. For example, *a lot of* in Present-Day English has similar grammatical functions as *many* and *much*, but this development required steps where the initial noun *lot* in Old English (*hlot*, an object, often a piece of wood, by which individuals were selected, e.g. for a draw) gained the grammatical function of denoting groups of objects, which then gained the further function of denoting quantity, which then further extended to abstract concepts as well (e.g. *a lot of courage*) (Traugott and Trousdale 2013: 23-26). This process involves a phenomenon known as semantic bleaching (Sweetser 1988), which comprises grammaticalizing items becoming semantically more general and abstract. Another well-known characteristic of grammaticalization is that transitioning items tend to lose phonological substance and become shorter. For example, many auxiliaries in English involve cliticization (e.g. *'m*, *'re*, *'ll*) or reduced forms (e.g. *gonna*, *wanna*). Again, this process does not happen instantly and the reduced forms require repeated use to become more entrenched. These features of grammaticalization are discussed in chapter 2.

Note that other ways of understanding lexical and grammatical categories have been proposed. For instance, Boye and Harder (2012) consider that the main distinction between lexical and grammatical material pertains to the ancillary status of the latter. Grammatical expressions are discursively secondary, whereas lexical ones have the potential to be discursively primary. This translates to lexical elements being more prominent in a discourse than grammatical elements. Discourse prominence denotes the fact that understanding

information requires prioritization and that some pieces of information are more crucial than others. Prominence is also a scalar concept where lexical material tends to involve higher degrees of prominence. Furthermore, prominence is mainly established by linguistic conventionalization and therefore has some arbitrariness to it. Boye and Harder (2012: 19-21) also discuss how certain elements from certain morphosyntactic categories (in particular adverbs, prepositions and several pronouns) may have members that belong to either grammatical or lexical categories. Therefore, while this dissertation treats categories as discrete in order to enable the operationalization of the data-driven approach (e.g. adverbs are categorized as lexical elements), especially when it comes to chapter 5, an important note is that such categorizations are sometimes more nuanced.

As noted earlier, there are different degrees of grammaticalization, where highly grammaticalized elements tend to display more grammaticalization features (e.g. semantic bleaching, phonological reduction) than weakly grammaticalized ones. The gradual nature of grammaticalization has often been observed through case studies (e.g. Patten 2010, de Smet 2010, Hilpert 2010, König 2011, de Mulder and Carlier 2011, Wiemer 2011, Krug 2011), but there have been few attempts to systematically investigate this from a broader perspective. The approach presented in the following chapters is an attempt at bridging this gap by proposing a quantitative grammaticalization score that ranges between 0 (highly lexical) and 1 (highly grammatical), and which can be applied to most linguistic items with phonological substance (an exception are affixes, which are not the object of the present study). This score is a direct answer to the first research question introduced above, as it provides a concrete value that can be used to rank items according to their degrees of grammaticalization.

While researchers on grammaticalization have a strong interest in the gradual nature of lexical and grammatical categories, there are other domains of research where this consideration is also relevant, even if less crucial. For example, when studying stylistic features of authors, such as measuring vocabulary size or extracting main themes of a discourse, researchers may need to take the distinction between lexical and grammatical items into account. This is the case with stylistic measures such as lexical density (Biber et al. 2002, Hewings et al. 2005), which consists in the ratio between the number of lexical elements and the number of tokens in a text. Texts that have a higher number of lexical elements are considered as more informative, since they involve more contentful items. These methods therefore rely on a way to distinguish between lexical and grammatical categories.

An approach often used is to take an established list of function words in English, such as the Linguistic Inquiry and Word Count Dictionary (Tausczik and Pennebaker 2010).

Another solution is to use an h-point measure, where elements are ranked according to their absolute frequency, from most to least frequent (e.g. Popescu 2009, Cech et al. 2015). In this situation, the h-point is the first point where frequency is equal to the rank (e.g. the 100th item has a frequency of 100). The idea is that all elements above this point (i.e. the most frequent) are grammatical elements, and all elements under this point are considered as lexical elements.

While measures such as lexical density focus on lexical elements, grammatical elements are also relevant in stylistic analysis. For instance, the use of pronouns might indicate stylistic preferences, such as using the pronoun *I* versus *we* (e.g. Savoy 2017 on the stylistic analysis of US presidential primaries). These stylistic analyses and related measures thus illustrate that the distinction between lexical and grammatical elements is also relevant to fields beyond grammaticalization studies, but also that gradualness might not necessarily be a central concern in these contexts. In this dissertation, a mixed approach is proposed, since it relies on using clear-cut categories as starting point (i.e. lexical vs. grammatical), which however results in a gradual classification (i.e. a score between 0 and 1).

Another question also central to grammaticalization research is whether some features of grammaticalization are more prominent than others, which leads to the second major research question of the dissertation:

RQ2: Which are the factors that lie behind different degrees of grammaticalization?

This is related to the question whether some grammaticalization features are specific to certain processes or stages of grammaticalization (Norde 2012). For instance, phonological reduction is not always observed, as there are highly grammatical elements that are not reduced, such as the conjunction *in order to*. This is relevant to the second research question, since features such as phonological attrition are not always observed, they may play a more secondary role in a model that aims at establishing a grammaticalization score. This observation of differences in relevance relates to Lehmann's theory of grammaticalization (2002: 149, 2015: 177) who suggests that there might be a possible hierarchy among his proposed parameters of grammaticalization. Developing a quantitative score based on a group of variables related to grammaticalization can help identify which of these variables are more relevant and how they relate to each other, highlighting which features are more common and shared among all grammatical elements.

As stated above, the main aim is to quantify grammaticalization through the development of empirical, corpus-based measures. These measures relate to previous works that have discussed the features of grammaticalization, such as Lehmann (1982, 2002, 2015) and Hopper (1991). The example of *back* above illustrates one of the possible quantitative ways to measure grammaticalization, namely to look at the diversity of items (i.e. collocate diversity) with which a given element can occur. Items that fulfil an increasing number of grammatical functions are expected to occur in a broader variety of contexts than they did before. Similarly, an item with an increasing number of grammatical functions should also be used more frequently overall (i.e. increase in token frequency) as it has more possible uses than it did before. Other relevant variables comprise colligate diversity (diversity of neighbouring morphosyntactic categories), orthographic length (as an approximation of phonetic length), and textual dispersion (whether an element appears regularly throughout a text or only in specific parts). These quantitative variables can be aggregated together to compute a grammaticalization score, where highly grammatical items get higher value given the characteristics they display (i.e. higher token frequency results in a higher grammaticalization score). The computation of the grammaticalization score is done by means of a binary logistic regression, which is further discussed in chapter 4.

Unlike previous approaches mentioned so far (and in chapters 2 and 3), the aim of the present dissertation is to find general trends of grammaticalization by using corpus data and by using minimal manual intervention. This has the benefit of minimizing personal interpretation and biases, as well as increasing the replicability of the study. Furthermore, in contrast to Petré and Van de Velde (2018) where the methodology is tailor-made for a specific grammaticalization process, the present approach is more readily applicable to a broader range of elements.

Another difference with existing research is that the features of grammaticalization are selected and motivated by previous research in chapters 2 and 3, but their actual links to grammaticalization are obtained on the basis of empirical data in chapters 5 and 6. For instance, while the assumption is that higher degrees of grammaticalization are linked to a higher diversity of collocates, the model features collocate diversity measures without such assumptions regarding the directionality of their relationship to degrees of grammaticalization. Similarly, no assumptions are made regarding the windows that should be used to determine collocate diversity, which is why different types of collocate windows are proposed in the following studies to search for potentially relevant effects.

One point of focus in the following chapters is whether there are differences regarding collocate and colligate diversity to the left versus to the right of lexical and grammatical elements. For instance, it has been suggested that grammaticalized elements have more constraints than lexical items regarding which elements they can appear with (section 2.2.1). To illustrate, an auxiliary verb such as *will* is followed by an infinitive verb in most cases, whereas a main verb such as *sing* can be followed by a broad variety of elements, such as *a song*, *loudly*, *along to the music*, *to you*. However, an open question is whether this idea applies similarly to the left and to the right of a given item. For example, a preposition tends to be the head of the prepositional phrase it forms (e.g. *on the roof*), which means that it will have dependent members of this phrase to its right. In contrast, there tends to be another phrase to its left, possibly a verbal phrase (e.g. *he is sitting on the roof*) or a nominal one (e.g. *the man on the roof*). This illustrates that holding a distinction between left and right can result in different situations when considering collocate and colligate diversity. The implementation of these variables is further discussed in chapter 3, and the models they are used in are presented in chapter 4.

As discussed earlier, what this quantitative approach can also highlight are the relationships between the different variables that are used to measure grammaticalization. For instance, token frequency is known to play a major role in language (e.g. Bybee and Thompson 1997, Bybee and Hopper 2001, Diessel 2007, Diessel and Hilpert 2016), in particular when considering usage-based approaches and grammaticalization (Bybee 2011). Investigating the coefficients in the models proposed in this dissertation is one way to discuss which variables are the most relevant ones and how they relate to each other. Similarly, the correlations between the variables (regardless of their coefficients in the model) can also be investigated from a statistical point of view. These correlations are investigated in chapter 5 from a synchronic point of view, and in chapter 6 from a diachronic one, in order to provide answers to the second research question above.

A large number of grammaticalization studies are based on detailed case studies. In addition, researchers tend to promote very different means of investigation, which makes comparison more difficult across these studies. This is why it has been rather difficult to provide robust support of more general claims regarding grammaticalization. An example of a general claim is that grammaticalization is mainly a unidirectional phenomenon (Hopper and Traugott 2003: 99-139, Börjars and Vincent 2011), where developments follow a trajectory

from less to more grammatical, while the reverse is uncommon (Norde 2009). This leads to the third major research question of this dissertation, namely:

RQ3: Can the proposed quantitative approach be used to test general claims such as the hypothesized unidirectionality of grammaticalization?

The observation that grammaticalization is mostly a unidirectional phenomenon has mostly been made on the basis of numerous case studies instead of a more global approach. Developing a quantitative approach to the question of grammaticalization is one way to address these limitations and to complement general observations made on the basis of case-study focused approaches.

As noted in many works such as Hopper and Traugott (2003: 99-139), Haspelmath (2004), Börjars and Vincent (2011), processes of grammaticalization tend towards increasing degrees of grammaticalization. Such examples involve forms that tend to become phonologically shorter or semantically more abstract. It is also often claimed that once such a change has taken place, it rarely reverts back to its previous state, even if it involves decreases in frequency (Bybee 2011). Since no research has used a quantitatively computed grammaticalization score so far, chapter 6 will try to shed light on this matter by looking at the evolution of such a score over two centuries for twenty different cases of grammaticalization. This will highlight whether degrees exclusively increase and whether some exceptions can be found. This also paves the way for future research that would involve even larger amounts of data or better measurements.

Another aspect of the unidirectionality of grammaticalization is that degrammaticalization (i.e. the process of grammatical categories developing into lexical categories) is also rarely observed (Norde 2009). While this dissertation does not study cases of degrammaticalization per se, the measure proposed in the following chapters could be used to investigate the matter. For example, Norde (2012) shows how Lehmann's (2002) parameters of grammaticalization may be applied to cases of degrammaticalization. While degrammaticalization is not strictly the opposite of grammaticalization (see section 2.5 on the question of degrammaticalization chains), it can be conceived as a gain of autonomy. Such a phenomenon should also translate to modifications regarding the types of collocates and colligate that appear in close proximity. A problematic aspect is token frequency, since degrammaticalization might also involve increases in token frequency, in a similar fashion to grammaticalization, which would signal that degrammaticalization is not the strict opposite of

grammaticalization. Claims that degrammaticalization involve collocate and frequency shifts could therefore also be investigated by an approach similar to what is proposed in this dissertation.

The endeavour proposed in this dissertation also takes inspiration from developments in the field of morphological productivity (Plag 1999, Bauer 2001). The study of morphological productivity is about the extent to which a morphological process is used to create new forms that were not listed in the lexicon. Earlier approaches to the concept of morphological productivity proposed fairly binary distinctions where a word formation process was either considered productive or unproductive (Booij 1977: 5, Zwanenburg 1983: 23). There were however also early approaches that considered a possible gradualness of morphological productivity, such as Aronoff (1976: 36) who suggested that it might be measured by looking at the ratio between the number of words formed by a specific word formation process and the number of potential words that could be created by the process. While this approach was difficult to operationalize since it is practically impossible to determine how many words could be formed by a certain process (and in theory, such possibilities are generally infinite), it provoked discussion that later turned into more concrete measures such as the ones proposed in Baayen and Lieber (1991) and Baayen (1992).

The measure that received most attention was probably "P" or "productivity in the narrow sense, which simply consists in dividing the number of words formed by a specific word formation processes that only appear once in a corpus (i.e. hapax), by the total token frequency of words formed by that word formation process. The reasoning is that words that only appear once in a corpus (i.e. hapaxes) are likely to be neologisms. Therefore, P gives an idea about the proportion of words that are brand new among a population of already existing ones formed by a specific word formation process. Other ways to measure morphological productivity have been proposed (e.g. Fernández-Domínguez et al. 2007) and other studies have also highlighted shortcomings of most of these measures by comparing them and also by proposing improvements (e.g. Evert and Lüdeling 2001, Pustylnikov and Schneider-Wiejowski 2009, Vegnaduzzo 2009).

While there are substantial disagreements in the field of morphological productivity regarding all the proposed measures, empirical studies have provided support to more theoretical dimensions. For example, it has often been argued that suffixes tend to be ordered by processing complexity in a word (Hay 2002, Hay and Plag 2004). Suffixes that are easier to process (i.e. easier to identify and tell apart from the stems/roots they are attached to) tend to be able to occur after another suffix that is less easy to process, but the reverse is rare. Plag

and Baayen (2009) were able to highlight another aspect of this phenomenon, namely that a high rank in the order of suffixes processing complexity correlates with productivity. In short, suffixation processes that are easier to process (or parse) tend to be more productive than those that do not. This concept is intuitive, since it is easy to argue that if a speaker cannot identify a suffix such as *-vore* in a word such as *carnivore*, it is going to be harder to coin new words using this suffix, such as *pizzavore* or *M&M'svore*. Plag and Baayen (2009) have provided empirical support for this hypothesis with the help of a measure of morphological productivity. In the same vein, a grammaticalization measure might lead to similar possibilities, in particular when dealing with the hypothesized unidirectionality of grammaticalization.

Seeing that morphological productivity is also highly focused on possible constraints or restrictions that determine the degree of productivity of a morphological process (Bauer 2001: 126-143), parallels go even further. In a similar fashion, grammaticalization might also have constraints which hinder or favour higher grammaticalization. For instance, could there be factors, such as the existence and prominence of similar words, which can result in different degrees of grammaticalization? Morphological productivity has the notion of token blocking (Aronoff 1976: 43), whereby the creation of new words is less likely to occur if there is an already existing synonymous word. For instance, there does not seem to be a need for a word such as *stealer* when the word *thief* exists. Yet, *stealer* may appear in terms such as *scene-stealer* (a charismatic character that gets all the attention in their scenes), showing that the notion of blocking is not an absolute one.

Similarly, the grammaticalization of lexical items into certain grammatical categories might be constrained by the existence of a similar already existing grammatical element, or by the prominence of the lexical source. Hopper (1991) introduced the notion of divergence (section 2.2.2), where a grammatical element might have its original lexical source still be used (e.g. *keep* as a main verb or as an auxiliary). A potential question is whether these cases of divergence where the original element is highly frequent tend to involve higher or lower degrees of grammaticalization. It is also possible that such cases require more time to reach higher degrees of grammaticalization. Similarly, grammaticalizing elements that have a competitor to their function might also grammaticalize to a lesser extent or take more time to grammaticalize. For example, the French negation marker *pas* (not) is an example of competition between many elements, although modern French only retains *point* as a less frequent competitor (Hopper and Traugott 2003: 32).

Another example of use of a grammaticalization measure is further investigation regarding the length of grammaticalization processes, which is an under-researched area according to Narrog and Heine (2011: 8). Being able to determine when grammaticalization processes reach their peak and to do so on large datasets would help investigate whether some cases are faster or slower, and also what the average duration might be. While these questions are not directly investigated in the subsequent chapters, they show that a quantitative measure of grammaticalization can open several doors for further research.

The following chapters provide answers to the three main research questions presented above and are organized in the following way. Chapter 2 presents the theoretical background of grammaticalization and discusses how this dissertation relates to it. It introduces the definitions of grammaticalization and its main characteristics. A section is dedicated to Lehmann's (1982, 2002, 2015) parameters of grammaticalization, as well as Hopper's (1991) principles. The notion of unidirectionality is also introduced, as it is one of the most relevant hypotheses regarding grammaticalization. Counterexamples to unidirectionality, known as instances of degrammaticalization, are discussed as well. Grammaticalization is also explored from broader cross-linguistic and construction-based perspectives at the end of the chapter.

Chapter 3 is concerned with the actual implementation of the variables used to measure grammaticalization. It presents how each variable is calculated from a more technical point of view, but also how it relates to previous research and to the main concepts introduced in chapter 2. The main variables are token frequency, orthographic length, collocate diversity, colligate diversity and textual dispersion. The diversity measures are also implemented in different ways to account for possible differences regarding left and right diversities. Limitations of these measures are also discussed, as well as possible alternatives and possible developments for future research.

Chapter 4 introduces the methodology used in the subsequent chapters, as well as the main corpora involved in retrieval of the data. A short introduction to statistical regression is offered, with a focus on binary logistic regression. Alternative methods such as mixed effect modelling are also briefly discussed. In addition, an introduction to principal component analysis is also proposed, as this method is rather useful to investigate links between variables. Finally, information regarding the British National Corpus and the Corpus of Historical American English is provided, as these two corpora constitute the main sources of data in the following chapters.

Chapter 5 discusses a synchronic model for the measurement of grammaticalization. A binary logistic regression was trained using 264 lexical and 264 grammatical elements with data taken from the British National Corpus. The chapter discusses how the list was established and how different implementations of the model change the results. The binary logistic regression attributes values ranging from 0 (highly lexical) to 1 (highly grammatical), and the quality of this classification process is evaluated. The correlations between the variables, as well as their prominence in the binary logistic regression model are also discussed. This chapter therefore contains the core of the quantitative approach to grammaticalization proposed in this dissertation and mainly deals with the first two research questions introduced above.

Chapter 6 presents a study that involves a diachronic approach to the measurement of grammaticalization and relies on twenty elements that have both a lexical and a grammatical counterpart in English, such as the verb *keep* which can be a main verb (e.g. *I will keep your jacket*) or have an auxiliary function (e.g. *You keep saying the wrong thing*). The chapter discusses data taken from the Corpus of Historical American English (1800-2009), and in particular how the variables from chapter 3 change over time for this set of twenty elements. Furthermore, the model presented in chapter 5 is also applied to this diachronic dataset, in order to see the evolution of the grammaticalization score over time. This chapter is therefore concerned with an empirical investigation of the unidirectionality of grammaticalization and deals with the third major research question presented above. Chapter 7 concludes by reviewing the main findings and by discussing ideas for future research.

2. Theoretical background

This chapter introduces the main theoretical aspects of grammaticalization. First, section 2.1 introduces the definition of grammaticalization used in this dissertation and discusses how it relates to main definitions in the literature. Section 2.2 presents research that is aimed at establishing the features of grammaticalization, and more specifically how they can be used to determine degrees of grammaticalization. These features will then be related to empirical, corpus-based parameters in chapter 3. These corpus-based parameters constitute the core of the empirical investigations presented in chapters 5 and 6, from synchronic and diachronic perspectives respectively. Section 2.3 discusses the cline-based view of grammaticalization, where grammaticalization processes are considered as gradual phenomena. This cline-based view also implies a directionality, as it is commonly observed that grammaticalization tends to involve a development from less to more grammatical. This is known as the unidirectionality hypothesis and is further discussed in section 2.4. However, there are also exceptions as there are cases that involve developments from more to less grammatical. These cases are known as degrammaticalization and are discussed in section 2.5. Since a large amount of research on grammaticalization is conducted on English, it may be of question whether observations regarding grammaticalization hold true in different languages. Section 2.6 briefly discusses this question by showing examples of similar grammaticalization processes across languages, but also by highlighting that there are languages where grammaticalization features that are common in English are not as prominent. Different linguistic frameworks can also lead to different takes on grammaticalization. This is why section 2.7 discusses the construction-based approach to grammaticalization and shows how several aspects of this view can complement more classical approaches. Finally, section 2.8 provides a brief review of the key concepts and issues introduced in this chapter and discusses how the studies presented in the subsequent chapters will contribute to these questions.

2.1 Defining grammaticalization

The definition used in this dissertation is based on Hopper and Traugott (2003: 18), where grammaticalization is defined as "the change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions and, once grammaticalized, continue to develop new grammatical functions." The appeal of this definition is that it is based on the commonly accepted notion that a distinction can be made between contentful elements words (i.e. lexical items) and functional elements (i.e. grammatical items). On the

one hand, lexical items mostly denote objects/entities (nouns), actions (verbs) and qualities (adjectives, adverbs). Grammaticalized items, on the other hand, mainly serve to indicate relationships and links between the elements of an utterance or a text. They include function words (e.g. prepositions, verbal auxiliaries, pronouns, conjunctions), clitics (e.g. *'m, 're, 'll*) and inflectional affixes (e.g. *-ed, -ing, -est*). The fact that lexical items can develop into grammatical items raises several points that are discussed in the subsequent sections. First, it highlights the fact that these categories do not have clear-cut boundaries and are therefore continuous rather than discrete (section 2.3). Second, it implies that this process is unidirectional (section 2.4), which is however challenged by the fact that a reverse process of degrammaticalization can sometimes be observed (section 2.5).

There is currently no consensus regarding the definition of grammaticalization. There are several explanations for this situation. Some researchers have pointed out that grammaticalization might in fact not be a distinct process in its own right, but rather a combination of independent linguistic processes (e.g. Newmeyer 1998, Campbell and Janda 2001, Joseph 2011). The term "epiphenomenon" is often used to describe grammaticalization in this context. Furthermore, grammaticalization has also been described as a "composite change" (Norde 2009, Norde and Beijering 2014) which consists of a cluster of small-scale processes, as opposed to one unified process.

There are other views that acknowledge grammaticalization as involving interlinked processes, but which also focus on the fact that their manifestations are language-dependent (section 2.6). In particular, grammaticalization works within the constraints set by the structure of the language in question (e.g. Fischer 2007, Bisang 2011). Therefore, it is advised that grammaticalization theory should also take into account existing systems and how they influence ongoing change (Breban 2009, Reinöhl and Himmelmann 2017). The existing structural features of a language make additional grammaticalization processes more or less likely. For instance, a language that already has modal auxiliaries is likely to grammaticalize new modals. Note that the notion of "existing systems" is not meant in a static way and therefore extends to ongoing and parallel changes. Thus, this type of approach is in line with an idea presented in section 2.7, namely that the development of a construction can be viewed as happening within a broader construction network.

In addition, the concept of grammaticalization has also undergone what has been called an "inflation" (von Mengden and Horst 2014: 353-355). The growing popularity of grammaticalization studies has led to more potential cases of grammaticalization being

investigated. As a result, the concept has become broader and tends to include a wider variety of items being added to what may count as grammaticalization. The fact that the number of researchers working on grammaticalization has increased is also a reason for this broadening, since most tend to focus on aspects that are relevant to their specific research questions. Von Mengden and Horst (2014: 355) advocate to keep the notion of grammaticalization "sufficiently sharp". In contrast, Wiemer and Bisang (2004) argue in favour of viewing grammaticalization from a broad perspective. One of their arguments is that certain main features of grammaticalization (section 2.2) are not necessarily prominent in languages that have been less extensively studied. For example, in East and mainland Southeast Asian languages (e.g. Bisang 2004), pragmatics plays a much greater role in the encoding of grammatical information, than more commonly studied criteria such as paradigmatization, or semantic bleaching (section 2.2.1).

Despite these divergences, at least two early definitions are highly present in the literature on grammaticalization. Both happen to be quite consistent and compatible with Hopper and Traugott (2003). The first comes from Meillet, who coined the term "grammaticalization" and who defined it as the "attribution of a grammatical characteristic to a word that was formerly autonomous"[1] (Meillet 1912: 131). Autonomy is a crucial feature of the parameters of grammaticalization proposed by Lehmann (1982, 2002, 2015) and is discussed in section 2.2.1. An example of an autonomy parameter is "syntagmatic variability", which is the ability for a sign to shift around its context and to appear in different positions in a sentence. According to Lehmann, a more autonomous sign tends to shift more easily, while a less autonomous one tends to be strongly constrained regarding the positions where it can appear. This is illustrated for instance by the fact that grammatical elements such as clitics and inflectional affixes are by definition bound morphemes and are therefore not autonomous.

The second prominent definition is from Kuryłowicz (1965: 69) who claims that "grammaticalization consists in the increase of the range of a morpheme advancing from a lexical to a grammatical or from less grammatical to a more grammaticalized status". This definition is rather close to that of Hopper and Traugott (2003) and also highlights the gradual aspect of grammaticalization. However, Hopper and Traugott's definition also includes constructions (see section 2.7) and is not restricted to morphemes, which makes the definition more encompassing. Furthermore, Kuryłowicz uses the concept of "range", which is related

[1] My translation from the French original "Attribution du caractère grammatical à un mot jadis autonome."

to Himmelmann's (2004) concept of host-class expansion. This increase in range can be understood as grammaticalizing elements increasing the range of elements they can appear with (i.e. an increase in the range of their collocations, see section 3.3).

A point of criticism of Kuryłowicz (1965) which also applies to Hopper and Traugott (2003) is that in both cases, the definition brings two separate phenomena under the umbrella of grammaticalization. On the one hand, the transition from lexical to grammatical, and on the other hand, the increasing grammaticalization of elements that are already grammaticalized. The term "primary grammaticalization" is often used to describe the former, while the term "secondary grammaticalization" is used to refer to the latter (e.g. Traugott 2002: 26-27, Norde 2012: 76). An example of primary grammaticalization in English is the development from the Old English main verb *willan* (to want/wish) into the future tense auxiliary *will* in Modern English. An example of secondary grammaticalization is the further development of *will* which cliticized into *'ll* in Modern English. Some argue against holding a distinction between primary and secondary grammaticalization (Breban 2014) and it has also been argued that this distinction requires some caution (von Mengden and Horst 2014) because it has been taken for granted so far in grammaticalization research. The broader question underlying this concern is whether there are features specific to early, as opposed to late, stages of grammaticalization. This question however is not easily answered at this point because the current understanding of the relevant features of grammaticalization is limited. For example, Lehmann (1982, 2002, 2015) proposes a parameter of grammaticalization called integrity (section 2.2.1), which comprises phonological attrition, as it is commonly observed that items undergoing grammaticalization tend to become phonologically shorter. A relevant observation is that not all parameters are always at play in specific cases of grammaticalization, especially when dealing with primary versus secondary grammaticalization. For instance, Norde (2012) notes that phonological attrition is rarely attested in cases of primary grammaticalization, using the development of Swedish *mot* from noun to preposition as an example where this feature is not observed. On the other hand, the development from the proto-Scandinavian demonstrative pronoun *(h)inn* (this) into the Modern Swedish suffix *-en* (the) only displayed phonological attrition in its late (i.e. secondary) grammaticalization stages. These examples illustrate that there might indeed be differences between primary and secondary grammaticalization when it comes to the grammaticalization features that are most prominent during these stages. However, the situation is more complex as there are several other parameters at play. The next section introduces these parameters and also discusses some of their limitations.

2.2 Grammaticalization features

While to my knowledge no previous attempts have been made to develop a grammaticalization measure per se, there have been attempts to identify aspects of grammaticalization that reflect the extent to which a form is grammaticalized. Lehmann (1982, 2002, 2015) proposes six parameters that are presented in section 2.2.1, while Hopper (1991) focuses on five principles of grammaticalization that are introduced in section 2.2.2. There is a certain degree of overlap between these two approaches. Furthermore, concrete corpus-based measures that relate to these parameters are introduced in chapter 3.

2.2.1 Lehmann's parameters

According to Lehmann, the main way to gauge grammaticalization is to look at the autonomy of a sign, since the "autonomy of a sign is converse to its grammaticality" (Lehmann 2002: 109, Lehmann 2015: 130).[2] This approach is therefore strongly rooted in the definition of grammaticalization proposed by Meillet (1912), where autonomy is a key element. Lehmann offers parameters of grammaticalization, which consist of three factors that can be used to determine the autonomy of a sign. Each factor can be viewed from a paradigmatic or syntagmatic point of view, thus resulting in six parameters that are summarized in Table 1. They can correlate positively or negatively with grammaticalization. The three factors are weight (the property of a sign that makes it distinct from the other members of its class), cohesion (the relations between a sign and the other signs), and variability (the possibility for a sign to change positions with other signs).

	Paradigmatic	**Syntagmatic**
Weight	Integrity	Structural scope
Cohesion	Paradigmaticity	Bondedness
Variability	Paradigmatic variability	Syntagmatic variability

Table 1. Lehmann's parameters of grammaticalization (Lehmann 2002: 110, 2015: 132).

The paradigmatic view of weight is called "integrity" and corresponds to the substantial size of a sign, both phonologically and semantically. The phonological integrity of a word consists of its phonemic or syllabic length, whereas its semantic integrity consists of its number of semantic features. Reductions in integrity (phonological attrition and

[2] Note that Lehmann uses the term *grammaticality* when discussing the degree of grammaticalization of an element (see Lehmann 2002: 8 for more details).

desemanticization) correlate with higher grammaticalization. The reduction in semantic weight of a sign is often a case of semantic generalization (Hopper and Traugott 2003: 100-103). Desemanticization is also commonly called semantic bleaching (Sweetser 1988). An example of both phonological and semantic reduction is the future tense construction *be going to* + infinitive. It is often shortened to *gonna* in spoken language, and can be used to describe any future action, which is an evolution from its original specific use as a motion verb (i.e. *I'm going to do my homework* vs. *I'm going to the train station*). Desemanticization and phonetic erosion are also two major aspects of grammaticalization discussed in Heine and Kuteva (2002: 2).

The syntagmatic view of weight is called "structural scope" and refers to the extent of the construction in which a sign enters or is part of. A decrease in structural scope means higher grammaticalization. For instance, auxiliary verbs such as *be* or *have* have a limited structural scope, because they work at the level of the VP they are part of, but not beyond that. In contrast, main verbs function at the level of the clause. Affixes also have a greatly reduced scope as their scope only consists of the element that they are attached to. The view that increasing grammaticalization leads to scope reduction differs from the earlier definition by Kuryłowicz (1965: 69) who posits that grammaticalization entails an increase in the range of a morpheme. It is therefore subject to debate whether grammaticalization entails expansion or reduction in scope (e.g. Tabor and Traugott 1998). In section 3.3, the adopted view is that grammatical elements involve a larger diversity of collocates, which would be in line with scope expansion instead of reduction.

"Paradigmaticity" is the paradigmatic view of the cohesion factor and refers to the fact that a sign can be part of a paradigm or independent from it. The size of the paradigm matters as well, and increased grammaticalization tends to lead to smaller, more tightly integrated paradigms. Grammaticalization also leads to closed-class paradigms, which is reflected in the fact that content word classes such as verbs or nouns tend to be open-ended, as opposed to function words such as articles and pronouns, which are closed classes in many of the world's languages. In addition to the size of the paradigm, its homogeneity might also be a factor, as highly grammaticalized concepts tend to belong to more homogenous paradigms. For instance, interrogative pronouns in English tend to be *wh-* words (with the notable exception of *how*).

The syntagmatic view of cohesion is called "bondedness" and corresponds to the degree to which a sign depends on, or attaches itself to, other signs. Increase in bondedness means less autonomy and thus means a more grammaticalized sign. This is consistent with the cline

presented in section 2.3, where clitics and affixes are at the highly grammaticalized end. Furthermore, bondedness between elements can also be conceived as the possibility to insert something in between. For instance, it is rare to find insertions between a pronoun and a clitic such as *'ll*. To illustrate this point, there are very few occurrences of sentences such as *I never'll wear it again* in the Corpus of Historical American English, and almost none at all in other copora such as the British National Corpus or the Corpus of Contemporary American English. Similarly, one can argue that the bondedness between *going* and *to* in the *be going to* + infinitive future construction is strong as there are rarely any insertions between them (e.g. *I am never going to talk to you again* vs. * *I am going never to talk to you again*). This means that bondedness can also be viewed as a parameter strongly related to phonological reduction, as illustrated by the existence of the reduced form *gonna*.

"Paradigmatic variability" is the freedom with which a language user can choose a sign from a range of options. A reduction in paradigmatic variability corresponds to a higher degree of grammaticalization. A common example of a reduction in paradigmatic variability is the grammaticalization of the negative marker *pas* in French which has suppressed all its competitors, with the sole exception of *point* which is still currently used in Modern French, but much more rarerly. Obligatoriness is the converse equivalent to paradigmatic variability and is an extreme situation where there is no freedom to choose. Lehmann (2002: 125) gives the example of the development of articles in Romance languages. He notes that in most Romance languages, it is in most contexts impossible to use a singular noun without specifying its definiteness and using an article, whereas Latin involved more variability in that regard. This situation is also similar in Germanic languages, including English. An example of context where such articles are not obligatory would be announcements (e.g. *Hat for sale*) or news headlines (e.g. *Tiger found in restaurant*). However English also does not require articles with certain concepts that are also singular nouns (e.g. *Grammaticalization is a complex topic*). This illustrates that paradigmatic variability can also be thought in terms of obligatoriness instead of freedom to choose. This parameter is rather problematic because it depends on context and cannot be isolated from the other factors of grammaticalization (Lehmann 2002: 123-127) and it is currently unclear how to quantify it. Also, such freedom of choice is highly difficult to observe, because there is no obvious way of finding natural data regarding what speakers might have said instead of what they actually said. The same comment applies to the notion of obligatoriness.

Finally, "syntagmatic variability" is the ease with which a sign can be shifted around its context. It decreases with increasing grammaticalization. For example, an adverb like

yesterday has some degree of syntagmatic variability, since its position in a given sentence is not necessarily fixed (e.g. *Yesterday, I saw a dog* or *I saw a dog yesterday* are both valid sentences). Lehmann (2002) mentions the development from adverbs to adpositions as an illustration of reduction in syntagmatic variability due to grammaticalization. Lehmann (2002: 142) also notes that "while the syntagmatic variability of the grammaticalized item decreases, its bond with a particular class of words which it comes to modify grammatically becomes tighter. That is, whenever such positional adjustments occur, they will produce an order in which the grammaticalized item is adjacent to its lexical support." This relates the parameter of syntactic variability to bondedness in an important way. Earlier, the example of the bond between *going* and *to* in the *be going to* + infinitive construction was discussed, but of course the position of *to* relative to *going* is also a matter of syntagmatic variability. However, it is not straightforward that grammaticalizing elements involve a decrease in syntagmatic variability. For instance, Heine and Kuteva (2002: 2) suggest that grammaticalization involves extension (or context generalization), where grammaticalized elements might be used in new contexts (see also Heine 2002, Heine 2003). Grammatical elements might therefore appear in a broader variety of contexts, which suggests that they might also be more easily shifted around. This idea is further explored in section 3.3, where collocate diversity is proposed as a way to assess syntagmatic variability, as grammaticalized items which are shifted more flexibly should also display a higher diversity of collocates as a result.

These six factors are interconnected and the main idea that links them is that of autonomy. There are however three important elements to take into account when discussing those parameters. First, when grammaticalization takes place, it does not mean that this will be reflected in the six parameters. For example, there are cases of grammaticalization that do not involve any phonological attrition. An illustration would be the verb *keep* which currently only has one pronunciation, whether it is used as a main verb (e.g. *I'll keep this hat*) or as an auxiliary verb denoting ongoing action (e.g. *I keep falling into the same trap*). This was also pointed out earlier in the discussion of Norde (2012) where cases of primary grammaticalization tend to feature phonological attrition less prominently.

Second, the correlation between the parameters is not necessarily obvious and Lehmann (2002: 152) even states that there is "at present no explanation for a lack of correlation among the grammaticalization parameters." He illustrates this with the example of *to* plus infinitive in English and the fact that according to most parameters, *to* is highly grammaticalized, but not when it comes to the parameter of bondedness. The construction of *to* plus infinitive is considered loosely bonded by Lehmann because in a sentence fragment such as *to describe*

and explain, there is no need to use *to* two times, unlike German for instance, which would require to do so (*zu beschreiben und zu erklären*). The idea proposed earlier of inserting elements is also possible (e.g. *to clearly describe and explain*), and further testimony to the lesser degree of bondedness of the English *to* plus infinitive construction. However, there are cases where the correlation between parameters seems straightforward. Correlations were discussed earlier between reductions in syntagmatic variability corresponding to increasing bondedness (i.e. they both deal with the relative positions of elements, especially within a construction) and also between bondedness and phonological reduction (e.g. strong bondedness between *going* and *to* and the existence of the reduced form *gonna*). A further instance of correlation between the parameters is the hypothesis that phonological reduction and semantic reduction occur in parallel (Bybee et al. 1994), which has been substantiated by empirical data (e.g. Rhee 2003).

Third, some changes in those parameters can be observed that are not due to grammaticalization, but to other processes instead. For example, phonological attrition (decrease in integrity) may happen in other areas of language change. Lehmann (2002: 113) gives the example of the reduction of Latin *aqua* (water) to French *eau* (pronounced /*o*/) as an illustration which involves lexical items that clearly do not show any signs of grammaticalization. Another illustration would be phonological shortenings in slang words such as *narc* (*narcotics*) which are unrelated to grammaticalization. Furthermore, it seems possible for grammaticalization and other phenomena to have a joint influence on cases of phonological attrition. As such, the reduction of *because* as *coz* or *cuz* might be an effect of grammaticalization as the term does have a common grammatical function, namely expressing causality. However, there is also a notion of register at play, since *coz* or *cuz* mostly occur in informal settings. Therefore, a speaker might use *coz* or *cuz* in order to fit better into a certain social setting and thus contribute to the emergence of those shortened versions in a way that does not directly stem from grammatical function.

In addition to those shortcomings, it is unclear whether some parameters are more relevant (or prototypical) than others. It may be observed that the parameter of integrity has received more attention (see von Mengden and Horst 2014: 349-350) than the others. One explanation is that it is a parameter that is directly observable and rather concrete when it comes to its operationalization (phonological attrition in particular). Other parameters such as paradigmaticity are not easy to deal with because measuring the size of a paradigm is not necessarily obvious. Another criticism of the parameter of integrity is that it encompasses two

different processes that might be independent. This is a valid concern, and thus phonological and semantic weight will be treated as two separate variables in the subsequent chapters.

What Lehmann's parameters show is that grammaticalization is a complex process that cannot be pinned down to one single parameter. Chapter 3 deals with the operationalization of those parameters and their quantification. Because of their complexity, it seems that a quantitative grammaticalization measure should in fact be computed on the basis of several other measures. This is why the term "score" will be preferred when discussing the grammaticalization measure introduced in chapter 5. The term "index" is also a viable alternative. Chapters 5 and 6 will shed light on the relationships between some of these parameters, and also how they relate to token frequency. Some of those parameters are easily quantifiable, while some seem more problematic.

2.2.2 Hopper's principles

Hopper (1991) offers five principles of grammaticalization that complement Lehmann's parameters. In a similar fashion to Lehmann's parameters, the principles can also be observed in relation to other phenomena than grammaticalization. The first parameter is called "layering" and states that new layers are continually emerging within a functional domain (e.g. tenses) and that the old layers may not be discarded but keep co-existing. An illustration would be the English past tense, which retains vowel alteration in some cases (e.g. *drive/drove*), as opposed to the new layer of *-ed* past tense. The second parameter, "divergence" is presented as a special instance of layering, and refers to the fact that a lexical item can grammaticalize into a grammatical item, a clitic or an affix, while its original form still remains as an autonomous element that can undergo the same changes that other lexical items do. For example, while French *pas* has grammaticalized into a negation marker, it is still used as a lexical item (*step*) as well. The earlier example of *keep* as a main verb or as an auxiliary would also be an illustration of divergence.

The third parameter, "specialization", refers to the narrowing of choices that characterize an emergent grammatical construction. When a form is in the late stages of grammaticalization, it may eventually become obligatory, as illustrated again by French *pas*, which has become almost the only possible negation marker. This closely parallels Lehmann's parameter of paradigmatic variability (the freedom to choose a sign instead of another) and therefore shares the same operational difficulties; there is no obvious way of finding out whether speakers have used a certain sign instead of another one.

The fourth parameter is "persistence" and accounts for the fact that a grammaticalized item may retain some meaning from its lexical stage. This may be opaque, especially in later stages of grammaticalization, but it can be observed with historical data. For instance, the future tense auxiliary *will* still retains some connection to its original lexical Old English main verb *willan* with the notions of intention and willingness. An utterance such as *I will help you* still conveys these more lexical notions, as opposed to other examples such as *We will probably die tomorrow*. This principle seems to be closely related to Lehmann's parameter of integrity, and more specifically to desemanticization. Persistence can be viewed as an additional dimension to the process of desemanticization, namely that while a sign may become more general and lose some of its semantic features, it may retain some that are still connected to its original meaning.

Finally, the fifth principle, "de-categorialization", is the idea that an item undergoing grammaticalization will lose the morphological markers and syntactic properties of full categories (noun and verbs) and show attributes of secondary categories. This principle therefore supports taking morphosyntactic categories into account as an aspect of grammaticalization. An example of grammaticalizing items losing the morphological markers of their main category are modal verbs, since they gain the morphological markers of auxiliaries, such as cliticized negation (e.g. *shouldn't, wouldn't*). This principle is also mentioned in Heine and Kuteva (2002: 2).

The next section is concerned with the fact that gaining these new functions entails a gradual development from the more lexical categories to the more grammatical ones. This is known as the cline-based view of grammaticalization.

2.3 The cline-based view

Grammaticalization is commonly viewed as a gradual phenomenon and can be thought of as a cline. The cline-based view of grammaticalization is described by Hopper and Traugott (2003: 7) as follows:

content item > grammatical word > clitic > inflectional affix

On the left end of the cline are content (lexical) items, which mostly denote objects/entities (nouns), actions (verbs) and qualities (adjectives, adverbs). Grammaticalized items, on the other hand, are at the other end of the cline and include function words, clitics and inflectional affixes. In this dissertation, the focus is mostly on content words and grammatical words

(primary grammaticalization), since they are more similar to each other and therefore more easily comparable. Clitics and inflectional affixes are bound morphemes and belong therefore to a category that is quite different from content and grammatical words. Constraining the study this way makes it possible to address the earlier criticism that primary grammaticalization might be a different phenomenon from secondary grammaticalization.

Note that these changes are considered as gradual in the sense that they involve small incremental steps where developing items gain new functions and start displaying the characteristics of grammaticalization discussed in section 2.2. For instance, semantic bleaching does not happen in a single discrete step, but rather through repeated new uses of a grammaticalizing item. However, Traugott and Trousdale (2010: 20) maintain that the steps themselves, when considered individually, can be viewed as a type of discrete (or "abrupt") change. Reanalysis is often mentioned as one of the mechanisms by which grammaticalization is made possible. Reanalysis is the fact that speakers can interpret the structure and/or the meaning (including the syntactic category) of a given item differently from its original ones.

For instance, the auxiliary use of the verb *need* is often described as a marginal modal (Palmer 2013: 3, Krug 2000: 199-203), which means that it is a particular case in Present-Day English. For example, it can be used to express lack of obligation such as *he need not go there*. It can also be cliticized as in *he needn't go there*. Therefore, speakers could analyse this verb as a regular modal and use it accordingly by producing an utterance such as *he need go there*. Regardless whether or not such a novel utterance might be considered ungrammatical from a prescriptive point of view, it illustrates the idea of an abrupt change because the verb is used in an unexpected or unusual way (see also the notion of extravagance in section 2.4). However, the gradual aspect comes when considering auxiliary uses of *need* in general, and thus whether it can be used similarly in other utterances. Forming questions is also a current possible use of *need*, with questions such as *Need we worry about this*. One could view this situation as *need* having gained new grammatical function and additional modal-like features. The major point is that each separate novel use is an abrupt change, while the overall change is a gradual one. In addition, the spread of this type of change is also gradual as individuals are unlikely to adopt the novel use simultaneously.

The grammaticalization cline can be understood synchronically and diachronically. A synchronic interpretation of the cline means that, at a given point, some elements are more grammaticalized than others. For instance, the grammatical element *the* can be considered as more grammaticalized than the lexical element *cat*. The word *cat* refers to a concrete entity in

the real world, whereas *the* is a determiner and therefore fulfils a grammatical function, namely to indicate that the noun that it precedes is definite. These two examples are obvious cases because they belong to different morphosyntactic categories, but since grammaticalization is conceived as a cline, this also means that the elements within the same category can have different degrees of grammaticalization. Thus, some grammatical elements might be regarded as more prototypical examples of their category (Taylor 2003: 215-220). In addition, it signals that the category boundaries are fuzzy, because the grammatical elements that are less grammaticalized will be closer to some of the lexical elements in terms of degree of grammaticalization. The notion of degree of grammaticalization in synchrony is referred to as "gradience" in Traugott and Trousdale (2010). Chapter 5 will try to highlight empirical, quantitative ways of measuring gradience and give empirical support to the idea that *the* is more grammaticalized than *cat*.

A diachronic interpretation of the cline is that a content item can develop, over time, into a grammatical word, which can then further develop into a clitic or an inflectional affix. The diachronic perspective is probably the one that has received the most attention in linguistic research. As noted above, this type of change happens in a succession of small, incremental steps. This phenomenon is called "gradualness" in Traugott and Trousdale (2010). Chapter 6 will use the measure developed in chapter 5 in a number of diachronic case studies in order to try to capture those small incremental steps.

A relevant question is to what extent the gradience and gradualness of grammaticalization can be investigated by similar means. As shown above, they are different phenomena, where gradience is mainly concerned with grammaticalization as an outcome and gradualness is concerned with grammaticalization as a process. While these two phenomena are strongly linked, it remains to be determined whether they can be investigated using the same methods. For instance, persistence has been discussed previously (section 2.2.2) as one of the characteristics of grammaticalization, where grammaticalized elements tend to retain some of the semantic features of their lexical sources. However, there are cases where these features are lost, sometimes due to the emergence of a new competing element from the same source (Dahl 2000: 10-11). In this situation, the outcome of grammaticalization has different features from the ones that were involved in its development. Therefore, the distinction between gradience and gradualness is relevant when considering the results of the synchronic approach proposed in chapter 5, and those of the diachronic approach proposed in chapter 6.

2.4 The unidirectionality hypothesis

The diachronic interpretation of the grammaticalization cline implies a direction, as illustrated by the > symbols between points on the cline proposed by Hopper and Traugott (2003: 7), which is known as the unidirectionality hypothesis. Grammaticalization is often described as a unidirectional process. Elements tend to develop from less grammaticalized to more grammaticalized, and the converse is rarely observed (Hopper and Traugott 2003: 99-139, Haspelmath 2004, Börjars and Vincent 2011). Grammaticalization is known to involve phonological reduction (sections 2.2.1 and 3.2) and semantic bleaching (sections 2.2.1 and 3.3), which are processes that tend to be unidirectional as well. The present section discusses possible explanations in favour of the hypothesis that grammaticalization is primarily a unidirectional phenomenon.

Two decades ago, Haspelmath (1999: 1049) noted that very few explanations for this hypothesized unidirectionality had been put forth. His proposal draws on Keller's application of the invisible hand theory to linguistics (Keller 1994), which posits that "language change is shown to result from the cumulation of countless individual actions of speakers, which are not intended to change language, but whose side effect is change in a particular direction" (Haspelmath 1999: 1043). According to Haspelmath (1999), the key concept that explains the unidirectionality of grammaticalization is the maxim of "extravagance". The idea is that speakers want to stand out and be noticed, and that in order to do so, they need to innovate. This leads individual speakers to come up with novel ways of expressing common grammatical concepts.

An example given by Haspelmath (1999: 1057) is using *by means of* instead of *with*. In this framework, being innovative and noticed is equated with being socially successful. An important note is that this notion of extravagance is not necessarily specific to grammaticalization (and neither is the rest of Keller's (1994) invisible hand theory). Extravagance itself is not sufficient to lead to change, and extravagant forms must be adopted by other speakers and spread. This is why frequency of use is also a required criterion for the raise of new grammatical forms. More recent work by Bybee (2011: 77) views increasing frequency of use as the central explanation for the unidirectionality of grammaticalization. Increases in the usage frequency of a linguistic element tends to be linked to phonological reduction (see sections 3.1 and 3.2 on frequency and phonological reduction). According to Bybee (2011), when frequency of use ceases to increase (or even decreases), the changes that have taken place do not revert back to their former states.

Several scholars have also used Chomsky's (1995) Minimalist Program in order to explain the unidirectionality of grammaticalization. Roberts and Roussou (2003: 202-218) claim that grammaticalization involves the "upwards" reanalysis of lexical material in a hierarchically organized syntactic structure. This means that the type of reanalysis involved in grammaticalization concerns items that move upwards in syntactic trees. This direction is explained by the need for structural simplification and the avoidance of syncretism, which is related to language acquisition and the learning of grammar by new generations (Roberts and Roussou 2003: 202-205). Newer generations do not necessarily have access to the former lexical interpretation of an item that has grammaticalized and viewing it as exclusively belonging to the higher categories (in the sense of elements higher in the syntactic structure) is a simpler interpretation. The idea of syntactic direction is also explored in van Gelderen (2004: 17-37) who offers "economy principles" that describe syntactic change in a similar fashion, where lexical items tend to move up the syntactic tree by becoming syntactic heads. Further discussion and limitations of these generative approaches can be found in Börjars and Vincent (2011: 174-175).

Jäger and Rosenbach (2008) have proposed another explanation for the unidirectionality of grammaticalization which relies on the cognitive mechanism of asymmetric priming. Asymmetric priming is the phenomenon by which one idea strongly primes (i.e. evokes) another idea, but not vice-versa. For instance, the word *hunter* strongly primes the word *rifle*, but conversely *rifle* does not necessarily strongly prime *hunter*. According to Jäger and Rosenbach (2008), less grammaticalized items tend to prime the more grammaticalized ones, but not vice-versa, which explains why developments tend to go from lexical to grammatical. However, Jäger and Rosenbach (2008) is a programmatic paper and empirical investigation of the phenomenon has highlighted unexpected results.

Hilpert and Correia Saavedra (2018) compared the priming of lexical and grammatical counterparts of the same words. For instance, the word *back* can refer to a part of the body (i.e. it's lexical counterpart) or be used in a more abstract way to mark opposite directionality when used with verbs such as *come back* (i.e. it's grammatical counterpart). Hilpert and Correia Saavedra (2018) investigated whether such lexical uses primed the grammatical ones within the same sentence (e.g. *Yesterday my back was hurting when I came back home*), and vice-versa (*He was turning back and fell flat on his back*), using a maze-task design (Forster et al. 2009). Their results show that lexical elements do not prime their grammatical counterparts as one would expect given Jäger and Rosenbach's (2008) predictions. In fact, the opposite effect is observed, namely that lexical elements hinder the processing of their

grammatical counterparts (i.e. negative priming). This might be attributed to a phenomenon known as "horror aequi" (Rohdenburg and Mondorf 2003), which refers to the fact that speakers tend to dislike the use of the same form within a short time, especially when it has different meanings. Despite these findings, more empirical work needs to be done to support those results, and asymmetric priming remains a possible explanation for the unidirectionality of grammaticalization. Indeed, Hilpert and Correia Saavedra (2018) relied on comprehension only and did not involve production, which could show different tendencies. Further discussion of the asymmetric priming hypothesis can be found in Baumann and Sommerer (2018), who developed a mathematical population-dynamic model to investigate links between asymmetric priming and grammatical change.

In sum, there are various different explanations regarding the reasons why grammatical changes tend to be unidirectional.[3] It seems plausible that several of these explanations can be correct at the same time, as there is no reason to believe that unidirectionality has to be accounted for by only one. Furthermore, most of these explanations acknowledge the central role of speakers in the processes of grammatical change, at the level of the individuals (e.g. extravagance, priming mechanisms), or at a more collective level (e.g. frequency of use, generational learning). While the strong tendency for grammatical change to be unidirectional is rather commonly accepted, there is more divergence regarding the existence of change going in the opposite direction, namely the phenomenon known as degrammaticalization.

There are researchers who are critical of the hypothesized unidirectionality of grammaticalization (e.g. Campbell and Janda 2001, Campbell 2001b, Janda 2001, Joseph 2001, Joseph 2011). One of their criticisms is that unidirectionality is "built into the definition of grammaticalization" (Campbell 2001: 124), since it is defined as a development from lexical to grammatical (or from grammatical to more grammatical). A consequence is that any observation that does not fit within this definition can be considered as falling outside the scope of grammaticalization. A grammatical element developing into a lexical element would be viewed as an instance of lexicalization and thus be dismissed as not being a counterexample to the unidirectionality hypothesis. According to this view, the problem is that unidirectionality is "not an empirical hypothesis that can be tested, but an artifact of the definition itself" (Campbell 2001: 124). While this criticism is certainly valid to some extent, it mostly relies on the idea that definitions of grammaticalization where unidirectionality is "built in" treat it as an absolute property, which is often not the case. It is possible to

[3] A notable different approach to the question is Kiparsky (2012) who proposes a new definition of analogy which includes grammaticalization. This leads to other explanations that account for unidirectionality.

subscribe to the definitions presented in section 2.1, such as Hopper and Traugott's (2001), while at the same time acknowledging the existence of cases of degrammaticalization as presented in section 2.5. The view adopted in this dissertation is that unidirectionality is a robust tendency as there are more concrete examples of developments from less to more grammatical (e.g. Givón 1971, Heine and Kuteva 2002) than the reverse.

Developing a way to quantify grammaticalization is a first step in addressing some of these criticisms, because it can be used to test whether there are tendencies for increases or decreases as will be shown in chapter 6. This would provide additional empirical support for unidirectionality, the lack of which was severely criticized in Campbell (2001). In addition, it also provides a way to analyse larger datasets, which could be used to provide more empirical data regarding the relationship between the different aspects of grammaticalization and how they relate to each other. In chapter 5, the links between the different grammaticalization variables involved in the models will also be discussed, which may also substantiate the claim that grammaticalization is a composite change, instead of an epiphenomenon as discussed in section 2.1.

Another point of criticism is that the claim that cases of grammaticalization are more common than cases of degrammaticalization has not been formally investigated empirically. Researchers assume that lexical to grammatical developments are more common, simply because they are more studied than their counterparts. This is how Janda (2001: 292-304) argues that counterexamples to unidirectionality are in fact not as infrequent as many scholars believe. There is presently no systematic study that tries to count whether there are more grammaticalization versus degrammaticalization cases, which means that these criticisms cannot be entirely dismissed. While this particular issue will not be considered in the subsequent studies, finding ways to quantify grammaticalization from a more computational and systematic way is a first step in doing larger-scale studies to address these issues regarding grammaticalization and degrammaticalization.

2.5 Degrammaticalization

This section provides a more detailed discussion of these cases of degrammaticalization. While some early works refuted the existence of degrammaticalization (e.g. Lehmann 1982, Heine et al. 1991, Haspelmath 1999), it is currently more accepted that it does indeed exist (Ramat 1992, van der Auwera 2002, Hopper and Traugott 2003: 130-139, Haspelmath 2004, Norde 2009). Researchers tend to disagree with regard to their respective definitions of degrammaticalization, which can explain why they also disagree regarding its existence.

Some of these controversies surrounding the notion of degrammaticalization have been discussed in Norde (2010). In earlier work, Norde (2006) defined degrammaticalization on the basis of Hopper and Traugott's (2003: 7) cline, which was introduced in section 2.4 and which is repeated here:

content item > grammatical word > clitic > inflectional affix

The basic definition is that degrammaticalization consists in shifts from the right to the left of the cline. However, Norde (2010) proposes a more specific definition of the phenomenon, where it is described as "a change whereby a gram in a specific context gains in autonomy or substance on one or more linguistic levels (semantics, morphology, syntax, and phonology)" (Norde 2010: 126). In this definition, the term "gram" refers to a grammatical morpheme (e.g. function word, clitic, affix) as introduced in Bybee et al. (1994). This definition also uses the notion of autonomy, as did Meillet's (1912) original definition of grammaticalization discussed in section 2.1.

Another important note is that in the same way that there is "primary" (i.e. lexical to grammatical) grammaticalization and "secondary" (i.e. less to more grammaticalized) grammaticalization, there are also two types of degrammaticalization. The first type is primary degrammaticalization, which involves the shift from a grammatical element to a lexical element, and the second type is secondary degrammaticalization, which consists in a grammaticalized element (in particular a bound morpheme from the right end of the cline) becoming less grammaticalized. Norde (2010: 136) notes however that unlike grammaticalization, where an element might move all the way from content item to an affix, and therefore undergo primary and secondary grammaticalization, cases of degrammaticalization generally do not progress in the opposite direction as part of a chain. A potential counterexample to this claim is the suffix *-taga* in Saami which developed into a clitic, then into a postposition and then into an independent adverb (Janda 2001: 127). It should be noted however that in such an example, the element in question technically first undergoes "secondary" degrammaticalization, and then "primary" degrammaticalization, which shows that the terms primary and secondary are somewhat tricky. This is therefore a characteristic specific to degrammaticalization, where primary and secondary degrammaticalization are rather separate phenomena.

While the basic definition of degrammaticalization according to which items move from right to left on the cline seems rather obvious, there is in fact much disagreement when it

comes to specific examples of degrammaticalization. If one considers the use of a grammatical item as a lexical one as an example of degrammaticalization, then any case of conversion from a grammatical to a lexical category would be considered as an example of degrammaticalization. For instance, the conjunctions *but* and *if* can be used as nouns in a sentence such as *I sense an if is coming*. They also have common nominal characteristics, such as having a plural form (e.g. *there were two buts in your sentence*). According to Haspelmath (1999: 1064, footnote 1) these cases are metalinguistic uses of these function words and are not instances of degrammaticalization.

Norde (2009: 3), building on Andersen (2006), proposes the term "degrammation" to refer to instances of degrammaticalization where "a function word is reanalysed as a content item, often as a result of pragmatic inferencing".[4] However, Norde (2009) also distinguishes between degrammation and lexicalization. When grammatical elements are taken "out of their context" to be used as lexical elements (e.g. *to up* and *to down*), this pertains to lexicalization and not degrammaticalization. It should also be noted that the boundary between lexicalization and grammaticalization is sometimes blurry and that the issue also has to do with how these concepts are defined (Brinton and Traugott 2005).

There are two other types of degrammaticalization that are proposed in Norde (2009) that complement degrammation, since the latter only accounts for primary degrammaticalization. The two other types are called "deinflectionalization" and "debonding". Deinflectionalization concerns inflectional affixes that become less bound and debonding involves bound morphemes (thus including derivational affixes) becoming free. As stated earlier, the research presented in the following chapters is not focused on affixes and the "right end" of the grammaticalization cline and as such, deinflectionalization and debonding will not be further discussed here.

In addition to the examples above, there is also a process known as insubordination (Evans and Wantanabe 2016) which can be thought of as an instance of degrammaticalization. It consists in the reanalysis of a subordinate clause as an independent one. This process differs from the examples presented previously because it does not entail the creation of a new lexical item and is more concerned with syntactic role. For instance, *If you could shut up* can be a sentence of its own, instead of being a subordinate clause in more complex sentences such as *It would be great if you could shut up* or *If you could shut up, I would be able to focus*.

[4] Willis (2007: 273) also proposes pragmatic inferencing as a criterion.

In this situation, there is a clear gain in autonomy (i.e. becoming a sentence of its own), which is in line with the concept of degrammaticalization.

To summarize, section 2.4 has discussed the idea that unidirectionality is a strong tendency of grammaticalization and further empirical investigation of this idea will be presented in chapter 6. However, the current section has also argued in favour of the existence of degrammaticalization, which means that there are exceptions to the notion that grammaticalization is unidirectional. Examples have been discussed to illustrate what is considered as degrammation in the present dissertation. The next section discusses whether grammaticalization is universal among languages and whether it occurs similarly across languages.

2.6 Grammaticalization across languages

The studies presented in chapters 5 and 6 use English datasets and most of the discussion and examples provided in this dissertation are focused on English cases. An important question is the extent of the universality of the grammaticalization features discussed so far, given that languages involve major typological differences. For example, the highly grammaticalized end of the cline introduced in section 2.3 involves cliticization and affixation. Yet a traditional distinction between languages is whether they are synthetic (where one word can be made of multiple morphemes) or analytic (where one word tends to consist of one morpheme only).[5] Therefore, it is to be expected that advanced grammaticalization processes in analytic languages display different features from those displayed by the synthetic ones, given that cliticization and affixation are rarer (or non-existent) in the former.

In a similar fashion, subject-verb-object word order is also a main way to categorize languages. Section 2.2.1 introduced the parameter of syntagmatic variability (i.e. how an element can be shifted around its context) where more grammaticalized items tend to have less syntagmatic variability. Since word order can also be relatively fixed or free from one language to another, this may also translate differently when it comes to grammaticalization processes and syntagmatic variability.

The view adopted in the following chapters is that the parameters of grammaticalization discussed in sections 2.2.1 and 2.2.2 are relevant to most major languages. The task of comparing grammaticalization across all languages is complicated for obvious reasons, but a

[5] Note that the distinction between synthetic and analytic is gradual, since languages can have both synthetic and analytic ways of expressing certain concepts. For example, comparison in English can be expressed in an analytic manner (e.g. *more intelligent*) or in a synthetic one (e.g. *smarter*). This is why the distinction between synthetic and analytic involves fuzzy boundaries.

notable work is Heine and Kuteva (2002) which contains a list of grammaticalization processes from a broad variety of languages. As noted in Narrog and Heine (2011: 13), studies of grammaticalization have mostly been focused on major languages of the world, in particular English and European languages.

There are examples of grammaticalization that are fairly similar among European languages, such as future auxiliaries developing out of motion verbs. These illustrate a move from concrete meanings (e.g. physical motion) to more abstract ones (e.g. future tense), which is a known characteristic of grammaticalization (sections 2.2.1 and 3.3). A straightforward example in English is the motion verb *go*, which developed into the *be going to* + infinitive future construction, which can be used to express future actions that do not necessarily involve physical movement (e.g. *I'm going to sing*). The same use exists in French and Portuguese with the verbs *aller* and *ir* (*to go*) which can form sentences such as *je vais chanter* and *Vou cantar* (*I'm going to sing*) which do not imply physical motion. In fact, most Romance languages also have this use of the motion verb *go*. Despite such similarities, there are also notable exceptions that question to what extent grammaticalization "behaves" in the same way in different languages, in particular when considering different language families.

In section 2.1, phonological attrition was presented as a rather common feature in secondary grammaticalization processes, with mention to Norde (2012) and the example of the development of the possessive pronoun *(h)inn* into the suffix *-en* in Swedish. There are also straightforward examples in English such as the development of *will* into *'ll*, or that of *going to* into *gonna*. However, Chinese is an example of a language where grammaticalization involves little form change (Bisang 2009), which means that phonological reduction is not a prominent parameter at play in this instance.[6] Furthermore, many other East and mainland Southeast Asian languages also have highly grammaticalized items that do not involve phonological attrition (Bisang 2011: 110-113). Similarly, in a broad study involving 19 languages from genetically unrelated families, Schiering (2006) shows that in the case of clitization, phonological attrition is not a defining property of grammaticalization in all the languages involved in that study. Another salient feature of grammaticalization known as semantic bleaching (sections 2.2.1 and 3.3) is also problematic in Chinese as it involves many cases of grammaticalization that show an accretion of more meaning instead (e.g. Xing 2015).

This shows that while some aspects of grammaticalization might seem universal, there can be exceptions in other languages. The parameters used in the subsequent English studies

[6] Note that reductions in syllable duration and vowel quality/duration are attested, as reported in Ansaldo and Lim (2004) who discuss isolating tonal languages (Sinitic languages in particular).

are presented in chapter 3 and there is no doubt that their prominence changes from one language to another. However, the models introduced in chapter 4 will show that the relevance of these parameters can be calculated on the basis of empirical data, a concrete application of which is presented in chapters 5 and 6. This is one of the advantages of these models because it means that when applied to a language different from English, the parameters themselves can be recalculated for that language specifically. This also enables comparison between different languages as the prominence of specific parameters of the models can be compared. This possibility will not be explored in the present dissertation, but might be a fruitful endeavour for future works.

Attention has also been given to processes of grammaticalization in situations of language contact. For example, the development of creoles and pidgins is of interest since creoles in particular are known to have grammars in which complexity is greatly reduced (McWorther 2001, Hopper and Traugott 2003: 212-230). Another potential source of grammaticalization processes is bilingualism in situations of language contact (Matras 2011). While these specific situations are not dealt with in the present dissertation, they illustrate that there are situations in which grammaticalization processes must inherently be studied by taking different languages into account.

When discussing grammaticalization across languages, some researchers have undertaken the task to compare similar processes across a set of languages, often from the same family, and to consider their respective degrees of grammaticalization. For instance, Lamiroy and de Mulder (2011) discuss auxiliaries and other products of grammaticalization across three Romance languages, namely French, Italian and Spanish. They state that French is more grammaticalized than Italian, which in turn is more grammaticalized than Spanish. They therefore rank entire languages across a grammaticalization cline. This approach is rather uncommon, as noted by Narrog and Heine (2011: 14). However, this shows that there are different interpretations of what constitutes a grammaticalization cline and what the elements on that cline can be. The next section deals with a similar concern as it discusses which elements are undergoing grammaticalization by presenting the construction-based view according to which these elements are constructions, as opposed to single lexical or grammatical items.

2.7 The construction-based view of grammaticalization

Before discussing how construction grammar relates to grammaticalization, a brief introduction of its key concepts is necessary. Construction grammar (Croft 2001, Langacker

2005, Goldberg 2006, Hilpert 2014) is a theory according to which all linguistic knowledge boils down to knowledge of constructions. Constructions are the minimal units of knowledge and they consist of a linguistic form paired with a meaning. The concept of linguistic form is to be understood broadly. A construction can consist of a single word, such as *cat*, but can also consist of more complex patterns such as *the X-er and Y-er* (e.g. *the more the merrier*). More abstract patterns such as the ways the comparative is expressed in English (*more X*, *most X*, *X-er* and *X-est*) also constitute constructions. The COMPARATIVE itself (as a grammatical concept) is also a construction (and is written in capitals when used as such).

The view that constructions are the basic and only unit of linguistic knowledge is complemented by the related assumption that constructions are thought to be organized in a network called the construct-i-con. This network is the mental representation of constructions that individual speakers have and therefore involves a certain organization and links between these constructions. For instance, the substantive lexical construction *more X* is linked to a more abstract COMPARATIVE construction. Constructions can be ranked according to their degrees of schematicity and abstractness. Thus, the COMPARATIVE construction is more schematic and abstract than the *more X* construction. The terms *macro-*, *meso-* and *micro-constructions* are sometimes used to label constructions, from more to less schematic/abstract (Traugott 2008: 236).

For the present purpose, the definition of what constructions are for the present purposes is taken from Goldberg (2006: 5):

> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency.

The first part of this definition posits that a construction must not be fully predictable from its component parts (non-compositionality), which is a logical consequence of the assumption that constructions are the minimal units of knowledge. This is how *more intelligent* is not an example of a construction, because it can be fully predicted on the basis of the two constructions it is made of, namely *more X* and *intelligent*. *More intelligent* is called a construct, which denotes an instantiation of a more general construction (i.e. *more X*). The second part of the definition points to an exception to non-compositionality. Highly frequent elements, such as *I don't know*, are also considered as constructions, despite having a

meaning fully predictable on the basis of the more general constructions it involves. This is because frequent elements are conventionalized and tend to replace possible alternative. If a speaker is asked a question to which they don't know the answer, they will likely respond with *I don't know*, instead of *This information is unknown to me*, or *I lack this piece of knowledge*. This is why *I don't know* is likely stored as a construction in its own right in the construct-i-con. The inclusion of frequency into the definition also implies that highly frequent elements such as common plurals (e.g. *cats*) can also be regarded as constructions if frequent enough. This blurs the line between the concept of construct and construction because the earlier example of *more intelligent* could be considered as rather frequent enough to be considered as its own construction. This particular question regarding how frequent a construct needs to be in order to be considered as a construction goes however outside the scope of the present discussion (see Bybee 2003, 2013).

Language change can be studied from a construction grammar point of view and the terms "constructionalization" and "constructional change" are often used to describe this approach (e.g. Traugott and Trousdale 2013). Constructionalization is the creation of new form-meaning pairings, while constructional changes are changes that affect the features of an existing construction. Constructional changes are believed to precede and succeed most cases of constructionalization. The creation of new form-meaning pairings (i.e. constructions) is mainly made possible by prior changes in an already existing construction.[7]

For example, the Modern English construction *a lot of* developed from the Old English word *hlot* which used to refer to a concrete object, mainly a piece of wood to select an individual, in the context of elections and of drawing lots. The word *hlot* was also used to refer to a share or unit of something gained by such drawing lots. In Middle English, partitive uses of *lott* are attested, along with the preposition *off*, such as *aniȝ lott Off Moysœsess lare* (any part of Moses' teaching), as well as uses with a meaning close to a "group of", such as *Summ lott off gode sawless* (certain groups of good souls), as reported in Traugott and Trousdale (2013: 23-26). In those examples, *lott* is a head, followed by a modifier (*off* NP) and it denotes units that are part of a larger whole. These uses of *lott* are still lexical as they refer to a concrete unit, but since this unit is part of a larger whole, *lott* also has a relational (i.e. partially grammatical) quality. It is only in later attestations from the early 19th century that *lot of* can be regarded as fulfilling the grammatical function of quantifier, similar to Present-Day English, meaning "a large quantity of" and where it is no longer a head, but a

[7] Traugott and Trousdale (2013: 29-30) distinguish between gradual constructionalization and instantaneous constructionalization. However, instantaneous constructionalization is rarer and is therefore not discussed here.

modifier instead. This change has been described by the following schema which shows the reanalysis of the structure of the construction (i.e. its form) and its associated meaning:

> [[N_i [of N_j]] ↔ [$part_i$ - $whole_j$]] > [[N of] N_j]] ↔ [large quant - $entity_j$]]
> (Traugott and Trousdale 2013: 25)

This illustrates that constructionalization (i.e. creation of a new construction) may require constructional changes to take place first. Partitive uses of lexical *hlot/lott off* paved the way for structural reanalysis and for uses as a quantifier. These changes are also followed by host-class expansion, such as current uses of *a lot of* with much more abstract concepts that are not inherently quantifiable (e.g. *a lot of courage*). This shows that constructional changes can also occur after constructionalization has taken place.

Constructional change overlaps with grammaticalization, although the two concepts are distinct. An obvious example are semantic changes within lexical categories, such as the narrowing of *hound* which currently refers to a specific type of dog in English, but used to refer to any kind of dog in the past. These changes fall under constructional changes (i.e. changes in the form-meaning pair), but do not involve grammaticalization. In contrast, grammaticalization can involve highly abstract developments that are not necessarily constructions themselves. For instance, does the development of modal auxiliaries in English entail the existence of the highly general construction MODAL AUXILIARIES as well? It is unsure that most speakers do this generalization and indeed have a MODAL AUXILIARY construction in their construct-i-con. This shows that some processes can pertain to grammaticalization, but not necessarily to constructional change (see Hilpert 2013: 9-13 for more details).

Despite these examples, there is a significant overlap between constructional change and grammaticalization. As discussed earlier, construction grammar proposes an organization of the contruct-i-con based on degrees of schematicity and abstractness, which is reminiscent of degrees of grammaticalization, especially when considering the increase in generality of meaning (and abstractness) often involved in grammaticalization (section 2.2.1). There are at least two main ways in which adding a construction grammar point of view enhances the understanding of grammaticalization processes that are relevant in the present dissertation.

First, construction grammar posits a clear minimal unit (also called atomic units) of study, the construction. As discussed earlier, constructions can consist of single words (e.g. *cat*), or more complex patterns (e.g. *the X-er the Y-er*). While the grammaticalization

continuum presented in Hopper and Traugott (2003: 7) involves the terms "lexical item" and "grammatical word", they should not be taken to necessarily mean single items/words but also to include more complex constructions. In fact, Himmelman (2004: 31) explicitly states that grammaticalization applies to "constructions and not isolated lexical items". He also points out that the grammaticalizing element itself has been given more importance as opposed to its syntactic context, giving the example of the focus on *go* in the future *be going to* + infinitive construction, to the detriment of the context around it which is highly relevant. Bybee (2003: 602) advocates for a similar position where these developments can only be studied "in the context of the construction in which the grammaticalizing element occurs", which is also supported by Traugott (2003). This will be relevant in the subsequent chapters, in different ways. Chapter 5 deals with the synchronic measurement of grammaticalization using lists of lexical and grammatical elements. While many of these elements involve single words, they also involve more complex patterns, such as complex prepositions (e.g. *as long as*, *in order to*). In addition, in chapter 6, other complex patterns involving auxiliaries such as *be going to* + infinitive and *Keep* V_{ing} are also included. Information regarding the parameters of grammaticalization discussed in chapter 3 will be collected using the entire pattern, instead of its main grammaticalizing elements. A limitation however is that changes within the patterns themselves will not be taken into account, as they are not the focus of the study. Instead, the focus will be more on the syntactic context around these patterns, in the form of collocate diversity (section 3.3).

Second, since constructions are part of the construct-i-con, the links they have with more general constructions can be explored as well. This was shown to be useful for example in Hilpert (2013: 79-109) who notes that the development in English from *mine* to *my* is extremely similar to that from *thine* to *thy*, on the basis of diachronic corpus-based data. Given that these two developments are fairly similar, Hilpert argues that these two elements might in fact instantiate a single construction that is a generalization of these two. This might be relevant in Chapter 6 as it involves complex prepositions which might instantiate a higher schema. As illustrated in Hoffmann (2004, 2005), complex prepositions can involve similar patterns. For example, *with regard to* and *with respect to* are instances of the pattern *with X to*. The empirical investigations in chapter 6 might also highlight that similar patterns display similar developments.

Two aspects of construction grammar, namely that constructions are the minimal unit of study and that these units are linked to each other in a construct-i-con, have been highlighted

as highly relevant to grammaticalization and further discussion on this matter can be found in Hilpert (2013), Trousdale (2010), Gisborne and Patten (2011), and Fried (2013). The main defining traits of grammaticalization have been introduced in the previous sections and a summary is offered in the next section.

2.8 Summary

The previous sections have introduced the core concepts of grammaticalization. Several definitions of grammaticalization were introduced, among which Hopper and Traugott's (2003: 18), for whom it consists in "the change whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions and, once grammaticalized, continue to develop new grammatical functions." This is the central definition used in this dissertation. Along with this definition, Hopper and Traugott (2003: 7) propose the following cline of grammaticalization:

content item > grammatical word > clitic > inflectional affix

This cline implies a directionality from less to more grammaticalized, which is known as the unidirectionality hypothesis of grammaticalization. In addition, the changes along the cline are considered as gradual, which means that they happen through small incremental steps. When an element grammaticalizes, several changes may take place. The most commonly discussed ones are phonological attrition and semantic bleaching. The next chapter will introduce ways to measure some of these characteristics by quantitative means.

It has furthermore been argued that the prominence of these characteristics can vary from one language to another. While the studies presented in the following chapters only use English datasets, there is no obstacle to their replication in other languages since they rely on corpus-based variables that are not language-dependent. Chapter 5 focuses on the computation of a grammaticalization score for a given set of lexical and grammatical elements and tries to determine which variables play the biggest role in doing so. This means that were the model to be applied to another language, one could compare the prominence of these variables in their respective languages. This would allow for more quantitative cross-linguistic studies of grammaticalization in synchrony.

Several shortcomings of grammaticalization theory were discussed, such as the lack of large-scale empirical support in favour of the unidirectionality hypothesis. A common criticism that was discussed is the fact that there are examples of degrammaticalization and

that they have received less attention in research, which means that they were potentially overlooked and that the unidirectionality hypothesis of grammaticalization might not be as robust as commonly thought. Chapter 6 proposes an empirical investigation of the tendency for unidirectionality.

Finally, since grammaticalization involves several distinct features, their relationships are still relatively under-investigated, in particular from a general point of view that involves larger datasets. Chapters 5 and 6 investigate the links between the variables used in their grammaticalization models. One of the objectives is to determine whether some characteristics are more prominent and how they correlate with higher degrees of grammaticalization. Computing these variables involves several technical steps that are presented in the following chapter. Furthermore, this chapter also discusses the link between these quantitative variables and some of the main parameters discussed in section 2.2.

3. Grammaticalization parameters quantified

The previous chapter discussed grammaticalization from a theoretical perspective, in terms of several underlying processes and concomitants. The present chapter will turn to the business of making these concepts measurable by focusing on quantitative, corpus-based variables. Each measure is presented and discussed in relation to the parameters presented in section 2.2.1. The aim of this chapter is to motivate the choice of these quantitative measures, but also to explain how they are implemented in the statistical models that will be presented in the subsequent chapters. The limitations of these measures are also discussed. This chapter therefore contains both theoretical and practical aspects. The parameters that are presented in this chapter constitute an open-ended list. There might be other relevant quantitative operationalizations of grammaticalization that are not discussed here.

For example, section 2.2.1 introduced the notion of paradigmaticity, which consists in the fact that grammaticalized elements tend to become part of smaller closed-class paradigms. The size of such paradigms can be considered as a relevant quantitative factor. However, defining paradigms is not always straightforward and determining whether a grammaticalizing element is part of such a paradigm tends to require more qualitative investigation. In a similar manner, it is difficult to evaluate paradigmatic variability, which is the freedom to select an option among others, since it requires comparing elements to their possible alternatives. This again requires more manual and qualitative intervention, but could technically be implemented by comparing frequency ratios between different alternatives. Adding new parameters or finding better alternatives to the current ones is therefore certainly a possibility for future research.

The chapter is structured as follows. Token frequency is presented first in section 3.1 as it has been the most discussed variable in previous research. Section 3.2 focuses on letter counts, which can be used to approximate phonological reduction, which is another well-known characteristic of grammaticalization. Section 3.3 introduces the concepts of collocate diversity and of collocational windows. Section 3.4 introduces the concept of colligate diversity, which focuses on the parts of speech of collocates. Section 3.5 discusses the dispersion of elements in a text as a possible measure of grammaticalization. It presents a specific way to measure textual dispersion called deviation of proportions (Gries 2008). Finally, section 3.6 proposes an overview of corpus-based works on grammaticalization.

3.1 Token frequency

Token frequency is the number of occurrences of a word in a corpus. For example, the word *give* occurs 43'488 times in the British National Corpus. Naturally, *give* also has other forms such as *gives*, *gave*, *given* and *giving*. When measuring token frequency, a choice has to be made whether the frequencies of all these forms are aggregated or whether they are kept separate. Aggregating different forms of the same word is a process that is known as lemmatization and the element under which they are grouped is called a lemma. When *give* is lemmatized so that all of its word forms are counted together, its token frequency goes up to 123'617.[8]

Lemmatization was not used in the subsequent studies, since there is a clear need to keep most entries separate. For example, *go*, *went* and *going* fulfil quite different grammatical functions. These differences could not be studied on the basis of lemmatized data. Furthermore, different word forms of the same lemma can show different degrees of grammaticalization. This is related to the principle of divergence (Hopper 1991) discussed in section 2.2.2, where a grammaticalizing form can have a lexical counterpart still in use. Similarly, while a verb form might grammaticalize, other verb forms of the same lemma might not participate in this process. To illustrate, *going* mostly occurs as an auxiliary verb, while *go* and *went* tend to be main verbs. It might be argued that there are cases where lemmatization might have been required, as for example *a* and *an* indeed have the same grammatical function and are just two allomorphs constrained by phonological criteria. However, these cases are not so common and choosing which items should be lemmatized would involve much manual processing and highly subjective decisions which might be problematic, in particular when it comes to the reproducibility of the study.

In addition to lemmatization, another important practical aspect when measuring token frequency is how the corpora involved are parsed. It is common to use regular expressions to find words in a corpus, which can involve several small decisions such as which characters can constitute words (e.g. numbers, accented letters such as *à* or *é*). Corpora also tend to involve spelling mistakes and encoding errors, such as missing blank spaces between words or punctuation. Decisions can be made to try to minimize these by adding extra steps to automatically avert some of these shortcomings. These aspects can therefore change the results when measuring the token frequency of given items. This is why the regular expressions involved in chapter 5 are explicitly listed in section 5.3 and those in chapter 6 in

[8] These numbers were obtained by using the BNC-BYU online interface (Davies 2004).

section 6.2. Different preprocessing decisions also explain why different interfaces and different researchers might not obtain exactly the same numbers when computing token frequency values, despite using the same corpora.

The relevance of frequency in grammaticalization can be approached from a naive angle, by making the observation that function words are generally more frequent than content words in most languages. This can easily be observed by looking at the most frequent words in any corpus, all of which happen to be grammatical elements. The five most frequent words in the British National Corpus are *the, of, and, to* and *a*. More generally, the majority of the top 100 most frequent words in this corpus are grammatical elements.

While token frequency is known to play a role in grammaticalization, there is some divergence when assessing the extent of this role. For example, Hopper and Traugott (2003: 126-127) acknowledge that frequency "has long been recognized informally as a concomitant of grammaticalization", which is a prudent assertion. On the other hand, Narrog and Heine (2011: 2-3) note that "in some of the definitions provided, frequency is portrayed as one of the driving forces, or *the* driving force of grammaticalization". This is especially the case with Bybee (2011: 77), who considers increases in frequency to even explain the unidirectionality of grammaticalization (section 2.4): "as long as frequency is on the rise, changes will move in a consistent direction." This is because in Bybee's view, "when a grammaticalizing construction ceases to rise in frequency, various things happen, but none of them is the precise reverse of the process" (Bybee 2011: 77). This shows that there are views of grammaticalization in which frequency occupies a core position. This is further illustrated in Diessel and Hilpert (2016).

An example of an effect of high token frequency is phonetic reduction, which is further discussed in the following section. Bell et al. (2009) have highlighted that there is a strong correlation between frequency of use and phonetic reduction, where frequent elements are more likely to be reduced in speech production. However, they report differences when it comes to lexical and grammatical items. Lexical items seem to display a straightforward correlation, where more frequent items are shorter, whereas grammatical items do not display such correlation when controlling for frequency and predictability. Grammatical items tend to be reduced when they are highly predictable from the linguistic context. For instance, *of* often appears after *kind* or *sort* in the *kind of* and *sort of* constructions (i.e. there is strong mutual information between these elements), which are subject to phonological reduction (*kinda/sorta*). Lexical and grammatical elements might therefore be processed differently in

speech production, and frequency may play a different role in both cases as pointed out by Bell et al. (2009).

Measures based on token frequencies have also been used to show the emergence of syntactic patterns, such as the grammaticalization of verb-object word order in English (Fries 1940) or of epistemic parentheticals (Thompson and Mulac 1991). Furthermore, token frequency has also been used to investigate the percentages of different uses of elements undergoing grammaticalization. For instance, Barth-Weingarten and Couper-Kuhlen (2002) used the percentages of different uses of *though* to illustrate that it is developing a new discourse marking function in Present-Day English.

It should be noted that in the same way that the Lehmann parameters (section 2.2.1) are not relevant in all cases of grammaticalization, something similar can be said about token frequency. There are examples of low frequency grammaticalization (e.g. Hoffmann 2004, Brems 2007). For example, there are complex prepositions that are relatively infrequent in English such as *by dint of*. Hoffmann (2004) argues that such low frequency complex prepositions can still involve high degrees of grammaticalization. There are cases where low frequency complex prepositions can grammaticalize by analogy to similar prepositions that are more frequent.[9] Therefore, since *by dint of* has a similar structure and meaning as *by means of*, it may display similar grammaticalization features despite a much lower frequency.

Furthermore, token frequency values have their limitations. For instance, certain concepts are rarer and so is the need to express them. If such a concept can only be expressed by one specific construction, then this construction is relatively prominent within this specific context, regardless of the overall frequency of the concept. This relates to Lehmann's parameter of paradigmatic variability (section 2.2.1), where highly grammaticalized elements involve less freedom to choose alternatives. Hoffmann (2004: 191-193) mentions the preposition *in front of* which became a prominent way to express a specific spatial relationship in Present-Day English, in contrast to the preposition *before* which can also be used to express a similar locative meaning.

For example, sentences such as *he is sitting before the fire* and *he is sitting in front of the fire* are both possible in Present-Day English and have similar meanings. However, from a diachronic perspective, locative *before* was the only of the two that was used to express this concept in the 17th Century. The preposition *in front of* developed later on, and has a frequency close to locative *before* in the British National Corpus, given that around 6'000

[9] Note that Brems (2007: 296-298) is critical of this particular explanation.

occurrences of *in front of* versus 7'000 of locative *before* are reported in Hoffmann (2004: 192). While the preposition *in front of* does not have a larger frequency than locative *before*, it is currently one of the main ways to express this specific concept of spatial relation. In short, whether a construction is a prominent way to express a specific concept (i.e. in comparison to similar realizations) is also a relevant factor that goes beyond mere token frequency.

This is why chapter 6 also includes some low frequency elements. The existence of such elements is not problematic because it just means that higher token frequency is not an obligatory aspect of grammaticalization. However, it remains a central one, especially since increases in token frequency can also lead to phonological reduction, which is discussed in the next section.

3.2 Letter counts

Phonological reduction (also called attrition), a common process in grammaticalization, affects elements that are highly frequent (Bybee 2001, Bybee and Thompson 1997). The idea is that frequent forms tend to erode at a faster rate than elements that are less frequent. This is illustrated by many frequent contracted forms in English, such as '*ll* and '*re*. It is also observed with the coalescence of the infinitive marker *to* in constructions such as *going to, got to, want to*, which are shortened to *gonna, gotta* and *wanna* respectively. Furthermore, less frequent elements might also not show any reduction at all, as evidenced by developing auxiliaries such as *keep* in the *keep* V_{ing} construction which do not display such reduction at all.

This process of reduction is present as Lehmann's parameter of phonological integrity (section 2.2.1). Grammaticalizing forms are phonologically reduced as they gain in frequency, but it remains to be formally answered whether this is a mere frequency effect, or whether there is an additional effect of grammaticalization so that grammatical words generally tend to be shorter than lexical ones. This does not need to be formally shown when taking a diachronic perspective, since there are numerous accounts of phonological attrition in a specific form. However, from a synchronic perspective, while it seems intuitively plausible that grammaticalized forms are shorter, it is necessary to support this idea with evidence. For example, an obvious observation is that many grammatical elements such as determiners or pronouns are generally very short, but it is less apparent whether this holds true in a large dataset that also includes longer grammatical elements (e.g. *throughout*, *in accordance with, in conjunction with*) and short lexical elements (e.g. *see*, *put*, *let*). Furthermore, as discussed

in the previous section, Bell et al. (2009) found different correlations between frequency and phonetic reduction in lexical and grammatical elements.

In the subsequent studies, this phonological parameter is approximated by letter counts, as the corpora that are used do not have a phonemic transcription. Automated approaches can use letter counts in place of phoneme counts in successful ways, as illustrated for example in morphological segmentation (Goldsmith 2001). While counting the number of letters is rather straightforward and does not need complex explanations, it should be noted that a decision has to be made regarding small constructions. For example, should *as long as*, be counted as an eight letters element, or as a four letters element, since its main component is the word *long*? Most of the constructions involved in the database in chapter 5 follow rather established patterns, such as *as X as*, *in X of,* or *in X with*. It is therefore highly complicated to determine how relevant the surrounding prepositions are and whether they should be counted, since they mostly consist in a fixed set of short prepositions (e.g. Hoffmann 2004: 182). In the subsequent studies, only the letters of the main component were counted.

From a practical point of view, the BNC-BYU interface for the British National Corpus (Davies 2004) only displays the main element of such constructions. For instance, the construction *as long as* is displayed as *long* when using a query to find all conjunctions in the corpus (i.e. [c*]), which is why such constructions were initially included in the database in chapter 5 using only their main element. A valid criticism is that constructions such as *as long as* can be shortened into *slong as*, especially in informal contexts. In this case, the reduction does not happen at the level of the main element of the construction. However, the purpose of the letter count variable is to be an approximation of phonological length, which implies that precision is not a central concern in the present case.

An additional concern is that not all constructions are affected similarly by this issue. For instance, there are 11 letters in the construction *depending on*, but 9 of them are accounted for by the main element. There are obvious cases where this is not true, such as *on top of*, where the main element does not even account for half of the letters in the construction. Therefore, constructions are affected differently when it comes to the approach taken to count their letters. Constructions with multiple elements represent about 10% of the items in the database presented in chapter 5, which is why this particular issue is considered to be minor for the subsequent studies. However, in the context of studies that would require more precision and involve more constructions, it seems advisable to measure phonological length using whole constructions instead of just their main element.

3.3 Collocate diversity

A collocate is simply a word that occurs near another word. What exactly "near" means is up to individual interpretation. The chosen context within which two elements are said to be near each other is called a window. For example, let us look at the verb *ate* in sentence (1). This verb has two direct neighbours, the noun *cat* on its left, and the determiner *the* on its right. These two collocates therefore occur within a symmetrical window of one element.

(1) My old cat ate the dog's homework

			window				
3	2	1	0	1	2	3	4
My	old	cat	ate	the	dog	's	homework

Table 2. Collocate window around the verb *ate* in sentence (1).

If the window is extended to two words instead of one, then the adjective *old* and the noun *dog* are also considered as collocates of the verb *ate*. Windows can be asymmetrical if one is only interested in the left or right collocates of a given element. For instance, if one decides to select an asymmetric window of three elements to the right, then the collocates *the*, *dog* and *'s* would be selected. Table 2 provides a visual representation of a collocate window.

Collocates give relevant information about the meaning of a word or construction around which they occur. For instance, if one looks at all the collocates around the verb *eat* in a corpus, it is likely that many of these collocates are going to be food-related items, especially to the right. Similarly, left collocates will mostly comprise agents, such as people or animals. Therefore, collocates can inform about a link between the verb *eat* and entities that can eat, as well as things that can be eaten. This lead Firth (1957: 11) to coin his famous slogan: "You shall know a word by the company it keeps".

Computational linguistics makes use of collocates to establish semantic profiles of given elements. These profiles can then be used in word-sense disambiguation tasks or information retrieval in general (Turney and Pantel 2010). For instance, the word *bank* is ambiguous because it can denote a financial institution or the bank of a river or lake. However, it is evident that when the word refers to the financial institution, most collocates will include finance-related words, whereas the other meaning will involve water-related words. This is how collocates can be used to gain information about a word's semantics, and therefore also about the kinds of contexts in which it occurs.

The fact that collocates can be used to investigate semantics is relevant for grammaticalization because it has often been observed that grammaticalized elements can be used in a wider variety of contexts as they become semantically vaguer or broader. The notion of semantic bleaching in grammaticalization (Sweetser 1988) is a generally well-attested phenomenon (Hopper and Traugott 2003: 127) that has been observed with modals (Krug 2001), auxiliaries (Heine 1993) and many other types of elements. It should also be noted that grammatical elements becoming semantically broader has also been described as a result of increases in frequency (e.g. Haiman 1994).

While Lehmann (2002) includes semantics as an aspect of the integrity parameter (i.e. semantic integrity), collocates also relate to some of his other parameters, such as bondedness and syntagmatic variability. Indeed, if a sign is constrained with regard to the other signs it can be associated with, then this should be reflected empirically in its collocate profile as well. A highly constrained profile would mean an overall lower number of possible collocates, whereas a largely unrestricted profile would translate into a wider range of possible collocates. This is why a measurement of collocate diversity might be useful to measure degrees of grammaticalization. In fact, Hilpert (2008: 13-48), as well as Torres Cacoullos and Walker (2011), have already illustrated that collocates of grammatical constructions can be used to show degrees of grammaticalization for specific cases such as modals, future tense markers and complementizer *that*. Furthermore, the idea that collocations are relevant to grammaticalization has also been discussed in other works, such as Himmelman (2004: 31-34) and Gisborne and Patten (2011: 96-98), although not necessarily in direct relation to gradualness.

In the subsequent studies, collocate diversity is measured in the following way for a given element. First, all the possible collocates are found within a given window. Three different windows are used. One window consists in a symmetric four words window (i.e. eight words in total). The other two consist in a one word window to the right, and another one word window to the left.

Second, the number of different types among these collocates are counted. This simply consists in the number of different forms that are found. For example, if ten collocates were found, where five instantiate the word *dog*, three the word *cat* and two the word *nice*, this means that three different types were found. In a similar fashion to section 3.1, no lemmatization is used for the list of collocates, which means that *dog* and *dogs* count as separate types.

Third, the number of different collocates (i.e. the number of types) is then divided by the token frequency of the item, in order to take into account that items with higher token frequencies are necessarily going to have more different collocates as a result of their high frequency. This results in a number that ranges between zero and the window size. A brief explanation using a window size w and a frequency n is that the maximum number of different collocates corresponds to w*n. This corresponds to the situation where all collocates are unique and not a single one is repeated. If this maximal number (w*n) is further divided by the frequency (n), then trivially the maximum diversity is equal to w. Conversely, the theoretical minimum diversity (where all the collocates are in fact the same one) corresponds to 1/n, which will approach zero when n is large.

The three different windows proposed above reflect different linguistic aspects. The larger symmetric window of four words arguably gives more information about the meaning of a specific element. Collocate-based research for semantics-related purposes (e.g. word disambiguation, grouping words by semantic similarity) tends to be better served by broader windows (e.g. Turney and Pantel 2010: 170-171). While there does not seem to be a recommended number as this may depend on the task at hand, grammatical elements occur rather close to the elements that they are used in conjunction with, which motivates using four words rather than a larger number to both sides. Another note is that studies that use collocates to retrieve information about meaning tend to focus on lexical meaning and therefore remove words that are called "stopwords", which mostly consist of function words. In the present case, since grammatical elements are the focus of the study, such words are also included, and no stopwords are removed from the elements found within the chosen windows. In the sense that the four-four window is more closely related to semantic aspects, it is connected to Lehmann's notion of semantic integrity (section 2.2.1).

The smaller windows of just one word are expected to give more information on morphosyntactic constraints, as they pertain to elements that are directly adjacent. Adjacency is related to Lehmann's notions of bondedness and syntagmatic variability (section 2.2.1), as both these parameters are concerned with how a given item attaches itself to another elements. For instance, prepositions tend to be heads of prepositional phrases, which means that most of the time, they will have another phrase to their left (e.g. *he was on the roof*). In contrast, it is likely that to the right of the preposition, one will find the second element of that prepositional phrase. Directionality is therefore a additional aspect that will be investigated in the subsequent studies. The remainder of this section presents several examples in order to illustrate the computation of the collocate diversity measure.

Consider for instance the case of a window of one element to the right. If one applies this window to an element that occurs 6'000 times in a given corpus, then there are going to be 6'000 collocates to the right of this item.[10] If all these 6'000 elements are unique types, then dividing this number by a frequency of 6'000 results in a maximum diversity of 1 (6'000/6'000). In contrast, if there is only one type repeated 6'000 times, this results in a diversity of almost zero (1/6'000). In a more plausible case, maybe there will be 1'250 different types among these collocates, which results in a value of 0.208 (1'250/6'000), which can also be interpreted as 20.8% of the collocates being unique types.

If the window comprised two elements instead, then there would be 12'000 collocates, which can result in a possible maximum collocate diversity of 2 (12'000/6'000), which corresponds to the size of the window in that case. Therefore, a further step is to divide this value by the window size to always get a value between zero and one (and that can therefore be interpreted as a percentage as well, as shown in the previous paragraph), but this step is not required in the subsequent studies, as other standardization processes are undertaken (section 4.1.2).

To give an actual example from the subsequent studies, the collocate diversity measures from section 5.4 for the clitics *'re* and *'m* are given in Table 3. As expected, they show a very low diversity when using a window of one element on the left (Colloc_L1), which is to be expected since the left element is almost always going to be *I* and *you*. The right diversity (Colloc_R1) is also relatively low and the reason is that many instances include *not, going* and *a*. On the other hand, the four-four window (Colloc_4-4) shows higher values, which is to be expected since once one looks beyond direct adjacency, many other elements are possible. This illustrates why using different window sizes can lead to different results.

	Colloc_4-4	**Colloc_L1**	**Colloc_R1**
<w V**>'M	0.486	0.001	0.066
<w V**>'RE	0.543	0.001	0.097

Table 3. Collocate diversity measures for the clitics *'m* and *'re* (data taken from Section 5.4).

[10] To keep the discussion simple, elements that occur where the corpus begins and ends are simply ignored. These cases are special in the sense that for instance the very first element in a corpus does not have left collocates at all. However, these cases are marginal and the way they are handled pertains more to practical matters.

The way collocate diversity is computed here is close to a type-token ratio, which is a measure of vocabulary diversity (or richness)[11] of a text, or of a speech sample. This measure has been particularly prominent in child language research for a long time (e.g. Richards 1986). The type-token ratio of a given text is obtained by dividing the number of different words (i.e. types) by the total number of words (i.e. tokens). However, a limitation of such a measure is that it is dependent on text length (e.g. Tweedie and Baayen 1998, Covington and McFall 2010). It is therefore problematic to compare the richness of texts of different lengths using a type-token ratio. In the present case, collocate diversity is obtained by dividing the number of different types by the token frequency of an element (which therefore corresponds to text length in the case of a type-token ratio). While it is possible to address this shortcoming of type-token ratios, for instance by using text samples of similar sizes instead of the whole texts, in the present case this cannot really be achieved because the data concerns the collocates of a given element. The extent to which this limitation is problematic can be investigated by empirical means, namely by checking how the collocate diversity measures correlate with the token frequency measure. Sections 5.4.1, 5.5.1, and 5.6.1 deal with this issue by showing that while token frequency and collocate diversity are indeed correlated to some extent, their correlation is relatively small.

3.4 Colligate diversity

A colligate is the same concept as a collocate, however instead of considering the collocate as a full element, only its part of speech is considered. To come back to sentence (1), and using a symmetrical window of two words around the verb *ate*, the collocates *old, cat, the* and *dog*, one would be replaced by the part of speech they represent, namely, *adjective*, *noun*, *determiner*, and *noun*.

Colligate diversity consists in the diversity of possible parts of speech that may come after or before an item. It cannot be operationalized, in the present case, in the exact same way as the collocate diversity measure presented in the previous section, because the number of possible types is rather limited. Indeed, the list of possible colligates for any item in the British National Corpus amounts to around 50 different tags (section 4.2.1). This is why a slightly different approach is taken, in which the percentage of the most frequent colligate is taken as a measure of diversity. The reasoning is that grammatical elements might be more

[11] Note that this is often referred to as lexical diversity (or richness), but here the term vocabulary diversity is preferred to avoid confusion since most of the text discusses lexical and grammatical counterparts. Furthermore, grammatical elements can be part of vocabulary diversity, so using the term lexical can be even more misleading in this specific context.

constrained in the range of possible parts of speech that comes after or before them. As an illustration, a determiner is quite often going to have a noun (or maybe an adjective) as a right colligate.

Two measures are used in the subsequent studies; one uses a window of one colligate to the right, and the other uses a window of one colligate to the left. Both measures calculate the percentage of the highest colligate of each item. To illustrate, let us look at an example taken from the data used in section 5.4. The verb *will* is followed by an item tagged as <*w VVI*>, which marks infinitive verbs, 64.4% of the time. Therefore, there is a highly predictive constraint on what might come after this element. Furthermore, its shortened form *'ll* gives a similar result, but with a higher percentage of 79.9%. Note that with this measure, higher percentages correspond to less diversity. The measure could more accurately be called "colligate constraints" to be more consistent with this fact, as a higher percentage corresponds to more constraints. However, it has been called colligate diversity here in order to mirror collocate diversity as introduced in section 3.3.

A small additional note is that depending on the level of detail in the tagging of a corpus, the colligate diversity measure can produce different results. For example, if the corpus has detailed verbal tags, such as tags for different tenses or specific tags for key verbs (e.g. *be* or *have*), this will be reflected in the diversity measure. To give a concrete example, let us imagine that a given item has twenty different verbs that occur to its right and that these constitute all its right colligates. If the corpus does not have detailed tagging and uses a general verbal tag, these twenty instances will be counted as the same category and yield a colligate diversity score of 100%. On the other hand, if the corpus has detailed tagging, it may highlight that fifteen of these verbs are infinitives, while the other five belong to other categories, such as past participles, instances of *have*, and so on. In this second case, the diversity would be of 75% instead. The corpora used in the subsequent studies use claws4 and claws7 tagging which do provide some level of details, including special tags for key verbs. Additional information about tagging and the corpora used in the subsequent studies is provided in section 4.2.

As illustrated above, using the percentage of the most frequent colligates of an item might be useful in determining its degree of grammaticalization. Indeed, grammaticalized items are less autonomous and might be expected to have higher constraints regarding the parts of speech that may occur after or before them. Similarly to the previous discussion regarding collocate diversity (one word windows), this relates to the concepts of bondedness

and syntagmatic variability introduced in section 2.2.2, where grammaticalized elements tend to be more strongly attached to other signs and moved around less freely.

3.5 Dispersion

Dispersion refers to the way an element is distributed within a text. It can have an even distribution and occur regularly throughout the text, or it can have an uneven distribution and occur in specific parts of the text only. There are different ways to measure how a word is distributed in a text. The main approach used in this dissertation is called deviation of proportions (Gries 2008: 415-419). The general idea is that the actual distribution of the word in a text is compared to its theoretical uniform distribution in that text. This is achieved by splitting a corpus into a certain number of chunks. The complete algorithm to compute the deviation of proportions of a given item is as follows:

1. Divide a corpus into n chunks (n is chosen by the researcher). Determine which percentage of the corpus each chunk represents. These percentages are the expected percentages for a given item in a uniform distribution.
2. Find the actual frequencies of the desired item in each chunk and express them as percentages. These are the actual percentages of a given item.
3. Calculate the absolute difference between the actual percentages in each chunk and the expected percentage in a uniform distribution. Add all these differences up.
4. Divide this result by two.

To give a concrete example, let us imagine a corpus of 10'000 words. In this corpus, the word *cat* occurs 240 times. One can decide to divide the corpus into ten 1'000 words chunks. Thus, each chunk represents 10% of the corpus. It is then necessary to check how the word *cat* is distributed within these ten chunks. Table 4 shows how many instances of *cat* occur in each chunk. These are the actual instances and they should be expressed as percentages.

To compare the actual distribution with the theoretical uniform distribution, a subtraction must be made for each chunk. Therefore, the difference between the actual distribution and the uniform distribution in chunk seven is 19.167% (29.167 – 10). Only the absolute values matter (i.e. the sign of the result of the subtraction is ignored). These differences are computed in Table 5. Once these differences are obtained, the percentages need to be added up, which shows how much the actual distribution differs from the theoretical uniform distribution. In the present case, the total amount in Table 5 is 99.168%. This percentage must then be divided by two, to get the final value of 49.584%. Normally,

this is expressed using the decimal notation 0.495 (i.e. the percentage notation divided by 100).

Chunk 1	Chunk 2	Chunk 3	Chunk 4	Chunk 5	Chunk 6	Chunk 7	Chunk 8	Chunk 9	Chunk 10
0	2	0	65	55	23	70	25	0	0
0%	0.833%	0%	27.083%	22.917%	9.583%	29.167%	10.417%	0%	0%

Table 4. Distribution of the word *cat* in an imaginary corpus. Each chunk contains 1'000 words. The theoretical uniform distribution is 10% per chunk.

Chunk 1	Chunk 2	Chunk 3	Chunk 4	Chunk 5	Chunk 6	Chunk 7	Chunk 8	Chunk 9	Chunk 10
24	22	24	41	31	1	46	1	24	24
10%	9.167%	10%	17.083%	12.917%	0.417%	19.167%	0.417%	10%	10%

Table 5. Absolute difference between the theoretical uniform distribution and the actual distribution of the word cat.

Deviations of proportions range between 0 and 1. Indeed, if in the previous *cat* example the word occurred exactly 24 times per chunk (10%), then it means that the total of the absolute difference would be 0, which divided by two is still 0. To give an example of another extreme difference, if all 240 instances of *cat* occurred in the tenth chunk, then this would result in a 10% difference in the nine first chunks, and a 90% difference in the last chunk. This adds up to a total of 180%, which yields a deviation of proportions of 0.9. Note that in this specific example, the maximum value is 0.9, but it is easy to imagine that if there were more chunks, for example 100 chunks, then each chunk would represent one percent. As a result, if all occurrences of *cat* were in the last chunk in that example, then there would be 99 chunks where the difference is one percent and one where the difference is 99, which would result in a deviation of proportions of 0.99. This shows that 1 is only a theoretical maximum and that this maximum can vary depending on the number of chunks. This has actually been addressed by Lijffijt and Gries (2012). The authors propose a way to normalize the deviation of proportions (DP) measure in the following way:

$$DP_{norm} = \frac{DP}{1 - \frac{1}{n}}$$

To go back to the previous example where the maximal deviation of proportions was 0.9, using 10 chunks, one gets 0.9 divided by 0.9, which amounts to 1, the actual maximal deviation of proportions. This is however a minor concern that is mostly important from a

theoretical perspective. In the subsequent studies, what matters is to compare different values of deviation of proportions that are all computed on the same corpora, using the same chunks. The normalization process was therefore not necessary, as all values under comparison have the same potential maximum.

In chapters 5 and 6, chunks have the size of 1'000 words (0.001% of a one hundred million word corpus), which provides a rather detailed chunking. Furthermore, as most corpora are divided into smaller files, one has to decide whether chunking of the corpus takes this into account or not. In the present analysis, the chunking did not take the file structure of the corpus into account. For example, if a chunk consists of 500 items at the end of a file, then the first 500 items of the next file are going to be included into this chunk. This seems to be the best option, since deciding to end chunks with the end of files could theoretically result in several very small chunks that could be as small as a single word, which is not optimal.

To give an actual example taken from the British National Corpus written data discussed in section 5.4, the articles *the* and *a* have the most even distributions with deviation of proportions of 0.117 and 0.125 respectively. On the other hand, the maximal deviation of proportions value (0.998) is obtained by low frequency elements such as the pronoun *ya*, which is rarer in the written portion of the corpus and which can be expected to occur specifically in more informal contexts.

Note that deviation of proportions is one way of measuring dispersion and that other researchers might disagree that a uniform distribution such as the one presented above applies to natural languages. Indeed, it has been suggested that many words tend to appear in bursts in certain contexts (Altmann et al. 2009) and that all words are bursty to some extent (Pierrehumbert 2012). An example would be the pronoun *she*, which can be expected to appear in a very even fashion, but will in fact mostly appear in parts of a text where there is a female referent. Therefore, the claim is that no actual word has a perfect even distribution as shown above and that alternative methods where bursts are also taken into account should be used. This issue illustrates that dispersion can be measured in other ways than the one adopted in the subsequent studies.

Dispersion relates to grammaticalization because it can easily be observed that grammatical elements are more evenly distributed than lexical elements (Hilpert and Correia Saavedra 2017a). From an intuitive perspective, it seems plausible to assume that grammaticalized words are generally more evenly dispersed, since they are not related to a specific topic or domain and are thus more likely to occur in all parts of a text.

This has also been indirectly investigated by Pierrehumbert (2012), who studied the burstiness of words based on their semantic class. Four semantic classes were distinguished, ranging from less to more abstract: entities (e.g. *Bible, Africa, Darwin*), predicates/relations (e.g. *blue, die, in, religion*), modifiers/operators (e.g. *believe, everyone, forty*), and higher level operators (e.g. *hence, let, supposedly, the*) (Pierrehumbert 2012: 5). In terms of grammaticalization, these four classes would rank in the same way (although classes two and three might be shifted around), from less to more grammaticalized. Pierrehumbert's main finding was that the more abstract forms tend to be less bursty. Therefore, it seems plausible that grammaticalized forms would also be generally less bursty, and therefore more evenly dispersed.

It therefore seems worthwhile to investigate whether dispersion can be used to measure the degrees of grammaticalization of a given item. As shown above, the even distribution of grammatical elements might be due to semantic criteria as well. Furthermore, as pointed out in Hilpert and Correia Saavedra (2017a), where dispersion was modelled using polynomial regression, frequency outscales semantic generality as a predictor in such a model. The even distribution of grammatical elements might therefore also be a frequency effect. Under this assumption, grammatical elements are evenly distributed because they are generally highly frequent and not because of their grammatical function resulting from the process of grammaticalization. This shows that the subsequent chapters need to carefully consider the relationships between all the parameters under discussion. These relationships are discussed from a synchronic perspective in chapter 5, while they are discussed from a diachronic perspective in chapter 6.

3.6 Corpus-based works on grammaticalization

This section offers a brief selection of studies that have quantified grammaticalization and also highlights which insights can be gained by quantitative approaches to this phenomenon. The aim of this section is threefold. The first aim is to establish which grammaticalization phenomena have been studied from quantitative perspectives, with particular emphasis on studies that use linguistic corpora and that focus on English. The second aim is to give a brief overview of the methods that were used in those studies. The third aim is to determine how the results of these studies have contributed to further grammaticalization research and theory.

A variety of grammaticalization phenomena have been investigated by quantitative means. The grammaticalization of auxiliaries has been a major topic of interest (Heine 1993), with modal auxiliaries in particular having been the object of quantitative research (e.g. Krug

2000, Krug 2001, Leech 2003, Tagliamonte 2004). The grammaticalization of prepositions has also been approached from quantitative perspectives (Company 2002, Rhee 2003), including complex prepositions (Hoffmann 2004, Hoffmann 2005). Discourse markers have received similar attention (Koops and Lohmann 2015), as well as articles (Sommerer 2018). From a broader perspective, various types of constructions have also been the object of quantitative studies, such as get-passives (Hundt 2001), small size nouns (e.g. *a flicker of, a speck of*) (Brems 2007), the way-construction (*he made his way through the crowd*) (Israel 1996, Perek 2018), and the *be going to* + infinitive future construction (Mair 1997, Tagliamonte et al. 2014, Budts and Petré 2016). While this list is not exhaustive, it shows that the types of elements under study belong to a broad range of grammatical categories.

Many of these quantitative approaches involve the retrieval of all instances of a given grammatical pattern in a corpus and an analysis of their frequencies. For instance, the functions of modal auxiliaries have often been studied from distributional perspectives, with emphasis on competing forms (e.g. *have to, need to, must*). Studies of the frequencies of such auxiliaries, as well as the distributions of their variants (e.g. their negative variants) has helped with the categorization and definition of auxiliaries (Krug 2011). For instance, Krug (2000, 2001) used corpus-based data to illustrate the emergence of modal auxiliaries that had previously often been considered marginal members of that category. Similarly, Hundt (2001) used corpus-based data to determine whether certain patterns of get-passives were frequent or marginal, to better assess their relevance in this grammaticalization process. Frequency of use is a simple empirical way to determine whether a linguistic item can be considered as marginal and also gives more fine-grained information about this marginality. It can therefore help with classification and with the identification of prominent patterns.

Changes in complementation patterns (Rudanko 2006, Vosberg 2006, de Smet 2013) have also been investigated by similar methods, with a focus on relative frequencies. For example, many English verbs can take complements and it is often observed that the relative frequencies of these complements can change over time (Hilpert and Mair 2015: 182-186). Verbs such as *start* and *begin* sometimes take a to-infinitive (e.g. *I'll start to work*) as complement, whereas they can sometimes take an ing-clause (e.g. *I'll start working*). Mair (2002) and Hilpert (2011) are examples of studies that rely on relative frequencies to determine which alternatives are preferred in cases where there are competing verbal complements. While verbal complementation might not always be regarded as grammaticalization per se, it can also reflect changes in the functions and semantic aspects of the verbs that they complement, which is often related to grammaticalization.

Studying relative frequencies can show how a pattern can take over another one and can also give information regarding the period and duration of the process. Relative frequencies can help identify whether change is fast- or slow-moving, for instance by showing that an alternative may only be preferred in 20% of cases, but that this number increases to 60% ten years later. In addition, the quantitative nature of the approach makes it possible to use statistical means to determine whether the phenomenon is linked to specific parameters. For instance, an alternative may be more preferred in certain contexts (e.g. formal/informal, written/spoken) and statistical analysis can be used to determine whether relative frequencies differ significantly from one of these contexts to another. Similarly, Gries and Hilpert (2010) have investigated the preference for *-(e)th* and *-(e)s* third person singular present tense inflection between the Late Middle English to Early Modern English period. One of the relevant variables that was highlighted is the type of verb (lexical versus grammatical), as more grammatical verbs (e.g. *do*, *have*) had a higher proportion of *-(e)th* inflections (e.g. *doth*, *hath*).

Examples of studies involving token frequency were discussed in section 3.1. However, type frequency has also been used to investigate how grammatical patterns change over time. For example, Israel (1996) investigated changes in the way-construction (e.g. *He pushed his way through the crowd*) by counting the different possible verbs that can occur in such a construction, on the basis of the Oxford English Dictionary. Increases in the number of possible verbs generally denotes a development of the construction and the diversity of associated verbs can also indicate further grammaticalization (see section 3.3 on collocate diversity). While the number of verbs involved in the way-construction increases over time, these verbs also tend to involve more semantic diversity (Perek 2018). The initial way-construction mostly involves verbs of motion (e.g. *to claw your way out of a hole*), but more recent uses involve semantically broader examples (e.g. *to cheat your way through college*). Note that there is disagreement regarding the grammatical status of the way-construction and whether it should be considered as an instance of grammaticalization. As mentioned in section 2.1, grammaticalization is considered by some researchers as a broad concept that comprises many different phenomena, whereas other researchers will advocate for a narrower view. This issue goes outside of the scope of the present section and detailed discussion can be found in Noël (2007: §3).

Corpus-based approaches of this type of phenomenon benefit from the fact that using algorithms to retrieve such instances will generally highlight many more examples than a

manual search. Furthermore, quantitative methods such as semantic vector space modelling (Turney and Pantel 2010) can be used to establish semantic links between words using collocates (i.e. words that have similar meanings tend to have similar collocates). This type of approach is a replicable way to formally determine semantic similarity or diversity for a given cluster of elements and is also less reliant on personal interpretation. An example of using semantic vector space modelling for the study of grammaticalization is Hilpert and Correia Saavedra (2017b), which tested the hypothesis according to which asymmetric priming is an explanation for the unidirectionality of grammaticalization (section 2.4).

The semantic aspects of grammaticalization have also been investigated by means of questionnaires and asking respondents to evaluate the degree of semantic complexity of given linguistic items. For instance, Rhee (2003) asked participants to rate the complexity of use of certain prepositions, as well as the diversity and clarity of their meanings (e.g. is it hard to pinpoint the specific meaning?). Another way to measure semantic generality introduced by this study is the use of dictionaries and counting the number of semantic designations for each entry of the prepositions under investigation. A relevant finding was that the judgement of the participants was consistent with the dictionary-based approach. There are therefore alternative methods to the use of corpora for the quantitative study of meaning and grammaticalization.

Collocation patterns have also been extensively used to study the development of grammatical categories. In the framework of constructional grammar (section 2.7), this type of approach is usually called collostructional analysis (Stefanowitsch and Gries 2003) and focuses on the attractions/repulsions between specific elements and the constructions that they are part of. This type of analysis uses measures based on mutual information, where frequency of co-occurrence is also weighted using the individual overall frequency of the elements involved. This can highlight changes in patterns of associations over time, which can be used to investigate changes in different ways from the relative frequencies discussed above (e.g. Hilpert 2008: 34-48, Torres Cacoullos and Walker 2011).

The studies mentioned so far mainly deal with the quantification of grammatical change and of certain instances of grammaticalization. However, when it comes to developing a measure of degrees of grammaticalization, few attempts have been made. A notable study is Petré and Van de Velde (2018) who propose a measure of grammaticalization that is tailor-made for the *be going to* + *infinitive* construction and that can be applied to single attestations. To determine the degree of a given attestation, the proposed approach is to check whether the attestation displays one of eight symptoms (or features) of grammaticalization of this specific construction. These eight symptoms are divided into syntactic (4 symptoms) and

semantic (4 symptoms) categories. The resulting score corresponds to the sum of all the displayed symptoms, which means that the maximum score is eight (i.e. displaying all possible symptoms). Once a score has been established for each individual instance, then an average grammaticalization score can be computed. The overall logic that more features correspond to higher degrees of grammaticalization is in line with works such as Lehmann (2002) or Boye and Harder (2012: 33). Note that the grammaticalization score is standardized (section 4.1.2 gives an example of the relevance of standardization processes) in order to make the final score independent of the number of symptoms. In short, if one wants to use the same approach with a grammaticalization process that only has six symptoms, then standardizing this score makes comparison possible with the current process and its eight symptoms.

The eight grammaticalization symptoms were mainly established on the basis of previous research. An example is the symptom of adjacency, which is a syntactic feature related to Lehmann's (2002) parameter of bondedness (section 2.2.1). Adjacency is considered as a feature of more advanced grammaticalization in the same way that Lehmann (2002) considers stronger bondedness as correlating with higher degrees of grammaticalization. For instance, an utterance such as *I'm going to the cinema* does not involve any insertion between *going* and *to*, which is a case of adjacency. This utterance would therefore receive +1 to its score for the symptom of adjacency. In contrast, an utterance such as *I'm going now to the cinema* has an element between *going* and *to*, which means that it would be considered as a case of non-adjacency and would therefore get +0 to its score. Considering adjacency as a feature of higher degrees of grammaticalization is easily motivated in the case of the *be going to + infinitive* construction, given the existence of *gonna*, which is an instance of coalescence between *going* and *to*. Adjacency can be regarded as "paving the way" for this coalescence.

This approach is however fairly different from what is proposed in this dissertation, since it is tailor-made for a specific construction, whereas the present aim is to have a general approach that works for most situations. For instance, one of the semantic symptoms of the grammaticalization of the *be going to + infinitive* construction is whether it involves motion or not. There are many cases of grammaticalization where motion is not a consideration at all, which shows that the approach needs to be tailored for each case separately. A further note is that Petré and Van de Velde (2018) also use their measure to determine degrees of grammaticalization across different authors over time (in the context of the study of grammaticalization at the level of the individual), which is a different goal from the present

endeavour. Another difference is that in this dissertation, the links between each variable and degrees of grammaticalization is obtained empirically (chapters 5 and 6), instead of being determined beforehand.

The studies presented above use a broad range of corpus-based methods, but what does their quantitative aspect bring to grammaticalization research? As pointed out at the beginning of this section, the works by Krug (2000, 2001) have quantified the marginality of modal auxiliaries in English, as well as studying their different uses. This has resulted in a clearer classification of what may be regarded as an emerging modal. Krug (2001: 309) has proposed that given the similar properties of the emerging modals under investigation, they should constitute an intermediate category on the verb-auxiliary cline. This illustrates that frequency-based approaches are particularly useful when dealing with gradual concepts, and can also help with classification. This is why chapter 5 focuses on the notion of grammaticalization as a gradual concept that can be approached by quantitative means. The aim is to have an empirical way of stating that an element is more grammaticalized than another, which can in turn also be used to classify elements as highly or lowly grammaticalized.

The use of large datasets is also a common feature of many of the studies mentioned in this section. Since the term large is somewhat subjective, it should be clarified that in the present case, it is used to describe datasets that could not be fully analysed manually and require automated means to be investigated. This type of data is particularly useful when testing general hypotheses regarding grammaticalization. Generalizations tend to be a delicate matter, since it is often easy to find counter-examples. Linguists sometimes use the term regularities instead (e.g. Traugott and Dasher 2001: 81-88, Diewald and Smirnova 2010). Testing generalizations is, to a large extent, testing whether they hold true for a large number of elements, and whether exceptions are in fact rather marginal. An example discussed previously was the investigation of the asymmetric priming hypothesis (section 2.4). Another example is the parallel reduction hypothesis (Bybee et al. 1994) which states that phonological reduction tends to occur in parallel to semantic reduction. Rhee (2003) was able to provide support for this hypothesis by using quantitative measurements of semantic generality, phonological reduction and grammaticalization (where degrees of grammaticalization were approximated by token frequency) for a set of 80 common English prepositions. This constitutes a larger set of elements than what could be dealt with using more qualitative means and is a more robust way to test generalizations such as the parallel reduction hypothesis (although this particular study tested this hypothesis on prepositions

exclusively). The present dissertation attempts to take a step in this direction with the hypothesized unidirectionality of grammaticalization (section 2.4) in chapter 6.

Most of the corpus-based studies discussed in this section use token frequency and patterns of collocations to investigate grammatical change, which further motivates the use of such variables in the present dissertation. In addition, many of these studies have investigated links between other quantifiable variables and grammatical change, which shows that having a quantitative measure of grammaticalization could facilitate such investigations. The next chapter illustrates how the links between the variables in the subsequent studies can be established by statistical means. The main methodological tools used in the dissertation are introduced, as well as the corpora from which the data has been retrieved. The main objective of the following chapter is to explain how the variables introduced in the present chapter are used to compute a grammaticalization score.

4. Data and methods

This chapter introduces the main statistical methods of the dissertation (section 4.1), as well as the corpora from which the data is retrieved (section 4.2). It offers a detailed explanation of each method and includes examples that illustrate how they can be used. The first sections focus on binary logistic regression, which is central to the studies conducted for this dissertation. Section 4.1.1 starts with the basic notion of what a regression analysis is, using a simple linear regression as an example to introduce the key concepts of coefficients and significance tests. Binary logistic regression is then introduced as an extension that relies on the same key concepts in section 4.1.2. The equations involved in these types of analysis are also presented, but not in a formal mathematical way. The idea is to show exactly how the different values are calculated, as this is of practical interest, but not so much how the equations are motivated, as this is more relevant to mathematicians than to linguists. Mixed effect models are also discussed as an alternative in section 4.1.3. Then, section 4.1.4 introduces the notion of principal component analysis and how it can be used to better understand the relationships between the variables of a dataset. Finally, section 4.2 provides information regarding the two main corpora used in the subsequent studies, namely the British National Corpus (BNC) and the Corpus of Historical American English (COHA).

4.1 Statistical methods

4.1.1 Linear regression

A regression analysis is a process by which the relationship between several variables is studied (Baayen 2008, Gries 2013, Levshina 2015). The aim is usually to predict how a variable changes when other variables also change. The variable which we are trying to predict is called a dependent variable, while the other variables are called independent variables. These other independent variables are also called predictor variables. Their relationship is described by parameters (henceforth marked as a_i) that are used in a regression equation. The most basic equation for a linear statistical regression using one dependent variable and one predictor variable is as follows:

$$y = a_0 + a_1 * x$$

In this notation, y is the dependent variable, x is the predictor variable, a_1 is the regression coefficient of x, and a_0 is a constant. In such a model, one can predict the value of y when given the value of x, using the parameters a_0 and a_1. These parameters are calculated on the basis of empirical data. They are calculated using the least squares method, which finds an equation that minimizes the difference between the values predicted by the model and the actual values that are observed empirically. A number of additional independent variables can be added to the model (x_1, x_2, …, x_i), each with their own coefficient, resulting in the following formula for a multiple linear regression:

$$y = a_0 + a_1*x_1 + a_2*x_2 + ... + a_i*x_i$$

Let us look at a concrete example. We can suppose that there is a link between how fast someone can type text on a smartphone and how many hours a day they spend using their device. It seems probable that more time spent results in a faster typing speed, which can be empirically investigated. We will measure typing speed by counting the number of characters that someone can type in 30 seconds. It may be possible to predict with a certain degree of accuracy how fast someone can type based on the time they spend on their phone on a daily basis. In this instance, typing speed is the dependent variable (y), whereas the number of hours is the predictor variable (x). Table 6 presents a fictional dataset of 30 people who had a spying app installed on their phone which monitored their typing speed and how much time they spent on their phone every day. The dataset was designed in a way to make sure that there would be a significant link between the two variables. As such, a list of 30 random hours between two and twelve was drawn. The typing speed was computed by multiplying the number of hours by 5, and by adding a random number between 5 and 10 to add some noise to the data. Figure 1 shows the dataset from Table 6 plotted onto a two-dimensional graph. It is apparent that there is a linear relationship between typing speed and the number of hours spent on one's phone. The linear regression model attempts to find the best way to draw a line that passes through the dots as accurately as possible. This is done by finding the (vertical) line for which the space between the line and each dot is minimal. This is achieved through a mathematical method called least squares, which finds the most optimal solution to minimize the squared difference between the observed values (the dots) and the ones predicted by the model (the line). The difference between an observed value and the one predicted by the model is also called a residual. It should be mentioned that the dots might not always form a

line and that another kind of relationship may be at play (e.g. polynomial, logarithmic, exponential), which requires a transformation of the basic regression equation.

ID	TypingSpeed	NbHours	ID	TypingSpeed	NbHours
Person_1	65	12	Person_16	37	6
Person_2	65	12	Person_17	34	5
Person_3	63	11	Person_18	33	5
Person_4	61	11	Person_19	35	5
Person_5	57	10	Person_20	28	4
Person_6	58	10	Person_21	25	4
Person_7	50	9	Person_22	25	4
Person_8	50	9	Person_23	25	4
Person_9	51	9	Person_24	25	4
Person_10	45	8	Person_25	27	4
Person_11	47	8	Person_26	20	3
Person_12	48	8	Person_27	23	3
Person_13	43	7	Person_28	19	2
Person_14	40	7	Person_29	19	2
Person_15	37	6	Person_30	17	2

Table 6. Typing speed and daily number of hours spent on phone of 30 imaginary participants.

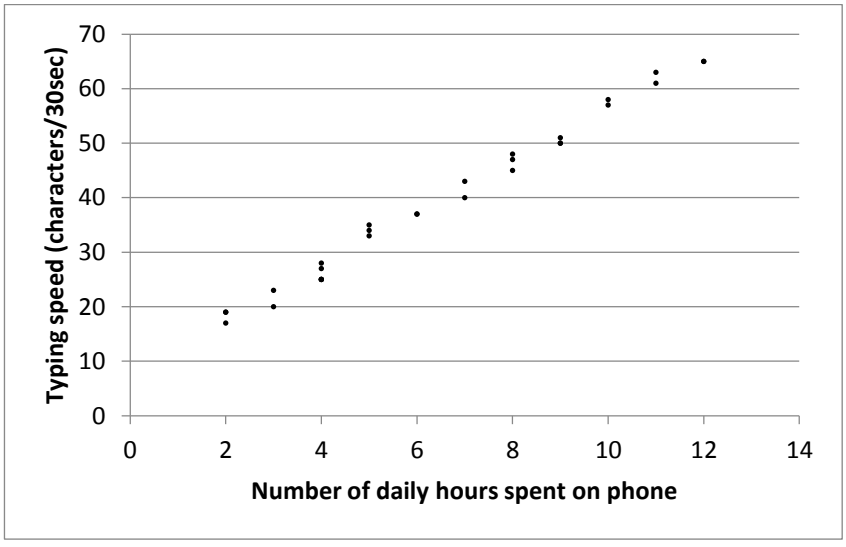


Figure 1. Typing speed and daily number of hours spent on phone.

A linear regression was conducted on this dataset and the main results are reported in Table 7. The coefficient (column B) that corresponds to the daily number of hours is 4.860. The

independent variable is thus positively correlated with the dependent variable, which means that more time spent on the phone is associated with faster typing speed, and vice-versa. This type of relationship is a linear one, which means that one additional hour of phone use corresponds to an increase of 4.860 characters per 30 seconds. This coefficient corresponds to the slope of the line obtained by the least squares method. The constant corresponds to the intercept of that line (i.e. where this line cuts the typing speed axis) and therefore to the basic typing speed of someone who does not use a phone at all. It was expected that the coefficient would be close to 5, given the way that the dataset was created (i.e. by creating a typing speed based on the number of hours multiplied by 5, plus a random number between 5 and 10).

The S.E column corresponds to the standard error of each coefficient. The model can only estimate the coefficients (versus their actual, true value) and the standard error tells us the margin of error of this estimate. The significance test, which corresponds to a t-test in the case of a simple linear regression, is computed on the basis of this standard error. It determines whether the standard error is small enough to guarantee that the coefficient is different from zero. This is crucial because if the coefficient is equal to zero, its associated variable will not matter at all in the equation, since it is multiplied by this coefficient. This is known as the null hypothesis, which corresponds to the situation in which there is no link between the dependent variable and the predictor variable. The p value found in the Sig column corresponds to the probability that the coefficient is actually equal to zero and that therefore the null hypothesis holds true. In the present case, this probability is lower than 0.001, which means that it is highly unlikely that the coefficient associated with the number of hours is equal to zero, and by extension that there is no link between this variable and typing speed. Note that p values must be multiplied by 100 to be interpreted as a percentage, so a p value lower than 0.001 means a probability lower than 0.1%. It should also be noted that choosing which minimal p value to use in order to reject the null hypothesis is a subjective choice. Common thresholds are 10%, 5%, 1% and 0.1%.

	B	S.E	t	Sig.
NbHours	4.860	0.092	52.957	0.000
Constant	7.635	0.657	11.626	0.000

Table 7. Coefficients of the linear regression between typing speed and daily number of hours spent on phone.

These p values indicate which coefficients are relevant to the model, which however does not inform us regarding its overall quality. A general way of assessing the quality of a regression model is to look at the R^2 score, which in the present case amounts to 0.990. This is a very high score as the maximum value is one. The R^2 corresponds to the amount of variance in the dataset that can be explained (i.e. predicted) by the model. Here the model can almost perfectly predict the actual values. The R^2 is calculated using the residuals of the model, namely the difference between the predicted and the actual values. For example, using the linear regression equation, we can predict the score of Person 1, who spends 12 hours a day on their phone, in the following way:

$$y = 7.635 + 4.860*12 = 65.955$$

Therefore, this prediction of 65.955 is different from the actual score of 65, and the difference between the two (65.955-65=0.955) is its residual. As discussed earlier, residuals corresponds to the distance between an observed data point and the regression line, and the model is computed in a way that minimizes the sum of the (squared) residuals. These residuals should be distributed in an even way, as the model may otherwise be problematic. For example, it could happen that the model regularly overestimates (i.e. makes higher predictions than the actual values), which is why it is often necessary to look at the distribution of the residuals to assess the quality of the model. However, this is not necessary in a binary logistic regression and will not be discussed further here.

4.1.2 Binary logistic regression

The basic linear regression model presented in the previous section is appropriate for a simple case of numeric variables which display a linear relationship. However, as there are different types of variables and relationships, there are also different types of regressions. There are three main types of variables: numeric, categorical and ordinal. Numeric variables correspond to quantitative values and are represented by real numbers. They are described as continuous variables, because they can take an infinite amount of possible values. Speed and time measurements are examples of such variables. Categorical variables, on the other hand, can only take a fixed number of values that can be assigned on the basis of qualitative criteria. They are considered to be discrete variables. Eye colour is an example of such a variable, as it can only take a set number of values (e.g. brown, blue, green, grey), although taking heterochromia into account would make for a larger number of possibilities. Ordinal variables

are somewhat in between numeric and categorical variables in that they correspond to categorical variables that can be ordered. For example, a satisfaction variable that can take the values “very good”, “good”, “bad” and “very bad” is ordinal because they imply a notion of order.

In the typing speed example above, the variables are numeric and display a linear relationship, which means that a basic linear regression model can be used. In this dissertation however, the dependent variable is a categorical one, namely whether an item is grammatical or lexical. This is why the type of regression that has to be used is a binary logistic one. A binary logistic regression is a type of regression in which the dependent variable only has two possible outcomes that are also mutually exclusive, such as pass/fail, alive/dead or win/lose. The aim is to predict which of these two outcomes is the most likely on the basis of a set of predictor (independent) variables. Note that these predictor variables can be a mixture of numeric and categorical variables. The outcome of the model is a score between zero and one, which depends on the initial encoding of the variables. Therefore, if pass is encoded as one and fail is encoded as zero, then the score will correspond to the likelihood of passing. A cut-off point of 0.5 can then be established, as any value above this point is more likely to be a pass. In terms of equation, the basic principle is the same as the linear equation, namely that it involves a dependent variable, (at least) one predictor variable and a corresponding number of parameters that are determined empirically. However, because a binary logistic regression deals with a categorical dependent variable with two outcomes and their probabilities, it uses the following logistic equation:

$$\mathrm{p}(1|\mathrm{X}) = \frac{e^{a_0+a_1*x_1+a_2*x_2+\ldots+a_i*x_i}}{1+e^{a_0+a_1*x_1+a_2*x_2+\ldots+a_i*x_i}}$$

The element p(1|X) reads as “the probability of 1 given X”, in which 1 corresponds to the element encoded as 1 (e.g. pass). This equation will always result in a value between zero and one, as it is a property of the logistic function. As one can notice, the elements in the exponent (i.e. $a_0 + a_1*x_1 + a_2*x_2 + ... + a_i*x_i$) correspond to the basic regression equation introduced earlier. Therefore, the bigger coefficients have more weight in the final score.

To illustrate how a logistic regression can be used to investigate a categorical variable, a simplified example will be discussed. Let us say that we are interested in investigating the differences between academic and non-academic texts. In this example, we assume a binary categorical variable, i.e. the type of text, which can only take two mutually exclusive values,

namely academic and non-academic. A binary logistic regression can help us predict the type of a text based on a number of predictor variables. In the present case, three predictor variables are included in the model: type-token ratio, average sentence length and average word length. Type-token ratio is a textual measure obtained by dividing the number of different types by the total number of tokens in a text (Baayen 2008). A higher type-token ratio corresponds to a more diverse vocabulary (see also section 3.3). Average sentence length consists in the average number of words per sentence, and average word length consists in the average number of letters in each word. This is a toy example and a real study of such a question would obviously require more variables and a more thorough discussion.

A simulated dataset was constructed in order to illustrate what a binary logistic regression can do. This imaginary dataset contains fifteen academic texts, and fifteen non-academic ones. The values for the type-token ratio were made deliberately higher for the academic texts, while the two other variables were randomized. This was done so that the type of text can mostly be predicted on the basis of its type-token ratio, while the other variables are virtually useless. Table 8 shows this simulated dataset in detail.

A binary logistic regression was conducted on the data presented in Table 8, using the type of text as dependent variable. Note that academic texts were encoded as one, whereas non-academic texts were encoded as zero. Type-token ratio, average sentence length and average word length were used as predictors. The binary logistic regression can calculate a regression score for each text, which ranges from zero to one. Texts closer to one (i.e. above 0.5) are more likely to be an academic text, whereas texts closer to zero are more likely to be non-academic.

Table 9 shows the regression coefficients for each variable (Column B), the standard error of these coefficients (Column S.E), their Wald coefficient (Column Wald) and their p value (Column Sig). The result is that one of the three predictor variables emerges as highly significant, namely type-token ratio, as it has a p value of 0.003. It means that there is only a 0.3% chance that this coefficient is actually equal to zero and therefore irrelevant to the model. This is an expected result, as this variable was designed for this purpose. The other predictor variables are not significant, with p values that are quite high ($p>0.2$). The p values are computed on the basis of the Wald score, where a higher Wald score means a more significant p value. The Wald score is used in the same way as the t-test for a linear regression (section 4.1.1). It also uses the standard error of the coefficient to determine whether its actual value is different from zero.

ID	Type of text	Type-token ratio	Average Sentence Length	Average Word Length	Binary Logistic Regression Score
Text_1	Academic	0.640	9.480	4.390	0.996
Text_2	Academic	0.610	9.140	4.320	0.988
Text_3	Academic	0.590	8.990	5.290	0.923
Text_4	Academic	0.590	8.970	5.580	0.888
Text_5	Academic	0.580	9.460	5.310	0.949
Text_6	Academic	0.580	8.440	5.940	0.644
Text_7	Academic	0.580	8.690	5.790	0.761
Text_8	Academic	0.560	8.850	5.960	0.672
Text_9	Academic	0.550	8.080	4.110	0.850
Text_10	Academic	0.530	9.720	5.130	0.917
Text_11	Academic	0.530	9.950	4.110	0.983
Text_12	Academic	0.510	8.340	4.210	0.746
Text_13	Academic	0.510	9.930	4.900	0.927
Text_14	Academic	0.420	9.780	5.930	0.257
Text_15	Academic	0.310	8.690	4.180	0.051
Text_16	Non-Academic	0.550	8.780	5.610	0.698
Text_17	Non-Academic	0.530	9.190	5.430	0.773
Text_18	Non-Academic	0.480	8.810	5.600	0.331
Text_19	Non-Academic	0.460	8.150	4.870	0.232
Text_20	Non-Academic	0.420	8.270	4.240	0.246
Text_21	Non-Academic	0.400	8.760	4.680	0.195
Text_22	Non-Academic	0.390	9.260	5.740	0.093
Text_23	Non-Academic	0.380	8.860	4.760	0.138
Text_24	Non-Academic	0.370	9.510	4.470	0.327
Text_25	Non-Academic	0.350	8.310	5.600	0.012
Text_26	Non-Academic	0.350	9.310	5.250	0.077
Text_27	Non-Academic	0.340	9.140	4.460	0.125
Text_28	Non-Academic	0.330	9.770	5.070	0.117
Text_29	Non-Academic	0.310	8.430	4.240	0.032
Text_30	Non-Academic	0.300	9.820	5.310	0.050

Table 8. Imaginary dataset of academic and non-academic texts.

Type-token ratio is the only relevant variable in the model, which means that analysing its coefficient will inform us about its relationship with the dependent variable. In the present case, the coefficient is positive (22.855), which means that an increase in the type-token ratio tends to increase the likelihood of a text being categorized as academic by the model. The specific regression score for any given text is computed by using the binary logistic regression equation presented above. For example, the data for Text_4 results in the following score:

$$p(1|\text{Text_4}) = \frac{e^{-17.517+22.855*0.590-1.291*5.580+1.484*8.970}}{1+e^{-17.517+22.855*0.590-1.291*5.580+1.484*8.970}} = \frac{e^{2.075}}{1+e^{2.075}} = 0.888$$

As can be observed from the formula, knowing which coefficient has the most weight is not straightforward in the present situation. Indeed, while the type-token ratio coefficient is the biggest one (22.855), type-token ratio values range from 0.300 to 0.640, which means that this will result in a smaller value after multiplication. Conversely, average sentence and word lengths have small coefficients (1.484 and -1.291, respectively), but they multiply with bigger values ranging from 4.110 to 9.950. A way of getting rid of this problem is to standardize the dataset. Standardization is a process by which variables are re-scaled so that their values are expressed as the number of standard deviations above (or below) the mean. This allows comparison between variables that use different measurement units. Standardizing the dataset and running the binary logistic regression again results in the coefficients shown in Table 10. The first observation is that the Wald score and the p values are not affected by standardization. The second observation is that the type-token coefficient is the one with the largest magnitude, as it is approximately three times bigger than the other coefficients.

	B	S.E	Wald	Sig.
Type-token ratio	22.855	7.764	8.666	0.003
Average Sentence Length	1.484	1.217	1.486	0.223
Average Word Length	-1.291	1.050	1.512	0.219
Constant	-17.517	11.190	2.451	0.117

Table 9. Coefficients of the binary logistic regression model conducted on Table 8.

	B	S.E	Wald	Sig.
Type-token ratio	2.432	0.826	8.666	0.003
Average Sentence Length	0.831	0.682	1.486	0.223
Average Word Length	-0.811	0.659	1.512	0.219
Constant	0.109	0.573	0.036	0.849

Table 10. Standardized coefficients of the binary logistic regression presented in Table 9.

The scores calculated on the basis of these coefficients may be used to discuss the degree to which a text is academic or non-academic. For example, Texts 1, 2, 3, 5, 10, 11 and 13 are extreme examples of academic texts, as they get values above 0.9. However, some texts do not get scores that correspond to their actual categories. Indeed, Texts 16 and 17 have

higher scores than other non-academic texts, with 0.698 and 0.773 respectively. They both have a type-token ratio value that is higher than the other non-academic texts, which explains why the model gives them higher scores.

There is therefore a continuum of "academicness" on which each text can be placed. This continuum consists in scores between zero and one. These scores are a fine-grained type of information, as they can take an infinite amount of possible values in this given range. The initial dataset used a coarse-grained type of information, namely the type of text, which could only take two possible values, i.e. academic or non-academic. The binary logistic regression makes it possible to use this coarse-grained information to compute a more fine-grained one, and also to highlight the more extreme members of the initial categories. This is why this method is useful when trying to place an item on a lexical to grammatical continuum.

These fine-grained values can be, if one wishes to, classified back into coarse-grained categories. A binary logistic regression calculates, as stated earlier, the probability of a pass/fail event. Therefore, a score above 0.5 is an above-chance of a pass, whereas a score under this threshold is a fail. In terms of "academicness" score, it means that a text above 0.5 is classified as an academic text, and a text under this score is classified as a non-academic one.

From a practical point of view, this information can be used to test the correct classification rate of the binary logistic regression. Indeed, it is possible to compare the category attributed given the regression score with the actual type of text. Table 8 shows that there are two academic texts (14 and 15) that have a score under 0.5, which means that the model classifies them as non-academic. There are also two non-academic texts (16 and 17) that have a score above 0.5, which means that the model classifies them as academic. These four texts therefore get a classification by the model that is different from their actual type. It means that, when it comes to predicting the type of text, the model has a correct classification rate of 26/30 (86.67%). These classifications are usually displayed in a classification matrix (also called a confusion matrix), which is shown in Table 11.

There are also general indicators of the quality of the model. The first one is a chi-square test of the coefficients of the model, which in the present case corresponds to 19.034 ($DF=3$, $p<0.000$), which means that there is indeed a link between the type of text and the variables under investigation. Furthermore, the R^2 (percentage of variance explained by the model) can also be computed for a binary logistic regression model, however it is only an approximation. The two main ways to calculate an approximation of the R^2 score for such a model are the Cox & Snell R^2 (Cox and Snell 1989) and the Nagelkerke R^2 (Nagelkerke

1991). They amount to 0.470 and 0.626, respectively. They range from zero (poor) to one (excellent), so the values observed here are rather average. An important note is that there is a lot of debate regarding the accuracy and relevance of these indices (e.g. Menard 2000, Smith and McKenna 2013). But these values are not necessarily critical as one can also look at the classification rates to have some idea of the quality of the model. In the present case, the model does have a high correct classification rate.

		Predicted type		
		Academic	Non-Academic	
Actual type	Academic	13	2	86.67
	Non-Academic	2	13	86.67
Total percentage of correct classification				86.67

Table 11. Classification matrix, observed types of texts versus predicted types of texts by the binary logistic regression.

One of the limitations of the basic binary logistic regression model presented so far is that the two variables that do not have significant scores are useless to the model. There are cases where some variables may even worsen the model. This is due to a situation known as multicollinearity, which happens when predictor variables are strongly correlated with each other. In such a case, a strong correlation means that these predictors are redundant to the model, as they bring the same kind of information to it. This situation can result in a major reduction of the significance scores of some of the variables in the model. This is why it is often useful to conduct a stepwise binary logistic regression.

In a stepwise regression, the variables are added to the model (or removed from it) in several steps. After each step, a test is made to determine whether the new iteration of the model is better than the previous one, until the best fit is found. There are different stepwise options for a binary logistic regression (conditional, maximum likelihood, Wald) which are all equally reliable (Hosmer et al. 2013: 125). In the present case, the Wald stepwise method is used, as the Wald score determines the significance score of the parameters of the model. This will be useful for the main study of the dissertation, since determining which parameters are the most relevant ones is a central question. The stepwise method can use either the ascending or the descending order, which means that it can either start with one variable and add new ones based on the Wald score, or instead start with all variables and remove the ones that have

the lowest scores in each step until the model does not improve anymore. Let us look at the previous example using a stepwise binary logistic regression to illustrate these concepts.

The data from Table 8 was used in a stepwise binary logistic regression using the ascending Wald method. The results are reported in Table 12. First, only the type-token ratio variable was added to the model. None of the other variables were considered as having a Wald score high enough to be added to the model in a useful way. The stepwise regression does therefore not add unnecessary variables to the model. Second, the p value of the type-token ratio variable has slightly changed from 0.003 to 0.002, which means that this version of the model is more confident that the link found between type of text and type-token ratio is significant. Third, the coefficient has also changed, from 22.855 to 18.373, but this is also a minor change as this coefficient is of similar magnitude and has a positive sign. An additional remark is that using the descending Wald method leads to the exact same result as Table 12. This is because in the present case, there are only three variables which can be added or removed from the model. An important note though, is that the descending Wald method gets to this result by starting with all the variables, and then determines that the average sentence and word lengths variables should be dropped.

	B	S.E.	Wald	Sig.
Type-token ratio	18.373	5.962	9.496	0.002
Constant	-8.679	2.901	8.951	0.003

Table 12. Coefficients of the stepwise binary logistic regression, using the ascending Wald method.

This simple example investigating the link between type of text and three predictor variables relates closely to the idea of using lexical and grammatical items in order to determine a grammaticalization score. Indeed, while there are highly lexical and highly grammatical items, some items are somewhere in between. The score attributed by a binary logistic regression can help us identify these individuals and to place them somewhere along a continuum, as was done with the academic and non-academic texts presented in Table 8. The main difference has to do with the selection of the relevant variables, as the example presented above is an imaginary dataset with a limited number of constructed variables. Since the main study of the dissertation contains a larger number of variables and their relationships are not straightforward, a general method of investigating these relationships is presented in section 4.1.4.

4.1.3 Mixed effects modelling

The linear regression presented in section 4.1.1 and the binary logistic regression presented in section 4.1.2 are types of regression that use the same slope and intercept for all individuals in the experiment. For instance, the fifteen imaginary participants who had their typing speed and number of hours spent on a phone recorded in Table 6 all displayed a similar relationship between these variables. However, the situation could have been less clear-cut. In a real dataset, it would be very likely that some participants are outliers, for example by having a high typing speed despite spending very few hours on their phones, and vice-versa. As it is impossible to study full populations, researchers work on samples. For instance, psycholinguistic studies are trying to investigate mental representations and processing of language, and they can only do so by looking at a restricted set of the human population.

Since there can be a lot of individual variation between different people, new statistical methods have been developed to account for this aspect. This phenomenon is not limited to studies that involve human participants, but any kind of "subject" in the statistical sense. Words can be the subjects of a study and also have their own idiosyncratic behaviours which need to be taken into account. One of the methods that deal with these issues is called generalized linear mixed effect models (Baayen 2008, Jäger 2008). The aim of this method is to take subjects' idiosyncrasies into account by featuring them as "random effects" in the model. An important note though is that random effects can sometimes be superfluous. For instance, there are cases where including a random speaker effect does not improve a generalized linear mixed model (e.g. Baayen 2008: section 7.4). Mixed effect models will be presented as an alternative statistical analysis to the datasets presented in chapter 5.

4.1.4 Principal component analysis

When dealing with a large number of variables, a principal component analysis can gauge the redundancy of a dataset and can try to summarize several variables into a smaller number of components. The method starts with a correlation matrix between all the variables and consists in finding out which ones are strongly correlated together in order to create a new set of variables (components) that replaces those variables. The purpose of the method is to make a complex dataset more easily comprehensible by compressing it into a lower number of components than the number of initial variables.

This principle is particularly efficient when variables strongly correlate. Let us imagine a questionnaire which asks respondents to agree or disagree with several statements by giving

a score ranging from 1 (strongly disagree) to 5 (strongly agree). The following five statements are presented:

1. You prefer to stay at home over going on a night out.
2. You can easily talk to people that you don't know.
3. You like raclette.
4. You are embarrassed to sing karaoke with your friends.
5. You like fondue.

It is very apparent that there are two sets of questions that are related. On the one hand, statements (1), (2) and (4) relate to social settings and introvert/extrovert behaviours and we can assume that respondents will tend to give similar answers to these three questions. On the other hand, the other two relate to liking a certain type of cheese-based meal and the assumption is that someone who likes raclette will also be likely to like fondue. A principal component analysis can investigate whether responses to (1), (2) and (5) tend to be correlated together, while (3) and (5) are also correlated together, but that the two sets do not necessarily correlate with each other. The principal component analysis can show that (1), (2) and (4) are well summarized by one component, while (3) and (5) can be summarized by another. Researchers can name and interpret the components as they like.

Let us now look at this situation from a more practical point of view and consider actual numbers. Again, fifteen imaginary participants had to respond to the five statements and their imaginary results are reported in Table 13. A principal component analysis provides a lot of output to analyse and only the outputs used in chapter 5 are going to be discussed here. The first most basic element is the initial correlation matrix (Table 14), which is itself not an output of the principal component analysis but rather its starting point. A correlation matrix is a symmetrical matrix, since the correlation between variable Q1 and variable Q2 is the same as the correlation between variable Q2 and variable Q1. This is why only one half of the matrix is filled with the correlation values, which makes it easier to read as appears from Table 14. Furthermore, the correlation of a variable with itself is one, as can be seen in the diagonal of the matrix. Correlations range from zero (no correlation) to one (strong correlation). The sign of the correlation indicates positive or negative relationships. For example, Q1 and Q2 have a strong negative correlation (-0.827), which means that people who agree with statement (1) tend to systematically disagree with statement (2). The strongest

correlations in Table 14 are in bold and one can already observe that Q1, Q2 and Q4 are strongly correlated together, while Q3 and Q5 are also strongly correlated together.

Participant ID	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**
Participant_1	1	5	1	2	2
Participant_2	5	2	2	5	2
Participant_3	1	5	1	2	1
Participant_4	1	4	2	1	1
Participant_5	3	1	5	4	5
Participant_6	4	3	4	3	5
Participant_7	4	1	5	5	4
Participant_8	4	2	4	4	4
Participant_9	5	2	5	4	4
Participant_10	5	1	3	5	3
Participant_11	1	4	1	2	1
Participant_12	4	1	2	5	1
Participant_13	1	4	5	3	5
Participant_14	2	5	5	2	5
Participant_15	4	1	2	4	2

Table 13. Participant responses to each statement, with scores ranging from 1 to 5.

	Q1	Q2	Q3	Q4	Q5
Q1	1	**-0.827**	0.317	**0.866**	0.237
Q2		1	-0.280	**-0.891**	-0.133
Q3			1	0.296	**0.924**
Q4				1	0.192
Q5					1

Table 14. Correlation matrix of the dataset from Table 13. Strongest correlations in bold.

We may now move to the actual results of the principal component analysis. The first output is presented in Table 15 and shows the percentage of variance explained by number of components. What this output shows is how well the dataset can be summarized by a certain number of components. Here, 60.895% of the information in the dataset is still retained if instead of all five variables, only one component is used. Similarly, if two components are used, then 93.163% of the information in the dataset is retained. In short, using two components almost gives the same information as using all five variables. This is a form of data compression. The rest of the table shows the percentages for three, four and five components. It stops at five components since this is equal to the number of variables at the

start. In a principal component analysis, one needs to decide how many components to choose. There are several ways to make this decision, and statistical software such as SPSS offers a variety of options. The default option is to only retain so-called eigenvalues that are superior to one. Eigenvalues are obtained by a process that is called spectral decomposition that operates on the correlation matrix. This process will not be explained in detail here, as it is a rather complex algorithm and is outside the scope of this discussion. What matters is that in the present case, this process would yield the selection of two components. Deciding how many components to keep can also be done intuitively in the present case, since adding a third component only adds 3.602% of information, which is much lower than the previous values.

In chapter 5, two components will generally by retained because this makes it possible to give a clear visual representation of the dataset, as illustrated in Figure 2. This representation is based on Table 16 which is a component matrix (also called a saturation matrix). It shows how the components relate to the initial variables and works like correlations (from zero to one, negative or positive). For instance, Table 16 shows that the first component mostly represents Q1, Q2 and Q4 (values in bold), while the second mostly represents Q3 and Q5 (values in bold). This is how one can decide to name component 1 “social profile” and component 2 “cheese affinity”. Communalities indicate whether a given variable is well-represented by the two components together. In the current situation, all variables are well-represented, as the smallest one is 0.886.

Component	Initial Eigenvalues		
	Total	% of variance explained	% cumulated
1	3.045	60.895	60.895
2	1.613	32.268	93.163
3	0.180	3.602	96.765
4	0.105	2.099	98.864
5	0.057	1.136	100

Table 15. Percentage of variance explained by number of component.

Regarding the interpretation of the visual representation in Figure 2, what is of interest are the angles that a pair of variables forms with the origin (center) of the graph. Right angles (90°) tend to correspond to correlations close to zero. Small angles (<45°) correspond to positive correlations, while wide angles (>135°) correspond to negative correlations. Variables Q3 and Q5 form a very small angle and are therefore strongly positively correlated.

The same can be said about variables Q1 and Q4. On the other hand, Q2 forms a very large angle with Q1 and Q4, which means that it is strongly negatively correlated with them.

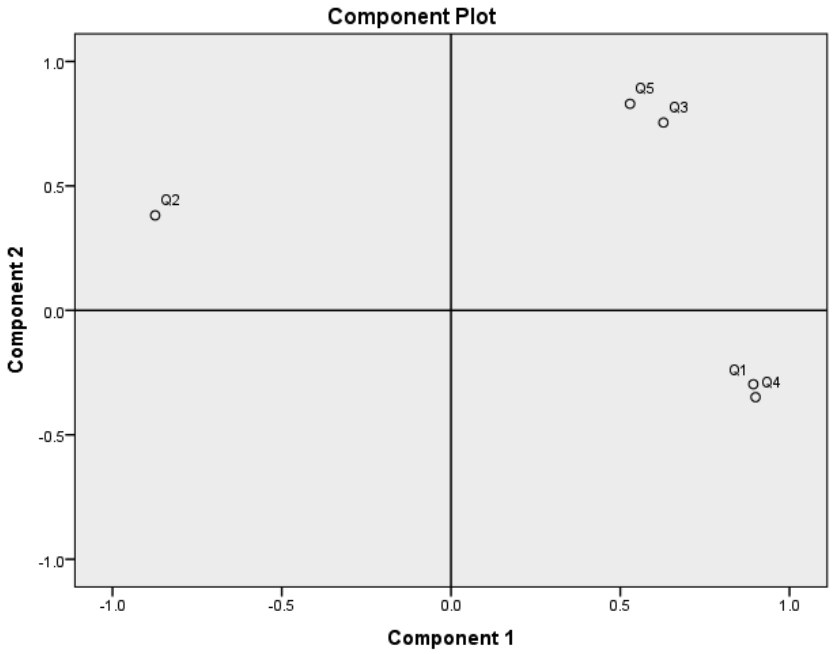


Figure 2. Component plot, using two components

	Component 1	Component 2	Communalities
Q1	**0.893**	-0.297	0.886
Q2	**-0.874**	0.382	0.910
Q3	0.628	**0.755**	0.963
Q4	**0.900**	-0.349	0.931
Q5	0.529	**0.830**	0.968

Table 16. Component matrix and communalities. The left columns show how much each variable contributes to each component. The communalities column shows the percentage of variance of each variable that is accounted for by the model.

While the situation presented here is rather clear, real datasets tend to have components which are less easily interpretable. The main interest of principal component analysis for the datasets presented in chapter 5 is to determine to what extent the variables correlate. It is important to know, for example, if collocate and colligate diversities correlate with frequency, and if so, to what extent. Indeed, there is a difference between weak and strong correlations. The percentage of variance explained by a number of components (Table 15) is a good way to

evaluate redundancy in a dataset, as it can show how many components are required to reach a similar amount of information.

A further note is that researchers often apply principal component analysis in order to then use the components themselves as substitutes for the variables. This is not the preferred solution in the subsequent studies because components, as shown above, may summarize several variables together, but not always in a clear-cut way. For instance, Q3 plays a role in both components one and two at the same time, which makes its individual contribution hard to isolate. Therefore, principal component analysis is primarily used in the subsequent studies to better understand the relationships between the variables themselves, instead of using the components as substitutes for the variables.

4.2 Corpora

The aim of this section is to provide information regarding the corpora used in the subsequent studies. Information will be provided regarding several key aspects, namely the constitution of the corpus (i.e. which types of texts do they involve), its tagging, and other various technical details. The first sub-section introduces the British National Corpus, that is prominently featured in chapter 5. The second sub-section introduces the Corpus of Historical American English, which was used for the study presented in chapter 6. Links to the official website of these corpora are provided under the "corpora" sub-section of the references.

4.2.1 The British National Corpus (BNC)

The British National Corpus is a 100 million words corpus that was compiled from 1991 to 1994. It mostly contains data from 1975 onwards, although there were several texts from the 1960s. It was designed as a representative corpus of British English, which means that it includes a broad variety of texts. The main distinction is between its written and spoken components, which amount to 90% and 10% of the corpus, respectively. The written component mostly comprises books and periodicals. They consist of imaginative texts (e.g. novels), but also informative texts on various topics (e.g. social sciences, arts, natural sciences, politics). They key point is that many different domains are represented. The spoken component is mostly sub-divided into two main categories, the "demographically sampled" part (4 million words), and the "context-governed" part (6 million words). The former consists of randomly selected British English speakers in the United Kingdom who recorded their everyday conversations and the latter consists of more formal speeches. Examples of these speeches involve lectures, business meetings, political speeches, sermons, parliamentary

proceedings, legal proceedings, sports commentaries. Again, a broad range of contexts is covered.

The original BNC was tagged using an automatic tagger called CLAWS4 (Garside 1987) and an additional tagging tool called Template tagger (Fligelstone et al. 1997). The "BNC Basic Tagset" contains 57 tags, and 4 additional punctuation tags. Note that the tagset itself is called "C5", which stands for CLAWS5. For example, the tag <*NN1*> marks singular nouns, the tag <*NN2*> marks plural nouns, and the tag <*NP0*> marks proper nouns. Automatic taggers make mistakes, which is why it is necessary to estimate the error rate based on a manually-analysed sample. In the case of the BNC, the tagger has a precision of 96.25%, which means that it only attributes a wrong tag 3.75% of the time. In addition, the tagger may use ambiguity tags, which further increases the accuracy of the corpus. Indeed, when the tagger is unsure which tag to attribute, it can offer two tags, such as <*VVG-AJ0*> which denotes ambiguity between a gerund or an adjective.

Another important aspect of the BNC is that different versions have been released. The subsequent studies use the first version that was released in 1994. A second version called "BNC World" was released in 2001 which was fairly similar and mostly improved upon the tagging of the corpus, using an enhanced version of CLAWS4. The BNC reference guide (Burnard 2007: section 6.1) notes however that "in most respects the word class information provided by the corpus now is identical to that provided with the first release of the BNC in 1994." Using the first or second version therefore does not differ greatly. Furthermore, a third version was released in 2007, which is known as the "BNC XML edition". XML is a formatting language that is particularly used with corpora. This third edition mostly differs from the previous ones in terms of formatting, especially when it comes to the tagging of multiword elements (e.g. *as well as*), but not so much in terms of content. In addition, the BNC-BYU interface developed by Davies (2004) constitutes yet another version of the BNC, because the corpus was entirely re-tagged using the CLAWS7 tagger. This tagger is the same one used by the Corpus of Historical American English, which is presented in the next section.

4.2.2 The Corpus of Historical American English (COHA)

The Corpus of Historical American English (COHA) is a 400 million words diachronic corpus of American English compiled by Davies (2012), between 2008 and 2010. It contains data that ranges from 1810 to 2009. In a similar fashion to the BNC, the COHA is a general corpus and includes a large number of genres. However, these mostly consist of written

genres. The COHA has four main genres, namely fiction, magazines, newspapers, and non-fiction. Fiction is the most represented genre as it accounts for approximately 50% of the dataset, while magazines are second and represent around 25% of the dataset. Non-fiction comes third with 15%, and newspapers is the least-represented genre with around 10%, which is also partly due to the fact that newspapers only appear from the 1860s onwards. It is therefore debatable whether the COHA is a balanced corpus, especially in the early years. This is why the data presented in chapter 6 will generally discard these early years for the interpretation of the results.

Another important aspect of the corpus that makes the early years more difficult to analyse is that the amount of data per decade is not the same from one decade to another. In particular, the early decades are much smaller than the rest, as can be observed in Table 17. Figure 3 provides a clearer visualization of the situation and shows that the first two decades have much less data than the others. Significant increases can be observed between 1860 and 1880, but also between 1910 and 1920. These aspects are fairly relevant in sections 6.3 and 6.4, since the frequency of several items is studied. An easy fix to the varying sizes of the chunks is to calculate the frequency "per million words" for a given element.

Decade	**Number of words**	**Decade**	**Number of words**
1810s	1'181'022	1910s	22'655'252
1820s	6'927'005	1920s	25'632'411
1830s	13'773'987	1930s	24'413'247
1840s	16'046'854	1940s	24'144'478
1850s	16'493'826	1950s	24'398'180
1860s	17'125'102	1960s	23'927'982
1870s	18'610'160	1970s	23'769'305
1880s	20'872'855	1980s	25'178'952
1890s	21'183'383	1990s	27'877'340
1900s	22'541'232	2000s	29'479'451

Table 17. Number of words in each decade of the COHA.

For example, if a word occurs 5'000 times in the 1810s, one needs to divide this number by 1.181, as there are 1.181 million words in that chunk. This results in a frequency of 4'234 per million words. If this word occurred 5'000 times in the 1820s, then its frequency per million words would be much lower, namely 722. Despite this adjustment, smaller collections of texts are less representative overall as they involve a smaller variety of texts, which can skew the frequency of a given word. This was briefly touched upon earlier by mentioning that

the first decades do not include newspapers for example. This is why expressing the words as frequency per million words is only a partial solution to the problem.

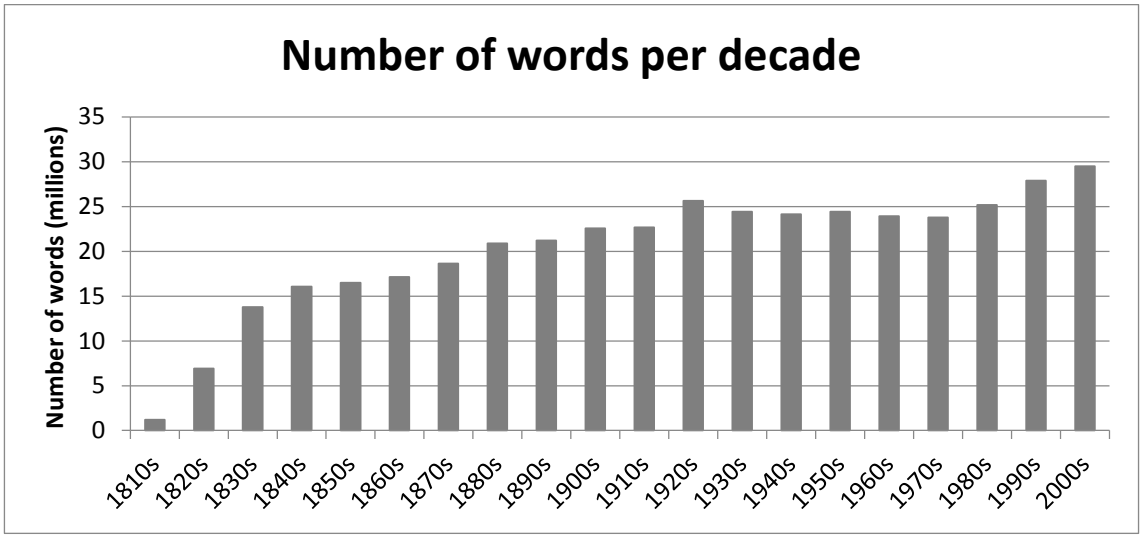


Figure 3. Number of words in each decade of the COHA.

The tagging of the corpus is done using CLAWS7. It is relatively similar to CLAWS4 and CLAWS5, which were used with the original BNC and introduced in the previous section. The main difference are the codes used for parts of speech. For example, the codes *AJ0* (unmarked adjective)*, AJC* (comparative)*,* and *AJS* (superlative) were used for adjectives in the CLAWS5 tag set, whereas the CLAWS7 codes for these adjectives are *JJ*, *JJR,* and *JJT* respectively.

While the texts involved in the BNC can be accessed in full, the COHA includes copyrighted texts, which therefore means that they cannot be legally distributed in full. To circumvent this problem, the texts were transformed by deleting ten words every 200 words. This means that it is much more difficult to read those texts in full, but that at the same time most of the data is unaffected (95% of the data is still there). Deleted words are replaced by the symbol @ in the corpus. Also, given that these deletions occur regularly and indiscriminately in the text, it affects all words equally. These deletions are therefore unproblematic for the broad quantitative analyses presented in chapter 6.

Note that the COHA comes in different formats when one wants to access its full content. The study conducted in chapter 6 uses the "wlp" format of the corpus, which stands for word, lemma, part of speech. In this format, each word of the corpus is presented on one line that consists of three elements separated by tabulations. The first element is the word

itself as it appears in the text, the second element is its corresponding lemma, and the third is its CLAWS7 part of speech. Also, a sample of the COHA that uses the same format is available freely and consists of 3.6 million words, which roughly corresponds to one percent of the whole corpus. This sample was used to conduct some of the case studies presented in section 6.2, because its smaller size makes processing easier.

Now that the main tools have been introduced, a concrete application regarding grammaticalization is presented in the next chapter. This chapter discusses a synchronic approach to measuring grammaticalization. A binary logistic regression model, as presented in section 4.1.2, is used to compute a grammaticalization score on the basis of the parameters introduced in chapter 3. This score ranges from zero (highly lexical) to one (highly grammatical). The model relies on data taken from the British National Corpus, which was introduced in section 4.2.1 above.

5. Quantifying grammaticalization in synchrony

5.1 Introduction

This chapter deals with the notion of grammaticalization from a synchronic perspective. The focus is therefore on the gradience of grammaticalization, and the main purpose is to support the idea that the degree of grammaticalization of different words or constructions can be quantified and compared. However, as discussed in chapter 2, there is currently no systematic way to quantify these degrees of grammaticalization. The aim of this chapter is to propose a quantitative approach that fills this gap. In chapter 3, several variables that are known to play a role in grammaticalization were introduced. The purpose of the current chapter is to build a model that uses those variables in order to compute a grammaticalization score. To achieve this goal, this chapter discusses a binary logistic regression model (section 4.1.2) that is trained using a synchronic dataset of 264 lexical and 264 grammatical elements.

The aim of a binary logistic regression is to predict the probability that an element belongs to one of two categories under observation, using a score that ranges between 0 and 1. These categories have to be dichotomous and mutually exclusive (e.g. *pass/fail*). The appeal of the binary logistic regression is that the score it computes will show whether an element is an extreme member of its category (i.e. score close to 0 or 1) or whether it is instead somewhere in between. In the present case, lexical and grammatical are considered as the two possible categories. The idea is to train the model using a list of lexical and grammatical elements (section 5.2), where grammatical elements are encoded as 1, and lexical elements as 0. Therefore, when the model computes a score for a given element, a higher score corresponds to a higher degree of grammaticalization.

In addition to a formal way of attributing a grammaticalization score to a given element, another, more theoretical, aim of constructing such a model is to assess how well the variables that were described in chapter 3 allow us to differentiate between lexical items on the one hand and grammaticalized items on the other. The model can shed light on the variables that are relevant predictors of grammaticalization by determining the main prototypical characteristics of each group (e.g. grammatical elements have higher frequencies). It should be pointed out that while the initial classification of items is limited to two mutually exclusive categories (i.e. grammatical vs. lexical), this is not in contradiction with the cline-based view of grammaticalization (section 2.3) because the fitted values are on a cline between 0 and 1.

Therefore, the initial classification is a coarse-grained approach to the phenomenon, but it results in a fine-grained score when used to train the binary logistic regression model.

As this approach is rather novel, it is necessary to try different implementations. Three main versions are presented in this chapter. The default version uses the written portion of the British National Corpus (BNC) and excludes punctuation from the analysis. Since collocate diversity (section 3.3) is considered as a potentially informative variable of grammaticalization, including punctuation or not might change the results significantly. This is why another version of the study also uses the written portion of the BNC, but includes punctuation. In addition to these two versions, a third one that uses the spoken portion of the BNC (excluding punctuation) is also presented in this chapter. Spoken data includes more informal variants, some of which are quite common in speech, but rare in written form. For example, using the BNC-BYU interface, 16'050 out of 16'676 occurrences of *cos* (*because*) occur in the spoken BNC. Also, the spoken portion of the corpus is smaller (10 million words vs. 90 million words), which is why it is important to treat it separately.

This chapter is structured in the following way. Section 5.2 explains how the 264 lexical and 264 grammatical elements were selected in order to constitute a database that was used to train a binary logistic regression model. Section 5.3 discusses how the relevant measures were retrieved from the original BNC files. Sections 5.4, 5.5 and 5.6 present the three different versions of the binary logistic regression model. All three versions are structured in the following way. They start with a sub-section that discusses the relationships between all the variables using a principal component analysis (section 4.1.4). Then, the second sub-section focuses on the parameters of the binary logistic regression. The third sub-section is a discussion which provides explanations for these results and how they further our understanding of grammaticalization. Since section 5.4 presents the first version of the study, it also includes an additional section (section 5.4.4) that discusses the inclusion of mixed effects into the binary logistic regression. Section 5.7 discusses shortcomings with the classification of certain elements, in particular modal verbs. Section 5.8 then provides an overview of the results by comparing the three versions. Section 5.9 concludes this chapter by discussing the relevance of these results for research on grammaticalization.

5.2 Data

This section discusses how a dataset of lexical and grammatical elements was established using the first version of the British National Corpus (section 4.2.1). The grammatical

elements were selected first, since there are fewer types of grammatical items as compared to types of lexical items. The initial selection method made use of the part of speech tagging of the BNC in order to classify items into the two categories, as summarized in Table 18. Reliance on part of speech tagging is mostly motivated by the fact that, as discussed in section 2.1, the lexical end of the grammaticalization cline (i.e. content words) mostly consists of nouns, verbs, adjectives and adverbs. Similarly, function words are located on the grammatical end of the cline and include prepositions, articles, determiners, conjunctions and pronouns. To get this part of the dataset, the interface of the BNC (BNC-BYU) developed by Davies (2004) was used. The reason is that the BNC-BYU interface offers a quick and effective way to retrieve the most frequent elements for a given part of speech tag. The queries that were used to gather the data are referenced in the BNC-BYU search column of Table 18. It should be noted that all instances of *be*, *have*, *go*, *do* and *will* were classified as grammatical elements given the frequent role of these verbs as auxiliaries.

Part of speech tags			
Lexical	**BNC-BYU search**	**Grammatical**	**BNC-BYU search**
Nouns	[n*]	Prepositions	[i*]
Verbs (**be/have/go/do/will*)	[v*]	Articles	[at*]
Adjectives	[j*]	Determiners	[d*]
Adverbs	[r*]	Conjunctions	[c*]
		Pronouns	[p*] and [app*]
		Negations	[xx*]

Table 18. Categorization of the part of speech tags and their respective BNC-BYU search queries. Note that the verbs *be*, *have*, *go*, *do*, and *will* are not included in the lexical category, given their common function as auxiliaries. Instead, all their forms were counted as grammatical items.[12]

In order to be included into the database, items had to have a minimal frequency of 1'000 in the whole BNC, and they also had to appear at least once in the spoken portion. The aim of this study is to observe the features of grammaticalization, and as such, it seems appropriate to select items that are common, in order to gather more data. Furthermore, items with low frequencies might be spelling mistakes or parsing errors. For example, the word sequence *state-of-the* is sometimes considered as a single word in the BNC. The sequence is

[12] Note that since the data was collected in 2016, the BNC-BYU interface has changed. The new search terms are however very similar. For instance, instead of *[v*]* for all verbs, the new syntax is *_v**.

wrongly tagged as a determiner, and it occurs only four times in the corpus. Including forms like these into the dataset would thus be misleading.

To make processing easier, words that included unusual characters or punctuation were not included. There were only three cases omitted due to this restriction: *and/or*, *v.* and *no-one*. Additionally, the conjunction *given* was also not included in the dataset because it is not directly tagged as a conjunction.[13] This resulted in the selection of 264 grammatical items. The same number of lexical elements was then selected, using a frequency ranking. The most frequent nouns, verbs, adjectives and adverbs from the BNC were ranked according to their token frequency, and only the first 264 of these items were kept. Again, the motivation is to include widespread and common items in the database. The total number of items in the database is therefore 528. Part of speech distribution is shown in Table 19. Nouns constitute the majority of lexical elements, while a large number of grammatical elements are prepositions. The complete list of the 528 items can be found in appendix A1.

Part of speech	**Tags**	**Numbers**	**Total**
Nouns	N** / CRD	101 / 1	102
Verbs (lexical)	V**	66	66
Verbs (grammatical)	V**	23	23
Adjectives	AJ*	34	34
Adverbs	AV*	62	62
Prepositions	PRP / PRF	105 / 1	106
Conjunctions	CJ*	41	41
Pronouns (+possessives)	P** / DPS	47 / 7	54
Determiners + articles	DT* / AT*	32 / 5	37
Negations	XX0	3	3

Table 19. Part of speech distribution in the database.

5.3 Retrieving the relevant measures from the British National Corpus

The previous section discussed how the dataset was compiled using the BNC-BYU interface. Now that the dataset has been established, information regarding each item must be retrieved from the corpus. All the variables introduced in chapter 3, such as frequency and collocate diversity must be measured. In order to achieve this, the procedures introduced in chapter 3 were used. The script that was used to retrieve data from the BNC can be found in appendix

[13] *Given* as a conjunction in the BNC is tagged as: *<VVN> given <CJT> that*. Including it in the dataset would have involved several modifications to the all the scripts that make use of the part of speech tagging. Since *given* was the only example of such a case, it seemed reasonable to simply drop it from the study altogether. It should be pointed out, however, that it is indeed returned by the *[c*]* query in the BNC-BYU interface.

D1. There are however a few differences between the original BNC files used by those scripts and the online BNC-BYU interface that need to be addressed.

First, the BNC-BYU interface uses a very general level of tag search with the queries presented in Table 18. For example, the preposition search *[i*]* can roughly be regarded as the equivalent of the original BNC tag <*w PRP*> (section 4.2.1 provides additional details regarding the tagging of the BNC). However, the correspondence is not perfect because *[i*]* will also return instances of the preposition *of*. The subtlety is that *of* in fact has its own special tag in the BNC, namely <*w PRF*>. Other examples of the differences between the BNC-BYU interface and the original tagging system are numbers such as *million*, which have a special BNC tag <*w CRD*>, whereas they are comprised under the BNC-BYU *[n*]* search query. The reason to mention this distinction between the actual BNC data and the BNC-BYU interface is because since the latter was used to gather the initial dataset, manual processing was required to correctly tag the items in preparation for the searches used in the scripts. For example the frequency script, which measures the frequency of the items of the database in the BNC files, only counts items if their tag matches. It seems logical that if one wants to measure the frequency of the noun *work*, then instances of the verb *work* should not be counted and vice-versa, making the use of the part of speech tagging crucial. As such, the items in the database were tagged according to the initial BNC-BYU query as shown in Table 20, along with several exceptions.

Part of speech	**BNC-BYU search**	**BNC tag**	**Exceptions**
Nouns	[n*]	<w N**>	Numbers: <w CRD>
Verbs	[v*]	<w V**>	
Adjectives	[j*]	<w AJ*>	
Adverbs	[r*]	<w AV*>	
Prepositions	[i*]	<w PRP>	Preposition *of*: <PRF>
Articles	[at*]	<w AT*>	
Determiners	[d*]	<w DT*>	
Conjunctions	[c*]	<w CJ*>	
Pronouns	[p*] and [app*]	<w P**>	<DPS> for possessives
Negations	[xx*]	<w XX0>	

Table 20. Conversion from BNC-BYU search queries to actual BNC tags. The * is a wildcard, so for example the category <*w N***> includes all the nominal tags from the BNC, namely <*w NN0*>, <*w NN1*>, <*w NN2*> and <*w NP0*>.

The BNC tag column in Table 20 shows the most general version of the tag. Therefore, while the BNC uses a variety of verbal tags, for example <*w VHI*> for the infinitive of *have*, or <*w VDD*> for the past tense form of *do*, the database does not encode this level of information and simply goes with <*w V***>, where the * is a wildcard (i.e. it can replace any letter or number).

Second, all the scripts use the same regular expression used when parsing the corpus into individual words. This regular expression is shown in (1) and can be read as follows: anything that contains <*w*, followed by at least one alphanumeric character, followed by >, followed by at least one letter/apostrophe/blank space. The reasoning is that all BNC tags used for words begin with <*w* as can be seen in Table 20. Any item preceded by one of the tags from Table 20 would be correctly detected as a valid word, as long as it is not made up of anything else than letters, an apostrophe or a blank space. Examples (2) and (3) are valid because they conform to the regular expression. Example (4) is invalid because the word itself can only consist of letters, an apostrophe or a blank space. The reason to only include these characters is to remove any odd or idiosyncratic item. Example (5) is invalid because it does not begin in <*w* and is, in fact, a punctuation tag. While removing punctuation was the default choice, section 5.5 presents a version of the study in which punctuation is included.

(1) <w \w+>[a-zA-Z']+

(2) <w NP0>London — Valid

(3) <w AV0>For example[14] — Valid

(4) <w NN1>H3llo — Invalid

(5) <c PUN>. — Invalid

Third, it should be noted that the BNC-BYU presents a shortened version of several constructions. For example, the *[i*]* query in the BNC-BYU returns the item *spite*. However, searching this item using the tagging <*w PRP*>*spite* will not work because the actual encoding is <*w PRP*>*in spite of*. This is rather relevant because this adds another layer of semi-automatic cleaning of the dataset. Therefore, some items that were obtained using the initial BNC-BYU queries had to be transformed into their longer forms, in order to be

[14] Note that the tagging of such elements depends on the version of the BNC being used. In the present study, tagged elements may include a blank space such as *for example*. In other versions of the BNC, such as the XML version, a multi-word element such as *for example* would be tagged as: *<mw c5="AV0"><w c5="PRP" hw="for" pos="PREP">for </w><w c5="NN1" hw="example" pos="SUBST">example</w></mw>*. Section 4.2.1 introduces the different versions of the BNC.

correctly found in the BNC's original files. The solution to this problem was to run the frequency script mentioned above and to find items that had unusually low frequencies, or that were not even found once and to correct them. Table 21 shows five examples of such modifications. To give another salient example of this issue, the BNC-BYU query *[c*]* returns the item *not* (with a frequency of 223 in the spoken BNC), which is in fact the construction *whether or not*. This issue only affected about 20 items in the database and was rather minor. Another minor problem related to this issue is that some items actually had double entries in the database. For example, *per cent* was present twice, due to the *[n*]* query returning an entry named *per* and another one named *cent*. A look at the concordances of those two entries shows that they are both instances of *per cent* and that they display the same sentences, which means that only one entry should be present. A similar issue was found with the constructions *rather than* (as a preposition), *for example* (as an adverb) and *in case* (as a conjunction).

Initial item	**BNC correct**
<w PRP>accordance	<w PRP>in accordance with
<w PRP>addition	<w PRP>in addition to
<w PRP>means	<w PRP>by means of
<w PRP>regard	<w PRP>with regard to
<w N**>cent	<N**>per cent

Table 21. Some examples of items where there is actually more than one word after the tag, in the BNC files. The left column shows the initial item with its converted tag according to Table 20. The right column shows the proper tagging that matches the way these items are actually encoded in the BNC (first version).

Fourth, the BNC offers several ambiguity tags, such as *<w AJ0-AV0>* which denotes an item that can be categorized both as a general adjective or as an adverb. An example is given in (6) where the word *pretty* gets such an ambiguous tag. Instances of such ambiguity tags were not counted, mostly to make processing easier. It should be noted that ambiguous cases are rather rare and mostly concern adjectives, adverbs and participles. According to the BNC's official website, among all the tags in the corpus, only 4.7%[15] are ambiguity tags.

(6) Thank you, and that's written by a thirteen year old <pause> <w AJ0-AV0>pretty good ah, let's have a male voice amongst all of this. (BNC, file F7R)

[15] This number can be found on this page: http://www.natcorp.ox.ac.uk/docs/gramtag.html

There are therefore several aspects regarding tagging that can affect measures such as frequency and collocate diversity. They are however relatively minor concerns overall. The next section presents the results of the three versions of the binary logistic regression model. As it contains a large number of variables, Table 22 offers a quick guide to each variable and their labels in the subsequent tables and figures.

Label	Details	Section
FreqWrit / FreqSpok	Token frequency of the item in the BNC written/spoken	3.1
Letter_Count	The number of letters of the item, used as an approximation of phonological length	3.2
Colloc_Div_W44	Number of collocate types of an item divided by its token frequency, window of 4 to the left and right	3.3
Colloc_Div_R1	Number of collocate types of an item divided by its token frequency, window of 1 to the right	3.3
Colloc_Div_L1	Number of collocate types of an item divided by its token frequency, window of 1 to the left	3.3
MaxColliPercent_R	The percentage of the most frequent colligate, window of 1 to the right	3.4
MaxColliPercent_L	The percentage of the most frequent colligate, window of 1 to the left	3.4
DevProp	The dispersion measure named deviation of proportions, which measures evenness of distribution (Gries 2008)	3.5

Table 22. Labels of the variables and their meaning.

5.4 Version 1: Written BNC (no punctuation)

This version is considered as the default study and will therefore contain more details than the others. In this version, the written portion of the BNC is used and the words are parsed according to the criteria introduced in section 5.3.

5.4.1 Principal component analysis

As introduced in section 4.1.4, a principal component analysis (PCA) can determine whether some variables are redundant and provide similar information. Let us first have a look at Table 23 which provides the correlation matrix between all the variables in the model. The first observation is that all the collocate diversity measures are strongly correlated together, with correlations of 0.667, 0.793 and 0.895. It means that in the present case, one gains almost the same information by simply looking at the diversity of one item to the left or right, as by

looking at a larger window of four items to the right and to the left. Since the smaller windows of one word to the right and to the left are included in the bigger window of four items, this is not necessarily an unexpected result. This is supported by the fact that the right and left windows have stronger correlations with the 4-4 window (0.793 and 0.895, respectively) than they do with each other (0.667). The second observation is that deviation of proportions is strongly correlated with frequency and all the collocate diversity measures. Its correlation with frequency is negative (-0.502) which is expected since frequent words tend to be more evenly distributed within a text. To sum up, this initial observation of the correlation matrix shows that there are strong correlations between some of the variables, but that these correlations are rather straightforward to detect.

	FreqWrit	Colloc_Div_ W44	Colloc_Div_ R1	Colloc_Div_ L1	MaxColli Percent_ R1	MaxColli Percent_ L1	DevProp	Letter_ Count
FreqWrit	1	-0.267**	-0.239**	-0.196**	0.042	-0.035	**-0.502****	-0.221**
Colloc_Div_ W44		1	**0.793****	**0.895****	-0.001	0.089*	**0.740****	0.237**
Colloc_Div_ R1			1	**0.667****	0.051	-0.038	**0.586****	0.241**
Colloc_Div_ L1				1	0.037	0.128**	**0.585****	0.241**
MaxColliPer cent_R1					1	-0.065	-0.076	-0.070
MaxColliPer cent_L1						1	0.079	0.002
DevProp							1	0.382**
Letter_Count								1

Table 23. Correlation matrix between all the variables involved in the model. The highest values are highlighted in bold (** $p<0.01$, * $p<0.05$).

Based on these observations of the correlation matrix, we can expect the PCA to summarize the collocate diversity measures into one factor. Table 24 shows the percentage of variance explained by the number of components in the PCA, and shows that component 1 can already explain about 43% of the variance. Table 25 shows that component 1 mostly summarizes the three collocate diversity variables, as well as the deviation of proportions variable. Component 1 also involves the frequency and letter count variables. This is consistent with the previous observations on the correlation matrix from Table 23. Further investigation of the data reported in Table 24 shows that component 2 adds around 14% of variance explained, and Table 25 indicates that this component mostly relies on the frequency, colligate diversity on the right, and letter count variables.

As discussed in section 4.1.4, there are different ways to decide how many components to keep. In the present and subsequent cases, only two components will be retained because this allows a two-dimensional representation, which is easier to interpret. As Table 24 shows, components 1 and 2 together account for 57% of the variance, but 43% comes from component 1, which shows that further components only provide smaller additions. The results can be summarized in a two-dimensional plot, which is shown in Figure 4.

Component	Initial Eigenvalues		
	Total	% of variance explained	% cumulated
1	3.45	43.093	43.093
2	1.15	14.313	57.406
3	1.04	12.989	70.394
4	0.9	11.195	81.589
5	0.79	9.807	91.396
6	0.33	4.067	95.463
7	0.31	3.819	99.282
8	0.06	0.718	100

Table 24. Total variance explained by number of component.

	Component 1	Component 2	Communalities
FreqWrit	-0.468	0.460	0.430
Colloc_Div_W44	0.937	0.195	0.916
Colloc_Div_R1	0.829	0.257	0.752
Colloc_Div_L1	0.858	0.238	0.793
MaxColliPercent_R1	-0.020	0.702	0.494
MaxColliPercent_L1	0.101	-0.273	0.085
DevProp	0.854	-0.194	0.767
Letter_Count	0.433	-0.410	0.356

Table 25. Component matrix and communalities. The left columns show how much each variable contributes to each component. The communalities column shows the percentage of variance of each variable that is accounted for by the model.

As stated earlier, the first component shows a clear negative correlation between, on the one hand frequency, and on the other hand the three collocate diversity measures, the letter count and the deviation of proportions variables. This is shown in Figure 4 as the frequency variable is to the left of the x-axis, while the others are far to the right. The negative

correlation between frequency and the letter count variable is expected and unsurprising, as most frequent elements tend to be shorter. Similarly, the negative correlation between frequency and deviation of proportions is not a surprise. A lower deviation of proportions means a more even distribution, and frequent items tend to be more evenly dispersed, as predicted (Hilpert and Correia Saavedra 2017a). What may be less obvious is the negative correlation between frequency and collocate diversity. It could be expected that frequent items tend to have more diverse types of collocates. However, it should be recalled that the collocate diversity measures were designed in a way to take frequency effects into account. For example, if an item occurs ten times in a corpus, it is likely that most of its collocates will all be unique, since there is not much room for repetition. However, an item that occurs a thousand times is more likely to have repetitions. Therefore, when the number of types of collocates is divided by the frequency of the item under observation (see section 3.3), highly frequent items are in fact strongly penalized by this process. This can mean that the operationalization of collocate diversity might have not been good enough, but it should be noted that an attempt at removing the frequency effect of collocate diversity was necessary, since otherwise those additional measures would have been too strongly correlated with frequency, and therefore not so useful for the models.

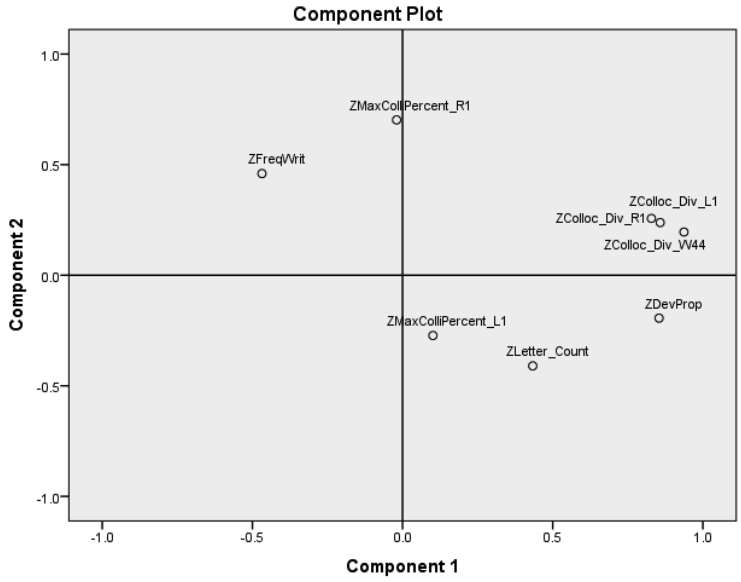


Figure 4. Component plot of the principal component analysis.

The second component (y-axis) shows a negative correlation between the two colligate based measures. Thus, more constraints on the right (translated here by a larger percentage of

the most frequent colligate), mean fewer constraints on the left, and vice-versa. There appears therefore to be a trade-off between using the left or right colligates. The reason behind this trade-off is not immediately obvious, and further investigation on how these colligate-based measures relate to grammaticalization is necessary. This issue is discussed in section 5.4.3.

5.4.2 Binary logistic regression

A stepwise binary logistic regression using the descending Wald method was conducted on the dataset presented in the previous sections. The original variables were kept instead of the PCA components in order to see their individual relevance to the binary logistic regression model. The parameters of this regression are summarized in Table 26. The first observation is that three variables were removed during the stepwise process, namely Colloc_Div_R1, DevProp and Letter_count. This is consistent with the results from the previous section which have highlighted that these variables were strongly mutually correlated, along with the other collocate diversity measures. Therefore, the regression model does not gain additional information by including these three variables. All the other variables have highly significant p values ($p<0.001$), except for MaxColliPercent_R1 ($p=0.071$). Since these p values are very small and almost identical, it is possible to look at the Wald values, if one wants to rank the variables by their significance score. The most significant ones are therefore the FreqWrit and CollocDiv_1L, which have the highest values. However, this is only partially important because, since all variables have p values under 0.001, it can safely be assumed that they are all relevant to the model.

What is of more importance are the coefficients of the model. As discussed previously in section 4.1.2, the dataset uses standardized values in order to directly compare the coefficients. What Table 26 shows is that FreqWrit is far above all the other variables with a coefficient of 10.245. This means that frequency is the most important predictor in the model and weighs about twice as much as all the other variables combined. Moreover, the positive value of the coefficient means that a higher frequency results in a higher grammaticalization score. In comparison, the collocate-based variables Colloc_Div_L1 and Colloc_Div_W44 are the second-most important variables in the model, with values of 2.218 and 2.074 respectively. Again, they are positive, which means that a higher collocate diversity value corresponds to a higher grammaticalization score. Despite a certain degree of correlation between the two variables, as discussed in the previous section, they are still both retained by the stepwise selection process.

The last two variables, MaxColliPercent_R1 and MaxColliPercent_L1 have the smallest coefficients, and therefore play a smaller role in the model. Interestingly, the coefficient of the variable MaxColliPercent_L1 is negative, whereas its right counterpart is positive. Since the colligate-based values are computed on the basis of the maximum percentage of the most frequent part of speech that occurs after a given element (section 3.4), a higher value corresponds to less diversity. As such, if the coefficient for such a variable is negative, the result is that less diversity corresponds to a lower grammaticalization score. In the present case, MaxColliPercent_1L and Collocate_Div_1L show similar trends, where more diversity translates into a higher grammaticalization score. A brief note of caution is that MaxColliPercent_R1 only has a p value of 0.071, which means that its relevance in the model cannot be ascertained with the same confidence as the other coefficients. This somewhat mirrors the situation with the collocate-based measures, as the left one is significant, whereas the right one is absent from the model. In fact, when using an ascending Wald stepwise selection method, MaxColliPercent_R1 is not retained in the final model at all, as shown in Table 27. These results indicate that left collocates and colligates are generally more relevant than their right counterparts when determining the grammaticalization score of an item and that in both cases, higher diversity is associated with a higher grammaticalization score.

	B	S.E.	Wald	df	Sig.
FreqWrit	10.245	1.464	48.955	1	0.000
Colloc_Div_W44	2.074	0.496	17.490	1	0.000
Colloc_Div_L1	2.128	0.366	33.875	1	0.000
MaxColliPercent_R1	0.222	0.123	3.260	1	0.071
MaxColliPercent_L1	-0.640	0.171	13.948	1	0.000
Constant	1.830	0.322	32.235	1	0.000

Table 26. Parameters of the binary logistic regression using the written component of the BNC, using the descending Wald method.

	B	S.E.	Wald	df	Sig.
FreqWrit	10.533	1.467	51.569	1	0.000
Colloc_Div_W44	2.125	0.502	17.906	1	0.000
Colloc_Div_L1	2.114	0.365	33.595	1	0.000
MaxColliPercent_L1	-0.654	0.172	14.485	1	0.000
Constant	1.891	0.324	34.114	1	0.000

Table 27. Parameters of the binary logistic regression using the written component of the BNC, using the ascending Wald method.

In addition to comparing the coefficients of the model, it is important to establish a general evaluation of the model. One way of evaluating is calculating the correct classification rate. The binary logistic regression has a correct classification rate of 85.0%, which is a high score, well above chance (50.0%). This classification rate is calculated by categorizing all the items with a score above 0.5 as grammatical, and the ones under 0.5 as lexical. The predicted classification is then compared with the actual initial classification. For example, the pronoun *she* has a score of 0.997 and is consequently classified as a grammatical element. Since *she* was indeed initially classified as a grammatical element, then this instance would be counted as a correct classification. In contrast, the adverb *today* got a score of 0.713, which is above 0.5 and means that it is classified as a grammatical item by the model. This counts as an instance of wrong classification, since *today* was initially categorized as a lexical element. Note however that temporal adverbs can be argued to belong to the lexical category (Boye and Harder 2012: 19), which seems to be supported by this result.

Furthermore, the Cox & Snell R^2 is 0.496 and the Nagelkerke R^2 is 0.662, which is in both cases a good fitness score of the model.[16] Another way to assess the quality of the model is to have a look at the items that receive the highest grammaticalization scores and the ones that receive the lowest ones, in order to see whether they correspond to what would be expected in those positions. Table 28 shows the items with the 100 highest and the 100 lowest grammaticalization scores. The first observation is that in both cases, the correct classification rate is particularly high, with 100% of correct classification for the top items, and 91% for the bottom ones.

[16] It should be noted that a binary logistic regression cannot be evaluated by an actual R^2 like a linear regression. Instead, a pseudo-R^2 measure is needed and there is no consensus regarding the best one for a binary logistic regression. However, the Cox&Snell R^2 and Nagelkerke R^2 are both widespread and the default options in SPSS, which is why they were chosen in the present case. A pseudo R^2 is interpreted in the same way as its regular counter-part, namely that it ranges from 0 to 1 and that it represents the percentage of variance explained by the model. A score of 1 means that all the variance is explained and that the model is an absolute perfect fit, which is hardly ever achieved when dealing with actual empirical data.

Top100	L/G	Score	Bottom100	L/G	Score
<w AT*>A	G	1.000	<w V**>MAKE	L	0.099
<w AT*>THE	G	1.000	<w AJ*>SURE	L	0.098
<w CJ*>ALBEIT	G	1.000	**<w V**>GO**	G	0.097
<w CJ*>AND	G	1.000	<w AV*>REALLY	L	0.096
<w CJ*>COS	G	1.000	<w N**>SYSTEM	L	0.095
<w CJ*>EVEN	G	1.000	<w AJ*>BEST	L	0.094
<w CJ*>LIKE	G	1.000	<w N**>BOOK	L	0.094
<w CJ*>RATHER THAN	G	1.000	<w N**>MOTHER	L	0.094
<w CJ*>THAT	G	1.000	<w N**>POWER	L	0.092
<w DPS>HIS	G	1.000	<w V**>GIVE	L	0.091
<w DT*>THIS	G	1.000	<w V**>TELL	L	0.089
<w P**>HE	G	1.000	<w N**>DEVELOPMENT	L	0.088
<w P**>I	G	1.000	<w N**>MONTHS	L	0.088
<w P**>IT	G	1.000	<w V**>GET	L	0.088
<w P**>YA	G	1.000	<w N**>PARTY	L	0.086
<w P**>YOU	G	1.000	<w AJ*>POSSIBLE	L	0.085
<w PRF>OF	G	1.000	<w N**>GOVERNMENT	L	0.085
<w PRP>ALONGSIDE	G	1.000	<w N**>LIFE	L	0.085
<w PRP>AMID	G	1.000	<w V**>KEEP	L	0.084
<w PRP>AS OPPOSED TO	G	1.000	<w N**>YEAR	L	0.083
<w PRP>AT	G	1.000	<w N**>POLICY	L	0.082
<w PRP>BUT	G	1.000	<w N**>NIGHT	L	0.081
<w PRP>BY	G	1.000	**<w DT*>SAME**	G	0.080
<w PRP>BY MEANS OF	G	1.000	<w V**>GIVEN	L	0.079
<w PRP>BY WAY OF	G	1.000	<w N**>CASE	L	0.076
<w PRP>DOWN	G	1.000	<w V**>CA	L	0.076
<w PRP>EXCEPT FOR	G	1.000	**<w V**>'LL**	G	0.076
<w PRP>FOLLOWING	G	1.000	<w AJ*>LITTLE	L	0.075
<w PRP>FOR	G	1.000	<w V**>KNOW	L	0.075
<w PRP>FROM	G	1.000	<w V**>USE	L	0.075
<w PRP>IN	G	1.000	<w V**>THINK	L	0.074
<w PRP>IN ADDITION TO	G	1.000	<w N**>AREAS	L	0.072
<w PRP>IN CONJUNCTION WITH	G	1.000	<w N**>WORDS	L	0.072
<w PRP>IN CONNECTION WITH	G	1.000	<w N**>LAW	L	0.071
<w PRP>IN LINE WITH	G	1.000	<w N**>WAR	L	0.071
<w PRP>IN RESPONSE TO	G	1.000	<w V**>FEEL	L	0.071
<w PRP>IN SPITE OF	G	1.000	<w N**>MAN	L	0.067
<w PRP>IN SUPPORT OF	G	1.000	<w N**>NAME	L	0.066
<w PRP>IN THE LIGHT OF	G	1.000	<w V**>BECOME	L	0.064
<w PRP>IN VIEW OF	G	1.000	<w N**>SIDE	L	0.063
<w PRP>INSTEAD OF	G	1.000	**<w V**>'RE**	G	0.063
<w PRP>MINUS	G	1.000	<w N**>INTEREST	L	0.061
<w PRP>ON	G	1.000	<w N**>MARKET	L	0.061
<w PRP>TO	G	1.000	<w N**>WEEK	L	0.060
<w PRP>TOWARD	G	1.000	**<w V**>GONE**	G	0.060
<w PRP>UNTIL	G	1.000	<w V**>ASKED	L	0.058
<w PRP>WITH	G	1.000	<w N**>NUMBER	L	0.057
<w PRP>WITH REGARD TO	G	1.000	<w N**>ROOM	L	0.057
<w V**>BE	G	1.000	<w AJ*>ABLE	L	0.056
<w V**>IS	G	1.000	<w N**>POINT	L	0.053
<w V**>WAS	G	1.000	**<w V**>AM**	G	0.052
<w CJ*>BUT	G	0.999	<w V**>FIND	L	0.052
<w CJ*>FOR	G	0.999	**<w V**>GOING**	G	0.051
<w CJ*>IN CASE	G	0.999	<w V**>TAKEN	L	0.051
<w CJ*>TILL	G	0.999	<w N**>HOUSE	L	0.050
<w P**>WHOEVER	G	0.999	<w N**>ROAD	L	0.050
<w PRP>CONTRARY TO	G	0.999	<w N**>WORK	L	0.050
<w PRP>DEPENDING ON	G	0.999	**<w V**>'M**	G	0.048
<w PRP>IN ACCORDANCE WITH	G	0.999	<w N**>COUNTRY	L	0.047
<w PRP>ON BEHALF OF	G	0.999	<w N**>WORLD	L	0.045
<w PRP>ON TOP OF	G	0.999	<w V**>HELP	L	0.045
<w PRP>PRIOR TO	G	0.999	**<w V**>DONE**	G	0.043
<w PRP>REGARDING	G	0.999	<w N**>COURT	L	0.042

<w PRP>TILL	G	0.999	<w N**>PROBLEM	L	0.041
<w PRP>UNLIKE	G	0.999	<w V**>KNEW	L	0.041
<w V**>ARE	G	0.999	<w V**>MEAN	L	0.041
<w V**>HAD	G	0.999	<w V**>HELD	L	0.039
<w V**>HAVE	G	0.999	<w N**>LEVEL	L	0.038
<w XX0>NOT	G	0.999	<w V**>SEEN	L	0.038
<w CJ*>NOW THAT	G	0.998	<w N**>CHILD	L	0.037
<w CJ*>OR	G	0.998	<w V**>LEFT	L	0.036
<w CJ*>PROVIDED	G	0.998	<w V**>NEED	L	0.032
<w CJ*>WHILST	G	0.998	<w V**>WANT	L	0.031
<w PRP>AHEAD OF	G	0.998	<w N**>USE	L	0.029
<w PRP>AMONGST	G	0.998	<w V**>MADE	L	0.029
<w PRP>VIA	G	0.998	<w V**>WORK	L	0.028
<w AT*>AN	G	0.997	<w V**>LIKE	L	0.026
<w CJ*>WHEREAS	G	0.997	<w N**>PLACE	L	0.025
<w DT*>NEITHER	G	0.997	<w N**>BODY	L	0.024
<w P**>SHE	G	0.997	<w N**>PERIOD	L	0.023
<w P**>THEY	G	0.997	<w N**>END	L	0.021
<w P**>YE	G	0.997	<w V**>TURNED	L	0.021
<w PRP>IN FAVOUR OF	G	0.997	<w N**>DOOR	L	0.020
<w DT*>WHICH	G	0.996	<w N**>FORM	L	0.020
<w PRP>IN CHARGE OF	G	0.996	<w N**>HAND	L	0.020
<w PRP>NEXT TO	G	0.996	<w N**>QUESTION	L	0.020
<w PRP>ON THE PART OF	G	0.996	<w N**>WAY	L	0.019
<w CJ*>EXCEPT	G	0.995	<w N**>EYES	L	0.018
<w P**>OURS	G	0.995	<w V**>LOOKED	L	0.018
<w CJ*>AS FAR AS	G	0.994	<w V**>FELT	L	0.017
<w CJ*>AS SOON AS	G	0.994	<w N**>FACT	L	0.016
<w PRP>APART FROM	G	0.994	<w V**>'D	L	0.016
<w PRP>IN RELATION TO	G	0.994	<w V**>THOUGHT	L	0.012
<w PRP>PLUS	G	0.994	<w N**>STATE	L	0.011
<w PRP>UP	G	0.994	<w N**>FACE	L	0.009
<w CJ*>AS LONG AS	G	0.993	<w N**>THING	L	0.009
<w PRP>BENEATH	G	0.993	<w V**>FOUND	L	0.009
<w CJ*>WHETHER OR NOT	G	0.992	<w V**>TOLD	L	0.007
<w PRP>IN RESPECT OF	G	0.992	<w N**>PERSON	L	0.005
<w CJ*>SUCH THAT	G	0.991	<w V**>GOT	L	0.003

Table 28. Items with the 100 highest and 100 lowest grammaticalization scores. Forms that are wrongly classified are marked in bold. Here, "wrongly classified" means that those items should have been under the 0.5 threshold if they were initially classified as lexical, and above it if they were initially considered as grammatical.

An examination of the problematic grammatical items that receive very low values shows that three of them are the clitics *'m*, *'re*, and *'ll*, as well as the first person verbal form *am*. The explanation is rather straightforward given the previous observations regarding left collocate and colligate diversity. Given the coefficients in Table 26, elements that have a high diversity of left collocate and colligate receive higher grammaticalization scores. However, for those four elements, the left diversity is very limited. Indeed, those clitics will almost always have a pronoun to their left, and *am* is usually preceded by *I*. Interestingly, the third person clitic *'s* is assigned a score of 0.611, resulting in a correct classification as a grammatical element. This is because it is common to use it with nouns, whereas the clitics

and verbs mentioned above are used with more restrictions regarding what appears to their left.

Also among the wrongly classified elements in the bottom 100 items are instances of the verbs that were chosen as exceptions in the initial classification, namely *be*, *have*, *go*, *do*, and *will*. In the present case, this corresponds to *go*, *gone*, *going* and *done*. As stated in section 5.2, all the instances of these exceptional verbs were initially categorized as grammatical. This choice is rather problematic because a form such as *go* or *gone* rarely has an auxiliary function. While there are some exceptions and cases in between, such as serial *go* constructions in *go get coffee* (Flach 2015), most instances of these forms are lexical uses. The reasoning to classify all those exception verbs from section 5.2 as grammatical elements was to use a rather strict and formal distinction. It was also done to avoid including a subjective aspect to the selection, when deciding on each item individually. Also, the initial categorization of all instances of *go* as grammatical elements was motivated by the fact that *go* has a much broader range of meanings than just physical movement, as for instance in semi-copular uses such as *he's gone crazy*. Nevertheless, the binary logistic regression correctly identifies the elements *go*, *gone* and *done* as being lexical rather than grammatical elements. As such, it is quite remarkable that the model detected how problematic the initial classification of *go* was. In the case of *going*, the situation is different, because it is one of the cases where classifying it as a grammatical element makes sense given the future *be going to* + infinitive construction. In the present case, it is identified as a lexical element, again because of the left collocate and colligate effect mentioned earlier. As *going* takes part in the future *be going to* + infinitive construction, it often has the verb *be* to its left, which results in a lower diversity of both collocates and colligates to the left. The future *be going to* + infinitive construction is further discussed in chapter 6.

In addition, one can also look at the average grammaticalization score of parts of speech in general to assess the quality of the model. Table 29 shows that nouns have on average the lowest grammaticalization scores, followed by lexical verbs. Adjectives and adverbs get higher values, which is in line with the notion that nouns and verbs constitute the major lexical category (Hopper 1991, section 2.2.2). Interestingly, adverbs being at 0.408 is consistent with the fact that some researchers tend to point out that adverbs are good examples of the gradation of grammaticalization because they can be partly lexical and grammatical (Traugott and Trousdale 2013: 73, Boye and Harder 2012: 19). Grammatical verbs are also close to the middle of the cline, which is due to the inclusion of several

predominantly lexical forms as discussed above (e.g. *done*). The other grammatical categories get high values as expected, with the exception of determiners, which get a similar score as the grammatical verbs. On average, Table 29 shows that no category is strongly misrepresented by the results of the model.

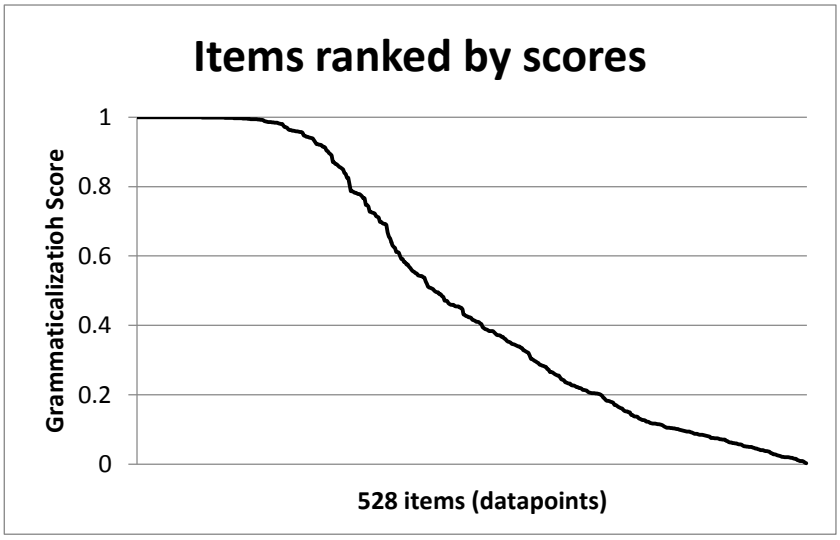


Figure 5. Distribution of the grammaticalization scores of the 528 items in the database.

Part of speech category	Tag	Average score
Nouns	N**	0.129
Verbs (lexical)	V**	0.175
Adjectives	AJ*	0.267
Adverbs	AV*	0.408
Verbs (grammatical)	V**	0.540
Determiners	DT*	0.606
Pronouns	P**	0.726
Possessive pronouns	DPS	0.811
Articles	AT*	0.829
Prepositions	PRP	0.866
Conjunctions	CJ*	0.867
Negations	XX0	0.870

Table 29. Average grammaticalization score according to part of speech.

Overall, the model offers a rather good fit, despite several classification errors. It has a particularly high correct classification rate when it comes to the items with the highest and lower scores, which supports the idea that the variables involved in the model are relevant to grammaticalization. An observation of the distribution of all the scores of the items in the

model shows that the binary logistic regression ranks them in a cline, as illustrated in Figure 5. The next section discusses these results and tries to address why the left collocate diversity is more relevant to the model than its right counterpart.

5.4.3 Discussion

The main results that emerged out of the previous section may be summarized as follows:

1. The model is able to successfully distinguish between lexical and grammatical elements and is also able to provide a score between 0 and 1 which corresponds to a given item's position on the cline.
2. Token frequency is the most prominent parameter in the model. A higher token frequency corresponds to a higher grammaticalization score.
3. Two collocate diversity variables were significant, namely the four words window on both sides, and the one word window on the left. In both cases, a higher diversity results in a higher grammaticalization score.
4. The one item to the left colligate diversity variable is also relevant, but has a smaller role than the others. A higher diversity also corresponds to a higher grammaticalization score. In addition, the one item to the right colligate diversity variable is also retained in the model, but with a much smaller p value and with an opposite relationship.
5. Deviation of proportions, number of letters, and collocate diversity using one word to the right were not relevant variables in the model.

The first result means that using the variables proposed in the model, it is possible to distinguish between lexical and grammatical elements with a rather high degree of accuracy. This confirms the general relevance of the variables used in the model and the possibility to quantify degrees of grammaticalization.

The second result is not surprising given that a basic observation about the most frequent words in any English corpus is that they tend to be grammatical elements such as *the*, *a, of*, and so on. Furthermore, frequency is probably one of the only quantitative parameters that has been systematically acknowledged as playing a role in grammaticalization, although its exact role is still subject to debate (section 3.1). Thus, having frequency as the most prominent parameter in the model is in line with those observations. However, it should be noted that the data selection procedure (section 5.2) uses the most frequent elements from both categories, which implies that other data selection methods might result in frequency

being featured differently in the model. For example, frequency can be removed as a factor if one compares pairs of lexical and grammatical items with similar frequencies (e.g. Hilpert and Correia Saavedra 2017a), which could be used to investigate whether the collocate and colligate diversity measures have a larger weight under those circumstances. Furthermore, section 5.6 will show that using a smaller corpus produces different results when it comes to the prominence of frequency as a parameter in the model.

What the third result highlights is that the different collocate windows do not have the same role in the model. The left-based measure appears to be more relevant than the right one. One reason why this finding is difficult to interpret is that the principal component analysis in section 5.4.1 has shown that there are strong correlations between the collocate-based measures. Therefore, one reason why the one word to the left collocate variable is retained in the model and not the right one could be that they provide the same information, and that the model determined that the left one gives slightly more information, resulting in the removal of the right one from the model. This would imply that if the left collocate measure was absent from the model, the right collocate measure could act as a substitute. Furthermore, since the four words window measure is also relevant and correlates with two other collocate-based measures, this means that when it comes to collocate diversity, the chosen window is not extremely crucial, as different windows yield similar information. In short, collocate diversity does matter to determine the grammaticalization score, but the chosen window seems rather flexible. Therefore, collocate diversity is a relevant parameter in the model which means that it is concomitant to grammaticalization. One explanation is that grammatical words tend to have more collocate diversity around them because they tend to occur in a broader variety of contexts (Heine and Kuteva 2002: 2, Heine 2002). It should be noted that this is not fully in line with Lehmann's parameter of syntagmatic variability (section 2.2.1) where an increase in grammaticalization tends to correspond to a decrease in the freedom to move an item in its context.

The fourth result deals with colligate diversity and the left-based measure is more relevant than the right one. Again, a higher diversity results in a higher grammaticalization score, meaning that overall grammatical elements have a wider range of parts of speech that occur to their left. In contrast, the right counterpart of the measure has the opposite tendency, namely that grammatical elements tend to have less diversity with regard to parts of speech to their right. A potential explanation for this tendency is that many grammatical elements are heads of syntactic phrases, and that English is generally right-branching (i.e. syntactic heads

tend to be on the left) (Berg 2009). For example, conjunctions such as *because*, *however* or *therefore* tend to have the end of a phrase to their left, whereas they are at the beginning of a new one to their right.

(1) I am happy because I ate the cake.
(2) He is not coming because he forgot his umbrella.
(3) The cat was mad at the dog because he heard him say something mean.

As examples (1), (2) and (3) illustrate, almost anything can occur before *because*. However, what comes after works in the same way that a new sentence would. Thus, many sentences simply start with a pronoun or a determiner, which can explain why overall one notices less part of speech diversity to the right of these conjunctions, as opposed to the left. Other types of grammatical elements such as prepositions (which are numerous in the dataset) also show the same behaviour when it comes to colligate diversity. A preposition like *on* will generally be followed by a noun phrase (e.g. *the roof*, *the rocks*), whereas it will have more diversity to the left in comparison. What this discussion highlights is that colligate diversity can be a characteristic of certain grammatical elements, but that this might only concern specific elements (such as prepositions) given that these colligate-based variables have smaller coefficients in the model.

Finally, the fifth result is that several variables were not retained by the stepwise selection process. In the case of deviation of proportions, a possible explanation is its strong correlation with many other variables, as discussed in section 5.4.1. It has also been shown that the fact that grammatical elements are more evenly distributed than lexical elements tends to be an effect of high frequency (Hilpert and Correia Saavedra 2017a). Therefore, it can be assumed that deviation of proportions is not a crucial variable because it does not bring a unique type of information, but is rather redundant with token frequency. Given the phonological integrity parameter (section 2.2.1), one could have expected the letter count variable to be somewhat relevant as grammatical elements tend to be shorter. However, the letter count variable was also not significant. A negative correlation with frequency can be observed in Table 23, which is consistent with the fact that high frequency elements (and therefore grammatical elements) are generally shorter.

In sum, these results show that frequency indeed plays a major role in grammaticalization. While this is not new knowledge, the importance of frequency is re-

assessed here through empirical means. The present analysis involves a large number of grammatical and lexical elements that belong to different morphosyntactic categories. In addition, it shows that collocate and colligate diversity are also important aspects, but that they yield smaller effects than frequency. To further support these results, the next section discusses an alternative way of conducting the binary logistic regression.

5.4.4 Including mixed effects

As discussed in section 4.1.3, using a mixed effect model to conduct the binary logistic regression is a possible alternative to what is presented above. Such an approach takes the idiosyncrasies of each item in the model into account by featuring the items as random effects in the model. When the binary logistic regression is computed using this approach, this results in the coefficients presented in Table 30.

	B	S.E.	t	Sig.
FreqWrit	1.263	0.486	2.598	0.010
Colloc_Div_W44	1.300	0.447	2.909	0.004
Colloc_Div_R1	-0.422	0.231	-1.823	0.069
Colloc_Div_L1	1.808	0.304	5.939	0.000
MaxColliPercent_R1	0.253	0.109	2.326	0.020
MaxColliPercent_L1	-0.562	0.145	-3.871	0.000
DevProp	-0.896	0.244	-3.667	0.000
Letter_Count	0.080	0.133	-0.603	0.547
Constant	0.271	0.139	1.948	0.052

Table 30. Coefficients of the Mixed effect binary logistic regression with items as random effects.

They are generally similar to what was presented in Table 26. First, the right collocate diversity measure and the letter count variables do not have highly significant p values ($p<0.01$) and therefore do not play a major role in the model.[17] Second, the signs of the coefficients that are featured in both tables are consistent. It was also the left colligate diversity measure that had a negative sign in Table 26. Third, the coefficients generally have similar values, with one major exception, namely frequency. Indeed, when using the mixed effect approach, frequency has a value similar to the other coefficients, and while some

[17] The generalized linear mixed effect procedure does not include the same stepwise process as used previously (i.e. descending Wald) in SPSS, where right collocate diversity and letter counts were removed from the model altogether.

variables have higher coefficients, the values are fairly similar as they range from 0.253 to 1.808. In contrast, Table 26 has values ranging from 0.222 to 10.245, where the highest coefficient was indeed the token frequency variable. Fourth, deviation of proportions has a highly significant ($p<0.001$) negative coefficient, whereas it was not featured in Table 26. Its negative sign goes in line with expectations, namely that grammatical elements have smaller deviation of proportions (i.e. more even distributions).

This model has a correct classification rate of 81.4%, which is lower than the previous method (85%). Furthermore, an SPSS warning appears which states that the validity of the results cannot be ascertained.[18] Discussing this warning in detail goes beyond the scope of our discussion and more information about this specific issue can be found in West et al. (2007). What matters is that the mixed effect procedure does not produce fully reliable results in the present case. One explanation is that taking the random factors into account (i.e. the items themselves) simply brings no additional value in that specific case. The lower classification rate certainly confirms this idea.

Therefore, these problematic aspects support the choice to use the simpler stepwise binary logistic regression presented above. This is particularly true since the results were similar, except for the size of the frequency coefficient and the relevance of deviation of proportions. It will be shown later in section 5.6 that using a smaller corpus, namely the spoken portion of the BNC which is about nine times smaller, results in a similar situation where the frequency coefficient also becomes much less prominent in the model when using mixed effects and where deviation of proportions is significant and gets a coefficient of -0.785. What this highlights is that frequency always plays a role in those models, but that methodological concerns can change how prominent this role is in comparison to other parameters.

5.5 Version 2: Written BNC (with punctuation)

The previous section has shown that collocate and colligate diversity play a relevant role in the binary logistic regression model. Grammatical elements have been shown to generally involve more diversity, and in particular to their left. A possible explanation that was put forth is that grammatical elements often occur close to syntactic boundaries. Including punctuation among the possible collocates and colligates might change how the diversity measures are

[18] This is the SPSS warning in full: "The final Hessian matrix is not positive definite although all convergence criteria are satisfied. The MIXED procedure continues despite this warning. Validity of subsequent results cannot be ascertained."

used in the model, given that punctuation tends to mark syntactic boundaries. The main purpose of this section is to determine whether the results differ significantly when punctuation is included.

This version of the study uses the same parsing method as presented in section 5.3. However, the regular expression used to detect words is changed from (1) to (2). The new expression also includes full stops, commas, semi-colons, exclamation marks and question marks. Punctuation in the BNC is marked by the tag *<c PUN>*.

(1) <w \w+>[a-zA-Z']+
(2) <[wc] \w+>[a-zA-Z'.,;!?]+

It should be noted that using regular expression (2) will also result in minor changes in token frequency. On a theoretical level, including punctuation should not change the frequency in the corpus of an item such as *cat*, but in practice, it does. The reason is that corpora, even of high quality such as the BNC, will always involve some tagging mistakes, especially when it comes to punctuation. The first regular expression (1) treats an occurrence such as (3) as a correct match and will return *<w NN1>cat*, but without the dot. Regular expression (2) however, will interpret this as *<w NN1>cat.* (note the dot here), which means that it will not be interpreted as equivalent to *<w NN1> cat* (no dot here) when counting all instances of *cat* in the corpus.

(3) <w NN1>cat.
(4) <w NN1>cat <c PUN>.

The correct tagging for (3) would be (4), where the punctuation is treated separately. This is a very minor point because those occurrences are somewhat rare. The only word in the database that is particularly affected by this is the title *Mr*. Regular expression (1) will detect both *Mr* and *Mr.* as the same thing, whereas regular expression (2) will notice that they are different and count them separately.

5.5.1 Principal component analysis

When it comes to the correlations between the variables in this new version, Table 31 is rather similar to Table 23. While it is straightforward that the collocate and colligate diversity measures are going to be different, it should also be pointed out that deviation of proportions

gets different values as well, because punctuation is now included into the chunks used by the method (section 3.5). The differences are generally small and the elements that get the highest correlations (bolded in Table 31) are the same as in the previous version. Therefore, there are no major changes in the way the variables correlate together.

Similarly, the percentage of variance explained by number of components (Table 32) and the component matrix (Table 33) are almost identical to their counterparts from the version that does not include punctuation (Table 24 and Table 25, respectively). It is therefore unsurprising that Figure 6 is also very similar to Figure 4. In sum, the same main observation can be made, namely that the three collocate diversity variables are strongly correlated together, along with deviation of proportions.

	FreqWrit	Colloc_Div_W44	Colloc_Div_R1	Colloc_Div_L1	MaxColliPercent_R1	MaxColliPercent_L1	DevProp	Letter_Count
FreqWrit	1	-0.267**	-0.202**	-0.189**	0.031	-0.039	**-0.514****	-0.221**
Colloc_Div_W44		1	**0.790****	**0.852****	-0.007	-0.019	**0.726****	0.244**
Colloc_Div_R1			1	**0.680****	0.155**	-0.098*	**0.511****	0.222**
Colloc_Div_L1				1	0.028	0.045	**0.543****	0.244**
MaxColliPercent_R1					1	-0.132**	-0.068	-0.055
MaxColliPercent_L1						1	0.046	0.006
DevProp							1	0.387**
Letter_Count								1

Table 31. Correlation matrix between all the variables involved in the model. The highest values are highlighted in bold (** p<0.01, * p<0.05).

Component	Initial Eigenvalues		
	Total	% of variance explained	% cumulated
1	3.361	42.018	42.018
2	1.26	15.749	57.767
3	1.003	12.536	70.303
4	0.873	10.91	81.213
5	0.785	9.813	91.025
6	0.343	4.291	95.317
7	0.297	3.707	99.024
8	0.078	0.976	100

Table 32. Total variance explained by number of component.

	Component 1	Component 2	Communalities
FreqWrit	-0.465	0.415	0.389
Colloc_Div_W44	0.933	0.128	0.887
Colloc_Div_R1	0.819	0.333	0.781
Colloc_Div_L1	0.847	0.153	0.741
MaxColliPercent_R1	0.021	0.676	0.458
MaxColliPercent_L1	-0.002	-0.563	0.317
DevProp	0.832	-0.243	0.751
Letter_Count	0.440	-0.322	0.298

Table 33. Component matrix and communalities. The left columns show how much each variable contributes to each component. The communalities column shows the percentage of variance of each variable that is accounted for by the model.

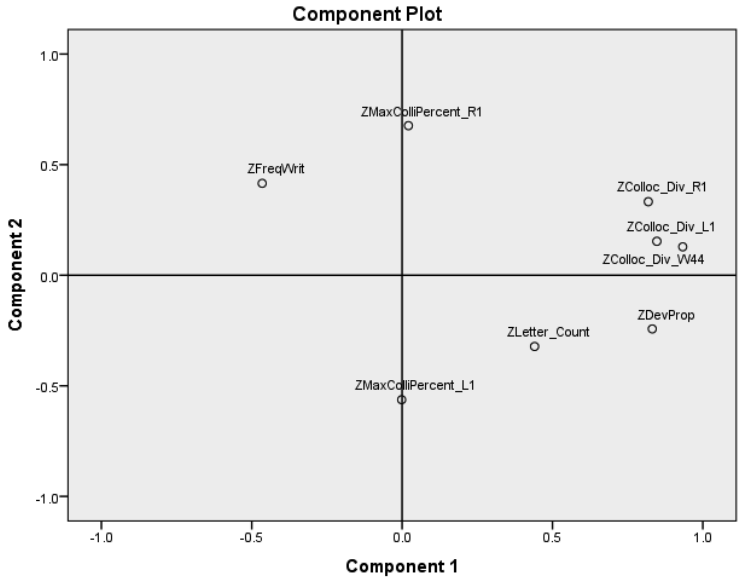


Figure 6. Component plot of the principal component analysis.

5.5.2 Binary logistic regression

A stepwise binary logistic regression was conducted using the descending Wald method, as in the previous version of the study. Table 34 shows the coefficients of the new model. The first observation is that the right colligate diversity measure (MaxColliPercent_R1) is now absent from the model. This is consistent with the fact that this variable had the least significant p value in the previous version of the study. The second observation is that the other coefficients are similar to what was observed before, with Colloc_Div_W44 being slightly bigger than Colloc_Div_L1 (as opposed to being almost equal in Table 26).

	B	S.E.	Wald	df	Sig.
FreqWrit	10.966	1.493	53.924	1	0.000
Colloc_Div_W44	2.573	0.470	29.921	1	0.000
Colloc_Div_L1	1.730	0.333	27.074	1	0.000
MaxColliPercent_L1	-0.903	0.176	26.414	1	0.000
Constant	1.898	0.324	34.231	1	0.000

Table 34. Parameters of the binary logistic regression using the written component of the BNC and including punctuation.

The correct classification percentage is of 85.8%, which is very close to the previous score of 85.0%, and even slightly better. The Cox & Snell R^2 is 0.500 and the Nagelkerke R^2 is 0.667, which is again very close to the previous observed values of 0.496 and 0.662 respectively. Because those results are similar, the analysis of the top and bottom 100 elements according to their grammaticalization score is also similar. Therefore, the table containing those elements is not reported here, but in appendix A2 instead. In a similar fashion, the distribution of the scores is identical to what is shown in Figure 5.

5.5.3 Discussion

It is rather clear that this version of the study does not yield major changes in the correlations between the variables or in the binary logistic regression model. The only main difference is that in this version, the right colligate diversity variable is not included into the model. In the previous model, this variable already had a small effect and had a low significance score, which means that overall, it should probably be excluded. Thus, collocate and colligate diversity to the right of an item do not appear to be necessary when computing the grammaticalization score in the present context.

The fact that including punctuation leads to similar results and even makes the model slightly more accurate gives credence to the explanation introduced in section 5.4.3 regarding syntactic boundaries. As stated earlier, one of the reasons why grammatical elements tend to have more diversity to their left is that they tend to occur at syntactic boundaries. When comparing Table 34 and Table 26, it appears that in Table 34, the coefficient of left collocate diversity measure is smaller than in Table 26 (1.730 vs. 2.128). An explanation is that now that punctuation is a possibility, many grammatical elements such as *because* or *however* now often have a comma or a full stop to their left, reducing the overall diversity. This makes left diversity less reliable as a predictor of grammaticalization, and a consequence is that the model now relies more on the coefficient of the four-four collocate diversity window as it

displays a higher coefficient than in the previous version (2.573 vs. 2.074). Despite these minor differences, including punctuation or not does not yield drastically different results. The fact that the model is slightly more accurate when including punctuation (85.8% vs. 85.0% of correct classification) is probably simply an effect of having more information overall. This is why including punctuation might be a better choice, albeit a minor one. The next section presents a version of the analysis that uses the spoken portion of the BNC, instead of the written one.

5.6 Version 3: Spoken BNC (no punctuation)

Using a spoken corpus might lead to different results for various reasons. Speech and writing are known to exhibit different characteristics (e.g. Biber 1991). For instance, spoken datasets tend to involve more informal situations which can result in a higher use of phonologically shortened forms (e.g. *cos*, *gonna*). This is why it is important to have a separate analysis for the spoken portion of the BNC. Furthermore, from a technical point of view, the spoken portion of the BNC is smaller than the written portion (10 million words vs. 90 million words), which is also something to take into account. In addition, while the spoken BNC does contain punctuation, it was not included at all in the datasets presented here. The reasoning is that punctuation is not always explicit in spoken contexts and is generally added later in transcription, which can be heavily dependent on personal interpretation.[19]

5.6.1 Principal component analysis

The correlations between the variables in the BNC spoken dataset are similar to that of the BNC written dataset, as can be seen in Table 35. Again, there is a strong correlation between all three collocate diversity variables, as well as the deviation of proportions. The situation resembles what was discussed earlier in section 5.4.1 and in Table 23. Again, the right and left collocate windows have stronger correlations with the 4-4 window than with each other (0.841 and 0.897 vs. 0.665). This supports the earlier idea that this correlation is due to the inclusion of the items of the smaller windows in the bigger one.

In the present case, the correlations are actually stronger overall, which can be observed in Table 36, which shows that two components can explain 62% of the variance, as opposed to 57% in Table 24. It means that the current dataset has redundancies that are easier to compress into two factors. Table 37 shows that those two components are similar to the ones

[19] A minor exception is that it is possible to talk about punctuation for emphasis, as in examples such as “He just can't manage money, full stop.” (BNC, file KD0)

presented in Table 25, where the first component mostly summarizes frequency (along with the collocate diversity measures, deviation of proportions and letter count), and the second component mostly summarizes the colligate-based measures. In the present case, the second component is much more clearly associated with the colligate-based variables, as the saturations of -0.685 and 0.810 reported in Table 37 are much bigger than the others (unlike what Table 25 shows regarding the BNC written dataset). This is also visually represented in Figure 7, which shows the relationship between the variables and the two components. Here, there are clearly four clusters of variables apart from each other, whereas they are much closer to each other in Figure 4 and Figure 6. What these results illustrate is that the principal component analysis on the spoken dataset is more efficient and is able to isolate the effect of the colligate-based measures more effectively.

	FreqSpok	Colloc_Div_W44	Colloc_Div_R1	Colloc_Div_L1	MaxColliPercent_R1	MaxColliPercent_L1	DevProp	Letter_Count
FreqSpok	1	-0.341**	-0.326**	-0.282**	0.041	0.128**	**-0.598****	-0.320**
Colloc_Div_W44		1	**0.841****	**0.897****	-0.037	-0.100*	**0.720****	0.380**
Colloc_Div_R1			1	**0.665****	0.141**	-0.211**	**0.646****	0.318**
Colloc_Div_L1				1	-0.070	-0.013	**0.621****	0.348**
MaxColliPercent_R1					1	-0.189**	-0.102*	-0.113**
MaxColliPercent_L1						1	-0.162**	-0.054
DevProp							1	0.463**
Letter_Count								1

Table 35. Correlation matrix between all the variables involved in the model. The highest values are highlighted in bold (** p<0.01, * p<0.05).

Component	Initial Eigenvalues		
	Total	% of variance explained	% cumulated
1	3.721	46.516	46.516
2	1.252	15.652	62.168
3	1.012	12.645	74.813
4	0.745	9.317	84.13
5	0.684	8.546	92.675
6	0.293	3.667	96.342
7	0.241	3.006	99.348
8	0.052	0.652	100

Table 36. Total variance explained by number of component.

	Component 1	Component 2	Communalities
FreqSpok	-0.572	0.042	0.328
Colloc_Div_W44	0.927	-0.004	0.859
Colloc_Div_R1	0.846	0.237	0.771
Colloc_Div_L1	0.843	-0.101	0.722
MaxColliPercent_R1	-0.042	0.810	0.658
MaxColliPercent_L1	-0.190	-0.685	0.506
DevProp	0.872	-0.064	0.764
Letter_Count	0.557	-0.232	0.364

Table 37. Component matrix and communalities. The left columns show how much each variable contributes to each component. The communalities column shows the percentage of variance of each variable that is accounted for by the model.

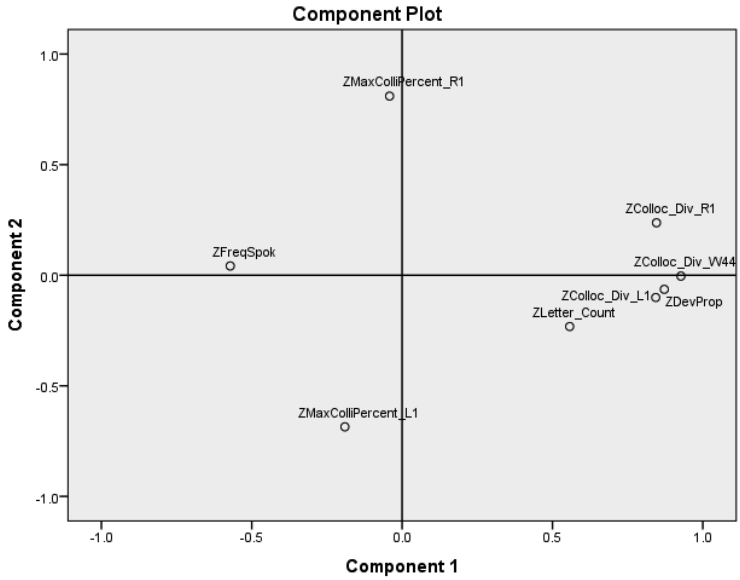


Figure 7. Component plot of the principal component analysis.

5.6.2 Binary logistic regression

A binary logistic regression using the stepwise descending Wald method was trained on the BNC spoken dataset and the results presented in Table 38. There are major differences with the models computed on the basis of the written dataset. The first major difference is that while frequency remains an important parameter, it does not have the highest coefficient anymore. Instead, Colloc_Div_L1 took its place. The second major difference is that deviation of proportions is now retained by the model, with a negative coefficient. It should be reminded that deviation of proportions ranges from 0 to 1, where 0 is the most even distribution. Therefore, the negative coefficient in the model means that even distributions

result in a higher grammaticalization score. The third major difference is the absence of Colloc_Div_W44 from the model. Interestingly, the main similarity between this model and the previous ones is that the colligate-based measures get similar coefficients. Indeed, MaxColliPercent_R1 gets a positive small coefficient of 0.262, whereas MaxColliPercent_L1 gets a negative, slightly bigger, coefficient of -0.759. This is really close to what was observed in Table 26 (section 5.4.2).

	B	S.E.	Wald	df	Sig.
FreqSpok	2.436	0.580	17.605	1	0.000
Colloc_Div_L1	3.500	0.340	106.069	1	0.000
MaxColliPercent_R1	0.262	0.115	5.200	1	0.023
MaxColliPercent_L1	-0.759	0.154	24.372	1	0.000
DevProp	-0.785	0.261	9.021	1	0.003
Constant	0.597	0.167	12.776	1	0.000

Table 38. Parameters of the binary logistic regression using the spoken component of the BNC.

The quality of the model is generally similar to its written counterparts, although slightly lower. The percentage of correct classification is of 83.5%, as opposed to 85.0% with the written dataset. The Cox & Snell R^2 is 0.470 and the Nagelkerke R^2 is 0.627, which are also lower than the written counterpart, namely 0.496 and 0.662 respectively. In order to further test the quality of the model, the top 100 and bottom 100 items according to grammaticalization score can be analysed. These items are summarized in Table 39. The correct classification rates are 98% for the top grammatical elements, and 92% for the top lexical elements, which are again quite high.

The lexical items that were "incorrectly" classified as grammatical were the construction *for example* and adverbial *however*. The reason why *for example* is in the initial lexical dataset is that *for example* is tagged as an adverb. However, one could argue that *for example* is a discourse-structuring element which works in a way similar to a connective. It is therefore not surprising to see this construction get a high grammaticalization score. Adverbial *however* also happens to get a high score in this situation, which is rather straightforward since *however* is also a conjunction and the line between its adverbial and conjunctive uses can be rather blurry.

Top100	L/G	Score	Bottom100	L/G	Score
<w AT*>A	G	1.000	<w V**>FELT	L	0.099
<w AT*>THE	G	1.000	**<w V**>DONE**	G	0.098
<w CJ*>ALBEIT	G	1.000	<w AV*>ALWAYS	L	0.096
<w CJ*>AND	G	1.000	<w N**>COMMUNITY	L	0.096
<w P**>I	G	1.000	**<w PRP>WORTH**	G	0.096
<w P**>IT	G	1.000	<w N**>COURT	L	0.094
<w P**>YOU	G	1.000	<w N**>THING	L	0.094
<w PRP>AGED	G	1.000	<w V**>LEFT	L	0.093
<w PRP>AMID	G	1.000	<w N**>MAN	L	0.092
<w PRP>BENEATH	G	1.000	<w V**>BECOME	L	0.092
<w PRP>BY MEANS OF	G	1.000	<w V**>USED	L	0.090
<w PRP>CONCERNING	G	1.000	<w N**>AREAS	L	0.089
<w PRP>FOLLOWING	G	1.000	<w N**>SERVICE	L	0.087
<w PRP>IN CONJUNCTION WITH	G	1.000	<w N**>SOCIETY	L	0.084
<w PRP>IN CONNECTION WITH	G	1.000	<w N**>PROBLEMS	L	0.083
<w PRP>IN RESPONSE TO	G	1.000	<w AJ*>BEST	L	0.079
<w PRP>IN THE LIGHT OF	G	1.000	<w V**>TURNED	L	0.074
<w PRP>ON THE PART OF	G	1.000	<w V**>SAID	L	0.073
<w PRP>TOWARD	G	1.000	<w AJ*>RIGHT	L	0.072
<w V**>'S	G	1.000	<w N**>TIMES	L	0.072
<w CJ*>EVEN	G	0.999	<w N**>LEVEL	L	0.070
<w CJ*>FOR	G	0.999	<w N**>SIDE	L	0.069
<w DT*>FEWER	G	0.999	<w AJ*>SURE	L	0.068
<w DT*>THAT	G	0.999	<w V**>GIVEN	L	0.068
<w PRP>ALONGSIDE	G	0.999	<w N**>HOME	L	0.067
<w PRP>DEPENDING ON	G	0.999	<w V**>USE	L	0.067
<w PRP>EXCEPT FOR	G	0.999	<w N**>DAY	L	0.066
<w PRP>IN	G	0.999	<w N**>LAW	L	0.065
<w PRP>IN ADDITION TO	G	0.999	<w N**>WORK	L	0.065
<w PRP>IN SPITE OF	G	0.999	<w V**>WORK	L	0.064
<w PRP>IN VIEW OF	G	0.999	<w N**>COUNCIL	L	0.059
<w PRP>UNLIKE	G	0.999	<w N**>HOUSE	L	0.058
<w PRP>UNTIL	G	0.999	<w V**>FIND	L	0.058
<w PRP>VIA	G	0.999	<w V**>TAKEN	L	0.058
<w CJ*>PROVIDED	G	0.998	<w N**>CAR	L	0.056
<w CJ*>RATHER THAN	G	0.998	<w N**>HEAD	L	0.055
<w P**>YE	G	0.998	**<w V**>GONE**	G	0.055
<w PRF>OF	G	0.998	<w V**>KNEW	L	0.054
<w PRP>AS OPPOSED TO	G	0.998	**<w V**>AM**	G	0.053
<w PRP>BY WAY OF	G	0.998	<w N**>NIGHT	L	0.052
<w PRP>DESPITE	G	0.998	**<w P**>ONE**	G	0.051
<w PRP>THROUGHOUT	G	0.998	<w V**>SAW	L	0.051
<w PRP>WITH REGARD TO	G	0.998	<w V**>SAYS	L	0.051
<w XX0>N'T	G	0.998	<w N**>BOOK	L	0.050
<w CJC>AS WELL AS	G	0.997	<w N**>LIFE	L	0.050
<w PRP>IN RELATION TO	G	0.997	<w N**>MONTHS	L	0.049
<w PRP>IN RESPECT OF	G	0.997	<w N**>WORDS	L	0.049
<w PRP>ON BEHALF OF	G	0.997	**<w DT*>LATTER**	G	0.048
<w PRP>REGARDING	G	0.997	<w N**>YEAR	L	0.048
<w CJ*>EXCEPT	G	0.996	<w V**>NEED	L	0.047
<w DT*>NEITHER	G	0.996	<w N**>BACK	L	0.046
<w PRP>ACCORDING TO	G	0.996	<w N**>HAND	L	0.046
<w PRP>AMONG	G	0.996	<w V**>KNOWN	L	0.046
<w PRP>INSIDE	G	0.996	<w N**>PERIOD	L	0.045
<w CJ*>THOUGH	G	0.995	<w N**>CENTRE	L	0.044
<w PRP>PRIOR TO	G	0.995	<w N**>WEEK	L	0.044
<w PRP>SINCE	G	0.995	<w V**>SEEN	L	0.044
<w PRP>AMONGST	G	0.994	<w N**>FAMILY	L	0.043
<w PRP>IN ACCORDANCE WITH	G	0.994	<w V**>FEEL	L	0.043
<w P**>WE	G	0.993	<w N**>POINT	L	0.040
<w CJ*>NOW THAT	G	0.992	<w V**>THOUGHT	L	0.040
<w DT*>WHOSE	G	0.992	<w N**>FORM	L	0.039
<w PRP>BESIDE	G	0.992	<w N**>CITY	L	0.037

<w PRP>IN SUPPORT OF	G	0.992	**<w XX0>N**	G	0.037
<w P**>NO	G	0.991	<w N**>END	L	0.036
<w P**>THEY	G	0.990	<w V**>MADE	L	0.036
<w PRP>INCLUDING	G	0.990	<w N**>CHILD	L	0.035
<w PRP>INSTEAD OF	G	0.990	<w N**>BODY	L	0.032
<w PRP>AHEAD OF	G	0.989	<w V**>HELP	L	0.032
<w PRP>CONTRARY TO	G	0.989	<w N**>DAYS	L	0.031
<w PRP>RATHER THAN	G	0.989	<w N**>PARTY	L	0.030
<w P**>HERSELF	G	0.988	<w N**>WORLD	L	0.029
<w AV*>HOWEVER	L	0.985	<w V**>LOOKED	L	0.029
<w PRP>OTHER THAN	G	0.985	<w AJ*>IMPORTANT	L	0.028
<w V**>IS	G	0.985	<w N**>WAR	L	0.028
<w DT*>LITTLE	G	0.984	<w N**>FACE	L	0.027
<w P**>HERS	G	0.984	<w N**>EYES	L	0.026
<w PRP>BELOW	G	0.984	<w N**>ROOM	L	0.026
<w PRP>ON TOP OF	G	0.984	<w V**>ASKED	L	0.026
<w V**>DO	G	0.984	<w V**>CALLED	L	0.026
<w PRP>SUCH AS	G	0.983	<w V**>FOUND	L	0.026
<w CJ*>WHETHER OR NOT	G	0.982	<w N**>ROAD	L	0.025
<w AV*>FOR EXAMPLE	L	0.981	**<w V**>GOES**	G	0.024
<w PRP>DUE TO	G	0.977	<w N**>GOVERNMENT	L	0.023
<w CJ*>WHILST	G	0.975	<w N**>PROBLEM	L	0.022
<w PRP>OUTSIDE	G	0.975	<w N**>QUESTION	L	0.021
<w P**>HIMSELF	G	0.974	<w N**>USE	L	0.019
<w PRP>BUT	G	0.974	<w N**>CASE	L	0.016
<w CJ*>NOR	G	0.973	<w N**>DOOR	L	0.015
<w PRP>IN LINE WITH	G	0.973	<w CRD>MILLION	L	0.013
<w P**>WHOEVER	G	0.972	<w N**>NAME	L	0.013
<w PRP>BEYOND	G	0.968	<w AJ*>ABLE	L	0.012
<w PRP>TO	G	0.968	<w N**>WAY	L	0.012
<w CJ*>WHEREAS	G	0.967	<w N**>COUNTRY	L	0.010
<w PRP>ABOVE	G	0.965	<w N**>FACT	L	0.010
<w PRP>NEXT TO	G	0.964	<w N**>PLACE	L	0.010
<w DT*>THIS	G	0.963	<w N**>PERSON	L	0.008
<w PRP>DURING	G	0.963	<w N**>STATE	L	0.008
<w PRP>AROUND	G	0.962	<w V**>TOLD	L	0.006
<w DT*>WHAT	G	0.961	<w N**>PER CENT	L	0.001

Table 39. Items with the 100 highest and 100 lowest grammaticalization scores. Forms that are wrongly classified are marked in bold. Here, "wrongly classified" means that those items should have been under the 0.5 threshold if they were initially classified as lexical, and above it if they were initially considered as grammatical.

Among the eight grammatical items that are classified as lexical items, four are the same elements discussed in section 5.4.2, namely instances of the exception verbs (e.g. *go*, *gone*). Among the other four items, the most specific to the BNC spoken is *n*. This is a rather peculiar item in the database, as it is actually a negation marker. Two examples of its use are shown in (1) and (2). In both cases, *n* is actually part of a larger group, respectively *in n it* (*isn't it*) and *du n no* (*don't know*). The latter accounts for most of the occurrences of this item and the explanation for its low score might simply be that it has almost no left or right collocate/colligate diversity. In addition, it is also a rather low-frequency element. Consequently, its final score is extremely low, resulting in its lexical categorization. This has

been highlighted earlier in section 5.4.2 with the example of *going* which also had a low score due to its neighbours being almost systematically *be* and *to*. Therefore, in the present case, the score is not entirely reliable, because instead of *n*, the elements under scrutiny should be *du n no* and *in n it*. This would be more consistent with how multi-word elements are treated in the database (e.g. *as well as* is treated as a whole).

(1) That's a long time in n it ? (BNC, file F7Y)
(2) I du n no if we'd er stayed there or (pause) and carried on with it. (BNC, file HMM)

Another item initially tagged as a grammatical element which appears in the bottom 100 item is *worth*, which is considered as a preposition in the BNC, as illustrated in (3). It should be noted that *worth* also occurs as a noun, as in (4), but that this case is rarer. There are 297 occurrences in the BNC spoken of *worth* as a noun, whereas there are more than a thousand cases where it is tagged as a preposition. In many cases, *worth* appears in the pattern *it is worth V-ing something*. So, *worth* usually comes after the verb *to be* and gives information regarding the worth of something, which can still be regarded as a lexical function and as really close to its other use as a noun.

(3) I didn't think it was worth it actually (BNC, file F7X)
(4) So I borrowed over what the house was worth (BNC, file KB7)

In the case of *latter*, which is tagged as a determiner in the BNC, the low score can be explained by the particular feature of this item, namely that it almost exclusively occurs after the determiner *the*. Therefore, its left collocate diversity is extremely small, and since left collocate diversity is the biggest parameter of the BNC spoken model, it is not surprising to see this case being classified among the lexical elements. To avoid this problem, the *the* should be skipped as a possible left collocate, which would correct the given value to this item. A further note is that elements such as *latter* are sometimes referred to as "postdeterminers" (Breban 2011), which suggests that they may form a different class of determiners.

Finally, the last grammatical element in the bottom 100 is the pronoun *one*. Looking at several sentences that contain this items shows just how broad these occurrences are. In cases like (5), *one* is a pronoun but also refers to a quantity, in cases like (6) it is also a pronoun, but not so much with a notion of quantity, and in cases like (7) it is used as a gender-neutral

indefinite pronoun. This last use can easily be replaced by *a person*, which certainly falls into the lexical category. These multiple uses can explain why this particular element was in the bottom items, although it remains a bit surprising that the score is this low.

(5) Er, one from Elizabeth and one from Ron (BNC, file D95)
(6) I think we weren't going to that one anyway (BNC, file D95)
(7) Well I think one does speak a bit louder in public meetings (BNC, file D97)

The distribution of the scores is again an s-curve as illustrated in Figure 5 and will therefore not be reported here again.

5.6.3 Discussion

The major differences between the spoken dataset reported in the previous section and its written counterpart can be summarized as follows:

1. Colloc_Div_1L has the highest coefficient and is bigger than frequency.
2. Colloc_Div_W44 is absent from the model.
3. Deviation of proportions is retained in the model.

The first difference is probably the most important one, since frequency had a much larger coefficient than any other variable in the written version. An explanation is that this is a corpus size effect. As the spoken portion of the BNC is nine times smaller than the written portion, the differences in frequencies between the words in the dataset are also smaller. Therefore, it means that the frequency variable contributes less information to the model. In such a case, the left-based collocate diversity takes a more prominent role in the model.

The second difference is the absence of the four words window collocate variable. In the spoken dataset, only knowing about the left collocate seems relevant. However, the results from section 5.6.1 need to be kept in mind, namely that all three collocate variables are strongly correlated together, which means that they generally provide similar information, with a preference for the left one. The spoken dataset therefore confirms the relevance of the left-based collocate measure.

Finally, the third difference is the presence of the deviation of proportions variable in the model. Again, the results from section 5.6.1 should be kept in mind, because deviation of proportions is strongly correlated with the collocate-based measures. The negative sign of the

coefficient is in line with our initial expectation that grammatical elements tend to have a more even distribution (section 3.5). The absence of deviation of proportions as a relevant parameter in the written model (section 5.4.3) was explained by pointing out that the evenness of the distribution of grammatical elements tends to be an effect of high token frequency, as evidenced by Hilpert and Correia Saavedra (2017a). However, that specific study used the whole BNC and it is quite possible that the spoken portion of the corpus presents different characteristics regarding this specific aspect. There can be multiple reasons why deviation of proportions is a significant parameter when using the spoken dataset as opposed to the written one. One possible reason is again the size of the corpus, which means that when the corpus is smaller, the frequency effect on deviation of proportions is less noticeable. Another possible reason is that in spoken language, repetitions and hesitations are more common, which can possibly lead to a different distribution of a given word.

5.7 Re-calibrating the study (modal verbs)

Attentive readers who looked closely at appendix A1 might have noticed a number of elements whose classifications are inaccurate. Specifically, the appendix shows that several modal verbs were wrongly classified into the lexical category in the datasets used in the previous studies. This mistake arose from a selection process that was done with minimal manual intervention. As a result, several modal verbs were wrongly classified. This section will address and resolve this issue.

However, instead of simply correcting this mistake outright, there is some interest in discussing here how the results that were presented above change when the mistake is rectified. The reasoning is that this will highlight the extent to which a single type of grammatical item may contribute to the coefficients in the models. There were seven modal verbs that were wrongly classified, namely *can, could, may, might, must, should* and *would*. The verb *need* could also be argued to count as a modal verb, however it is fairly apparent that the instances of *need* in the database mainly reflect its lexical use (e.g. *I need an apple*). Before changing the category of these modal verbs, a question that merits discussion is what score they received in the previous iterations of the model. Table 40 displays these scores using the BNC written corpus as presented in section 5.4. The first observation is that, as expected, *need* is strongly identified as a lexical element since it has a low score of 0.032. The second observation is that common modals such as *would*, *can* and *could* have the highest scores. The third observation is that the other modals do get a lower score than what would be expected of a grammatical item (i.e. below 0.5). However, an important note is that as shown

in Table 29 is that lexical verbs (as which these modal auxiliaries were wrongly classified) have an average score of 0.175, which indicates that the modals did in fact receive scores that are substantially above this average.

Modal	Score (section 5.4)	Corrected score	Frequency (BNC Written)
<w V**>Would	0.919	0.991	219’090
<w V**>Can	0.825	0.976	196’451
<w V**>Could	0.623	0.918	146’969
<w V**>May	0.421	0.847	107’016
<w V**>Should	0.328	0.697	98’270
<w V**>Must	0.228	0.630	65’521
<w V**>Might	0.225	0.615	52’495
<w V**>Need	0.032	0.040	29’151

Table 40. Grammaticalization scores of modal verbs, before and after correction.

The dataset was corrected by re-classifying the seven modal verbs *can, could, may, might, must, should* and *would* as grammatical elements. This means that the number of grammatical items went up from 264 to 271 and that the lexical category went down from 264 to 257. This slight imbalance is not a major problem. Having a dataset with the exact same number of items in each category was mostly to avoid over-representation of either of the two categories, but a small difference between them is acceptable. The binary logistic regression models presented in sections 5.4, 5.5 and 5.6 were thus re-computed using this new classification. The new data is different with regard to seven items out of 528, although the change more realistically corresponds to the double of that amount, since the items were moved from one category to the other. One may wonder how a small change can affect the final scores and the coefficients. It turns out that it does in fact result in slight changes.

First, let us discuss the new model that corresponds to the original one presented in section 5.4, which used the written BNC. Table 41 shows the new coefficients of the binary logistic regression, as well as the older ones. A first remark is that most of the coefficients that were prominent in the former version became even more prominent in the newer one. Frequency went from 10.245 to 12.566 and the collocate diversity measure using a four words window moved up from 2.074 to 3.048. The right-based colligate variable also increased from 0.222 to 0.538, which is unsurprising since modal and auxiliary verbs are commonly followed by an infinitive. The collocate diversity measure using one word to the left, as well as the colligate diversity measure using one word to the left remained fairly similar in size.

	B (new)	B (old)	S.E.	Wald	Sig
FreqWrit	12.566	10.245	1.838	46.731	0.000
Colloc_Div_W44	3.048	2.074	0.685	19.779	0.000
Colloc_Div_R1	-0.811	-	0.306	7.049	0.008
Colloc_Div_L1	2.080	2.128	0.381	29.819	0.000
MaxColliPercent_R1	0.538	0.222	0.130	17.084	0.000
MaxColliPercent_L1	-0.502	-0.640	0.171	8.663	0.003
Constant	2.420	1.830	0.405	35.753	0.000

Table 41. Coefficients of the binary logistic regression of section 5.4 (BNC written, no punctuation), after correction of the initial categories of modal verbs.

The major difference in this new version is the collocate diversity variable using a window of one word to the right, which was absent from the earlier model. In this new model, it presents a coefficient of -0.811. This seems to show that when modal verbs are included as grammatical elements, the elements in the right window become more relevant to the model, both because collocate diversity to the right is now present in the model, but also because the coefficient for colligate diversity to the right became bigger. The coefficient of -0.811 for the collocate diversity variable to the right is however small and is of similar size to the colligate diversity ones. This variable is therefore one of the least prominent ones, along with the colligate diversity ones.

Another observation is the significance of this coefficient exclusive to this new model, which amounts to 0.008 (i.e. 0.8% chance that the variable is in fact not relevant to the model). This is above the other variables which generally display smaller values. One may wonder how the earlier model from section 5.4 did not feature the right collocate diversity variable as a significant effect. The earlier stepwise regression model was computed in four steps and it turns out that collocate diversity using one item to the right was in fact present up to the third step, where it had a coefficient of -0.457, but a p value of 0.116 (11.6%). Because of this high p value, the variable was discarded and not featured in the final model.[20] Note that it displayed a similar coefficient, as it had a negative sign. The sign of that coefficient is also interesting, as it shows the same tendencies observed with the left-based variables. In both cases, the collocate-based measure has the opposite sign of its colligate counterpart. This is because, as discussed in earlier sections, a higher colligate-based measure corresponds to higher constraints (i.e. less diversity). Thus, both right-based collocate and colligate measures

[20] Note that a similar situation can be observed in section 5.4.4, where the mixed effect models (that does not involve a stepwise process), also had a coefficient of -0.422 and a p value of 0.069 for this variable.

display reduced diversity when it comes to grammatical elements. One explanation is that modals are grammatical elements that involve less diversity to the right as they have strong constraints to that regard. Indeed, modals tend to be followed by infinitive verbs or by pronouns when used in a question.

Regarding the percentage of correct classification, it remains similar as the new model has a correct classification rate of 84.7%, which is close to the previous value of 85.0%. The same can be said about the Cox and Snell R^2 and the Nagelkerke R^2 which both display similar values of 0.514 (vs. 0.496) and 0.686 (vs. 0.662). There are therefore no significant differences in the fitness of the model. Furthermore, Table 40 also displays the new grammaticalization scores that were computed by the model. It is apparent that the new model deals with modal verbs in a much more efficient way, as they all have values above 0.6, with the exception of *need* that was discussed earlier as being in fact lexical instances. The new model still keeps the same hierarchy, where *would*, *can*, *could*, and now *may* are above the others. This can be party attributed to their frequencies, as Table 40 shows that they are also the most frequent ones, although *should* has a lower score while almost as frequent as *may*, which shows that the other variables are also important in the model, despite the really big coefficient of frequency.

To further illustrate the points made in the previous paragraphs, Table 42 and Table 43 show the same process applied to the models introduced in sections 5.5 and 5.6. Overall, values are fairly similar between the new and old studies, but again with the addition of right-based collocate and colligate diversity variables that play more significant roles. The signs of these coefficients are also consistent with previous observations, namely a negative sign when it comes to collocate diversity (Table 43), and a positive one when it comes to colligate diversity (Table 42), which is in line with the earlier observation that grammatical items involve less diversity to the right.

In summary, rectifying the misclassification of the modal verbs highlights how the relative prominence of the coefficients can change in the model. In particular, modal verbs are characterized by reduced diversity to the right, which is reflected in the coefficients of the updated models. A remark is that if the model was further filled with other grammatical items, it is possible that the coefficients would change again, as the representation of different grammatical items would change.

	B (new)	B (old)	S.E.	Wald	Sig
FreqWrit	12.949	10.966	1.743	55.174	0.000
Colloc_Div_W44	2.626	2.573	0.475	30.505	0.000
Colloc_Div_L1	1.875	1.730	0.337	30.855	0.000
MaxColliPercent_R1	0.421	-	0.126	11.181	0.001
MaxColliPercent_L1	-0.656	-0.903	0.164	16.041	0.000
Constant	2.361	1.898	0.366	41.569	0.000

Table 42. Coefficients of the binary logistic regression of section 5.5 (BNC written, with punctuation), after correction of the initial categories of modal verbs.

	B (new)	B (old)	S.E.	Wald	Sig
FreqSpok	2.417	2.436	0.612	15.618	0.000
Colloc_Div_R1	-0.448	-	0.245	3.337	0.068
Colloc_Div_L1	3.527	3.500	0.339	108.502	0.000
MaxColliPercent_R1	0.572	0.262	0.129	19.716	0.000
MaxColliPercent_L1	-0.618	-0.759	0.154	16.129	0.000
DevProp	-0.611	-0.785	0.292	4.380	0.036
Constant	0.647	0.597	0.171	14.381	0.000

Table 43. Coefficients of the binary logistic regression of section 5.6 (BNC spoken, no punctuation), after correction of the initial categories of modal verbs.

However, the results shown in this section seem to support the idea that there is a core of coefficients that does not change between the different iterations of the study. The most important coefficient is frequency, which is consistently large and highly significant. Another core coefficient is the collocate diversity measure using a window of one item to the left. Beyond this core are more volatile coefficients that change depending on methodological choices, including the constitution of the database and whether it does feature certain grammatical categories or not. For instance, the four word collocate windows was not significant in the spoken model, whether modals are correctly included or not. It also appears that modal verbs in particular lead to higher coefficients that pertain to items to the right. This is a first step in addressing whether grammaticalization consists in increases or reductions in range (cf. sections 2.2.1 and 3.3), because it shows that the answer can depend on which types of language items are involved. However, the prominence of the left-based collocate window suggests that grammatical elements tend to involve more diversity in general. This is also supported by the written model, in which the four word windows is also prominent.

5.8 Overview

This section summarizes the main findings and discusses the differences between all the models. Let us first list the findings that were consistent across all versions of the study in the previous sections:

1. Token frequency and the left-based collocate diversity measure are prominent parameters in all versions of the study. Higher token frequencies and higher collocate diversity always correspond to higher grammaticalization scores.
2. The left-based colligate diversity measure is also a relevant parameter in all the versions, where more diversity to the left corresponds to a higher grammaticalization score, but it is smaller than the other highly significant coefficients.
3. There are strong correlations between the collocate-based measures and deviation of proportions. The colligate-based measures do not have this property.
4. The overall quality of the models is good, as the percentage of correct classification is rather high (generally 80% and above) in all versions of the study. The top and bottom scores are particularly accurate. The scores are distributed on a cline as illustrated by Figure 5.

When it comes to the main differences between the three versions, they can be summarized as follows:

1. Including punctuation in the written version of the study is generally beneficial as it increases the quality of the model overall, while not significantly changing anything else.
2. The role of frequency is more prominent in the written versions, whereas it is superseded by the left-based collocate diversity measure in the spoken version.
3. Deviation of proportions and the four words window collocate diversity measure are sometimes relevant, but not consistently across all versions.
4. The coefficients of the models can change depending on which grammatical elements are correctly represented in the database. Modals are an example where this leads to more prominent right-based coefficients.

5.9 Conclusions

All versions of the implementation of the binary logistic regression have resulted in rather good qualities of classification, which shows that the variables involved are useful in the distinction between grammatical and lexical elements. The fact that it produces a cline of values between 0 and 1, where the top and bottom elements are clear prototypes of their categories shows that it is possible to operationalize a quantitative take on grammaticalization, in particular when it comes to gradience.

All these studies have highlighted the prominence of token frequency as a parameter in the model. While this is not a novel finding per se, it is in line with previous research on the role of frequency in grammaticalization. In the present case, it is done from an empirical perspective on a large dataset of lexical and grammatical elements. Also, it has been shown that the prominence of token frequency can depend on methodological factors such as corpus size (section 5.6) or the statistical approach taken (section 5.4.4). As discussed in section 3.1, researchers have different opinions regarding the extent of the role of token frequency and these methodological considerations might also be relevant to that discussion.

The main novel finding is the major role of the collocate diversity variables. While all three collocate diversity variables are correlated with each other to some extent, the left one appears to be more easily retained in all versions of the model. The explanation that has been put forth so far is that grammatical elements have more diversity to their left because they tend to be syntactic heads, and therefore to be located at syntactic boundaries. This is an interesting feature of grammatical elements that might be worth exploring in future research on grammaticalization both from quantitative and qualitative perspectives. What encourages this conclusion even more is that the left colligate-based diversity measure has systematically shown the same effect across the different versions of the study, although sometimes with smaller coefficients. This is particularly important because unlike the collocate diversity variables, the colligate diversity variables do not show strong correlations with the other variables in the dataset. It means that diversity should be investigated both at the level of the collocates and at the level of their parts of speech.

The results of the principal component analysis have also shown that there is some redundancy in the dataset. This could be further explored by investigating whether some of the morphosyntactic categories under investigation are more subject to such redundancies. An example would be that since nouns and verbs are more frequent overall, they might also have stronger correlations with orthographic length, as opposed to adjectives and adverbs.

Investigating the morphosyntactic categories in general would also be beneficial for other reasons. For instance, is the left-based collocate diversity variable more prominent in the model if only prepositions are involved? This could be investigated by analysing whether the different morphosyntactic categories can be clustered on the basis of variables such as collocate diversity. More generally, clustering methods such as conditional inference trees and random forests (Levshina 2015: 291-300) could also be used as alternatives to the binary logistic regression approach proposed in this dissertation.

From a more methodological point of view, the findings have highlighted the limitations of the way the collocate diversity measures are computed. They tend to be strongly correlated, which means that they offer redundant and overlapping information. One way to deal with this fact would be to merge them into a single value, maybe by averaging their values together. Another way would be to use a component obtained by principal component analysis that summarizes these three variables specifically. However, these alternatives might not be optimal because the findings have shown that the left-based measures are more relevant in all versions of the study. Another possibility would be to use random sampling of fixed size, where n collocates are randomly selected. In this situation, the number of different types in the sample could be divided by n, which would ensure that frequency is not a factor. While this approach might provide more accurate results, there is still a limitation in the sense that lower frequency items are represented by a higher proportion of the total number of occurrences, as opposed to more frequent items. In the case of highly frequent elements, there can be cases where the chosen sample size might be too small to be truly representative, which can produce substantially different results when trying to replicate the study. Choosing an appropriate size for the sample is therefore a crucial issue when considering such an approach. This issue is particularly relevant when dealing with broad frequency ranges, as is the case in the present studies.

Other measures related to diversity, such as entropy measures (Shannon 1948, Bavaud 2009, Gries and Ellis 2015), could also be tested as possible alternatives to determine whether they display similar characteristics. Entropy is a measure of uncertainty that concerns itself with how predictable an outcome is. Higher unpredictability corresponds to higher entropy. For example, if a balanced dice has six faces, there are six outcomes, which translates to more uncertainty regarding the outcome than a balanced coin that only has two outcomes. However, if the six face dice is rigged to land on face number four 99% of the time, then that dice would have lower entropy than the balanced coin because its outcome is highly predictable (despite having more possible outcomes).

In the studies above, the colligate diversity measure uses the percentage of the most common part of speech. An entropy measure would deal with more information, because it would also take into account the other colligates, not just the most frequent one. For instance, if a word is followed by a noun in 60% of cases, the situation is different if 40% of the other cases are adjectives, as opposed to the situation in which the remaining cases are four tags that each represent 10%. The approach chosen for colligate diversity is therefore quite coarse as it does not make a distinction between these two situations, whereas an entropy measure could. Using entropy-based measures might therefore be a step towards a more fine-grained approach when it comes to the colligate diversity measures, which could potentially lead to a more prominent role in the model, since they currently display rather low coefficients.

Regarding the collocate diversity measures, entropy could be used to investigate additional aspects. For instance, an item can have many possible types (i.e. high diversity), yet a possible situation is that maybe one of these types is much more frequent than the others. On the other hand, maybe all these types tend to occur equally, which is a situation that could be argued to involve more diversity than the previous one. These situations could be investigated by using the concept of entropy, which could improve the models presented earlier. Again, investigating diversity in general as a characteristic of grammaticalization seems like a promising venue given the current results, especially since it can even supersede token frequency as observed with the spoken dataset.

The next chapter proposes further testing of the written model presented in section 5.4 from a diachronic perspective. Does the model work on diachronic datasets and is it able to detect increases and decreases in grammaticalization? Twenty words and constructions that have both a lexical and a grammatical counterparts are going to be investigated using the Corpus of Historical American English (COHA). The corpus contains data from 1810 to 2009. The aim is to see whether it is possible to empirically observe unidirectionality (i.e. increasing grammaticalization over time and not the other way around) by computing a grammaticalization score for each decade.

6. Diachronic study

6.1 Introduction

While the previous chapter has shown that the degree of grammaticalization of two elements can be quantified and compared in synchrony, this chapter explores the idea that these degrees of grammaticalization are susceptible to change over time. The focus of this chapter is therefore on the gradualness of grammaticalization. The main research question is whether it is possible to empirically observe a general increase in grammaticalization in a given set of items, which is expected given the hypothesis of unidirectionality in grammaticalization (section 2.4). To achieve this, twenty pairs of elements that have lexical and grammatical counterparts have been selected. An example is the verb *keep*, which currently has a main verbal use in English as in (1), but also has auxiliary-like function as in (2).

(1) I will keep your jacket with me.
(2) I keep bumping into doorframes.

Instance (1) is considered as a lexical use, whereas instance (2) is considered as a grammatical use (Quirk et al. 1985, Biber et al. 1999, Huddleston and Pullum 2002).

The data used in this chapter was retrieved from the Corpus of Historical American English, which contains data from 1810 to 2009. It should be noted that there is currently no clear answer regarding how long it takes for an element to grammaticalize. Narrog and Heine (2011: 8) remark that this question has not received enough attention and that it is currently unknown what the minimum or maximum amount of time is required for grammaticalization to take place. Quotative markers in Akkadian are an example where grammaticalization took place over around two millennia (Deutscher 2011: 651-654). In contrast, it has been suggested that grammatical change can happen during an author's lifespan (Petré and Van de Velde 2018), which supports the idea that grammatical change can take place over relatively short periods. The present analysis will not only focus on elements that are currently grammaticalizing, but also on such elements that can be considered to have undergone grammaticalization. To this purpose, the dataset for the present analysis also contains elements that have grammaticalized before the period under discussion (e.g. auxiliary *have*) in order to test whether they already display a high stable score over time, as one would expect. A grammaticalization score using the synchronic model presented in Chapter 5 will be

computed for both the lexical and the grammatical counterpart in each pair of elements, and for each decade in the dataset. This will result in a grammaticalization score over time of the lexical and grammatical counterparts which can be used to empirically test the hypothesis of unidirectionality.

In addition, several auxiliary questions will also be explored. The previous chapter has highlighted how all the variables in the model are correlated, such as the fact that three collocate diversity variables correlate strongly together (section 5.4.1). In this chapter, the general evolution of these variables over time is going to be discussed. For instance, is frequency more subject to change than collocate diversity? Another aspect that will be discussed is how the elements differ from one another and which methodological consequences arise. For example, does the model have the same accuracy when dealing with auxiliary verbs as it does with regard to prepositions? This leads to a more general question, which addresses whether grammaticalization can be generalized as a single concept if it encompasses very different processes. Do these processes have enough shared characteristics (e.g. increasing collocate diversity), to be considered as instances of the same phenomenon? The results of the model should show that this is the case.

The chapter is structured as follows. First, the dataset is presented in section 6.2. The motivations for the choice of the twenty elements are presented. Also, since the distinction between lexical and grammatical uses is not always straightforward, section 6.2 also discusses how the respective counterparts are retrieved from corpus data, using semi-automated means, and also some of the limitations of such an endeavour. Next, section 6.3 presents the main results of the diachronic approach by showing the aggregated results of all the twenty elements. It discusses whether unidirectional processes of change can indeed be observed in this dataset, and also how the variables of the model change over time. In contrast, section 6.4 focuses on the results of the individual elements in the database. Different types of behaviours are observed when it comes to grammaticalization scores over time, and this section tries to discuss the main observed behaviours: increasing, decreasing, stable and chaotic scores over time. The main conclusions are drawn in section 6.5.

6.2 Data

6.2.1 Selection of cases

Twenty pairs of lexical and grammatical elements were selected to constitute the dataset used in this chapter. These twenty elements are meant to display various cases of grammaticalization with different characteristics. There are low and high frequency elements, highly and weakly grammaticalized elements, and different types of grammatical domains involved. While the aim is to have some diversity regarding these aspects, these elements are not meant to represent the whole of English grammar, especially since they mostly consist of auxiliaries and complex prepositions. Instead, they were chosen because they form pairs where the lexical and grammatical uses currently co-exist in English. This allows for a comparison between lexical and grammatical elements at another level than what was done in chapter 5. The complete list can be found in Table 44, along with an example sentence for each use. The "domain" column indicates the grammatical domain of the grammaticalized counterpart. The "period" column contains the first attestations found in the Oxford English Dictionary online (OED online) and is based on the sense (or phrase) under the column "OED entry". Note that these are approximations, since it is sometimes possible to find earlier attestations using other data sources. The dates were also sometimes rounded up or adjusted, because the exact date of the first instance might not be fully accurate. For example, the first mention of *as a matter of fact* in the OED is from 1871, yet there is evidence in the data presented in this section that instances can be found earlier than this. In addition, the Early English Books Online (EEBO) database has instances of *as a matter of fact* from the late 17th century that also correspond to grammaticalized uses. When a data range has been adjusted on the basis of EEBO it is marked by a * symbol. The "Gzn Status" column indicates the current grammaticalization status of the item. Coarse distinctions are made, since the purpose of this indicator is to show that the selection for the following studies involves diversity when it comes to grammatical statuses. The term "advanced" is used when the form currently has a broad range of meanings/uses and includes common shortenings (e.g. *'ve, gonna*). Forms that still have a rather specific meaning, but have been present for a while are noted as instances of "low" grammaticalization. The term "recent" is used when the grammatical element is relatively recent (c1700 onwards). The following paragraphs discuss and motivate this selection.

First, complex prepositions (Hoffmann 2004, Hoffmann 2005) are strongly represented in this list as it is generally straightforward to distinguish between their lexical and

grammatical uses. It is often observed with complex prepositions that the meaning of their components parts is bleached, which is a prominent feature of grammaticalization (Sweetser 1988). There are also several auxiliaries (Heine 1993, Kuteva 2001), as these are very common objects of grammaticalization that can be easily detected. These two categories are frequent targets of grammaticalization, which motivates the reliance on a larger number of these constructions for the present analysis. It is quite difficult to include determiners, pronouns or conjunctions in such a dataset, since their lexical counterparts are not easily traceable. Despite the apparent lack of diversity in the types of grammatical items involved (i.e. prepositions and auxiliaries), Table 44 shows that they are involved in a broader variety of grammatical domains. The choice of elements for the present study draws to some extent on pairs of lexical and grammatical elements that have been used in a previous study that contrasted the way in which they are cognitively processed (Hilpert and Correia Saavedra 2018).

Second, the elements that were selected have grammaticalized during different periods in the history of English and thus display different degrees of grammaticalization. For instance, the grammaticalization processes of elements such as auxiliary *have* or the future *be going to* + infinitive construction have started a long time before the period under observation (i.e. 1810-2009). The reasoning is to include elements that have different profiles in order to see how these differences are reflected in the grammaticalization scores obtained by the model. As will be discussed below, elements that are known to be highly grammaticalized during the whole period should display a constant high score. The next section discusses how to tell apart the grammatical and lexical pairs, first from a linguistic point of view, and then from a computational point of view.

Item	Lexical source	Grammatical use	Domain	Period	OED entry	Gzn Status
back	He has a wart on his back.	He went back there.	Aspect	c1300	Senses I, II, III	Advanced
die	He was slowly dying.	She was dying to know their secret.	Modality	c1700	Sense 7b	Recent
event	This is a great event.	In the event of a fire, please leave.	Preposition	c1600	Sense 1a	Low
go	I am going to the park	He is going to win the game.	Tense	c1600	Sense 51a	Advanced
have	I have two cats.	I have found your wallet.	Tense	Early OE	Sense VI	Advanced
help	I needed to help him.	I couldn't help singing.	Modality	c1600	Senses 11a,b,c	Low
keep	You can keep the change.	We kept asking him to go.	Aspect	c1600*	Sense 40b	Advanced
long	This road is quite long.	As long as I can see him, I'll be fine.	Clause combining	Early OE	Sense 1	Low
matter	I have a matter to discuss	As a matter of fact, she knew.	Discourse	c1700*	Sense 1c	Recent
means	The ends justify the means	I'll reply by means of poetry.	Preposition	c1300	Phrase P1a.	Low
order	The order was never executed.	In order to succeed, get a boat.	Clause combining	c1600	Phrase P3b	Low
quit	They quit their job yesterday.	He wouldn't quit bothering me.	Aspect	c1700	Sense 13c	Recent
regard	He has no regard for human life.	His conduct with regard to this problem is childish.	Preposition	c1400	Phrase P1a	Low
respect	I have no respect for you.	My opinion with respect to this matter is important.	Preposition	c1550	Phrase 5a	Low
soon	I will leave soon.	As soon as he leaves, I will sing.	Clause combining	c1600	Sense 8a	Low
spite	We did it out of spite.	In spite of our argument, I like him.	Preposition	c1550	Sense 5a	Low
terms	I did not agree to these terms.	We can do more in terms of analysis.	Preposition	c1500*	Phrases P1b, d	Low
use	I have never used this bathroom.	I used to come here all the time.	Aspect	c1400	Senses 21b, c	Low
view	What a nice view !	In view of your efforts, I salute you.	Preposition	c1800	Phrases P1e, b	Recent
well	She sings well.	They were there, as well as Luke.	Preposition	c1200	Phrase P3	Low

Table 44. The twenty pairs of lexical and grammatical counterparts in the database, along with example sentences. The "period" dates correspond to the earliest attestations of the grammatical counterpart and are based on the Oxford English Dictionary (OED) online. The * symbol indicates a modified date from the OED, using the EEBO database.

6.2.2 Distinguishing lexical and grammatical counterparts

The basic distinction between a lexical and a grammatical element is that the latter has little meaning by itself, and its function is to provide extra information about the former. Grammatical functions can take many forms. Building on the definitions introduced earlier in section 2.1, and on Hopper and Traugott (2003: 18) common grammatical functions are:

- Indicating relationships and links between the elements in a sentence (e.g. prepositions, possessives)
- Linking and structuring different parts of discourse (e.g. conjunctions)
- Giving grammatical information about participants and elements of a discourse, for instance whether they have been identified previously (e.g. pronouns, determiners). This also comprises giving extra grammatical information such as tense or aspect regarding a main verb (e.g. auxiliary verbs).

From a qualitative point of view, these situations are generally easy to identify. However, in the present situation, the lexical and grammatical counterparts of the twenty elements in Table 44 need to be found in a 400 million word corpus, which means that manual processing is not possible. Instead, there is a need for a quantitative and automated way to tell lexical and grammatical uses apart. This is especially important since the elements in the dataset share the same forms.

Fortunately, for most complex prepositions, the tagging of the corpus is helpful. Indeed, a pair such as *view* and *in view of* is tagged differently. The lexical uses here are tagged as nouns, which is transcribed as *<nn1> view* or *<nn2> views*. In contrast, the grammatical uses are tagged as a complex preposition with the following sequential tagging: *<ii31> in <ii32> view <ii33> of*. However, it will be shown later that for individual cases, the tagging of the corpus is sometimes not fully reliable, which means that this approach has its limitations.

Regarding auxiliary verbs, the situation is quite different. The corpus does not have a special auxiliary tag for instances of *have* or the *be going to* future. As stated above, auxiliaries take a lexical verb as their complement, which means that when a verb is used as an auxiliary, there is a main verb in close proximity. This is a tangible way of automatically finding an auxiliary use, as one way to distinguish grammatical (auxiliary) *have* from lexical *have* is to check whether a participle is present somewhere to its right. Participles do have a special tag in the corpus, so this is feasible. What needs to be determined for every auxiliary

at an individual level is which non-finite verb form it is associated with, how many words tend to occur in between auxiliary and lexical verb, and how many wrong instances (false positives) are retrieved.

The solution adopted is to use a general algorithm that can be modified for each individual element. Once the algorithm has been applied to retrieve lexical and grammatical uses, they are analysed manually to assess the quality of the selection. There are two main implementations of the algorithm, one that uses tagging (complex prepositions) and one that uses the detection of a non-finite verb form in a given window (auxiliary verbs). The implementations work as follows:

Default implementation for complex prepositions:

1. All possible forms of an item are first catalogued. This is done on a case-insensitive basis, which means that these occurrences can be at the beginning of sentences.
Example: <*nn1*> *view,* <*nn2*> *views* and <*ii32*> *view*

2. The instances with nominal tags are classified as lexical elements, while the prepositional tags are classified as grammatical elements. This step is adapted depending on the morphosyntactic categories of the lexical and grammatical counterparts and follows the same logic as chapter 5 where nouns, verbs, adjectives and adverbs are considered as lexical elements.

3. A minimum of 25 items is randomly drawn from the grammatical and lexical selections, in order to manually check whether they correspond to such uses. This selection is retrieved from a default sample of the COHA[21] rather than the whole corpus, when the item under observation is common enough. This allows faster processing of the dataset. When the item is rare, bigger custom samples were made. Details are given in the individual sections where this alternative is used.

[21] The default COHA sample can be freely downloaded here: http://corpus.byu.edu/full-text/samples.asp

Default implementation for auxiliaries:

1. All possible forms of an item are first catalogued (case insensitive). Note that only the forms that are shared between the lexical and grammatical counterparts are kept. Therefore, the form *keeping* is not selected for the lexical counterpart of *keep* because its grammatical counterpart does not use it (**I'm keeping seeing you*).
Example: <*vv0*> *keep*, <*vvz*> *keeps*, <*vvi*> *keep,* <*vvd*> *kept* and <*vvn*> *kept*

2. If a participle is detected within an X words window to the right, that particular instance is classified as a grammatical instance. This step is adapted by looking for specific types of participles (e.g. *-ing* participles for grammatical *keep*, which are tagged as <*vvg*>, <*vbg*>, <*vhg*>, and <*vdg*> in the COHA[22]).

3. A minimum of 25 instances is randomly taken from the grammatical and lexical selections, in order to manually check whether they correspond to such uses. For each possible window, 25 additional instances are selected. This means that if one is using a 3-words window, 75 extra instances will be checked manually. A sample of the COHA is used in the same fashion as in the implementation presented above.

4. Manual analysis of the window sizes is performed in order to determine which window size is optimal, namely which size finds the biggest number of correct instances, while minimizing the number of wrong matches.

Two Python scripts are provided in appendix D2 as illustrations of the implementations above. An example of a wrong match in the grammatical instance of *keep* would be the sentence *a ruler he kept for measuring purposes*. This is detected as grammatical because *measuring* is tagged as a participle. A central aspect of these algorithms is therefore the manual evaluation of the selection. This process relates to two key concepts in pattern detection algorithms, namely precision and recall. Precision is the accuracy of the pattern, which can be measured by counting the number of correct items that were found. Recall assesses whether all findable items were indeed found. To give a concrete example, imagine a

[22] The tags are found in a separate column in the wlp format (word, lemma, part of speech) of the COHA, and they do not include the symbols < and >. However, these symbols were added by my own script for the processing of the data. The main reason is that I wanted to have the tags displayed in a clear way within the texts when manually controlling the lexical and grammatical instances. Also, this mirrors the BNC tagging.

robot that has to collect all the blue balls in a room. There are also decoy black balls that must not be collected. There are 10 blue balls and 10 black balls. If the robot collects 8 blue balls and 3 black balls, then it has a precision of 8 correct items out of 11 (72.7%), and a recall of 8 items found out of 10 (80%).

In the context of this study, recall cannot be formally evaluated because the total number of instances to find could only be determined through an exhaustive manual analysis of the entire corpus. In the absence of such an analysis, it is unknown how many grammatical instances of *keep* actually are in the corpus. An alternative would be to test recall on a subset of the whole corpus, but it would also require a large amount of manual analysis, since the chosen chunk of the corpus would still need to be of a certain size, especially for elements that have lower frequencies. Precision, on the other hand, can be calculated, as it is possible to look at a given number of items selected by the method discussed above and to count how many are actual grammatical or lexical instances, based on manual evaluation. In the subsequent sections, cases where precision falls below 75% are considered as not sufficiently reliable and are excluded from the selection.

A problematic aspect in deciding what counts as a lexical source of the grammatical counterpart is how to deal with the plurals of nominal lexical sources. For instance, is the plural *views* related to the complex preposition *in view of* and should it be included as its lexical counterpart? A complex preposition such as *with regard to* is occasionally written as *with regards to* in the COHA (only 25 occurrences, 12 of which are in the 2000's), but most complex prepositions do not have plural versions such as **in orders to*. The decision in the present case was to include plural nouns, if possible (e.g. *spite* does not really occur in plural form, and by contrast *means* almost exclusively occurs in plural form).

A basic search indicates that plural occurrences tend to be rarer than singular occurrences for most elements in the COHA. The singular and plural part of speech nominal tags can be used to calculate the ratio of singulars to plurals. Elements such as *regard* and *back* rarely occur in plural form, with 1% and 5% of plural occurrences respectively. Higher percentages can be observed with *respect* (21%), *matter* (22%), *view* (26%), and *order* (30%). The only elements that have a higher percentage of plural forms are *term* (57%) and *event* (58%). In the case of *term*, the grammatical counterpart does use the plural form (*in terms of*), whereas this is not the case for *event* (*in the event of*). It is therefore complicated to find a fully consistent selection method given these differences. In most cases, including plurals or not will not yield major differences given that they represent a smaller proportion of

occurrences. The decision was to include both singular and plural forms, with the exception of *in terms of*, which already has a plural form in its grammatical counterpart.

The next sub-sections contain details regarding the elements in the database on an individual basis. The aim is to highlight the technical difficulties regarding each element in the database. Tables containing the percentage of correct classifications for each element are also presented and discussed. Verbs with auxiliary or auxiliary-like functions (e.g. *keep, have, go*) are discussed first, as they have additional concerns, such as finding the most optimal window. All the raw data used for the evaluations of the classifications can be found in appendix B.

6.2.3 Keep

The grammatical use of *keep* consists of its auxiliary-like function in the *keep* + V_{ing} construction, such as *keep going* or *keep talking*. Its function relates to the category of grammatical aspect. More specifically, it can be seen as a continuative marker. Aspectual categories are particular as they can further develop into tenses (Hengeveld 2011: 588-592). The grammatical function of *keep* has been referenced in many studies (Hopper 1991: 23, Heine 1993: 66, Wischer and Diewald 2002: 24, Heine and Kuteva 2002: 185), which tend to include both *keep* + V_{ing}, but also *keep on* + V_{ing}. This is why the inclusion of the preposition *on* in the construction *keep on* V_{ing} is considered as the same grammatical use as *keep* V_{ing} and was not considered as an independent construction in the present study. The following forms were found in the default COHA sample:

<*vv0*> *keep*, <*vvz*> *keeps*, <*vvi*> *keep,* <*vvd*> *kept*, and <*vvn*> *kept*

These instances correspond to the possible forms that grammatical *keep* can take. Table 45 provides an example for each of these cases, since this is the first case presented.

Tag	**Example sentence**
<vv0> keep	I keep forgetting my keys!
<vvz> keeps	He keeps forgetting my keys!
<vvi> keep	She did it to keep us waiting.
<vvd> kept	You kept me waiting for two hours.
<vvn> kept	I was kept waiting for two hours.

Table 45. Examples of grammatical *keep*, by tag categories.

The *-ing* participle <*vvg*> *keeping* was not included as it is not used to convey grammatical meaning. In order to be counted as a grammatical *keep*, an *-ing* participle has to occur in a window of less than 5 words to the right. Items that do not fit this requirement are classified as lexical instances. To find these participles, the <*vvg*> and <*vvgk*> tags were looked for, as they correspond to the participles of lexical verbs and catenatives, respectively. Note that the verbs *be*, *have* and *do* have special tags for their *-ing* participles, namely <*vbg*>, <*vhg*> and <*vdg*>. To include all these tags, the following regular expression was used:

<.*v(v|b|h|d)g.*>

The expression selects elements that begin in < and end in >, and which contain *vvg, vbg, vhd,* or *vdg* somewhere in the middle of a potentially larger string.[23] Consequently, ambiguous tags were also included, such as the ambiguity tag <*vvg_nn1*>. This tag indicates that the item was most likely a <*vvg*> but that the parser also assigned a significant probability to its being an <*nn1*>. These tags are rather rare and manual inspection of the data reveals that they usually turn out to actually be a <*vvg*>. It should be noted that this regular expression could technically match an aberrant string such as <*helloworldvvgbyebyeworld*> because it does indeed contain a *vvg* string in the middle. However, such cases do not occur in the corpus, which is why a "greedy" regular expression like this one can be used.

25 items were randomly selected for each window size, in order to control the precision of the method for the grammatical items. The aim is to determine the window size after which the data contains too many wrong matches. The 25 sampled items were analysed manually for each window size and are summarized in Table 46. The full table, which includes the sentences in which those items appeared, can be found in appendix B1-A. As Table 46 shows, the precision strongly decreases at the third word to the right. The reason for this decrease is the high presence of the construction *keep from* V_{ing}, which often includes a pronominal object in between the verb *keep* and the preposition *from*, thus explaining why this case is frequent when the V_{ing} is in third position. This was not considered as a target grammatical use of *keep* because it does not convey a continuative meaning. Instead, it conveys a meaning that is roughly equal to *stop* or *prevent*. This idiomatic use was discarded

[23] This regular expression is only used to detect a pattern and not to retrieve it, as it would in fact only return *v, b, h* or *d* using this syntax. To return the full string, it would be necessary to rewrite it as <.*v[vbhd]g.*>. For the current purpose however, they are equivalent.

from the dataset. The average precision is calculated by weighting each category, with the weight reflecting the share of the data for that category. For example, the case with no words in between (V_{ing} is the 1st word on the right) has a precision of 25/25 and a weight of 0.493. A weighted average is calculated between all five cases, which results in a low average precision of 73.88%, which means that the method used for *keep* has to be re-worked to be more reliable.

The selection method was re-worked to remove any instance that contained a *from* in between *keep* and the *-ing* form. Furthermore, as Table 46 shows, going beyond the third word does not provide accurate results (i.e. 44% and below), meaning that the third word should therefore be taken as the maximal window. Also, there are only 18 matches for the fifth word to the right, which is much less than for the other categories. The results of the new method are much more reliable, as shown in Table 47. The actual sentences can be found in appendix B1-B. The final selection of data using up to three words has a weighted average precision of 98.8%, which is a satisfactory result. Unfortunately, it does mean a slightly lower recall as some instances are missed by ignoring the four and five words windows.

Window	Number of correct instances	Total number of instances
V_{ing} is the 1st word to the right	25/25 (100%)	224 (49.3%)
V_{ing} is the 2nd word to the right	20/25 (80.0%)	83 (18.3%)
V_{ing} is the 3rd word to the right	11/25 (44.0%)	82 (18.1%)
V_{ing} is the 4th word to the right	4/25 (16.0%)	47 (10.4%)
V_{ing} is the 5th word to the right	2/18 (11.1%)	18 (3.9%)
Average precision (weighted)	18.47/25 (73.88%)	454

Table 46. Precision of the detection of grammatical *keep*.

Window	Number of correct instances	Total number of instances
V_{ing} is the 1st word to the right	25/25 (100%)	224 (69.6%)
V_{ing} is the 2nd word to the right	25/25 (100%)	67 (20.8%)
V_{ing} is the 3rd word to the right	22/25 (88.0%)	31 (9.6%)
Average precision (weighted)	24.71/25 (98.80%)	322

Table 47. Precision of the detection of grammatical *keep*, automatically removing instances of *keep from* V_{ing}.

As stated earlier, anything that was not detected by the selection process was considered to be a lexical use. To control for possible shortcomings, 50 lexical elements were randomly

selected (out of 1'598) and a manual investigation was conducted in order to look for overlooked grammatical uses. However, no such uses were found. Thus, the precision for the selection of the lexical items is also maximal. These lexical instances are also reported in appendix B1-B.

6.2.4 Quit

Finding the grammatical instances of *quit* involves a similar process to that of *keep*. These instances correspond to the pattern *quit* + V_{ing}, as in *quit talking*. Similar to *keep* + *Ving*, this pattern pertains to the grammatical domain of aspect and its grammatical function has been mentioned in previous research (Heine 1993: 66), although less extensively than that of *keep*. Since this pattern is rarer than *keep* + V_{ing}, the default COHA sample is not large enough to provide enough data to control the selection. Instead, a random draw (without replacements) of 5'000 files from the complete COHA dataset was performed. The following instances were searched in the sample:

<vv0> quit, <vvz> quits, <vvi> quit, and *<vvd> quit*

These instances correspond to the possible forms that grammatical *quit* can take. Therefore, the forms *<vvg> quitting* and *<vvn> quit* were not included as they do not appear in grammatical uses. In order to be counted as a grammatical *quit*, the next word has to be an *-ing* participle. A difference with regard to *keep* V_{ing} is that *quit* and the *-ing* participle have a near-categorical tendency to be adjacent. This was verified by running a version of the script which included up to two words in between. As Table 48 shows, no correct instances of *quit* + V_{ing} were found when words occur in between. They are also very rare, as there were only four occurrences of one word in between, and three occurrences of two words in between, as opposed to 43 occurrences with no word in between. A closer inspection reveals that when one word occurs in between, this typically involves cases such as *When did you quit your peddling in Georgia?* This example is a result of the inclusion of ambiguity tags, because *peddling* was tagged as an ambiguous *<nn1_vvg>*. This is not a grammatical use and should therefore not be included in the dataset. When two words occur in between, it tends to be an unrelated *-ing* participle, for example in a sentence such as *Edward can not quit America without sacrificing his prospects*. The choice is therefore easy and consists in only taking instances where the *-ing* participle directly occurs after *quit*, as this results in a maximum

precision of a 100%. Instances of *quit* that are not followed by such a tag are considered to be lexical uses.

Again, 50 lexical uses were randomly selected (out of 225 instances) and manually controlled. There do not appear to be any intruders in the dataset, as can be seen in appendix B2, so the lexical items are also selected with a precision of 100%.

Window	Number of correct instances	Total number of instances
V_{ing} is the 1st word to the right	25/25 (100%)	43 (86%)
V_{ing} is the 2nd word to the right	0/4 (0%)	4 (8%)
V_{ing} is the 3rd word to the right	0/3 (0%)	3 (6%)

Table 48. Precision of the detection of grammatical *quit*.

6.2.5 Have

Only the most common grammatical use of *have* is considered in the present study, namely its perfect use (e.g. *I have been to France*), which concerns the grammatical domain of tense. Other uses such as *have to*, which can be used to express obligation (domain of modality) have been discussed in the context of grammaticalization (e.g. Tagliamonte 2004), but they are not considered here to preserve the general binary classification. Finding the grammatical instances is achieved by finding whether an instance of *have* is followed by a past participle, as in *have done*. The following instances were found in the sample corpus:

<*vh0*> *have,* <*vhz*> *has,* <*vhi*> *have,* <*vhd*> *had,* and <*vhg*> *having*

In addition, the corresponding clitics were also retrieved:

<*vh0*> *'ve*, <*vhz*> *'s*, and <*vhd*> *'d*.

They correspond to the possible forms that an auxiliary *have* can take. In order to be counted as a grammatical *have*, a participle has to occur in a window of less than five words to the right. Past participles are tagged as <*vvn*> and <*vvnk*> for lexical verbs and catenatives, respectively. The verbs *be*, *have* and *do* use special variants of these tags, namely <*vbn*>, <*vhn*> and <*vdn*>. These tags were detected using the following regular expression:

<.*v(v|b|h|d)n.*>

As with the <*vvg*> tags in the *keep* section, ambiguous tags were also matched by this regular expression and they were equally kept. The same comments regarding regular expressions (1) and (2) also apply here. Instances of *have* that are not followed by a past participle tag are considered to be lexical uses.

The results of the 25 randomly selected items for each window size are reported in Table 49. The first observation is that *have* has a strong tendency to be directly followed by a past participle, as there are 30'255 occurrences of such a case. There are also 6'132 occurrences in which one word occurs in between. The rest of the cases are rather marginal in comparison, as they represent less than 6% of the data. This observation, combined with the fact that Table 49 shows a clear decrease in precision after more than one word in between (which is below 75%), makes one word in between (i.e. <*vvn*> is the second word to the right) a good threshold. This also gives the maximum precision score of 100%, while only sacrificing a low number of actually correct matches that occur in the bigger window sizes.

Window	Number of correct instances	Total number of instances
vvn is the 1st word to the right	25/25 (100%)	30'255 (78.6%)
vvn is the 2nd word to the right	25/25 (100%)	6'132 (15.9%)
vvn is the 3rd word to the right	18/25 (72.0%)	1'391 (3.6%)
vvn is the 4th word to the right	18/25 (72.0%)	405 (1.1%)
vvn is the 5th word to the right	7/25 (25.0%)	315 (0.8%)
Average precision (weighted)	24.5/25 (98%)	38'498

Table 49. Precision of the detection of grammatical *have*.

A debatable case is causative *have*, which is exemplified by the construction *to have something done*, which is considered as a grammatical use by this method. For example, in the sentence *He had them killed*, the verb *have* is not used as an auxiliary in the same way as in *He has killed them*. Such instances can be considered as an intermediate stage of grammaticalization. However, such cases are rather rare and they are not expected to skew the data in a significant way.

13'377 lexical uses of *have* were found, which means that they amount to 26.2% of all instances. A sample of these uses was also manually processed to determine whether they correctly consisted of uses of lexical *have*. Out of these 50 occurrences, 49 (98.0%) were indeed lexical. The grammatical use that made its way to the lexical sample was an instance of *had been receiving* in which *been* was not tagged as a <*vbn*> but instead as <*nn1_jj*> (i.e. ambiguous singular noun or adjective). In conclusion, the lexical dataset is also rather

reliable, despite some minor flaws. The complete set of sampled sentences can be found in appendix B3.

6.2.6 Go

The grammatical use of *go* that will be examined involves the *be going to* future construction, which consists of the pattern *be going to* + infinitive. This construction pertains to the grammatical domain of tense, as it is used to express the future. This pattern has been the object of many studies (e.g. Brisard 2001, Tagliamonte et al. 2014, Budts and Petré 2016). The shortened form *gonna* is also included as the grammatical construction. Its lexical counterpart are instances of *going to* that are followed by a preposition and a noun phrase which refer to a movement (e.g. *to go to the bathroom, to go down the road*). This construction is more complex than the ones that were presented earlier, and therefore the algorithm is presented in full. Instances of *be going to* are detected using the following steps:

1. Instances of <*vvg(k)*> *going* and <*vvg(k)*> *gon* are found in the sample corpus (the form *gonna* is tagged as two separate elements, *gon* and *na* in the COHA). The *going to* construction is supposed to be tagged as <*vvgk*>, which corresponds to catenative uses of participles. However, a small proportion of <*vvg*> *going* are actually instances of the *going to* construction in questions, which is why it is included here.
2. In order to be counted as a grammatical use of *go*, the next word has to be the preposition *to* and the word after that needs to be an infinitive (<*vvi*> tag). Note that *gonna* is tagged as <*vvgk*> *gon* <*to*> *na*, therefore *na* counts as the preposition *to*.

Note that only infinitives directly adjacent to *to* were considered, when in fact there is a fairly low number of occurrences where an element appears between the *going to* and its infinitive. These are for instance quotation marks, and also the adverb *really*. However, there were only four occurrences in the COHA sample, which means that they are quite marginal and can be ignored to make processing more efficient.

3. In order to be considered a lexical instance of *go*, the next word has to be a preposition, and the word after that needs to be an article (<*at(1)*> and <*appge*> tags).

4. 25 grammatical elements were randomly drawn to check for the precision of the algorithm, while 50 lexical elements were also randomly drawn. The reasoning to draw more lexical elements here is that they are more problematic and include more wrong matches, which is why it is better to include more data in such a case.

As can be seen in Table 50, there are several wrong matches in the lexical category. Among these wrong matches are mostly metaphorical uses of *go* which use the same syntactic pattern. For example, *to go against someone's wishes, to go into a frenzy,* or *the spoils go to the winner.* Note that some metaphorical uses also convey a possible notion of going to a specific place, such as *we're going to the police* (which can both mean to contact the police or to go to a police station), which were counted as correct instances. These wrong matches are not a fundamental problem because they are not grammatical instances where *going* fulfils an auxiliary function. These sentences can be found in appendix B4.

Window	Number of correct instances	Total number of instances
Grammatical *going to + Vinf*	25/25 (100%)	1'085 (87.3%)
Lexical *go + prep + article*	46/50 (92.0%)	158 (12.7%)

Table 50. Precision of the detection of grammatical *going to* and its lexical counterpart.

6.2.7 Die

The construction *be dying to + infinitive* is the grammatical use of the verb *to die* and is relevant to the grammatical domain of modality. There do not seem to be previous studies on this specific construction that argue for its grammatical status. Generally, the construction is used to express desire, a desire which can even be close to obligation given that *dying to* expresses a great craving. Desire and obligation can be conveyed by modals in English, such as *wanna* or *gotta* (Krug 2000: 53-166). This is why *dying* to + infinitive can be considered as a marker of modality as well. Furthermore, *dying to* takes a non-finite verbal complement, as all the verbs discussed in the previous sections do, which is also indicative of grammaticalization. This construction can be detected in the data by finding the following instances:

<*vvg_jj*> *dying* and <*vvg*> *dying*

Note that grammatical instances *dying* + *to* are generally tagged as <*vvg_jj*>, which is an ambiguity tag between a progressive verb and adjective, with more likelihood of being the former. In order to be counted as a grammatical *die*, the next word has to be the preposition *to* and the word after that needs to be an infinitive (<*vvi*> tag). Other instances were categorized as lexical. This also means that only the progressive form *dying* was included in the lexical uses.

As done previously, 25 grammatical elements were randomly drawn to check for the precision of the method, while 25 lexical elements were also randomly drawn. The grammatical construction *dying to* + *infinitive* has a rather low frequency in the COHA, which means that the default COHA sample is too small to find enough instances to evaluate the selection method. There were only three occurrences in that sample. Therefore, a larger sample of the COHA consisting of 10'000 random files is used in the present case. Table 51 shows that the grammatical counterpart has a much lower frequency. Despite this, the precision of the selection process is fairly accurate. The entire set of sampled sentences can be found in appendix B5.

Element	Number of correct instances	Total number of instances
Grammatical <*vvg*> *dying* <*to*> *to*	25/25 (100%)	48 (6.1%)
Lexical <*vvg*> *dying*	24/25 (96.0%)	735 (93.9%)

Table 51. Precision of the detection of grammatical and lexical *dying*.

6.2.8 Help

The grammatical use of *help* is its auxiliary-like function in the *help* + V_{ing} construction, for example in the sentence *I can't help feeling that we almost had it all*. Similarly to *dying to* in the previous section, this construction is relevant to the grammatical domain of modality. Other grammatical uses of *help* have been studied, such as *help to* + *infinitive* (Mair 1995, Lohmann 2011), but they are not included in the present study. The *help* + V_{ing} construction is similar to *keep* + V_{ing} and *quit* + V_{ing}, but with the added dimension that it is common for *help* + V_{ing} to be preceded by a modal verb and a negation, such as *can't*, *couldn't* or *wouldn't*. Nevertheless, the detection of such a pattern can easily be done using the usual method that starts by finding the following relevant instances of *help* in the sample corpus:

<*vvi*> *help*

Since grammatical *help* is always an infinitive (tagged <*vvi*>), only infinitives are included as possible lexical counterparts. In order to be counted as a grammatical *help*, the next word has to be an *-ing* participle. Grammatical *help* and its *-ing* participle are mostly adjacent, but a window of up to two words in between was included during the evaluation of the selection process to make sure that other cases are either non-existent or rare. It would be possible, in theory, to imagine an utterance such as *I couldn't help always feeling bad*, although it might be considered ungrammatical by some speakers. As Table 52 shows, no correct instances are found when going beyond direct adjacency. Furthermore, the process is maximally accurate. Instances of *help* that are not followed by an *-ing* participle were classified as lexical instances.

50 lexical instances were randomly drawn (from a total of 648 lexical instances) to control for potential wrong matches. None were found. The complete random draws can be found in appendix B6.

Window	Number of correct instances	Total number of instances
V_{ing} is the 1st word to the right	25/25 (100%)	65 (86.7%)
V_{ing} is the 2nd word to the right	0/4 (0%)	4 (5.3%)
V_{ing} is the 3rd word to the right	0/6 (0%)	6 (8.0%)

Table 52. Precision of the detection of grammatical *help*.

6.2.9 Use

The grammatical instances of *use* consist of the construction *used to* + infinitive. *Used to* has been referenced in Heine and Kuteva (2002: 308) and described as a quasi-auxiliary in Neels (2015). It pertains to the grammatical domain of aspect as it is a marker of habituality. All other instances of *used* were considered lexical uses. The following instances were retrieved:

<*vvd*> *used,* and <*vmk*> *used*

Instances of <*vmk*> *used* are classified as grammatical instances if they are immediately followed by the preposition *to*. Note that adding an infinitive after the *to* as a requirement was not done, since *used to* tends to appear at the end of sentences as well (e.g. *he did not come as much as he used to*). Instances of <*vvd*> *used* are classified as lexical instances.

A random draw of 25 items from the grammatical and lexical selections was performed, in order to manually check whether they indeed correspond to such uses. Adjectival uses

(<*jj*> tag) were not searched for as they might be unrelated to the lexical source which the grammatical counterpart developed from, although they may share the same form (e.g. *I'm not used to doing that*). As with most other cases, instances of *use* which had an ambiguity tag were not included to maintain high accuracy. The manual analysis of the 25 grammatical instances and of the 25 lexical instances only highlights one problem in the lexical sample, which is in fact an adjectival use (**requiring such tanks to he used outdoors*), which is due to the fact that the *he* in this sentence is a misspelled *be*, which probably caused a verbal tagging instead of an adjectival one. These results are reported in Table 53. The complete dataset can be found in appendix B7.

Element	Number of correct instances	Total number of instances
Grammatical <*vmk*> *used + to*	25/25 (100%)	318 (69.1%)
Lexical <*vvd*> *used*	24/25 (96.0%)	142 (30.9%)

Table 53. Precision of the detection of grammatical and lexical *use*.

6.2.10 Long

The grammatical instances of *long* correspond to the conjunction *as long as*, which is tagged as <*cs31*> *as* <*cs32*> *long* <*cs33*> *as*, which pertains to the grammatical domain of clause-combining (Haiman and Thompson 1988). This corresponds to its conditional function similar to *if* or *on the condition that* (e.g. *you can play as long as you do your homework*). The following instances are found in the sample corpus:

<*cs32*> *long,* <*jj*> *long*, <*rr*> *long*, <*rr_jj*> *long*, and <*jj_rr*> *long*

Note that the ambiguity tags were also included in this case. They are quite frequent as they account for 824 lexical instances out of the 2'848 that were found. Also, distinguishing between adjectival and adverbial cases is not a central concern in the present case as they both count as lexical instances. Instances of <*cs32*> *long* are classified as grammatical instances if they are preceded by <*cs31*> *as* and followed by <*cs33*> *as*, whereas adjectival <*jj*> and adverbial <*rr*> instances are classified as lexical uses.

A set of 25 items has been randomly drawn from the grammatical and lexical uses to assess the quality of the selection. As this construction is rather common, the default sample version of the COHA was used for the random draw. The reliability of the tags is acceptable,

even though some problems can be pointed out. There are several instances of *as long as* not used as a conjunction among the grammatical instances (e.g. *You can sleep as long as you need*), because they were tagged as a conjunction. Other lexical instances are correctly tagged, such as *as long as possible*, which is tagged as separate propositions: *<rg> as <rr> long <csa> as <jj> possible*. Accuracy is maximal for the lexical items, but lower for the grammatical items (76.0%), as shown in Table 54. The details can be found in appendix B8.

Element	Number of correct instances	Total number of instances
Grammatical *<cs32> long*	19/25 (76.0%)	142 (4.7%)
Lexical *<jj> long, <rr> long*	25/25 (100%)	2'848 (95.3%)

Table 54. Precision of the detection of grammatical and lexical *long*.

A short additional note to consider is that there can be genuinely ambiguous uses of *as long as*, such as:

(1) You can stay as long as you want to.

There are two ways to interpret such a statement:

(2) You can stay for any indefinite amount of time you want.
(3) You can stay, provided that you want to stay.

Disambiguation under those circumstances will likely depend on pragmatic matters. For instance, if sentence (1) was prompted by a question such as "how long can I stay", then meaning (2) can unambiguously be inferred. On the other hand, if sentence (1) was uttered by someone who was going to start an activity as an offer to another person to stay and watch, such as rehearsing something, then it can be interpreted as (3), since the willingness of the guest to stay might be the element under question (instead of the duration of the stay). Traugott and König (1991: 201) note that concessives such as *as long as* can express a condition while serving both as temporal frames of reference and as grounds for a situation, which can explain this possible ambiguity. Therefore, such a use of *as long as* can be considered as being somewhere in between lexical (temporal) and grammatical uses (conditional).

6.2.11 Well

The grammatical instances of *well* are part of the construction *as well as*, when used in a similar fashion to *and* or *in addition to*. It therefore has a prepositional use that is not far away from a conjunctional use given its similarity to *and*. An important note is that *well* also grammaticalized as a discourse marker (Traugott 1995, Fraser 1999), but this grammatical use is not considered in the present study. Furthermore, comparative uses such as *he can play as well as a professional* were not counted as grammatical uses. The lexical instances consist of adverbial and adjectival uses of *well*. The following instances are first retrieved:

<ii32> well, <jj> well, and *<rr> well*

Unlike *long*, *<rr_jj>* and *<jj_rr>* ambiguity tags do not seem to occur. Instances of *<ii32> well* are classified as grammatical instances if they are preceded by *<ii31> as* and followed by *<ii33> as*, whereas adjectival *<jj>* and adverbial *<rr>* instances are classified as lexical uses. In the case of *well*, there are ambiguity tags for *as well as* that are used only for the *as* parts. Namely, *well* is still tagged as unambiguous *<ii32> well*, but the surrounding tags are labeled as *<rg_ii31> as* and *<csa_ii33> as*. These ambiguous tags are ignored as they are not really reliable. They are neither included in the grammatical nor in the lexical instances and are simply discarded. Note that they consist of about 120 items in the default COHA sample dataset (vs. around 300 grammatical and around 3'000 lexical items).

As Table 55 shows, the results of the random selection of 25 lexical and grammatical items results in a maximal accuracy, which suggests that this case is rather unproblematic. Complete sentences can be found in appendix B9.

Element	Number of correct instances	Total number of instances
Grammatical *<ii32> well*	25/25 (100%)	331 (9.8%)
Lexical *<jj> well, <rr> well*	25/25 (100%)	3'061 (90.2%)

Table 55. Precision of the detection of grammatical and lexical *well.*

6.2.12 Soon

The construction *as soon as* was considered as instantiating a grammatical use of *soon*. The adverbial and adjectival uses were considered to be lexical uses. The method is similar to that of *long* and retrieves the following instances:

<cs32> soon, <jj> soon and *<rr> soon*

There do not seem to be ambiguity *<rr_jj>* and *<jj_rr>* tags. Instances of *<cs32> soon* are classified as grammatical instances if they are preceded by *<cs31> as* and followed by *<cs33> as*, whereas adjectival *<jj>* and adverbial *<rr>* instances are classified as lexical uses. Similarly to *long*, ambiguity tags occur with *<cs32> soon*, but they are extremely rare, with only two such occurrences in the default COHA sample dataset. They were therefore discarded.

The default COHA sample dataset was used to test the selection method by retrieving 25 random grammatical and lexical items. The results of this selection are shown in Table 56. As with *well*, the precision is high, as the problematic cases are not numerous. Also, in a similar fashion to *long*, the case of *as soon as possible* is correctly tagged, which results in its classification into the lexical instances. Complete sentences can be found in appendix B10.

Element	Number of correct instances	Total number of instances
Grammatical *<cs32> soon*	24/25 (96.0%)	281 (20.8%)
Lexical *<jj> soon, <rr> soon*	25/25 (100%)	1’067 (79.2%)

Table 56. Precision of the detection of grammatical and lexical *soon.*

6.2.13 Regard

The grammatical use of *regard* consists of the constructions *with regard to* and *in regard to*. It should be noted that these two constructions are considered equivalent, and that while *with regard to* is more common in current American English, this is only a recent development. Indeed, searching for the most frequent left collocate of *regard to* returns 10’320 instances of *in* versus 5’071 instances of *with* in the COHA. Nominal uses of *regard* are considered as the lexical counterparts to these constructions. The following instances are first retrieved in the default COHA sample:

<ii32> regard, *<nn1> regard,* and *<nn2> regards*

Instances of *<ii32> regard* are classified as grammatical instances if they are preceded by *<ii31> with* or *<ii31> in* and followed by *<ii33> to*, whereas nominal *<nn1>* and *<nn2>* instances are classified as lexical.

As can be seen in Table 57, the selection process is accurate, despite a few mistakes in the 25 randomly selected elements. Among the four mistakes reported in the table, one is a verbal use that is mistagged as a noun. Another one is a mistagging of the construction *in regard to* (which is tagged as *<ii> in <nn1> regard <ii> to*. Finally, two of the wrong lexical matches are grammatical constructions with the omission of initial *in/with*. Since they represent a low percentage, they can be ignored. The complete sentences can be found in appendix B11.

Element	Number of correct instances	Total number of instances
Grammatical *<ii32> regard*	25/25 (100%)	153 (63.0%)
Lexical *<nn1> regard <nn2> regards*	21/25 (84.0%)	90 (37.0%)

Table 57. Precision of the detection of grammatical and lexical *regard.*

6.2.14 Respect

The grammatical instances of *respect* consist of the construction *with respect to*. As with the construction *with regard to*, there are several variations that need to be taken into account. First, the construction *in respect to* also exists as an alternative to *with respect to*. Therefore, instances of *in respect to* are counted as grammatical instances. It should be noted that this alternative is however much less frequent in the 20th than in the 19th century, and also that it almost never occurs from 1980 to 2010. Second, the construction *in respect of* is sometimes used, but it is much less frequent than the other ones and will therefore be ignored, as it represents less than 10% of all occurrences of *with/in respect to*. The following instances are first found in the default COHA sample:

<nn1> respect, *<nn2> respects,* and *<ii32> respects*

Instances of *<ii32> respect* are classified as grammatical instances if they are preceded by *<ii31> with*, and followed by *<ii33> to*. In addition, instances of *<nn1> respect* that are preceded by *<ii> in* and followed by *<ii> to* are also classified as grammatical instances. Finally, other instances of *<nn1> respect* and *<nn2> respects* are classified as lexical uses.

The construction *with respect to* is properly tagged with the sequential tags *<ii31>*, *<ii32>* and *<ii33>*. There is only one instance (out of 66 in the default COHA sample) where the wrong tags *<iw> with* and *<ii> to* are used. Furthermore, the construction *in respect to*

never uses the sequential tags, but instead uses the regular following tags: *<ii> in <nn1> respect <ii> to*. There are no instances of *<ii31> in <ii32> respect <ii33> to* (out of 44 in the default COHA sample). The default COHA sample was used to retrieve 25 random lexical and grammatical items and investigate to what extent the selection is correct. As Table 58 shows, all the instances are correctly selected. Details can be found in appendix B12.

Element	Number of correct instances	Total number of instances
Grammatical *<ii32> respect*	25/25 (100%)	110 (25.8%)
Lexical *<nn1> respect <nn2> respects*	25/25 (100%)	317 (74.2%)

Table 58. Precision of the detection of grammatical and lexical *respect*.

6.2.15 Spite

The grammatical instances of *spite* consist of the complex preposition *in spite of*, which is tagged as *<ii31> in <ii32> spite <ii33> of*. Lexical instances are tagged as the noun *<nn1> spite*. The instances that were retrieved from the default COHA sample are therefore simply:

<nn1> spite and *<ii32> spite*

Instances of *<ii32> spite* are classified as grammatical instances, whereas instances of *<nn1> spite* are classified as lexical uses.

Element	Number of correct instances	Total number of instances
Grammatical *<ii32> spite*	25/25 (100%)	156 (89.1%)
Lexical *<nn1> spite*	9/19 (47.4%)	19 (10.9%)

Table 59. Precision of the detection of grammatical and lexical *spite*.

The random selection of 25 items from the grammatical and lexical categories was performed on the sampled version of the COHA. As Table 59 shows, there are only 19 lexical instances, which means that the default sample version of the COHA might not be large enough to collect enough data in the present case. A manual inspection of the lexical instances shows that there are unfortunately some errors in the tagging of the corpus, as clear instances of the complex preposition *in spite of* appear in the lexical selection. The sentences that were manually inspected can be found in appendix B13-A. This is why, with the default searching

strategy, there are only 47.4% of correct lexical instances of *spite* (reported in Table 59). The precision of the method is therefore in need of significant improvement.

In order to improve precision, an additional means of detecting instances of the complex preposition *in spite of* was added. Furthermore, a random selection of 5'000 files from the complete COHA was taken in order to have a big enough sample dataset. With the current method, some instances of the complex preposition *in spite of* are included in the lexical results because *spite* is tagged as <*nn1*> instead of <*ii32*>. To confirm that an <*nn1*> *spite* is indeed a lexical instance, it is necessary to look whether there is an *in* preposition before, or an *of* preposition after it. If either an *in* or an *of* is found, then it is very likely that this <*nn1*> *spite* is actually a grammatical <*ii32*> *spite* that's wrongly tagged. It should be mentioned that the preposition <*io*> *of* is sometimes shortened as <*zz1*> *o*, in particular in the earlier texts. To make the selection of the lexical elements more accurate, the following rules are applied to <*nn1*> *spite* to decide on its lexical status:

A. There must not be an instance of <*ii*> *in* immediately to its left
B. There must not be an instance of <*io*> *of* or <*zz1*> *o* immediately to its right.

If any of these rules are broken, then the instance is not counted as a lexical instance. Instead it then falls into the "probably grammatical" category reported in Table 60.

The first observation is that the new method is much more precise for the lexical elements. Some of the lexical elements tagged as <*nn1*> were in fact verbal uses, such as *but merely to spite his English cousins.* There were four of such uses among the 25 elements tagged as lexical uses, which are reported in appendix B13-B, which results in an accuracy of nominal tags of 84%. This is not ideal, but still acceptable. These four are wrong matches due to a wrong tagging of the corpus. Verbal uses of *spite* tend to be rare (83 occurrences in the whole corpus), which may be because some are wrongly tagged as nouns. Verbal uses of *spite* should not be included in the selection because the grammatical counterpart *in spite of* is formed on the basis of *spite* as a noun.

The elements that were included in the category "probably grammatical" were not retained in the end because they represented only a small portion of the dataset (3.3%). Also, there are 10 of these 25 occurrences that are likely an archaic use of *spite of*, as in *They came, spite of the cold*. Since these grammatical instances are marginal, they were not included in the final dataset.

Element	Number of correct instances	Total number of instances
Grammatical <*ii32*> *spite*	25/25 (100%)	889 (90.2%)
Lexical <*nn1*> *spite*	25/25 (100%)	64 (6.5%)
Probably grammatical <*nn1*> *spite*	25/25 (100%)	33 (3.3%)

Table 60. Precision of the detection of grammatical and lexical *spite*, using extra steps to remove some wrong matches in the lexical category.

6.2.16 Terms

The grammatical instances of *terms* consist of the construction *in terms of*. On the other hand, the lexical instances correspond to nominal uses in plural form. The following instances were first retrieved:

<*nn2*> *terms* and <*ii32*> *terms*

Instances of <*ii32*> *terms* are classified as grammatical instances if they are preceded by <*ii31*> *in* and followed by <*ii33*> *of*, whereas instances of <*nn2*> *terms* are classified as lexical uses. Unlike some of the other items, the random selection did not highlight wrong tagging of the construction *in terms of*. As Table 61 shows, there do not seem to be instances of wrong classifications, so the tagging is unproblematic and reliable in that specific case. The complete datasets can be found in appendix B14.

Element	Number of correct instances	Total number of instances
Grammatical <*ii32*> *terms*	25/25 (100%)	127 (26.1%)
Lexical <*nn2*> *terms*	25/25 (100%)	360 (73.9%)

Table 61. Precision of the detection of grammatical and lexical *terms*.

6.2.17 View

The grammatical instances of *view* consist of the construction *in view of*. On the other hand, the lexical instances correspond to nominal uses. The method used is the same as the one for *spite* and starts with the following instances:

<*nn1*> *view,* <*nn2*> *views* and <*ii32*> *view*

Instances of <*ii32*> *view* are classified as grammatical instances, whereas instances of <*nn1*> *view* and <*nn2*> *views* are classified as lexical uses.

While *in spite of* was sometimes found to be incorrectly tagged, the problem does not seem to happen with *in view of*, as it is consistently tagged as <*ii31*> *in* <*ii32*> *view* <*ii33*> *of*. There might in fact be one wrong instance, but this results from the COHA's regular gaps where the words are deleted for copyright reasons. As Table 62 shows, there might be an accuracy problem with the lexical instances as there is only a correct rate of 72%. However, all wrong instances are due to a particular case, namely the part of the default COHA sample that includes data on the 1938 Tom Sawyer movie. This section of the corpus includes the sequence <*nn1*> *view* <*nn1*> *page* <*nn1*> *image* which seems to be textual mark-up, rather than actual text. Examples can be found in appendix B15, which contains the sampled sentences. Unfortunately, since the default COHA sample is rather small, this section has a lot of weight as it contains more than 200 occurrences of *view*. All seven wrong instances are from this section, which means that the accuracy reaches a hundred percent when discarding the specific wrong chain <*nn1*> *view* <*nn1*> *page* <*nn1*> *image*. Therefore, in the final data collection, instances of *view* followed by *page image* were not counted as valid lexical items. This was done in order to avoid a skewed dataset for the year 1938 (and subsequently for the 1930-1940 chunk), but also because other documents, movies transcripts in particular, might also contain the same sequence.

Element	Number of correct instances	Total number of instances
Grammatical <*ii32*> *view*	25/25 (100%)	76 (7.5%)
Lexical <*nn1*> *view,* <*nn2*> *views*	18/25 (72%)	940 (92.5%)

Table 62. Precision of the detection of grammatical and lexical *view*.

6.2.18 Event

The construction *in the event of* is considered as the grammatical use of *event*, whereas its nominal use is its lexical counterpart. A similar method was used as with *spite* and *view*, but with the notable difference that the grammatical instance does not have its own encoding using sequential tags. Instead, it is just the sequence of the words *in, the, event* and *of* with their own independent tags. This is why the following instances were first searched:

<*nn1*> *event* and <*nn2*> *events*

When preceded by <*ii*> *in* <*at*> *the* and followed by <*io*> of, these instances are counted as the grammatical use. When this condition is not met, then the instance is classified as a lexical use.

As Table 63 shows, there were only 20 occurrences of grammatical uses in the default COHA sample, which is low but sufficient to evaluate the accuracy of the method by drawing random elements. As with similar constructions, the accuracy is maximal. The details can be found in appendix B16.

Element	Number of correct instances	Total number of instances
Grammatical *in the event of*	20/20 (100%)	20 (3.5%)
Lexical <*nn1*> *event*, <*nn2*> *events*	25/25 (100%)	558 (96.5%)

Table 63. Precision of the detection of grammatical and lexical *event*.

6.2.19 Matter

Grammatical instances of *matter* are part of the construction *as a matter of fact*. It is used in a similar way to *in fact* (Traugott 1995). It is the only discourse marker among the twenty selected elements and is also one of the most recent, as shown in Table 44. Discourse markers have been considered as instances of grammaticalization by several researchers (Traugott 1995, Brinton 2001, Schiffrin 2001), although there have been arguments against this viewpoint (Heine 2013). Outside this construction, the noun *matter* is considered lexical. The following instance were first found:

<*jj32_nn132*> *matter,* <*nn1*> *matter,* and <*nn2*> *matters*

Note that the construction *as a matter of fact* is usually tagged as <*ii*> *as* <*at1*> *a* <*jj31_nn131*> *matter* < *jj32_132* > *of* < *jj33_nn133*> *fact* in the COHA. It means that *matter of fact* is tagged both as an adjectival and nominal construction. In addition to the 47 instances found in the default COHA sample dataset, there was only one instance in which the construction is tagged as <*jj31*> *matter* <*jj32*> *of* <*jj33*> *fact* for no apparent reason. Therefore, only the most common tagging was retrieved. Instances of <*jj31_nn131*> *matter* are classified as grammatical instances if they are preceded by <*ii*> *as* + <*at1*> *a* and followed by <*cjj32_nn132*> *of* + <*jj33_nn133*> *fact*. Nominal <*nn1*> and <*nn2*> instances are classified as lexical uses.

The default COHA sample dataset was used to test the precision of this selection method by random sampling. Table 64 shows that the precision is maximal. There do not seem to be any problematic cases. The complete sample sentence can be found in appendix B17. It should however be pointed out that the grammatical counterpart is significantly less frequent than the lexical one. To make sure that no instances were missed, an additional script was made to only use lemmas and discard part of speech tagging entirely. Apart from the one peculiar instance mentioned above *(<jj31-33> tags*), no other instances were found, which means that the method is fairly reliable.

Element	Number of correct instances	Total number of instances
Grammatical *<jj31_nn131> matter*	25/25 (100%)	47 (3.3%)
Lexical *<nn1> matter <nn2> matters*	25/25 (100%)	1'384 (96.7%)

Table 64. Precision of the detection of grammatical and lexical *matter*.

6.2.20 Means

Grammatical instances of *means* consist of the construction *by means of*, whereas its lexical counterpart is the noun *means*. The following instances are found:

<ii32> means and *<nn> means*

Instances of *<ii32> means* are classified as grammatical instances if they are preceded by *<ii31> by* and followed by *<ii33> of*, whereas nominal *<nn>* instances are classified as lexical. In the case of *means*, there are also a few sequences of *<ii> by <nn> means <io> of*, which are not tagged using the <ii31-33> label, but which are nonetheless correct grammatical instances. These instances were also searched for and added into the grammatical category.

As some instances of *<ii32> means* only come with either *by* or *of*, they were discarded. These cases are rare, as there are only three in the default COHA sample dataset. In two cases either the *by* or *of* are simply missing from the corpus, and in another case it is a wrong tag as it is actually a nominal instance (namely the phrase *systematic means of translation*). This is why the decision is to discard these few cases entirely.

The default COHA sample dataset was used to evaluate the quality of the selection, which is quite accurate as can be read in Table 65. There is only one wrong match in the

lexical category, in the following sentence: *I say, vat does " elopes " means?* Further inspection of the corpus reveals that this is from a work of fiction from 1877 in which a character with a (probably) French accent produces this utterance. This is therefore a verbal use that was not tagged properly. The sampled dataset can be found in appendix B18.

Element	Number of correct instances	Total number of instances
Grammatical <*ii32*> *means*	25/25 (100%)	118 (22.8%)
Lexical <*nn*> *means*	24/25 (96.0%)	400 (77.2%)

Table 65. Precision of the detection of grammatical and lexical *means*.

6.2.21 Order

The construction *in order to* is considered as the grammatical counterpart of the nominal use of *order*. This construction is a purpose clause (Schmidtke-Bode 2009) and pertains to the grammatical domain of clause-combining. The following instances were found:

<*bcl22*> *order*, <*nn1*> *order,* and <*nn2*> *orders*

Instances of <*bcl22*> *order* are classified as grammatical instances if they are preceded by <*bcl21*> *in* and followed by <*to*> *to*, whereas nominal <*nn1*> and <*nn2*> instances are classified as lexical. There are also a few sequences of <*ii*> *in* <*nn1*> *order* <*to*> *to*, which are not tagged using the <bcl21-22> label, but which are nonetheless correct grammatical instances. These are, however, quite rare, with only six instances found (out of 340) in the default COHA sample dataset.

A small number (56) of instances of <*bcl22*> *order* were not directly followed by a <*to*> *to*. They mostly consist of the construction *in order that*, which could be argued as an alternative to *in order to*. A quick inspection of the whole COHA shows that *in order that* was somewhat frequent in the late 19th century, but that it decreased significantly in the mid to late 20th century. Another part of these instances of <bcl22> order are simple instances of *in order* which would normally be tagged as <*ii*> *in* <*order*>, for example in a phrase such as *to put your house in order*. All these instances were discarded because they constitute a low number overall. Thus, a very low number of actual instances of *in order to* get discarded as well, since some of these are cases with an additional element in between. For example, there are a few cases such as *in order not to* or *in order for you to* where additional elements appear in the

middle of the construction. It should be made clear that these examples are removed here because their detection cannot be easily automated in a way that also consistently removes wrong matches. They are in fact a sign that the bondedness of the elements of *in order to* is not maximal (although quite high), which is interesting from a more qualitative point of view.

The accuracy of the selection method is unsurprisingly maximal given the discarded items, as shown in Table 66 which reports the results of the 25 randomly drawn lexical and grammatical elements. The entire selected dataset can be found in appendix B19.

Element	Number of correct instances	Total number of instances
Grammatical <*bcl22*> *order*	25/25 (100%)	340 (26.2%)
Lexical <*nn1*> *order* <*nn2*> *orders*	25/25 (100%)	960 (73.8%)

Table 66. Precision of the detection of grammatical and lexical *order*.

6.2.22 Back

The grammatical use of *back* corresponds to its prepositional-adverbial function, which is tagged as <*rp*> in the dataset. It acts as an aspectual marker. Note that *back* is described as an adverb in Heine and Kuteva (2002: 48-49). There might be some terminological difference between "adverb" and "prepositional adverb", the latter term being the one used in the tagging of the Corpus of Historical American English, but they are considered as equivalent in the present case. Adverbs were initially categorized as lexical elements in chapter 5, but they can display grammatical function as well (Traugott and Trousdale 2013: 73). On the other hand, the lexical use is found by searching for the <*nn1*> and <*nn2*> tags (singular and plural nouns, respectively). The following instances were retrieved:

<*rp*> *back,* <*nn1*> *back,* and <*nn2*> *backs*

Instances of <*rp*> *back* are classified as grammatical instances, whereas instances of <*nn1*> *back* and <*nn2*> *backs* are classified as lexical uses.

50 items were randomly drawn from the grammatical and lexical selections, in order to manually check whether they indeed correspond to such uses. A key point in this selection process is that ambiguity tags are not included. The reasoning is that in this situation, tags are central in the selection process. In the cases of *keep*, *quit* and *have*, ambiguity tags can be

included because there is an initial pattern which makes the ambiguous tag very likely to be the desired one (i.e. *vvg* or *vvn*). In the case of *back*, there is no such initial selection.

The sampled items were analyzed manually, and the complete sentences can be found in appendix B20-A. The results are reported in Table 67, which shows that the precision of the method is maximal for the grammatical items. This is mostly due to the absence of ambiguity tags. The precision is smaller when it comes to the lexical uses (88.0%), as there are in fact some prepositional-adverbial uses that are wrongly tagged as *nn1* in the corpus. Despite these errors, the overall precision is still quite acceptable.

A legitimate question, however, is whether many instances are not missed this way. First, it should be noted that two possible tags are not included in the selection process at all, namely verbal and adjectival tags. In the COHA sample, there are only three occurrences of strict matches of adjectival *back*, which correspond to *<jj> back*. The main reason for this low number of occurrences is that this tag is usually part of an ambiguity tag. These ambiguity tags are mostly used for compounds such as *back door* or *back seat*. They can be found with the following regular expression:

<.*jj.*> back$

There are 238 such instances in the COHA sample and most of them contain an *nn1* or *nn2* tag. Verbal tags (*vv0, vvz, vvi, vvd, vvn, vvg*), on the other hand, have 87 strict matches and 42 ambiguity tags. This is small in comparison to nominal uses of *back*, which is why they are not included. What is more relevant, however, is the number of ambiguous *nn1* or *nn2* tags.

In order to find ambiguous tags that include an *nn1* or *nn2* tag, the following regular expression can be used:

<.*nn(1|2).*> back$

This regular expression finds all the items in the corpus that end in *back*, and that are preceded by a < > tagset that contains an *nn1* or *nn2* somewhere in between. Note that it is important to use this regular expression as part of an if/elif structure which first detects the strict matches of *<nn1> back* and *<nn2> back*. The reason is that these strict matches are also detected by the regular expression. Therefore, the regular expression must only take effect after the strict matches have been found. Using such a regular expression provides 928 matches, which means that many ambiguity tags include a possible nominal use.

To test whether the ambiguity tags are reliable, 50 items were randomly selected. The complete sentences can be found in appendix B20-B. In an ambiguity tag, the possible tags occur in a sequence, and the first item is the most likely one. For example, in the ambiguity tag <*rp_nn1*> *back*, the most likely tag is <*rp*>. The 50 items were processed manually in order to determine whether the first element in the sentence was indeed the correct one. There are 41 cases in which the first element of the ambiguity tag is the correct one. It should be noted that among the 50 items, 37 were cases of prepositional-adverbial uses (74%), 9 were adjectival uses (18%), and only 4 of them were nominal uses (8%). Therefore, the instances that are missed by not including the ambiguity tags are mostly grammatical uses. Also, the distribution of these missed ambiguity tags is close to the percentages of grammatical and lexical items reported in Table 67. Removing the ambiguity tags is thus positive for the overall accuracy of the results in this specific case.

Window	Number of correct instances	Total number of instances
Grammatical <*rp*> *back*	50/50 (100%)	2’662 (82.3%)
Lexical <*nn1/nn2*> *back*	44/50 (88.0%)	574 (17.7%)

Table 67. Precision of the detection of grammatical and lexical *back*.

6.3 General results

The goal of this section is to show overall tendencies when all the twenty elements presented in the previous section are aggregated. First, the focus is on how each variable develops over time, and how the lexical and grammatical counterparts differ (section 6.3.1). This will lead to the main aim of this section, namely to observe if there is an increasing grammaticalization score on average (section 6.3.2), as would be expected under the unidirectionality hypothesis (section 2.4). As in chapter 5, the variables are calculated using the methods presented in chapter 3, and an example script is provided in appendix D3. The only difference is that values are calculated for each decade. Therefore, each item has twenty measures in total, as there are twenty decades between 1810 and 2009. What is discussed in this section are averages computed using all twenty items.

6.3.1 Individual variables

The first variable under discussion is token frequency, given its prominent role in the previous chapter and also in previous research. In the subsequent graphs, frequency is always the

number of occurrences per million words. This is necessary given that in the COHA, the different decades do not contain the exact same number of words (section 4.2.2). The most basic information is shown in Figure 8, namely the average token frequency over time for the average of the twenty elements. An important note that is relevant for all the figures discussed in this section is that the early years (from 1810 to about 1870) are difficult to interpret because they are smaller than the other chunks of the corpus and also have less genre variety. When a chunk is much smaller than the rest (as are the early years), even using a per million word ratio cannot be completely reliable.

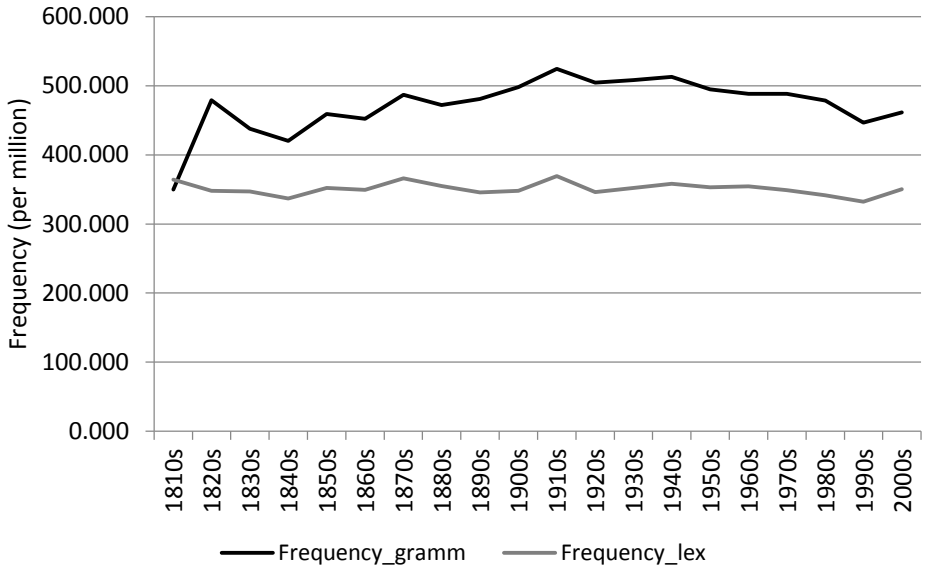


Figure 8. Average token frequency (per million) of the twenty elements in the database.

The first relevant information that is displayed in Figure 8 is that the grammatical counterparts are on average more frequent, as expected given the results obtained in chapter 5 where frequency correlated positively with grammaticalization. There is a difference of about 100 words per million that can be observed for the whole period. The second important element to notice in Figure 8 is that despite this difference, the frequencies of the lexical and grammatical counterparts seem to follow the same patterns of change. Small peaks can be observed in the 1870s, in the 1910s and at the very end in the 2000s. The general observation, then, is that when a grammatical counterpart becomes more frequent, so does its lexical counterpart. There might be a small effect due to the different chunk sizes of the corpus, in particular between the 1910s and 1920s, where there is a jump from 22'655'252 to 25'632'411 words (section 4.2.2), as well as a small difference in the balance of the genres

(fiction becomes less prominent from the 1920s onwards), but these are minor concerns. What is of importance is that the grammatical and lexical counterparts display the same tendencies over time, even if these tendencies might be slightly biased for specific decades.

To study frequency from a different perspective, it is possible to look at the average increases in frequencies, instead of the average frequency. Instead of doing a simple average of the frequency per decade as above, the increases in frequencies of each item from one decade to the next are averaged. This results in Figure 9, which shows increases and decreases in a clearer way.

To clarify why Figure 8 and Figure 9 are so different, it is important to understand that averages are sensitive to extreme measures. For example, if we have four items that have frequencies of 1, 1, 1 and 1'000'000, the average ends up being 250'000.75. This is a limitation in Figure 8 and the averages it illustrates. Figure 8 is dominated by elements such as *have* which have high frequencies. However, if one compares percentages in increases and decreases, which by definition range from 0 to 1, this effect is much more limited. In such a case, elements that are less frequent can be better represented. This is why Figure 9 can give a more representative idea of increases and decreases.

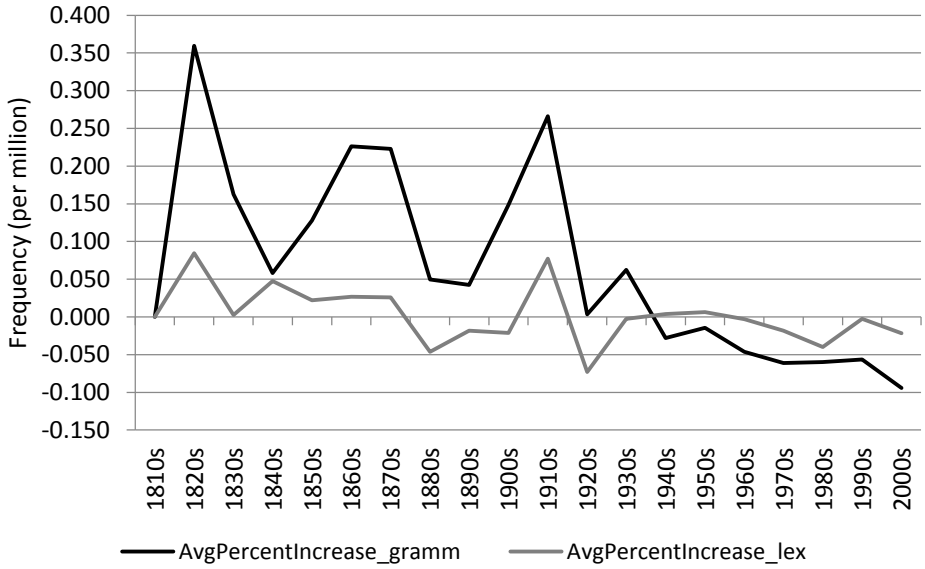


Figure 9. Average increase in token frequency of the twenty elements in the database.

The first observation regarding Figure 9 is that grammatical elements display more peaks and more fluctuation than their lexical counterparts. The lexical counterparts are in fact quite stable, as they stay between 10% decreases and increases. As observed in Figure 8,

there is also a common peak in the 1910s, which again might be due to the size and genre balance of that particular decade. What is more surprising in the present case is the common decrease in the latter years, especially for the grammatical elements. In fact, this is due to several of the complex prepositions in the database falling out of fashion, such as *with regard to*. The second main observation is that these fluctuations are mostly large increases instead of decreases, as one can see from the vertical axis in Figure 9 which shows peaks up to 35%, but not lower than -10%. To sum up, these figures show that grammatical counterparts are on average more frequent, and that they tend to show more fluctuations over time. On the other hand, lexical counterparts are generally less frequent and show fewer fluctuations.

The next variables to discuss are the collocate diversity variables, since their role in chapter 5 was also rather prominent. Given the way collocate diversity is computed (section 3.3), looking at the average over time is sufficient. The problem discussed with frequency above, where high frequency items tend to hide the others, is not really present with collocate diversity. It consists in the number of types divided by frequency, for a given window. As discussed in section 3.3, the maximum diversity value corresponds to the window size. For example, when using a symmetrical window of four items on the left and four on the right, with a given item with a frequency of n, there are 8*n collocates in total. The maximum value is obtained when all these collocates are different, which corresponds to 8*n types. In this maximal situation, dividing the number of types (8*n) by the frequency (n), results in 8 which is the size of the window. The minimal value consists in only one type divided by the frequency, which means that it corresponds to 1 divided by n (which is close to 0 unless n is small). Therefore, for a symmetrical four words window, the values range from 0 to 8. For a one word window, the values range from 0 to 1. This is why simply looking at the average collocate diversity measures over time is sufficient to understand what is happening. This comment also holds for the colligate diversity measures and for deviation of proportions.

Figure 10 shows the average collocate diversity over time for all three possible window sizes. It is quite apparent that the three situations are very similar. In all cases, the grammatical counterparts display higher diversity than the lexical counterparts. This is consistent with the observations in chapter 5 on the synchronic dataset. Another similarity with the synchronic dataset is that the three measures display the same tendencies overall. In the synchronic dataset, it was shown that these variables tend to strongly correlate with each other. These results therefore illustrate that there is consistency between the synchronic and diachronic studies.

These variables seem relatively stable over time as there are no drastic increases or decreases. There is however a clear increase in the latter years, but it is relatively small. Given the prominence of collocate diversity in the models from the previous chapters, this relatively stable situation might also be observed with the grammaticalization scores that will be discussed later in section 6.3.2, especially since the average frequency displayed in Figure 8 was also stable, despite a few small peaks.

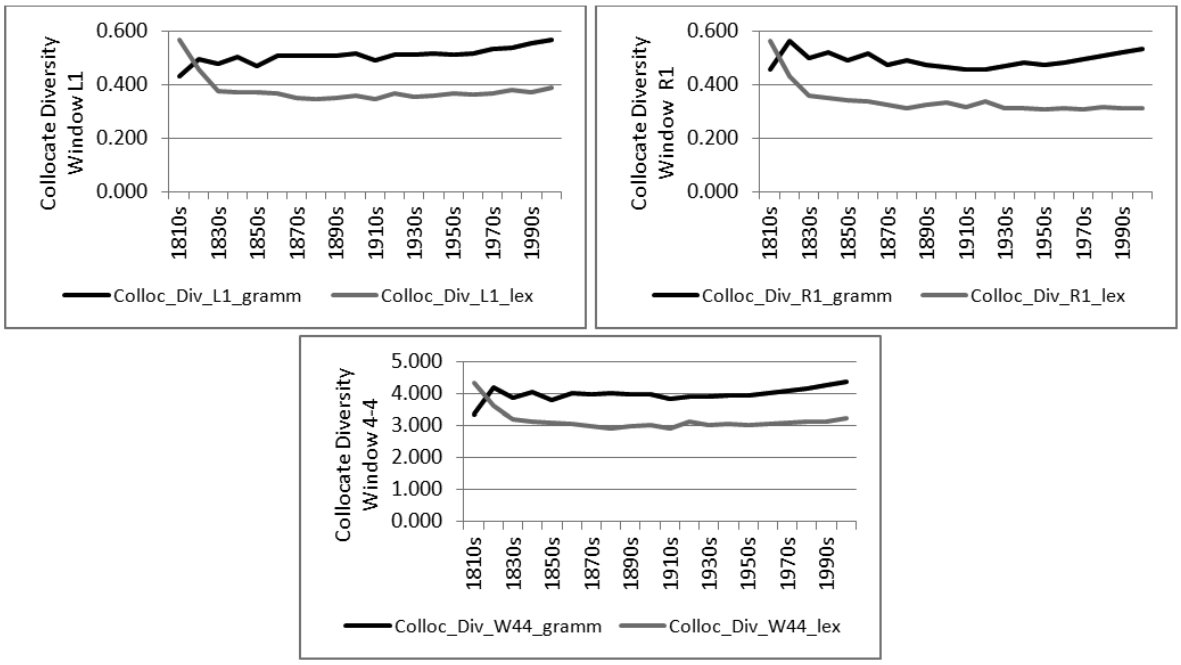


Figure 10. Average collocate diversity over time. The three different window sizes are shown.

The next set of variables are the colligate diversity variables, which consist in the percentage of the most frequent colligate (section 3.4). The average values for the one word to the left window and for the one word to the right window are displayed in Figure 11. In both cases, the values are relatively stable over time, as observed with the collocate diversity measures. However, the right colligate diversity measure shows a larger gap between the grammatical and the lexical counterparts. In this situation, the grammatical counterparts display higher values, which means that there is less colligate diversity. Again, this is consistent with the observations in chapter 5 where grammatical elements showed less colligate diversity to the right. The results for the left colligate diversity measure are more surprising since there is almost no difference between the grammatical and lexical counterparts. In the synchronic experiment, grammatical elements were shown to have more colligate diversity to the left. What is possible is that in the diachronic selection, there are

many elements in the database that involve strong constraints to their left, regardless of whether they are the grammatical or the lexical counterparts. Most verbs tend to have pronouns to their left and functioning as either a lexical verb or an auxiliary does not necessarily change this.

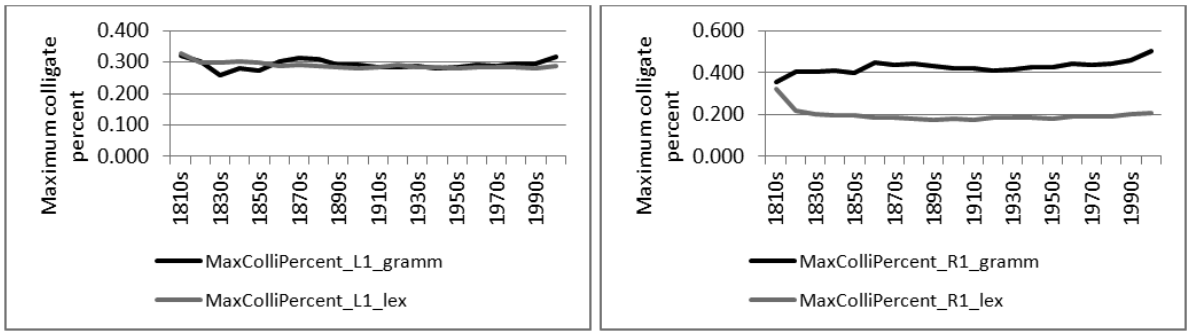


Figure 11. Average colligate diversity over time. The two different window sizes are shown.

Finally, deviation of proportions is the last variable in the model. Figure 12 shows how the average deviation of proportions of grammatical and lexical counterparts changes over time. The results are somewhat puzzling since grammatical elements have higher deviation of proportions, which means that they are less evenly distributed, which goes against the observations in chapter 5.

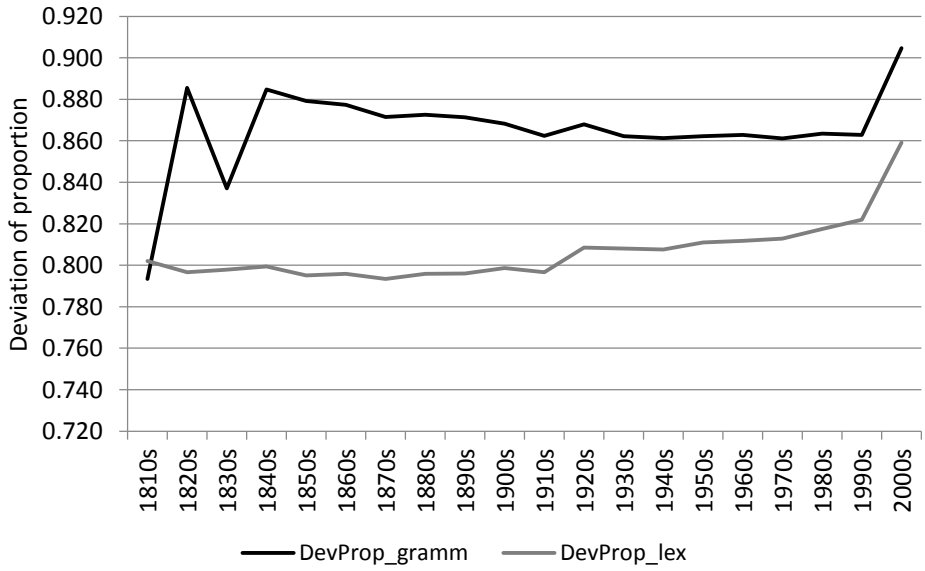


Figure 12. Average deviation of proportions over time.

One possible explanation is that it is difficult for some of the low frequency items in the database to have their deviation of proportions well-represented. For example, if a word only occurs 1'000 times in a decade that consists of 20 million words, using chunks of 1'000 words for the procedure (section 3.5), this results into 2'000 chunks, which means that the 1'000 words can distribute however they want, they will not be able to be in all 2'000 chunks to get a somewhat even distribution. The synchronic dataset did not have this problem because it mostly involved elements that are quite frequent. Here, some elements have grammatical counterparts which are actually low frequency elements (e.g. *can't help Ving*). This situation can also explain the burst in the last decade. As discussed earlier, some grammatical elements such as *with regard to* are less frequent in the late decades. In short, it means that deviation of proportions is not fully reliable in this particular situation. It would have worked better if the frequencies of the elements in the database were more similar. Low frequency elements are further discussed in section 6.4.

To summarize, the results are generally consistent with the synchronic analysis, as grammatical counterparts are generally more frequent, have more collocate diversity given all windows and have less colligate diversity to the right than their lexical counterparts. The notable difference is that grammatical counterparts do not show higher colligate diversity to the left. The variables overall tend to show little evolution over time and present stable values. However, the grammatical counterparts tend to display more changes in frequency from one decade to the other, as well as a slight increase in collocate diversity in the latter years. The next section focuses on a grammaticalization score computed on the basis of the synchronic model presented in chapter 5.

6.3.2 Grammaticalization score over time

Based on the variables discussed in the previous section, a grammaticalization score can be computed using the coefficients of the binary logistic regression trained on the synchronic dataset. Since the COHA is a written corpus, the model trained on the written portion of the BNC (section 5.4.2) was used. The diachronic dataset was standardized in a similar fashion to the synchronic one, although these standardization processes were performed separately. This results in the average grammaticalization scores shown in Figure 13. As can be seen, the grammatical counterparts receive, on average, higher grammaticalization scores than their lexical counterparts, which shows that the method produces coherent results. What can also be observed is that the grammaticalization score of lexical counterpart tends to be stable

(barring the early years). In contrast, the score of the grammatical elements shows more fluctuations and can be seen to increase in the late decades. Figure 13 therefore motivates the notion of unidirectionality from a perspective that is based on a set of twenty items. Therefore, future studies that involve a larger number of items could benefit from this method in order to further support the unidirectionality hypothesis.

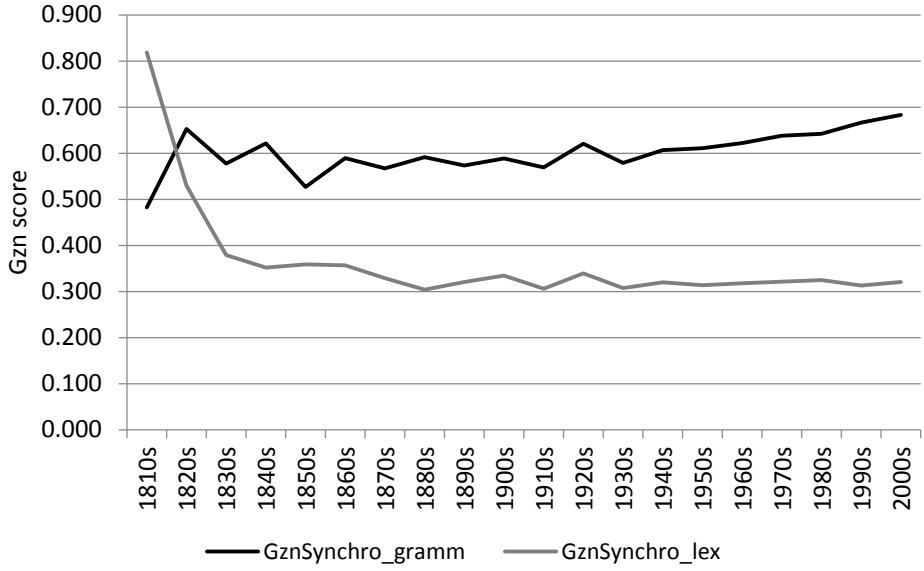


Figure 13. Average grammaticalization scores over time for the lexical and grammatical counterparts. The score is computed using the binary logistic regression trained on the BNC written dataset (section 5.4.2).

An additional observation is that the lexical and grammatical counterparts occasionally display similar increases and decreases, such as the peak in the 1910-1920 decades. This was also observed with their frequencies in the previous section (Figure 8 and Figure 9), although the peaks were in slightly different positions. A possible explanation is that this is an effect of corpus size, given that the decades of the COHA have different sizes, particularly when it comes to the early 20th century, as highlighted in section 4.2.2.

An alternative to compute the grammaticalization score is to train a completely new statistical model using the diachronic dataset only. However, this approach is presented here as an alternative because it involves a number of problems. The main problem is that while the BNC dataset uses 528 different elements, the diachronic dataset only has 20 elements, which does not allow generalizations to the same extent. A second problem is that the observations are auto-correlated, since for instance the frequency of an item in the 1920s

tends to be related to its frequency in the 1910s and so on. The interest in training a binary logistic regression on the diachronic dataset is to compare the coefficients of the model with the one that was introduced in section 5.4.2.

The coefficients are shown in Table 68. There are significant differences between the two models, as expected. If the diachronic dataset is used to train a binary logistic regression, the right colligate diversity measure is more prominent in the model, which is consistent as it was indeed one of the variables that showed the biggest gap between lexical and grammatical counterparts in Figure 11. What is also of interest is that despite major differences in the magnitude of the coefficients, the signs are the same. The difference between the left and right colligates, namely that grammatical elements have more colligate diversity to the left, but less to the right, is observed here too. However there is one exception, namely deviation of proportions, which has a positive coefficient. This is unexpected because grammatical elements tend to be more evenly distributed (i.e. smaller deviation of proportions value) and a negative coefficient would have been expected, as observed with the BNC spoken dataset (section 5.6.2). One explanation discussed earlier is that despite grammatical elements being more frequent on average, there are in fact many low frequency grammatical elements in the diachronic dataset. This also explains why the frequency coefficient is much lower than in the synchronic model. This particular situation will be further explored when the deviation of proportions variable will be discussed on an individual basis (section 6.4.1). In terms of the quality of the model, the diachronic training results in a very poorly fitted model, with a Cox&Snell R^2 of 0.296 and a Nagelkerke R^2 of 0.395. Despite the bad fit, the model trained on the diachronic dataset has a correct classification rate of 75.9%. This is nonetheless lower than what the synchronic training achieved on the synchronic dataset (85.0%).

	B (diachronic)	B (synchro_written)	S.E. (diachro)	Wald (diachro)	Sig. (diachro)
FreqWrit	0.759	10.245	0.222	11.683	0.001
Colloc_Div_W44	0.471	2.074	0.139	11.571	0.001
MaxColliPercent_R1	1.160	0.222	0.132	77.303	0.000
MaxColliPercent_L1	-0.483	-0.640	0.096	25.297	0.000
DevProp	0.731	-	0.220	11.032	0.001
Constant	0.075	1.830	0.091	0.681	0.409

Table 68. Coefficients of a binary logistic regression trained on the diachronic model only, with the synchronic (BNC written) coefficients for comparison.

Despite these limitations, it is still possible to compute a grammaticalization score and see how it evolves over time. Figure 14 shows the results of this process. What this figure shows is similar to Figure 13, where grammatical counterparts receive higher scores than their lexical counterparts, as expected. It also displays an increase in the later years. This is interesting because Figure 13 and Figure 14 are obtained by quite different means and yet display similar results. The models put emphasis on different variables, as shown in Table 68.

One way to interpret these results is to consider the synchronic model as the most general one, since it involves many more different items that belong to different morphosyntactic categories. On the other hand, the diachronic model contains less diversity (mostly auxiliaries and complex prepositions) and is therefore tailor-made to these specific cases. As a result, the variables involved in the model are similar, yet more emphasis is put on right colligate diversity. This is not surprising given that auxiliaries and prepositions tend to strongly constrain what may appear to their right. This is in line with what was observed in section 5.7 where the right-based coefficients became more significant when considering modal verbs. The diachronic training of the model therefore shows that the generalizations based on the synchronic training are rather robust.

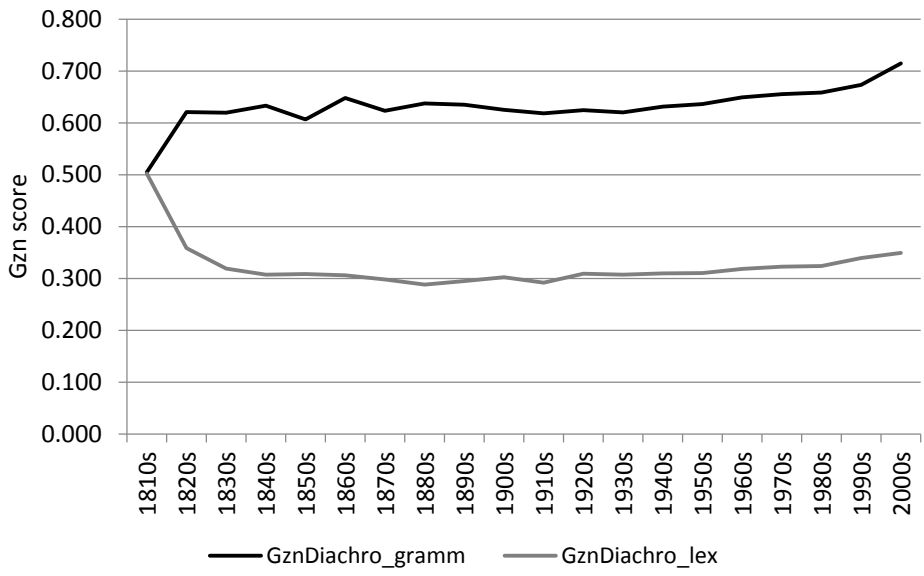


Figure 14. Average grammaticalization scores over time for the lexical and grammatical counterparts. The score is computed using a binary logistic regression trained on the BNC diachronic dataset (section 6.2).

6.3.3 Overview

The aim of the general investigation of the diachronic dataset was to show whether the unidirectional tendencies of grammaticalization can be shown empirically. Binary logistic regressions trained both on the synchronic and diachronic datasets show an increasing grammaticalization score over time, which supports the idea of unidirectionality when it comes to the selected elements. Further studies that involve a larger number of elements, and from a broader range of grammaticalization processes, could provide a more representative sample that would further support the unidirectionality hypothesis. A secondary aim was to see how the main variables (frequency, collocate diversity, colligate diversity and deviation of proportions) behave over time and how they relate to what was observed in the synchronic model. Many of these variables show rather stable values over time, but the differences observed between lexical and grammatical counterparts are consistent with what was observed in chapter 5. Grammatical counterparts tend to be more frequent, and involve more collocate diversity overall. They however display less colligate diversity to the right, which was observed in chapter 5 despite this effect not being systematically significant in all versions of the studies. A further interesting observation is that while frequency was generally decreasing in the late decades (Figure 8), grammaticalization scores associated to that period did not decrease and in fact display a small increase both for the synchronically trained (Figure 13) and diachronically trained (Figure 14) models during the same period. This is in line with Bybee (2011) who suggests that grammaticalizing constructions which no longer increase in frequency do not display a reversal of the process of grammaticalization.

However, these results are based on an average aggregation of the twenty elements in the database. The limitations of this approach have been highlighted previously, such as the fact that high values tend to have more influence on those averages. This is minimized by the fact that except for frequency, most values are constrained to specific ranges (such as percentages) and that in those cases, those limitations are less problematic. The analysis presented in the next section takes a different approach, and instead of doing averages, the results are shown on an individual basis.

6.4 Individual results

The previous section has shown the results based on an average of the twenty elements in the diachronic database. This section keeps the individual items separate and discusses specific cases. Figures will show all twenty elements together at the same time for comparison.

6.4.1 Individual variables

The twenty elements comprise very different frequency ranges as can be seen on Figure 15. The vertical axes show that there are low frequency elements that are below 100 occurrences per million words and high frequency elements that go up to almost 10'000 occurrences per million words. Many frequency ranges are therefore represented. What is of interest is that in many cases, the grammatical counterpart is in fact less frequent than the lexical counterpart. Indeed, the following elements have more frequent lexical counterparts: *die, event, help, keep, long, matter, means, order, quit, respect, soon, terms, view and well*. This is fourteen elements out of twenty, which means that in this dataset, this situation is the norm. This can explain why in section 6.3.2 the version of the binary logistic regression trained on the diachronic dataset gives frequency a very small coefficient, as opposed to its synchronic counterpart. This situation is in fact also a good way of testing the synchronic model. One possible criticism of the synchronic model (based on the written portion of the BNC) is that it gives frequency a fairly large weight in the binary logistic regression, when there are in fact examples of low frequency grammaticalization. This is why it will be interesting to see what kinds of scores the synchronic model gives to some of these low frequency elements. This is discussed in section 6.4.2, which covers grammaticalization scores.

Regarding the evolution of the frequency of these elements over time, several cases can be observed. There are traditional and expected cases, such as *go, back, keep, terms* and *use* where a clear increase in frequency can be observed for the grammatical counterpart. There are also elements such as *quit* that do display an increase, although very small. The other grammatical counterparts tend to display a rather stable behaviour, as observed with *die, event, have, help, long, matter, means, order, soon, view,* and *well*. There are some rare cases where the grammaticalized counterpart becomes less frequent, such as *regard* and *respect*. Finally, *spite* shows the most peculiar case as it first increases until the early twentieth century, but decreases in the later years.

Thus, the most common case is stability, while increasing tendencies are also fairly common. This is in line with previous observations that grammaticalization tends to entail increases in frequency. It also seems reasonable to assume that stability in frequency comes from the fact that specific elements simply stop grammaticalizing at some point. The decreases observed with *regard, respect,* and *spite* are probably due to concerns of style rather than grammaticalization. It is quite possible that writers simply tend to disfavour the use of these complex prepositions. Instead of using *with regard to* or *with respect to*, alternatives

such as *about* might have come to be preferred. Similarly, using *despite* might be a preferred alternative to *in spite of*. *About* and *despite* are indeed becoming more frequent as illustrated by Figure 16. However, investigating this situation would require taking all possible alternatives into account and also that there are different possible uses, which goes beyond the scope of the present discussion.

A brief additional methodological remark is that while several lexical elements are more frequent than their grammatical counterparts, the cases that involve a complex preposition and its nominal lexical source (e.g. *in order to / order*) might have shown a smaller difference between the counterparts if plurals had not been included (see section 6.2.2 on this issue). For instance, lexical *order* would have decreased by about 30%, which would have made its token frequency closer to its grammatical counterpart.

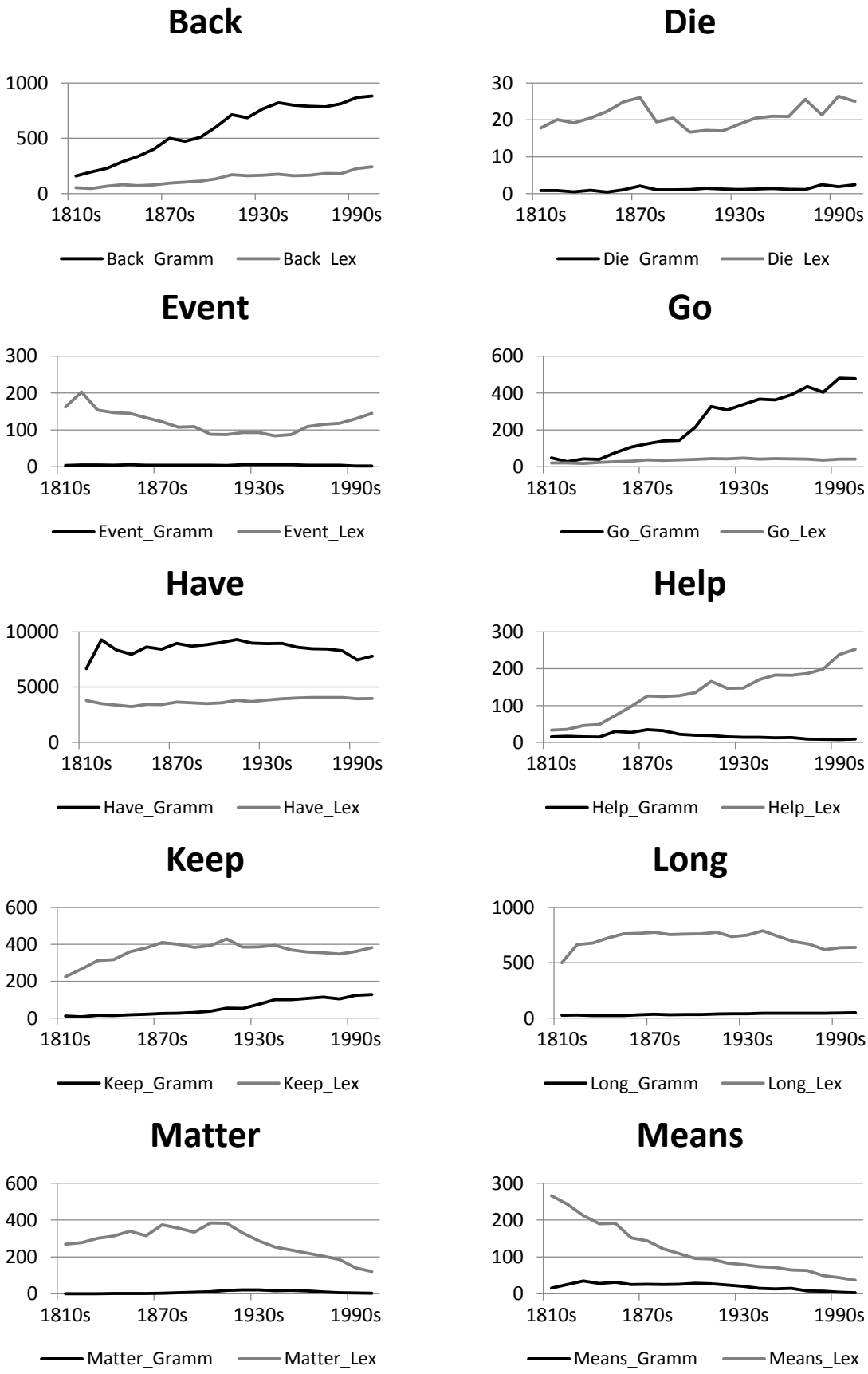


Figure 15. Frequency over time. The horizontal axis corresponds to each decade, while the vertical axis shows token frequency (per million words).

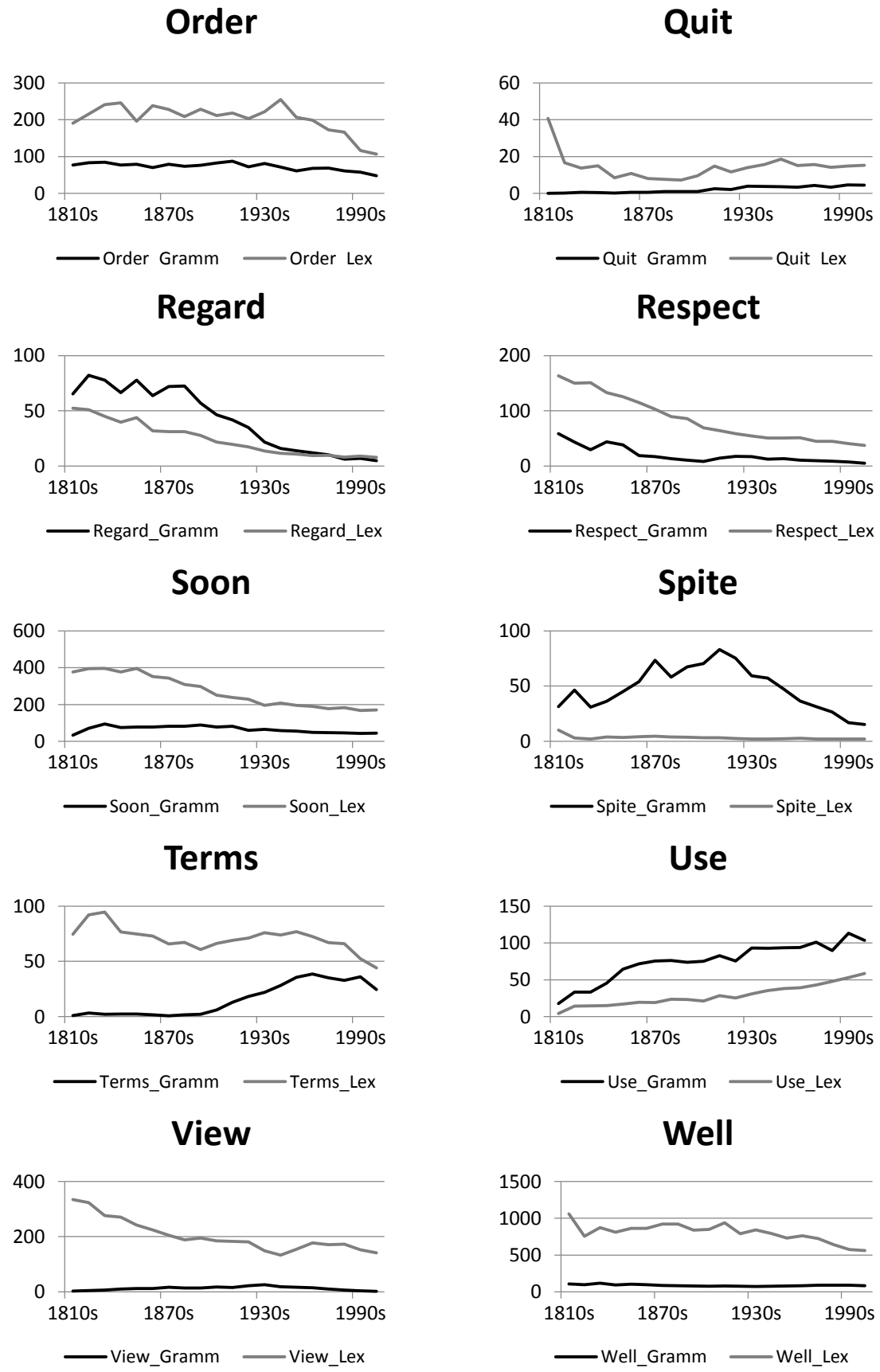


Figure 15 (continued). Frequency over time. The horizontal axis corresponds to each decade, while the vertical axis shows token frequency (per million words).

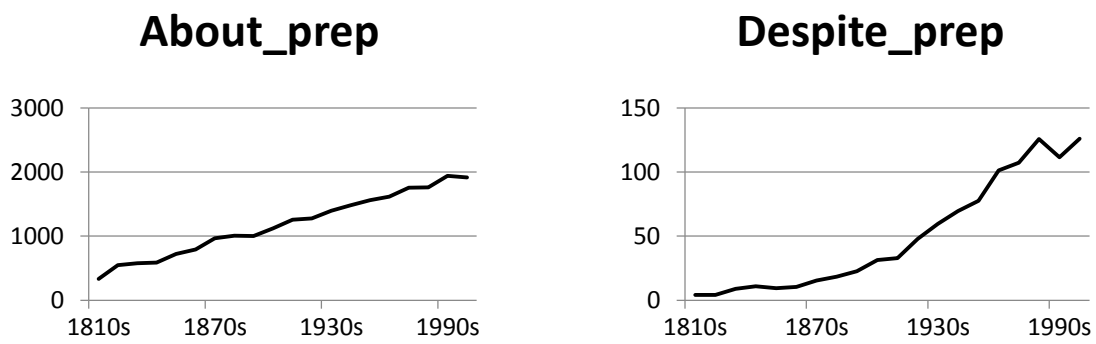


Figure 16. Token frequency (per million) of prepositional *about* and *despite* over time.

The next variable to discuss is collocate diversity. Rather than a redundant discussion of all graphs for all three windows I will only discuss the graphs regarding the 4-4-window. This window contains the smaller ones and that, as shown in chapter 5, these three variables tend to be strongly correlated. The graphs for the one word to the left and one word to the right collocate diversity can be found in appendix C.

The collocate diversity scores over time are shown in Figure 17. Note that the maximum possible score is eight. There are many cases where there is little difference in diversity between the lexical and grammatical counterparts. This can be observed with *back, have, help, regard, soon,* and *use*. In some cases, however, there is a clear difference in diversity. Many items display as expected a higher diversity when it comes to their grammatical counterparts, namely *die, event, keep, long, matter, means, order, quit, respect, terms, view,* and *well*. This therefore confirms the hypothesis that grammatical counterparts tend to involve more collocate diversity, which was also the outcome of chapter 5. On the other hand, two items display grammatical counterparts that have less collocate diversity, namely *go* and *spite*. This goes against expectations, but it should be noted that the difference is not that large. In the case of *spite* it is not necessarily surprising since it has been suggested earlier that *in spite of* might be falling out of use, which can therefore also explain why it gets used in a narrow range of contexts, which can in turn explain its low collocate diversity.

Regarding *go*, the situation is more difficult to evaluate and requires discussing the methodology more closely. Collocate diversity is measured by dividing the number of different types by the overall frequency of the item (section 3.3). This is necessary because otherwise, highly frequent elements would always have a larger number of different collocates, just because they happen to be frequent. One of the limitations of this approach is that high frequency elements tend to involve many repetitions, which decreases their diversity

score. A consequence of this design is that an element such as *go*, whose grammatical construction increases in frequency as shown in Figure 15, ends up having a much lower diversity than its lexical counterpart. In fact, looking at other elements that had a high frequency in Figure 15, namely *have* and *back*, the same tendency can be observed in Figure 17. *Have* and *back* were mentioned as showing little difference between the lexical and grammatical counterparts, but if one looks closely, the lexical counterpart does indeed display a higher diversity.

When it comes to changes in collocate diversity over time, the grammatical variant in most cases shows stable behaviour. This holds for *back, event, have, help, long, matter, means, order, regard, respect, soon, spite, terms, use, view,* and *well*. There are some examples where diversity increases over time, such as *die* and *quit*, which is what would be expected in case of an increase in grammaticalization. In contrast, some elements display a decreasing behaviour such as *go* and *keep*, which is not what is expected in case of increasing grammaticalization. However, this situation can be attributed to issues that have been discussed previously with the early years of the corpus. On the one hand, they constitute smaller chunks as opposed to the later years, and on the other hand they include less genre diversity by consisting mostly of fiction texts. Therefore, cases such as *keep* and *go* can be considered as stable situations, when the early years are discarded.

Colligate diversity also consists in two possible windows, but it was observed earlier in Figure 11 that the average colligate diversity of one item to the left does not show major differences between lexical and grammatical counterparts. These observations remain true at the individual level, which means that the individual graphs will not be further discussed. They can be found in appendix C for any curious readers. On the other hand, colligate diversity of one item to the right is shown in Figure 18.

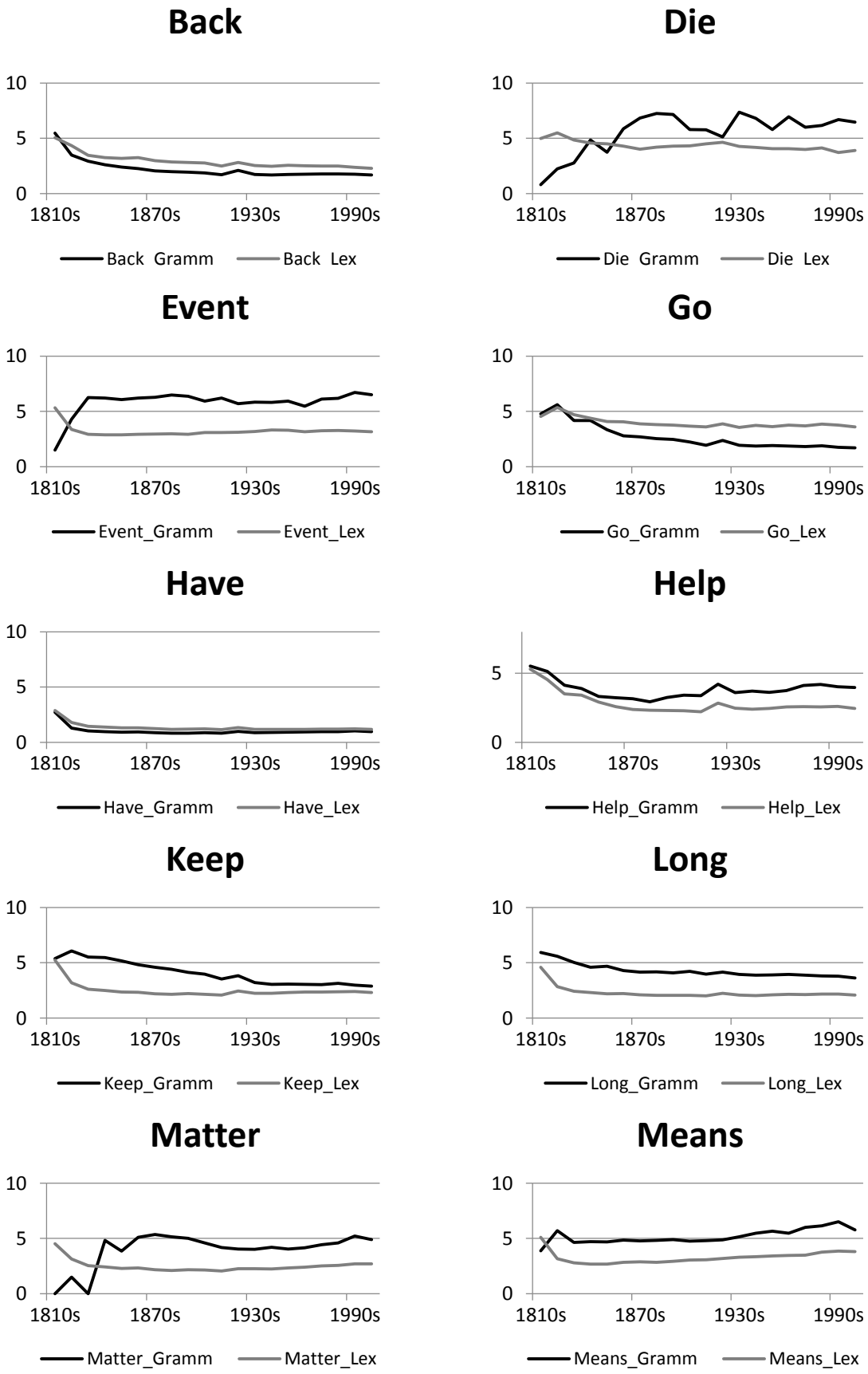


Figure 17. Collocate diversity (4-4 window) over time. The horizontal axis corresponds to each decade, while the vertical axis shows collocate diversity.

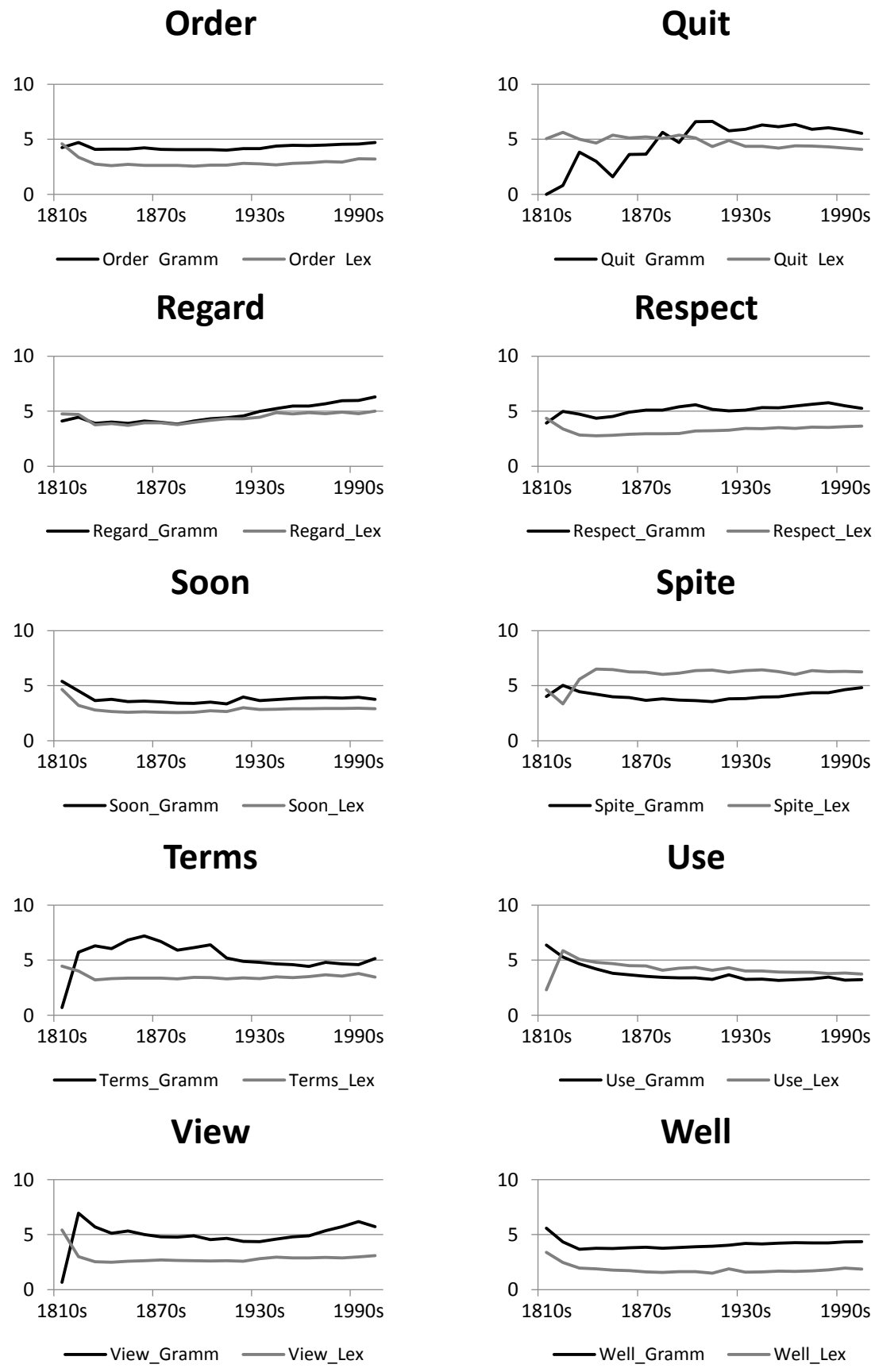


Figure 17 (continued). Collocate diversity (4-4 window) over time. The horizontal axis corresponds to each decade, while the vertical axis shows collocate diversity.

As observed earlier with average colligate diversity to the right (Figure 11), the grammatical counterparts tend to involve less colligate diversity, which is consistent with the observations in chapter 5. Figure 18 shows many grammatical counterparts confirming this, namely *die, event, go, have, help, keep, matter, means, order, quit, respect, terms,* and *use*. This is not surprising for some elements. Indeed, *in order to* is a preposition followed by infinitives. Similarly, *going to, dying to,* and *used to* also tend to be mainly followed by infinitives. As Figure 18 shows however, this is not always an absolute situation. In the case of *going to*, there is a clear increase which almost reaches 100%, which shows how bonded the infinitive is starting to become in the construction *be going to* + infinitive. In contrast, auxiliary *have* is often followed by a participle, but not systematically. A small note here is that since *be*, *have* and *do* have special verbal tags, so do their participles. Thus, the occurrences of *have been, have done,* and *have had* are not added to the total of colligate diversity, which means that the values in Figure 18 are probably a bit lower than they should.

Some elements do not display much difference between their counterparts, more specifically, *back, regard, soon, view,* and *well*. There are two cases where the lexical counterparts display less colligate diversity than their grammatical counterparts, namely *spite* and *long*. In the case of *long*, this difference is rather small and can be dismissed. Again, *spite* shows how special it is because a low colligate diversity on the right would be expected, since the determiner *the* to often appear after *in spite of*. However, manual inspection reveals that while *the* is a very frequent collocate, possessive pronouns (e.g. *his, her, their*) are in fact the most frequent types of collocates. Because there is a lot of competition, the score of the most common type (i.e. possessive pronouns) gets a smaller value in the end. Again this is more of a methodological observation to explain the odd one out, but as stated above, most elements conform to the expected situation where grammatical elements display less colligate diversity to the right.

Regarding the changes over time of colligate diversity to the right, stability is again the most common observation. This can be seen with the grammatical counterparts of *back, event, go, have, help, keep, long, means, order, regard, respect, soon, spite, use, view,* and *well*. Some grammatical elements display an increase (i.e. reduction in colligate diversity), namely *die, keep* and *matter* (omitting early years). The grammatical counterparts of *terms* and *quit* display a chaotic behaviour with increases and decreases which are hard to interpret. The colligate diversity measure is of limited use when it comes to these specific items.

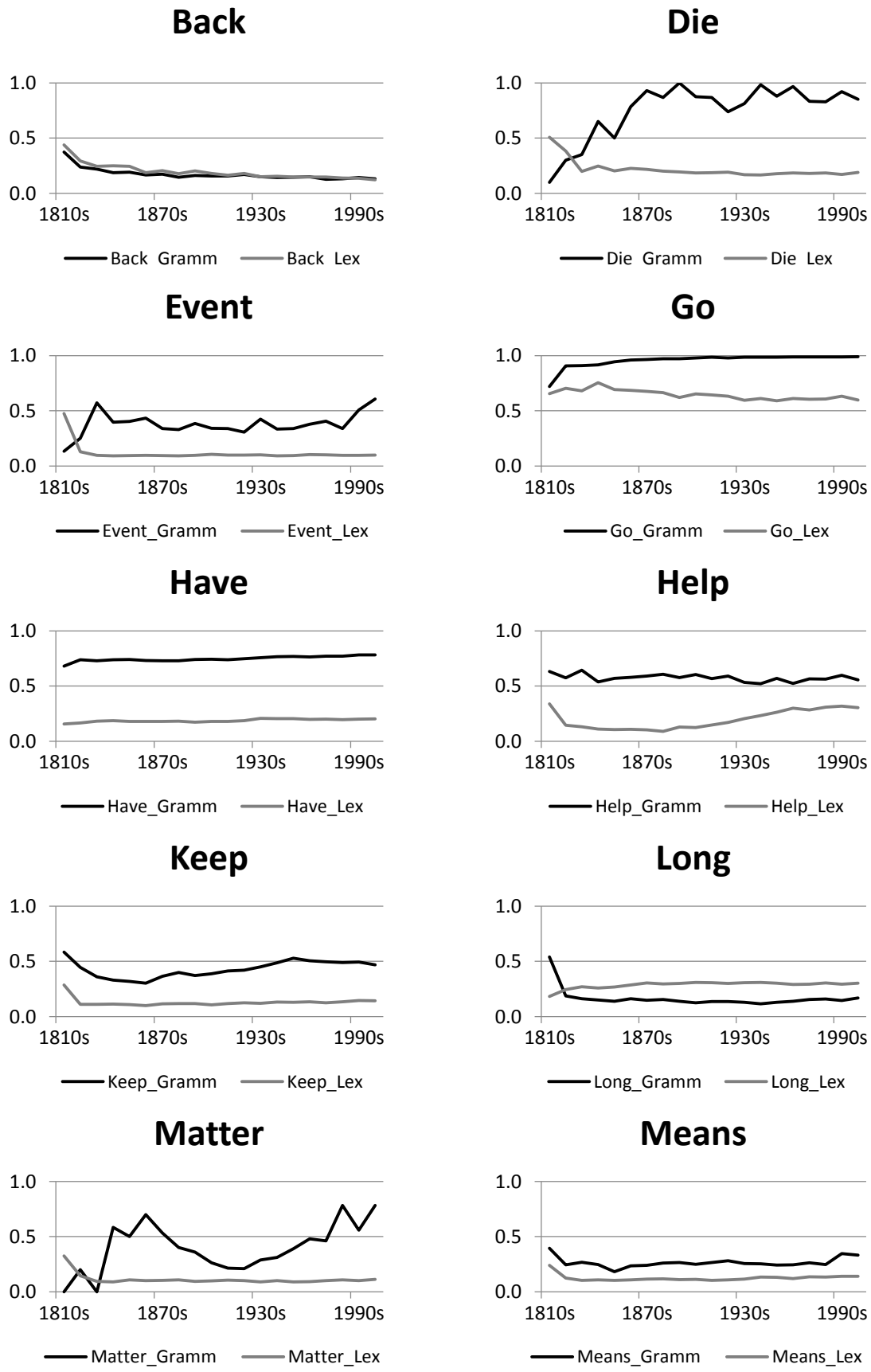


Figure 18. Colligate diversity (R1 window) over time. The horizontal axis corresponds to each decade, while the vertical axis shows colligate diversity.

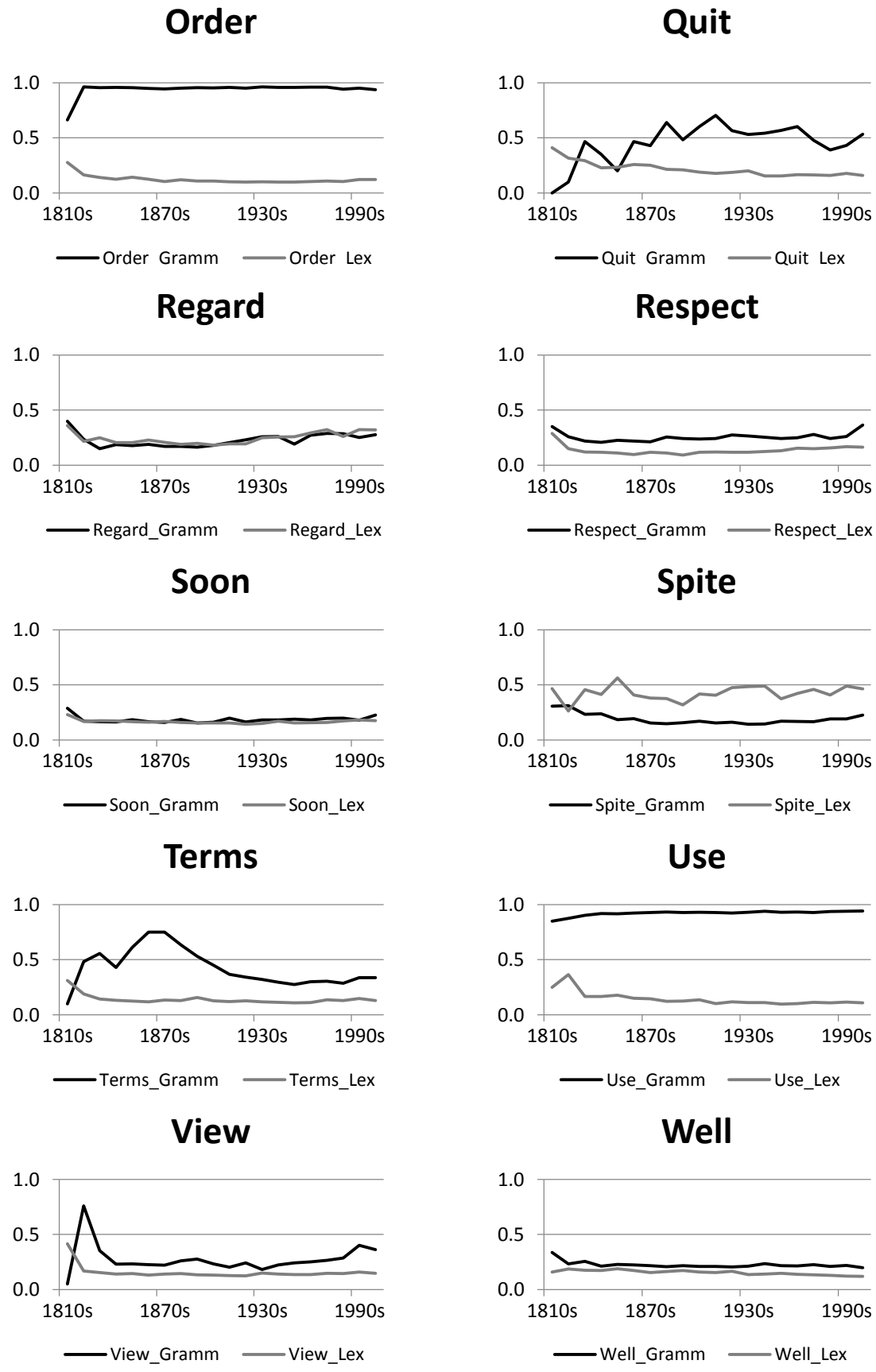


Figure 18 (continued). Colligate diversity (R1 window) over time. The horizontal axis corresponds to each decade, while the vertical axis shows colligate diversity.

The last variable to discuss is deviation of proportions, which has already shown surprising results earlier by having a less even distribution for grammatical counterparts in Figure 12 and by having a positive coefficient in Table 68. It means that the observed behaviour in the diachronic dataset is not similar to that observed in chapter 5. Figure 19 allows us to explore this by looking at individual behaviours. One explanation that was put forth in section 6.3.1 is that the diachronic dataset involves many low frequency elements, and that makes it difficult if not impossible for them to evenly distribute in the large chunks of the corpus. This explanation is still quite likely given Figure 19.

Most items do not have differences between the lexical and grammatical counterparts, such as *die, have, order, quit, regard, respect, spite, terms,* and *use*. The items that display a less even distribution for the grammatical counterparts are *event, help, keep, long, matter, means, soon, view,* and *well*. These are nine items that do not show the expected result. However, Figure 15 shows that all nine elements are low frequency as they are under 100 instances per million words. Therefore, overall low frequency of the grammatical counterparts is a likely explanation as to why their distribution is so uneven in the corpus. A good illustration is auxiliary *have*, which does have a high frequency in Figure 15 and which displays a very even distribution in Figure 19.

To sum up, investigating the items in the diachronic dataset on a more individual basis confirms the observations made in section 6.3.1 that presented the average results. It confirms that the variables in the diachronic dataset show a similar situation as the synchronic dataset. Grammatical elements tend to involve more collocate diversity overall, but less colligate diversity to the right. The main difference with the synchronic dataset is the role of deviation of proportions, as in the diachronic dataset, grammatical counterparts are less evenly distributed. This has been attributed to the fact that low frequency grammatical elements are more represented in the diachronic dataset. Investigation on an individual basis confirms that it is indeed the low frequency elements that display this unexpected behaviour.

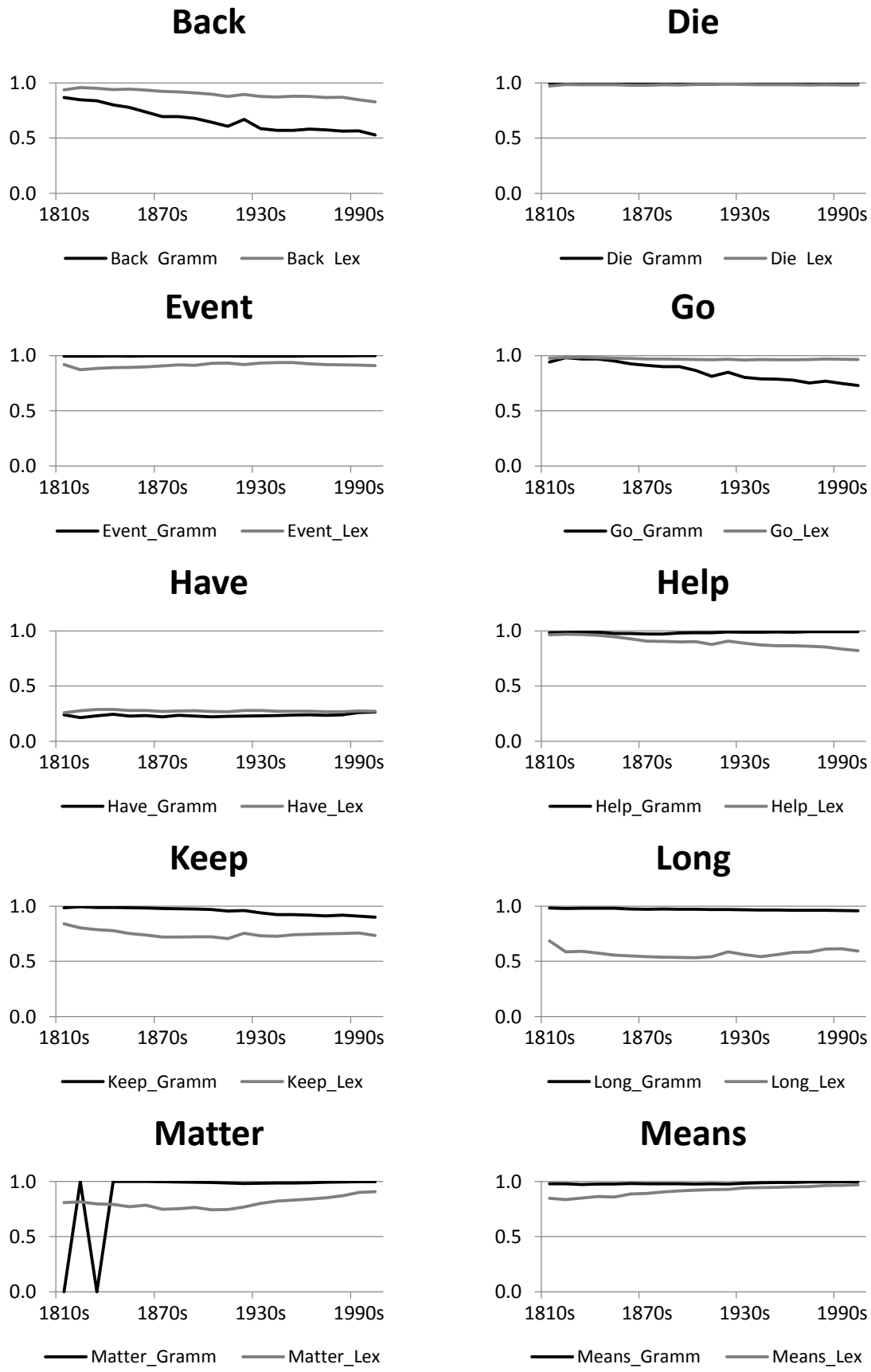


Figure 19. Deviation of proportions over time. The horizontal axis corresponds to each decade, while the vertical axis shows deviation of proportions.

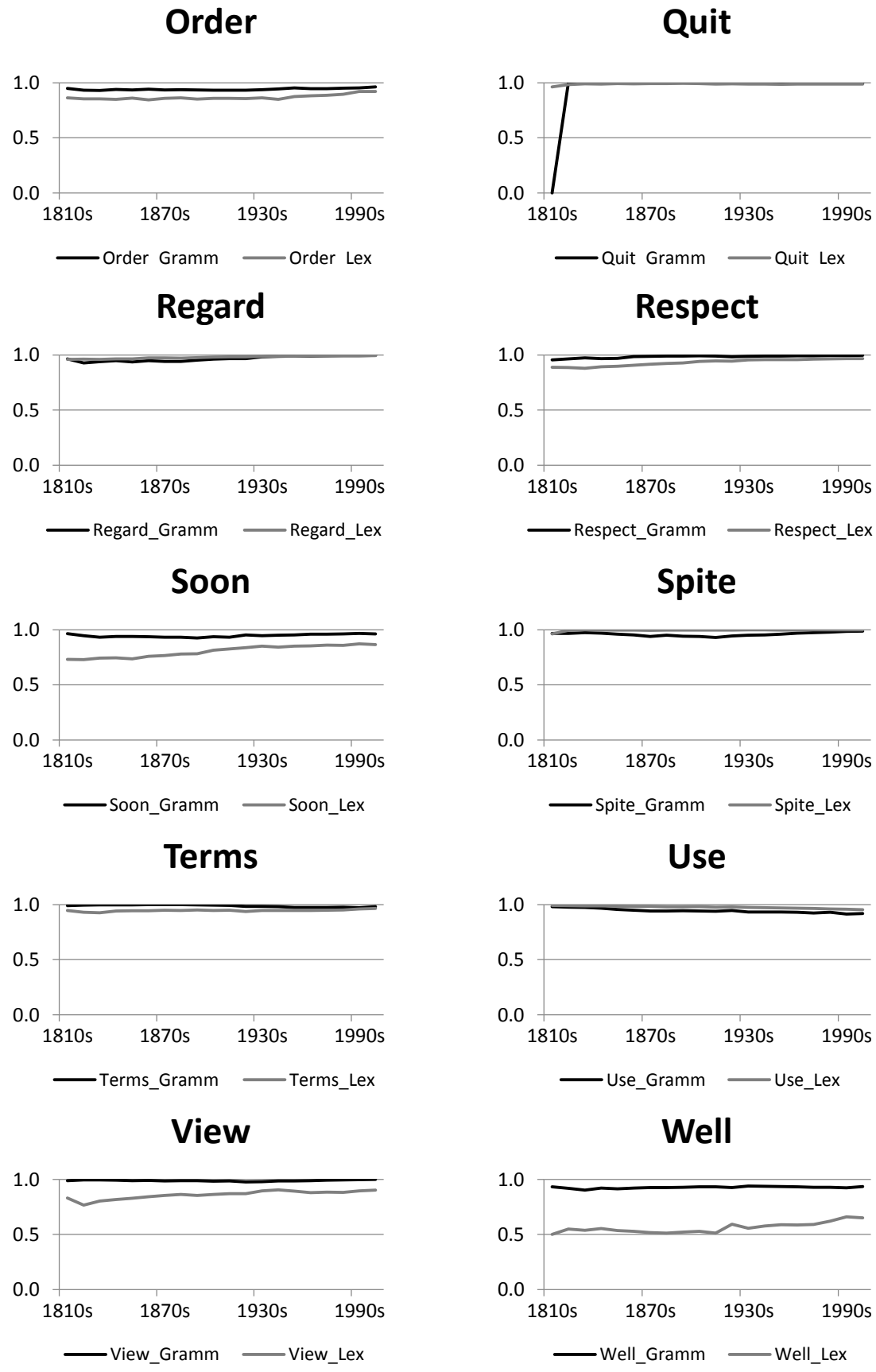


Figure 19 (continued). Deviation of proportions over time. The horizontal axis corresponds to each decade, while the vertical axis shows deviation of proportions.

6.4.2 Grammaticalization score over time

This section discusses grammaticalization scores computed on the basis of the synchronic model based on the written portion of the BNC introduced in section 5.4.2, as well as grammaticalization scores computed on the basis of the diachronic dataset itself, as was done in section 6.3.2.

Figure 20 reports the grammaticalization scores of the items over time, computed using the synchronic training of the model that uses the written BNC. There are eighteen cases where the grammatical counterparts have higher grammaticalization scores than their lexical counterparts, as expected. The two notable exceptions are *go* and *spite*. The case of *spite* is not surprising since it regularly displayed singular behaviour in the previous sections. A major problem was posed by *go*'s collocate diversity over time (Figure 17), which was discussed in the previous section. The explanation for its low collocate diversity was that it resulted from its higher frequency in the later years. This high frequency results in a large number of repetitions, which is even more salient in the left collocate diversity (appendix C2) as it mostly involves pronouns. Since collocate diversity is an important coefficient in the synchronic model (both for the 4-4 window and for the 1 word to the left window), it explains why it gives such a low value to grammatical *go*.

Let us first discuss the cases that show an increasing score over time, as one would expect given the unidirectionality of grammaticalization. This situation is observed with *back, die, means, order, quit, regard, respect, soon* and *well* (although *soon* and *well* display a minor dip at the end). There are therefore nine cases that show an increase over time of the grammaticalization score of their grammatical counterpart. Several grammatical counterparts show stable values over time, namely *event, have, help,* and *terms* (again it is sometimes necessary to ignore the first decades to reach this observation). With the exception of *help*, these cases exhibit a maximal score for the whole period. This is certainly what *have* is expected to display since it has been an established auxiliary in English for a long time. Combined, increasing and stable grammaticalization scores account for thirteen cases out of twenty.

Another type of behaviour is a decreasing grammaticalization score. The following grammatical variants show a decreasing score: *go*, *keep, long,* and *use*. These cases are somewhat problematic because a decrease is unexpected. A general explanation is that many of these cases are cases of grammaticalization of a main verb into an auxiliary function. The BNC model does not contain auxiliary verbs as a category of its own, but special verbs (*be*,

have, go, and *do*) are categorized as grammatical elements. However, these only represent a small portion of the 264 grammatical elements included in the dataset. Therefore, one could argue that the synchronic training of the model has an under-representation of auxiliaries. On the other hand, the diachronic dataset contains seven of such cases out of twenty: *die, go, have, help, keep, quit* and *use*. It happens that most of these show a decreasing grammaticalization score in Figure 20, or at best a stable value. Only *die* and *quit* seem to display a clear increase. Regarding the grammatical counterpart of *long*, the explanation for its decreasing grammaticalization score can probably be attributed to the small decrease in its collocate diversity shown in Figure 17 for its 4-4 window, and also its decrease in left collocate diversity shown in appendix C2. These decreases are not sudden, but rather small and steady over time. This corresponds to the decrease in grammaticalization score shown in Figure 20, which displays similar behaviour. It should be noted that a decreasing grammaticalization score does not mean that these are cases of degrammaticalization (section 2.5). In those cases, the decrease cannot be attributed to a new lexical use of a grammaticalized form. Instead, the decrease is mainly the result of methodological concerns as discussed above.

Finally, three elements display chaotic scores that consist both in increases and decreases, namely *matter, spite,* and *view*. Regarding *matter*, the first decades are the most problematic ones, as is the case with many other elements. What is more surprising is that there is a sudden burst in the grammaticalization score at the very end. Looking at appendix C2 points to left collocate diversity as the likely explanation. There is indeed a significant increase in the left collocate diversity score of grammatical *matter* in that period. Figure 15 shows that grammatical *matter* did not really increase in frequency, so it is in fact possible that *as a matter of fact* gets used in a broader variety of contexts in the most recent years.

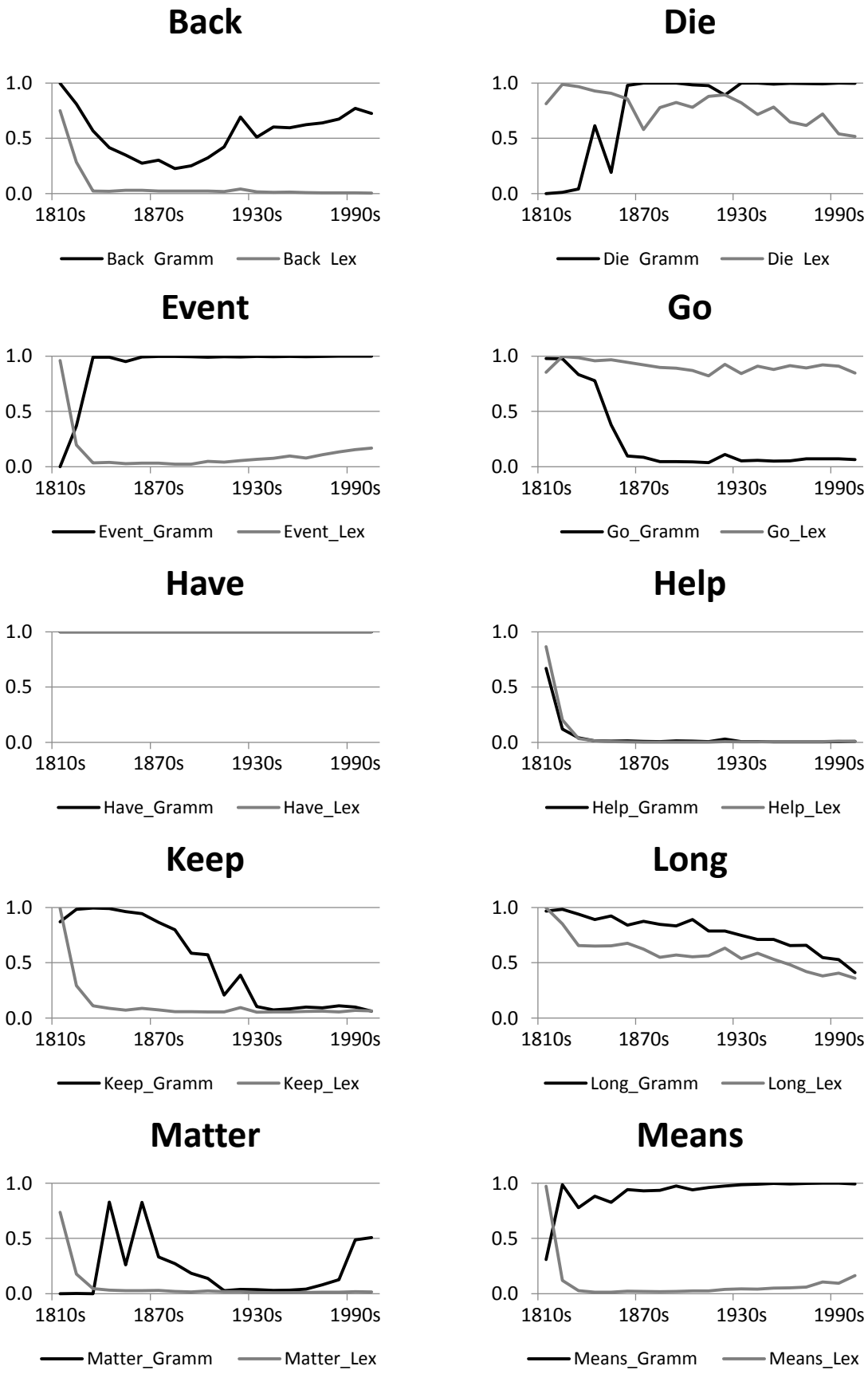


Figure 20. Grammaticalization score over time (training with data from section 5.4.2). The horizontal axis corresponds to each decade, while the vertical axis shows the score.

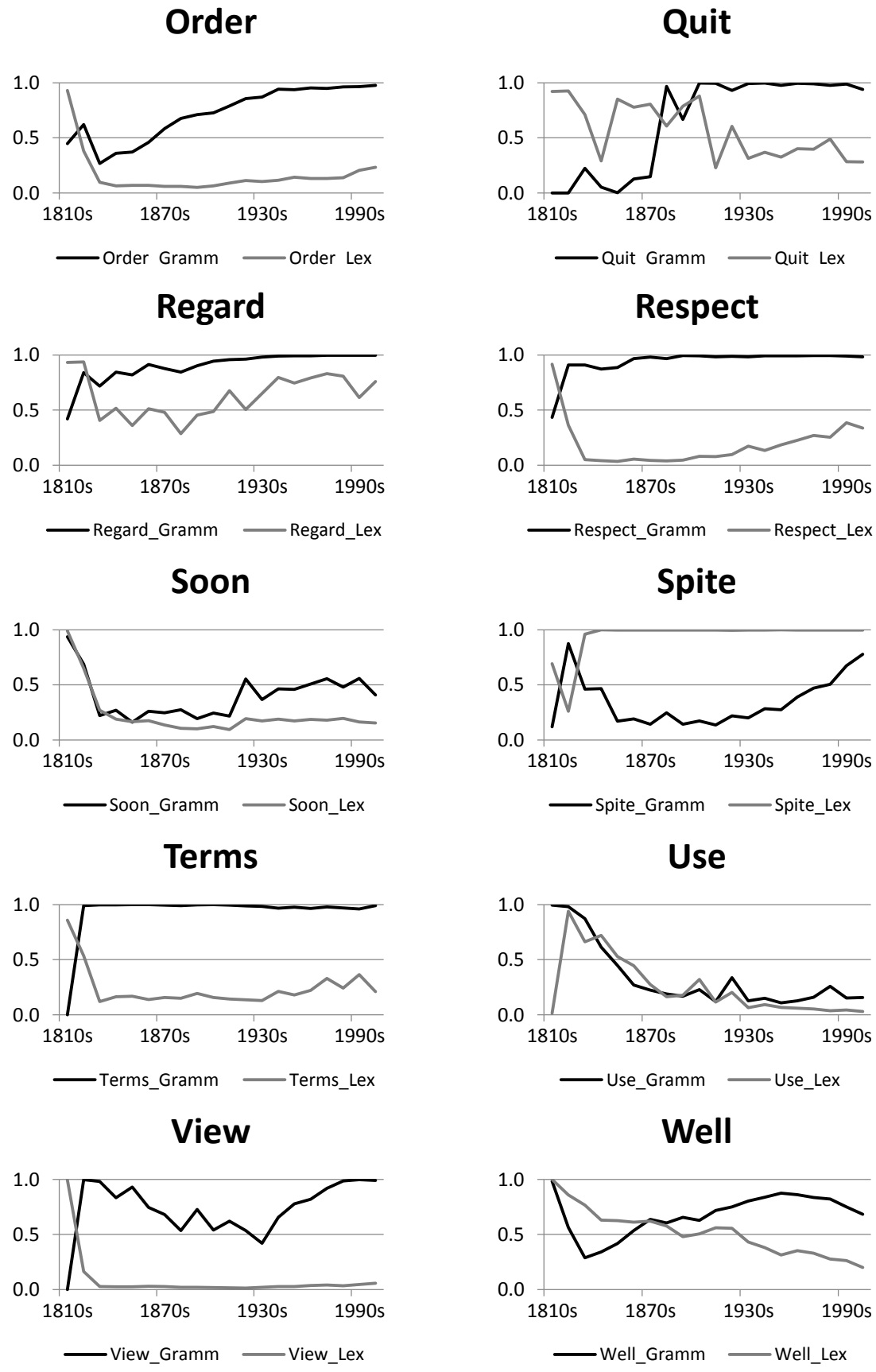


Figure 20 (continued). Grammaticalization score over time (training with data from section 5.4.2). The horizontal axis corresponds to each decade, while the vertical axis shows the score.

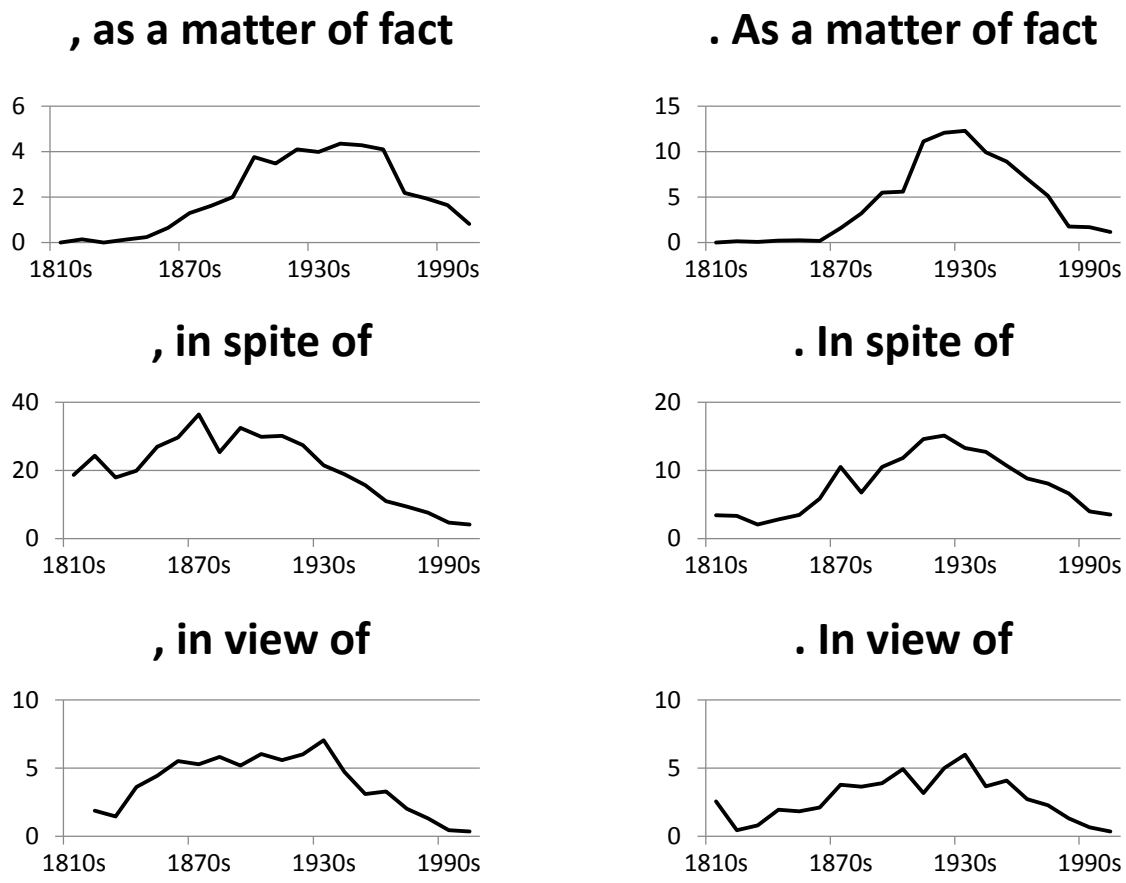


Figure 21. Number of occurrences (per million) of *as a matter of fact*, *in spite of,* and *in view of* preceded by a comma, and preceded by a full stop in the COHA.

There is also a methodological explanation that comes from the fact that punctuation is included as possible collocates in the diachronic corpus. Interestingly, *as a matter of fact* preceded by a full stop or a comma is becoming rarer in the most recent years, as shown in Figure 21. It means that *as a matter of fact* is becoming more used mid-sentence, without adding extra punctuation, which can point to an increase in its syntagmatic variability. This matter would require additional investigation to be more exhaustive, since the approach taken here is just to check for commas and full stops to the left of *as a matter of fact* and there are possibly other important collocates that could explain the shift in collocate diversity. This seems like a promising area to investigate in the future. In a similar fashion, one can also observe the same situation with *spite* and *view*, which are also shown in Figure 21.

Another approach to computing the grammaticalization score is to use the model trained on the basis of the diachronic dataset itself (Table 68). Remember that this model had generally low values of fitness. What is of interest here is to see whether the diachronic model offers a better distinction between lexical and grammatical counterparts. A limitation pointed

out earlier regarding the synchronic model is that it had difficulties with auxiliaries. Figure 22 shows the grammaticalization scores over time when they are computed on the basis of the model presented in Table 68.

The first observation is that only *back* and *spite* display a grammatical counterpart with lower scores. In the case of *spite*, this is consistent with all the previous sections where it showed unusual behaviours. Regarding *back*, the difference between the lexical and grammatical counterpart is not very large. The second observation is that most lines display stable behaviours and it is hard to notice increases are decreases. This is because the training of the model aims to distinguish lexical and grammatical counterparts, and the best way to do so is to use the right colligate diversity measure (Table 68). While the right colligate diversity measure is quite useful for this distinction, it does not display much change over time as seen in the average displayed in Figure 11.

Therefore, the advantage of the diachronic model is that it is quite good at telling counterparts apart, but the disadvantage is that it is not really able to highlight changes over time. This model might therefore be efficient with the measurement of the gradience of grammaticalization, but more limited when it comes to its gradualness. As discussed earlier, grammaticalization as a process involves small incremental changes, but these changes are not shown in a salient way by the present implementation. Again, methodological considerations are certainly a possibility; the twenty elements included in the present study likely have grammaticalized before the period under observation, and are therefore outcomes of grammaticalization, as opposed to an ongoing process.

A brief additional note regarding methodological concerns is that *go* does not display the same problems as earlier, because as shown in Table 68, the collocate diversity measures are either unimportant (4-4 window) or outright absent from the model (1L window). These variables were the ones causing a low grammaticalization score in the synchronic model. Instead, in the diachronic model it is the right colligate diversity that is a much more important coefficient and *go* does not have major problems in this regard.

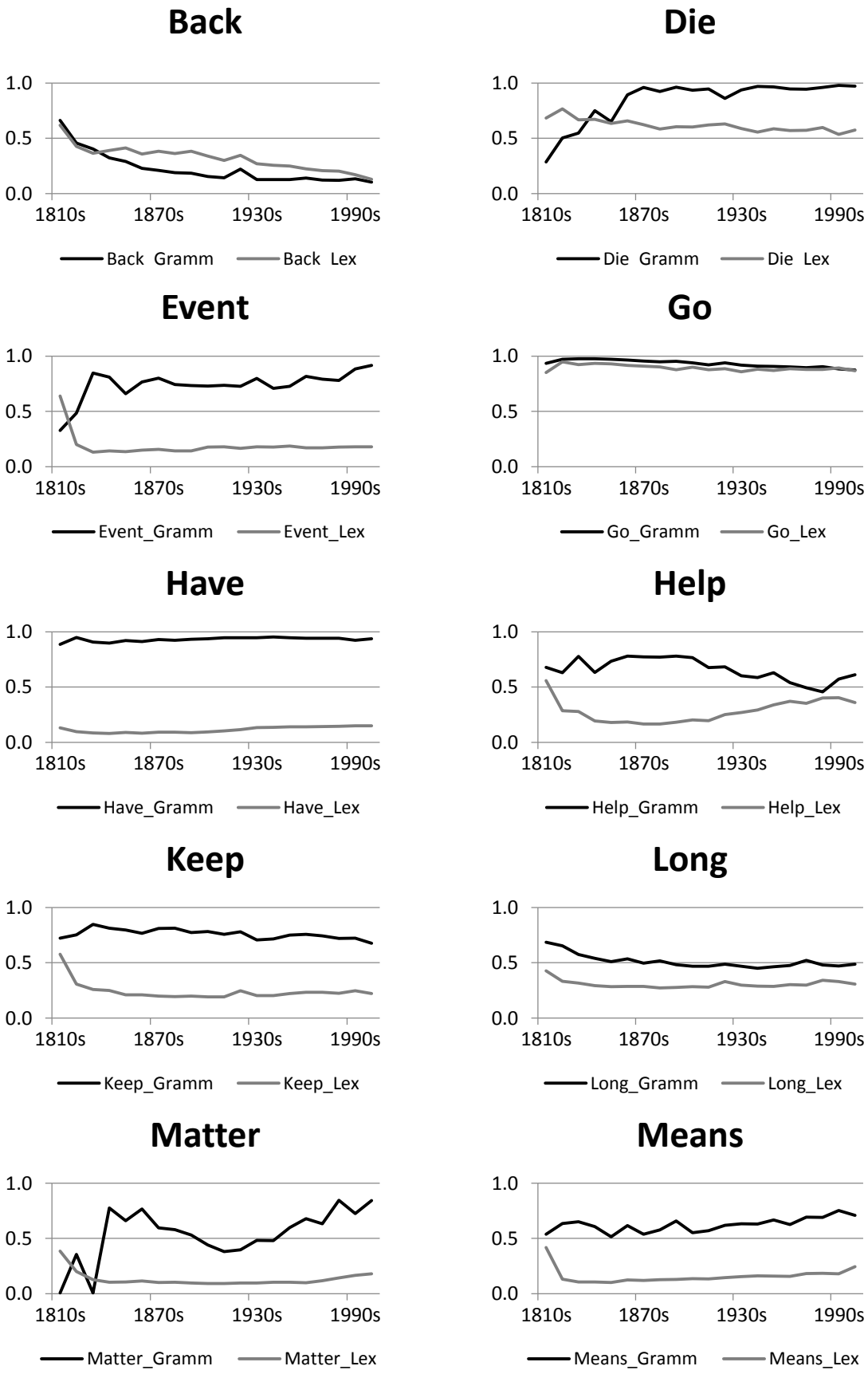


Figure 22. Grammaticalization score over time (training with data from section 6.2). The horizontal axis corresponds to each decade, while the vertical axis shows the score.

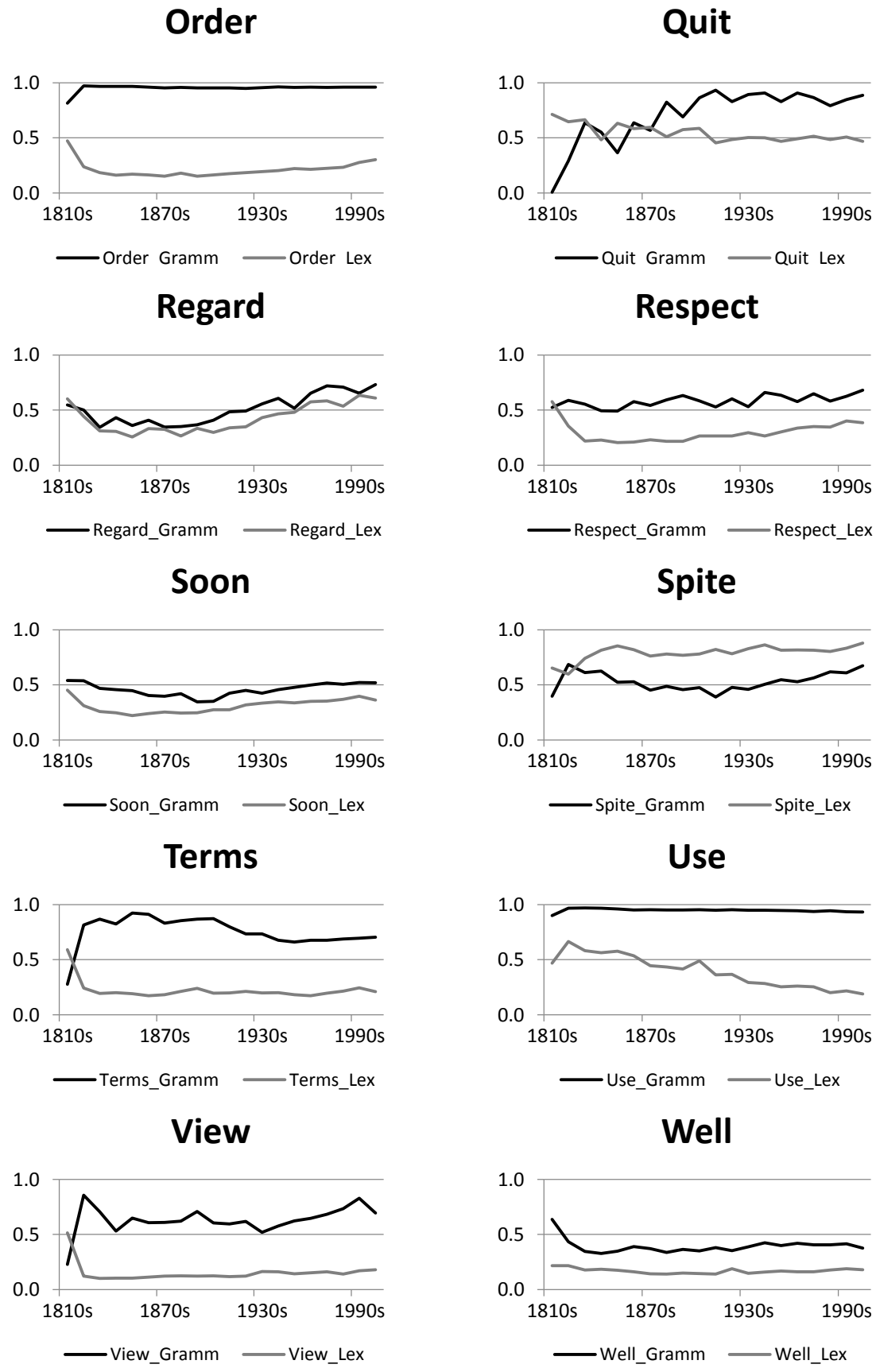


Figure 22 (continued). Grammaticalization score over time (training with data from section 6.2). The horizontal axis corresponds to each decade, while the vertical axis shows the score.

6.4.3 Overview

The aim of section 6.4 was to elaborate on the general findings from section 6.3, by keeping each element distinct from the rest instead of using averages. The average findings are in line with a closer observation at the individual level, but many specific cases have been highlighted. While grammatical counterparts are most frequent on average, it was shown that most of the grammatical counterpart elements in the diachronic database are actually low frequency elements. The average being so high is mostly the result of high frequency items such as *have* pushing the average upwards. Despite this, many of the grammatical counterparts do display increases in frequency over time, or simply a stable situation. Instances of the opposite trend are quite rare. Similarly, it was shown that collocate diversity tends to be higher for the grammatical counterpart, while lower colligate diversity was observed to the right. This is consistent with observations from chapter 5. The major difference is deviation of proportions, which has been shown to go against expectations. Indeed, grammatical counterparts tend to be less evenly distributed. While the average study has already shown this aspect, what the individual study has shown is that in most cases, deviation of proportions does not help distinguish the lexical and grammatical counterparts. An explanation comes from the fact that many elements have low frequencies and that this makes measuring their deviation of proportions quite problematic in a large corpus. In short, deviation of proportions is not a fairly reliable parameter in the diachronic dataset.

Section 6.4.2 then proposed to investigate grammaticalization scores over time at the individual level. The synchronic model was shown to provide results that confirm most expectations, despite its limitations with auxiliaries. This is quite remarkable because frequency was a major coefficient in the synchronic model and yet the diachronic dataset contained many low frequency elements. Despite this, the synchronic model displayed consistent higher grammaticalization scores for grammatical counterparts and also generally showed increases as one would expect given the unidirectionality of grammaticalization. There were a few notable exceptions for which the model performed poorly, such as *go* and *spite*. The explanation often stems from their collocate diversity measures. In the case of *spite*, its diversity may be strongly dependent on its location within the sentence and whether it appears next to commas or full stops. In the case of *go*, there is a more general problem related to auxiliary verbs, as the fact that they often involves pronouns and participles around them tends to reduce their overall collocate diversity, which is a situation that tends to be different with other types of grammatical elements.

6.5 Theoretical relevance of the findings and methodological concerns

The aim of this chapter was to propose a diachronic approach to the model presented in chapter 5 and to determine whether grammaticalization can be successfully measured on a diachronic dataset. The idea was that this should also constitute an empirical way to observe the unidirectionality of grammaticalization (section 2.4). The fact that the synchronic model from chapter 5 is able to successfully tell lexical and grammatical counterparts apart shows that this method of quantifying grammaticalization can also work with diachronic datasets. Furthermore, the model generally displayed increasing or stable grammaticalization over time, which is in line with the notion of the unidirectionality of grammaticalization.

While the diachronic study used the same methods as the synchronic one, it is relevant to discuss to what extent this approach is successful, especially since gradience and gradualness are usually considered as two distinct concepts (section 2.3). The synchronic studies model gradience in a consistent way, by having elements distributed along a continuum, as was shown in Figure 5. This continuum involves small differences from one ranked element to the next, in terms of their grammaticalization scores. In contrast, the diachronic study does not necessarily show consistent small incremental changes from one decade to the next. Therefore, the current approach is more limited when it comes to the modelling of gradualness. There are multiple methodological considerations that can explain this shortcoming.

The present study used a set of twenty elements that is not fully representative of all the grammaticalization processes that exist in English. First, the set contains elements which both have a lexical and a grammatical counterpart currently in use. This is a particular feature that was necessary to compare these uses and check if the model was able to distinguish between the two. However, many grammaticalized elements in English do not have a co-existence of both counterparts. The second reason why these elements are not fully representative of English grammar is that while they pertain to several grammatical domains, they mostly consist of complex prepositions and auxiliaries, as mentioned in section 6.2.1. Despite these limitations of the selected elements, the method is replicable with a larger number of elements.

One key problem that is apparent is that all the twenty elements required some degree of manual inspection and manual processing in order to retrieve the grammatical and lexical uses. In order to scale this study up, it would be required to come up with a way to select the grammatical uses of a large number of elements without needing a tailor-made approach as

the ones reported in sub-sections 6.2.3 to 6.2.22. While there are methods for sense disambiguation (e.g. Gallant 1991, Turney and Pantel 2010), these tend to be based on collocate profiles, which happen to be featured by the models that were used in the current chapter, and in chapter 5 as well. This is therefore a circularity problem, as the selection of the grammatical uses would be performed by similar methods as the one used to compute the grammaticalization score. In addition to providing an empirical way to measure grammaticalization and to support the unidirectionality hypothesis, many other aspects were also highlighted.

The synchronic dataset in chapter 5 and the diachronic dataset contain different types of grammaticalization. The synchronic dataset mostly contains unrelated lexical and grammatical elements, whereas the diachronic one contains grammatical elements for which there is a currently existing lexical counterpart in English. Despite this difference, the same observations were made regarding the relationship between the main variables and grammaticalization. A basic observation is the fact that since most variables are stable over time, the similarity of their behaviours are consistent with the results of the principal component analyses presented in chapter 5, where multiple correlations were observed between each variable. Another observation is that grammaticalization is associated with increasing frequency, despite the fact that many grammatical counterparts in the diachronic dataset are low frequency elements in comparison to their lexical counterpart. Grammaticalization is also consistently associated with increases in collocational diversity, in spite of several elements such as *go* or *in spite of* which involve additional methodological concerns because the way they are used tends to involve less diversity overall. Lesser right colligational diversity is also shown as a characteristic of grammaticalization in the diachronic dataset. There are therefore multiple observations that hold true both in the synchronic and diachronic datasets. The one exception is deviation of proportions which consists in a less even distribution of grammatical counterparts, but which has been explained as a methodological shortcoming of deviation of proportions when used on low frequency elements in large corpora.

Furthermore, it should also be noted that the types of items involved in the database also change the parameters of the binary logistic regression models. Indeed, if a binary logistic regression is trained on the diachronic dataset, the coefficients of the model differ from that of the synchronic model. This leads to a more general question regarding grammaticalization. Should one come up with a general model of grammaticalization that is based on a broad

range of grammatical elements, or should several case-specific models be used? This is a difficult aspect because it has been shown that the synchronic model does have some limitations when it comes to auxiliaries and also that collocate diversity with auxiliary verbs such as *go* tends to work differently from prepositions. However, what has been shown is that a general model such as the synchronic one still gives coherent results from a general point of view (as with the averages values presented in section 6.3) but is limited when observing certain specific cases (as with some items from section 6.4). Different types of grammatical elements might entail different characteristics of grammaticalization, which therefore means that their measurements could also differ. In short, it is difficult to find universal characteristics of grammaticalization that can be transcribed into specific quantitative variables and that work with all possible cases of grammaticalization. This is in line with an observation regarding the Lehmann parameters, namely that they do not always correlate with grammaticalization in a consistent way (section 2.2.1).

While some critics of grammaticalization would argue that these differences between different types of items is problematic because it illustrates that the term grammaticalization is a catch-all term that involves too many different things (e.g. Campbell and Janda 2001, Joseph 2011), the fact that the synchronic model is still able to show general tendencies that are in line with that of the diachronic dataset supports the idea that there are actually shared characteristics between very different processes. This why future research on grammaticalization should strive to find out whether features observed in specific cases of grammaticalization are common, or case-specific. Another question is also whether these features are relevant both in synchrony and in diachrony, given the differences observed between chapter 5 and chapter 6. Ideas for future research on grammaticalization as well as ways to improve the models presented in this dissertation are discussed in the final chapter.

7. Conclusions

Research on grammaticalization has produced a wealth of empirical observations that capture regularities with regard to the development of grammatical structures. It was the aim of this dissertation to build on these insights and to address a gap in the research literature, namely the question of how grammaticalization as a process can be quantified, so that different linguistic elements that are more or less grammaticalized can be effectively compared. The main contribution of the present work is a concrete proposal of how degrees of grammaticalization in linguistic elements and constructions from English can be measured in a quantitative way. In order to construct such a measure, the analysis has focused on variables that are known to be linked to grammaticalization. The measure was developed in order to account for degrees of grammaticalization in synchronic language use, which reflects grammaticalization as a product, and also degrees of grammaticalization that are changing over time, which reflects grammaticalization as a process. The synchronic approach was presented in chapter 5, while a diachronic perspective was offered in chapter 6. This concluding section first reviews the main findings and answers the research questions that were presented in the introductory chapter. Their relations to broader generalizations and theoretical concerns regarding grammaticalization are also explored. Challenges and limitations that were encountered in the previous chapters are then discussed, along with comments regarding possible improvements and ideas for future research.

7.1 Main findings

The first research question was to determine whether it was feasible to quantify the gradualness of grammaticalization. To answer this question, a database with lexical and grammatical elements was constructed with the aim of training a statistical model to distinguish between the two categories. The elements in the database were richly annotated, using several variables that theoretical accounts have presented as playing a role in grammaticalization (chapters 2 and 3). A binary logistic regression model (chapter 4) was then fitted to the data. In the analyses discussed in chapter 5, different implementations of the binary logistic model successfully assigned grammaticalization scores to the elements in the database in a consistent way, giving most lexical elements lower scores and most grammatical elements higher scores. The results were generally in line with previous research, as illustrated by the average grammaticalization scores for the major morphosyntactic categories (section 5.4.2).

Nominal and verbal categories received, on average, lower scores than adjectives and adverbs. Table 29 displayed the average scores of the main morphosyntactic categories, where nouns and lexical verbs had the lowest averages, putting them on the lexical-end of the cline, whereas adverbs had an average close to the middle of this cline. This is in line with Hopper (1991: 20) who considers nouns and verbs as "full categories", as opposed to adjectives and adverbs which are secondary and more grammaticalized according to his principles (section 2.2.2). This result is also consistent with claims regarding the particular status of adverbs that can both be lexical or grammatical (e.g. Boye and Harder 2012: 19, Traugott and Trousdale 2013: 73). While this is not new knowledge per se, it is an empirical way of supporting a theoretical conceptualization of the phenomenon that has mostly been substantiated by individual case studies so far. These results indicate that from a synchronic perspective, degrees of grammaticalization in English can be successfully measured using the quantitative approach proposed in this dissertation.

In addition, when all linguistic items under investigation were subjected to a discrete and binary classification into lexical and grammatical elements, all implementations of the model had high correct classification rates (above 80%). A manual evaluation of the items with the 100 highest and the 100 lowest scores also highlighted the accuracy of the model and the rare cases of wrong classifications were explained as methodological artefacts. These illustrate limitations of the approach that could, in principle, be remedied. An example of such a limitation is that in the synchronic implementation proposed in chapter 5, no distinction was made between lexical and grammatical uses of the same form, which can be observed with English verb forms such as *have*. As an unintended consequence of this, elements that both have lexical and grammatical functions are sometimes given scores that do not reflect their multiple uses. A more fine-grained model could take these distinctions into account and would achieve higher classification accuracy.

This was however not a concern in chapter 6, where lexical and grammatical counterparts of the same form were investigated. This chapter has shown that the model can distinguish lexical and grammatical counterparts in a reliable fashion, with several exceptions, such as the *be going to* + infinitive future construction, that will be further discussed in section 7.3 below. Chapter 6 also investigated whether the gradualness of grammaticalization could be measured from a diachronic perspective. Again, the model was able to do so, although increases and decreases in grammaticalization were not salient in all cases that were investigated. Most cases showed stability over time, which can be attributed to the selection

of grammaticalizing elements. Most of the elements under investigation started to grammaticalize in earlier historical periods than the ones that are represented in the Corpus of Historical American English, which can explain why many cases showed stability in their grammaticalization score. Furthermore, this stable situation is in line with the notion of unidirectionality in the sense that a grammaticalized element does not "lose" its grammaticalization features.

The second research question asked which factors lie behind different degrees of grammaticalization. From an empirical perspective, this was investigated by studying the coefficients of the variables involved in the different models, and also by measuring the correlations between those variables. A main result across all implementations is that token frequency plays a major role, which is coherent with approaches to grammaticalization that rely on this idea (e.g. Bybee 2011, Diessel and Hilpert 2016). As has been argued in various discussions, higher token frequency correlates with higher grammaticalization (Hopper and Traugott 2003: 126-130).

Other variables in the analysis focused on issues of distribution, specifically the occurrence of other linguistic elements to the left and right of the element under analysis. The variables of collocate and colligate diversity were found to be relevant, although less so than token frequency. Generally, more diversity corresponded to higher degrees of grammaticalization, which can be explained by the fact that grammatical elements are semantically bleached (Sweetser 1988) and may therefore occur with a broader variety of elements and in a broader variety of contexts. This goes against Lehmann's parameter of syntagmatic variability (section 2.2.1), where grammatical elements are thought to be less easily shifted around. A sub-question regarding these collocate-based variables was whether left and right collocates play different roles. A main finding was that more diversity to the left correlates with higher degrees of grammaticalization. However, the right-based variables were not systematically relevant. When they were, grammaticalized elements displayed less diversity to the right. One explanation that was put forth is that grammaticalized elements tend to be syntactic heads, and that they are therefore located at syntactic boundaries. Since English is mostly right-branching (i.e. syntactic heads to the left), then more diversity is likely to the left.

An interesting avenue for future research is to determine whether similar observations can be made for other languages, in particular those that are mostly left-branching, such as

Chinese (which is further discussed below in section 7.3).[24] The reason why such languages are of particular interest is that their investigation might further support branching as an explanation for the differences that were observed between the left and right collocate diversity measures in the case of a right-branching language such as English. However, it should also be kept in mind that some degree of correlation between the left and right collocate and colligate variables was reported, which shows that there is some overlap between them. The principal component analyses in chapter 5 have shown that the collocate diversity variables, along with frequency, could be summarized by one component, but that this component only accounts for about 40% of the total variance in the data. Furthermore, the principal component analyses showed that the colligate-based measures bring their own information to the table and that they are better summarized by a second component, which shows that they share less overlap with the collocate-based ones.

Some variables that were included in the model turned out not to allow a reliable distinction between lexical and grammatical elements. The dispersion and phonological length variables did not seem to play major roles in most implementations of the model, and the likely explanation that was put forth is their strong correlation with token frequency, which has been highlighted in other studies such as Bell et al. (2009) in the case of phonological reduction and Hilpert and Correia Saavedra (2017a) regarding dispersion. However, in the case of the dispersion variable, some of the shortcomings might be due to the chosen implementation (deviation of proportions), which has some limitations that were introduced in section 3.5 and which are further discussed in section 7.3 below.

The third research question was to determine whether the proposed quantitative approach could be used to test the hypothesized unidirectionality of grammaticalization. This question was investigated in chapter 6 by applying the grammaticalization measure to several known cases of grammaticalization that both have lexical and grammatical counterparts. As stated above, most cases showed stability instead of outright increases in grammaticalization scores. This result may seem unexpected, but it can be explained in terms of two different aspects of the unidirectionality hypothesis. On the one hand, the hypothesis predicts that grammaticalizing elements only undergo changes towards the more grammaticalized end of the cline. But on the other hand, the hypothesis also predicts that grammaticalized elements do not lose their grammaticalization (i.e. they do not change towards the less grammaticalized

[24] Branching direction is a broad categorisation and also involves gradience. A right-branching language can have left-branching structures and vice-versa. Therefore, English is not purely right-branching, and Chinese is not purely left-branching. However, these distinctions are fairly common in language typology (Berg 2009).

end of the cline). While this claim might seem rather trivial, it is in fact relevant when approaching this question from an empirical perspective. This has also been supported by Bybee (2011) who notes that grammaticalizing forms which no longer display increases in frequency do not undergo a reversal of their grammaticalization. In the situation presented in chapter 6, it is mostly this second aspect of unidirectionality that is supported (i.e. not "reverting" back to less grammaticalized), given that most elements under study have started to grammaticalize before the period under observation. Decreases in grammaticalization were very rare among the elements under study and were mostly due to methodological concerns. Therefore, unidirectional tendencies were observed in the selected cases of chapter 6, although future research might extend the number of elements under analysis and also use corpora that involve larger time periods. This illustrates that a quantitative approach to measuring degrees of grammaticalization is a fruitful way to investigate theoretical matters from an empirical perspective.

7.2 Generalizations regarding grammaticalization

The findings presented in the previous section pertain to the broader endeavour of investigating generalizations regarding grammaticalization. Generalizations are hard to substantiate on the basis of case studies, which is why larger-scale corpus-based studies can help complement these approaches. There are three types of generalizations that were discussed in the previous section: how certain variables correlate with grammaticalization, how whole morphosyntactic categories can be ranked on the grammaticalization cline, and whether grammaticalization is mainly a unidirectional phenomenon. The present results seem to indicate that while some generalizations can be substantiated, there are elements for which the situation is more nuanced.

The first type of generalization concerns the variables presented in chapter 3 and how they correlate with grammaticalization. For instance, it has been argued that frequency is "a concomitant of grammaticalization" (Hopper and Traugott 2003: 126), which is a prudent generalization. The results presented in chapter 5 support a more specific generalization, namely that grammatical elements have higher frequencies than lexical elements, given the significant weights of token frequency in the different models. However, chapter 6 proposes more nuanced results, where several grammatical elements were less frequent than their lexical counterparts (Figure 15). Generalizations regarding grammaticalization might therefore also depend on the view under consideration. In chapter 5, where a synchronic

approach is proposed, and where grammaticalization is studied as an outcome, there seems to be substantial evidence and support to the claim that grammaticalization and token frequency are positively correlated; grammatical elements involve higher token frequencies. However, chapter 6 highlights that in diachrony, when looking at grammaticalization as an ongoing process, this correlation becomes less obvious.

Another similar issue pertains to the notion of collocate diversity. There was strong evidence in chapter 5 that grammatical elements involve more diversity overall, but also that they specifically involve more collocate diversity to their left. What is of interest in this present case is that counterexamples to this claim can be very easily pointed out, such as certain clitics (e.g. *'m, 're*) where diversity to the left is almost non-existent since they are bound morphemes that occur after a pronoun. Such clitics received low grammaticalization scores in the model (section 5.4.2), when they would have been expected at the top according to the grammaticalization cline presented in section 2.3. While this situation does show a limitation of the model, it also shows that the current approach is successful in making generalizations while highlighting elements that do not conform to it. In fact, clitics constitute their own category on the grammaticalization cline, and the present results support this idea since they show different properties from most other grammatical elements involved in the database. Similarly, an issue that was highlighted in section 5.7 is that collocate diversity to the right might play a more substantial role for certain categories, where the inclusion of modal auxiliaries into the grammatical categories resulted in a more significant weight for the right-based collocate diversity variable. Furthermore, there have also been claims that different processes and stages of grammaticalization might entail different features (Norde 2012), which is again a generalization that can be investigated and substantiated by the approaches proposed in this dissertation.

The second type of generalization concerns the ranking of whole morphosyntactic categories on a grammaticalization cline. As discussed in the previous section, the present results are in line with previous research, where nominal and verbal categories are regarded as the major lexical categories (Hopper 1991: 20), in contrast to adverbs that have a more variable position on the grammaticalization cline (e.g. Boye and Harder 2012: 19, Traugott and Trousdale 2013: 73). Again, what the present approach offers is a way of computing averages and of looking into the question in a general way. However, it can also be used to discuss individual elements or paradigms. For instance, section 5.7 focused on modal verbs and their ranking by the model. Table 40 showed that certain modals (e.g. *would, can, could*)

get high scores and display the main features measured by the model, such as higher token frequencies and higher collocate diversity, in contrast to other modals where these features are less prominent (e.g. *should, must, might*). It can also show that if a modal is not explicitly distinguished from its lexical source (e.g. *need*), its score can be much lower than expected. The present results can therefore both support generalizations about morphosyntactic categories, and also be used to compare elements within those categories. This can highlight exceptions or particular situations, but it also relates to broader concerns regarding the category under observation, such as the decline in frequency of core modal verbs in Present-Day English (e.g. Leech 2013).

The last type of generalization that was investigated is whether grammaticalization is mainly a unidirectional phenomenon. The hypothesis according to which grammaticalization is mainly unidirectional, from less to more grammaticalized, would be better substantiated if a unique approach can be applied to a broad range of elements. Again, this can highlight exceptions and elements that display particular features, while supporting more general tendencies. While chapter 6 focused on a restricted number of elements due to technical concerns that are further discussed in the following section, it did support the generalization put forth by Bybee (2011) that grammaticalized elements do not revert back when their frequency ceases to increase, as discussed more extensively in the previous sections.

The approach proposed in this dissertation is therefore particularly well-suited for the investigation of generalizations regarding grammaticalization. It is also a way of highlighting exceptions and particular items. These generalizations can pertain to theoretical concerns regarding grammaticalization but also to issues that are specific to certain morphosyntactic categories. However, the present approach also has limitations that are further discussed in the following section.

7.3 Current and future challenges

This section summarizes the main challenges highlighted in the previous chapters, as well as improvements and ideas for future research. Chapter 5 has shown that elements can be convincingly ranked in a continuum which models the gradience of grammaticalization. The elements under investigation in that chapter were all highly frequent in English. In contrast, chapter 6 proposed to model twenty elements over time. The results mostly indicate stable situations, where small incremental changes that would be expected of ongoing grammaticalization processes were not readily observed. The gradualness of grammaticalization was therefore not captured by the model as efficiently as gradience. One

explanation that was put forth is that the twenty elements in the database are outcomes of grammaticalization processes that took place before the period covered by the chosen corpus. Interestingly, the diachronic model was still quite successful in distinguishing the lexical and grammatical counterparts of the twenty elements under investigation, which again shows its efficiency when dealing with the gradience of grammaticalization.

Another main problem encountered in chapter 6 was that certain complex constructions yielded fairly unexpected results. Constructions like *be going to* + infinitive or *keep* + V_{ing} make the collocate-based measures more complicated to implement because elements can be inserted within the construction itself (e.g. *I am never going to say this again*). This is related to Lehmann's parameter of bondedness (section 2.2.1), as the elements within a given construction can be loosely or strongly bonded together. The collocate diversity measure is thus limited, since it cannot both account for diversity outside and inside the construction at the same time. It is in fact difficult to find a general way to operationalize the diversity measure in such a manner that does not require some kind of manual specification when it comes to those constructions. For instance, the analyst has to decide whether or not the *be* and *to* in the *be going to* + infinitive future construction should be included in the collocate window that forms the basis for the analysis. This choice cannot be implemented on the basis of a single, general algorithm for all constructions, and as such some manual intervention is required, thus limiting the generalizing potential of the proposed approach.

One possible solution to this problem is that one should investigate degrees of bondedness to determine core parts of the construction and more peripheral ones. In the *be going to* + infinitive future construction, the core elements could be argued to be *going* and *to*. An argument is that insertions in between these elements are extremely rare (**I'm going never to talk to you again*) and that the form *gonna* exists, which illustrates coalescence between these two elements. Elements such as *be* and infinitive verbs can be argued to be more peripheral. Again, potential insertions are a good indicator of lower bondedness (*I am never going to do that*), as well as possible omissions (*You gonna pay for this*).

In the case of the infinitive verbs following the construction, the situation is more complex because one should also factor in whether collocate or colligate diversity is considered. In fact, the distinction between colligate and collocate diversity may highlight that bondedness might be a larger concept than previously thought. On the one hand, the *be going to* future construction is almost always followed by an infinitive, which points toward low colligate diversity and strong bondedness, and thus that the infinitive is part of the construction itself. On the other hand, there is a large number of possible infinitives that can

occur with this construction, which points to high collocate diversity (i.e. high paradigmatic variability in the infinitive slot) and therefore to lower bondedness. The fact that the verb *go* itself can occur with this construction (e.g. *I'm going to go to the airport*) further supports the idea that increasing grammaticalization involves higher collocate diversity. It has been argued that the possibility for a lexical source and its auxiliary form to occur together is a diagnostic of advanced stages of auxiliation (Heine 1993: 58), which is also in line with the idea of increasing collocate diversity (even though this specific case only adds one new possible collocate).

This discussion is also related to a larger concern, which pertains to how constructions should be identified. As mentioned in section 2.7, a linguistic pattern is considered as a construction if it has non-compositional meaning, or if it is sufficiently frequent. However, it is sometimes not obvious to decide which elements are part of a given pattern or not, and how similar patterns relate to each other. An example would be patterns such as *good for NP*, *good for VP* and *good to VP*, which can be regarded as instances of a more general *good XP* pattern (Zeschel 2009: 188-198). When it comes to the methods proposed in this dissertation, such cases are not straightforward to deal with. From an operational point of view, should the general pattern be *good for/to NP/VP*? Should the three specific patterns be considered as distinct elements? While these specific examples are not cases of grammaticalization, there are examples of grammatical elements that are involved in similar, yet formally different patterns, such as *with regard to* and *with respect to*. A question that was not discussed in the previous chapters is whether these patterns are instances of a more general *with X to* pattern, which would also comprise *with reference to*. The results from chapter 6 seem to indicate that *with regard to* and *with respect to* have fairly similar grammaticalization scores, as well as similar situations when it comes to their collocate-based diversity measures. While this specific question requires more investigation, it shows that the model proposed in this dissertation might also be used to provide empirical support regarding the similarities between such patterns and how they might relate to a more general one. This is of particular interest for construction grammar as well, since it can help investigate the organization of the construct-i-con (section 2.7).

A more problematic aspect in delimiting constructions is deciding whether an element is included into a syntactic construction, which is subjective to some extent. From an operational perspective, if measures such as collocate diversity are to be used to determine degrees of grammaticalization, there should be arguments to decide what to include in the

construction or not. Bondedness criteria were discussed above regarding the *be going to* + infinitive future construction. A possible implementation of this solution is to posit that elements that have very low colligate diversity to their left and right are strongly bonded together and therefore constitute a construction. A consequence is that the actual diversity required for the models in chapters 5 and 6 would then be measured outside of that construction.

Another solution is to use different diversity measures, using both an intra-construction diversity measure, as well as an extra-construction diversity measure. The intra-construction diversity measure should display low diversity for grammatical elements, given their increased bondedness according to Lehmann (2002, 2015). The extra-construction diversity measure should display higher diversity in the case of grammatical elements, as observed in chapter 5 where these elements were linked to higher diversity, with occasional exceptions when it comes to collocates to the right. What this discussion reveals is that the current implementation of the model has limitations when dealing with complex constructions and that it might be informative to investigate diversity within and without specific constructions.

A further limitation of the approach proposed in chapter 6 is related to the fact that distinguishing the lexical and grammatical counterparts requires using regular expressions along with some manual evaluation. While this works on a case by case basis, it would be beneficial if this distinction could be made in a fully automated way. This would make it possible to have more individual elements in the diachronic study and to enable broader generalizations to be made. In particular, it would help confirm the unidirectional tendencies observed on the elements presented in chapter 6 by being able to deal with significantly more elements.

An additional concern with the choice of these elements was that a decision had to be made whether to only take counterparts that shared the same form (i.e. *going* versus *going* and not *go/went* etc…), or to take all possible forms into account. There is no straightforward answer to this problem. It can be argued that all forms of the verb *go* might contribute to its grammaticalization and that forms such as *went* are also part of the lexical sources out of which the future *be going to* + infinitive construction developed. It could be argued that more abstract uses such as *he went crazy* contribute to make overall representations of the verb *go* more abstract, which then might turn it into better source material for grammaticalization. While this position has not been adopted in the present endeavour, it is possible that results regarding the lexical counterparts in chapter 6 would have been different if all forms had been

included. As shown in section 6.2.2, the inclusion of singular and plural occurrences of the lexical sources of some of the complex prepositions was also a relevant issue. Further empirical investigation of this question might show whether the inclusion of such forms fundamentally changes the results, or if it does not have much weight overall.

Given that lexical and grammatical counterparts were selected using relatively strict regular expressions, polysemy tends to be overlooked in the current approach. For instance, a verb like *have* has multiple meanings, such as concrete and abstract possession (*I have a book*, *I have a headache*), obligation (*I have to do this*), causation (*I had my nails done*), and auxiliary uses (*I have done my homework*). These meanings range from lexical to more grammatical, but the present approach does not really reflect this dimension. Future work could focus on sets of polysemous elements and investigate to what extent they differ when it comes to the variables discussed in the previous chapters. This would highlight how each use relates to the more general aggregation of lexical and grammatical meanings in the present studies. It could also demonstrate whether considering each use as a distinct item leads to major changes to the present findings.

Other methodological concerns pertain to the chosen measure of dispersion, namely deviation of proportions (Gries 2008). One limitation that has been discussed is that elements with lower frequencies are more problematic when it comes to determining their deviation of proportions. If they are very rare in a given corpus, then it is almost impossible for them to be distributed evenly. It is especially true if the words occur significantly less often than the number of chunks that is selected by the analyst. A word that only occurs ten times in a total of ten thousand chunks will be considered to have a very un-even distribution, regardless of its actual distribution in the chunks. Because of this limitation, the role of dispersion might be overlooked in some cases. To further investigate this question, alternative measures such as burstiness (Altmann et al. 2009, Pierrehumbert 2012) might be used to further investigate the relevance of dispersion in the present model. Similarly, other measures of diversity could also be investigated, such as entropy-based measures, as suggested in section 5.9.

Investigating such alternatives could also be done on the level of the database that was used in chapter 5. The constitution of the database might change the parameters of the model, in particular when it comes to the left versus right collocate and colligate diversity measures. For instance, section 5.7 showed that modal verbs give more weight to the right collocate diversity measure and contribute to link higher grammaticalization to lower diversity on the right. This is supported by the fact that modals have strong constraints to what appears to their

right (i.e. mostly an infinitive). The choice made in chapter 5 was to include as many grammatical elements as possible, because the aim was to uncover the general tendencies. As discussed above, this leads to some challenges when dealing with certain grammatical items. Thus, a strategy for further research could be to conduct similar quantitative studies, while restricting the scope to a specific type of item in order to study features that are specific to certain categories (e.g. auxiliaries having less diversity on the right). This would complement the general features that were highlighted in the present dissertation.

Furthermore, the possibility to establish general features as opposed to category-specific ones is also a crucial endeavour to address the fact that grammaticalization is sometimes criticized for being too broad (section 2.1). If such features can be established and measured by quantitative means, then there is an empirical way to determine whether a given category displays general features of grammaticalization and can be regarded as an instance of this phenomenon. Moreover, processes that display more general features might be regarded as more prototypical instances of grammaticalization. For instance, given the results from the previous chapters, a process that does not involve more diversity to the left would be considered as a less prototypical instance of grammaticalization than one that does. Whether this specific feature is truly representative of grammaticalization remains to be further investigated, but finding such features would help deal with the criticized broadness of grammaticalization by providing more organization within this category of processes.

An additional common criticism of grammaticalization is the fact that it involves different phenomena and might itself be an epiphenomenon of independent processes (Campbell 2001). The results from the previous chapters illustrate that the variables under investigation sometimes correlate together, which suggests that it is not always possible to isolate each phenomenon. For instance, phonological length (approximated by letter counts) and token frequency have consistent strong correlations across all experiments in chapter 5. Certain aspects of grammaticalization are thus strongly tied together, which supports the idea that it is a composite change that involves several phenomena at the same time. It also supports that they are linked, instead of being independent. Future research could further explore these links, especially by using more accurate phonetic representations, as well as other diversity and dispersion measures as suggested above. This would corroborate that those links are not simply due to methodological concerns.

Another relevant question is to what extent the quantitative approach proposed in the previous chapters really captures most aspects of grammaticalization. As highlighted in

section 2.2, not every single grammaticalization parameter that has been proposed in previous research has been implemented in the current approach. Parameters such as paradigmaticity (i.e. grammatical elements tend to be part of closed-class paradigms) are difficult to implement, since determining a paradigm and measuring its size requires qualitative interpretation. As the current aim was to use corpus-based variables that require little to no intervention, such parameters are difficult to include.

Beyond the technical difficulties of including such parameters, there is also the fact that not every single parameter is agreed upon and that their roles are subject to debate (section 2.2.1). In addition, there are also diverging definitions of grammaticalization, as well as critical opinions of the phenomenon as a whole (section 2.1). Thus, it seems unlikely that a single grammaticalization measure could deal with all concerns at once. This parallels some developments in the field of morphological productivity that were mentioned in chapter 1, where multiple measures are proposed and can be regarded as measuring different aspects of the phenomenon. Therefore, the grammaticalization measure proposed in the previous chapters might be better regarded as mainly focusing on one specific aspect of grammaticalization, namely "gradience" or "grammaticalness versus lexicalness". The broader endeavour of measuring grammaticalization might therefore require developing multiple complementary measures, which would be in line with the view that grammaticalization involves multiple interlinked processes (section 2.1).

Finally, another promising area of investigation is to compare different languages using the same quantitative approach. As discussed earlier, applying the same quantitative approach to a left-branching language could further support whether branching is a likely explanation for the differences that were observed between the left and right collocate and colligate measures in English. Sun and Correia Saavedra (submitted) is a study that applies the methods proposed in this dissertation to Chinese data from the Lancaster Corpus of Mandarin Chinese (LCMC). A synchronic grammaticalization model of Chinese is built in a similar fashion to the models proposed in chapter 5. While Chinese has certain grammatical categories that English does not have (e.g. numeral quantifiers), and that conversely, English has categories that Chinese does not (e.g. articles), the approach does not change in any major way.

The results of that study highlight that token frequency plays a major role in the Chinese model, which is similar to the current observations on English. This is a rather expected result given the prominence of frequency in grammaticalization (sections 3.1 and 3.6). Similarly,

the overall collocate diversity measure (which uses a symmetric window of four items) is also relevant in the Chinese model and has the same type of relationship with degrees of grammaticalization, where grammatical elements tend to involve more collocate diversity overall. However, when it comes to the left and right collocate and colligate diversity measures, the Chinese model comes with an interesting twist; Chinese grammatical elements tend to have less colligate diversity to their left, while they tend to have more collocate diversity to their right. This is different from English, where grammatical elements generally have more diversity to the left, and less to the right. A brief note is that collocate diversity to the left was not a significant parameter in the Chinese model, which interestingly mirrors the absence of collocate diversity to the right as a significant variable in section 5.4.2 for the English model. Thus, English and Chinese seem to display opposite behaviours when it comes to the relationship between left/right diversity and grammatical elements. This further supports branching as a possible explanation regarding the observed higher and lower diversities observed to the right and left of grammatical elements. A follow-up would be to consider other right- and left-branching languages and determine whether the present observations are supported. This illustrates how applying the quantitative approach proposed in this dissertation can also be used to investigate typological differences between languages and how they relate to grammaticalization.

7.4 Concluding remarks

This dissertation has tried to show that quantifying degrees of grammaticalization is a viable endeavour. It is a fruitful way to provide extra empirical support to theoretical claims and also to study larger datasets, going beyond case studies. The crucial role of token frequency in grammaticalization was further supported, along with some new insights regarding the empirical testing of the notion of unidirectionality. This approach also provides a way to look at grammaticalization from a more general perspective and to highlight tendencies such as increased collocate and colligate diversity (in particular to the left, in the case of English) of grammaticalized elements, which would benefit from being further investigated. Being able to study grammaticalization from such a perspective is also of particular importance to support the idea that the term is not too broad to be useful and that the many different cases of grammaticalization that are being investigated have enough in common to be considered as instances of the same phenomenon.

References

Monographs and articles

Aarts, B. and Meyer, C.F. (eds). 1995. *The Verb in contemporary English. Theory and description*. Cambridge: Cambridge University Press.

Altmann, E.G., Pierrehumbert, J.B. and Motter, A.E. 2009. *Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words*. PLoS ONE 4:e7678.doi:10.1371/journal.pone.0007678.

Andersen, H. 2006. Grammation, regrammation and degrammation: Tense loss in Russian. *Diachronica 23(2)*. 231-258.

Ansaldo, U. and Lim, L. 2004. Phonetic absence as syntactic prominence. In: Fischer, O., Norde, M. and Perridon, H. (eds). *Up and down the Cline – the Nature of Grammaticalization. Typological Studies in Language 59*. Amsterdam: John Benjamins. 345-362.

Aronoff, M. 1976. *Word Formation in Generative Grammar*. Cambridge, Mass.: MIT Press.

van der Auwera, J. 2002. More thoughts on degrammaticalization. In: Wischer, I. and Diewald, G. (eds). *New Reflections on Grammaticalization*. Typological Studies in Language 49. Amsterdam/Philadelphia: John Benjamins.

Baayen, H.R. 1992. Quantitative Aspects of Morphological Productivity. In: Booij, G.E. and Van Marle, J. (eds). *Yearbook of Morphology 1991*. Dordrecht/Boston/London: Kluwer. 109-149.

Baayen, H.R. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Baayen, H.R. and Lieber, R. 1991. Productivity and English Word-formation: a corpus-based Study. *Linguistics 29*. 801-843.

Bauer, L. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.

Baumann, A. and Sommerer, L. 2018. Linguistic diversification as a long-term effect of asymmetric priming: an adaptive-dynamics approach. *Language Dynamics and Change 8(2)*. 253-296.

Barth-Weingarten, D. and Couper-Kuhlen, E. 2002. On the development of final though. A case of grammaticalization? In: Wischer, I. and Diewald, G. (eds). *New Reflections on Grammaticalization*. 345-362.

Bavaud, F. 2009. Information Theory, Relative Entropy and Statistics. In: Sommaruga, G. (ed). *Formal Theories of Information*. *Lecture Notes in Computer Science 5363*. Berlin/Heidelberg: Springer. 54-78.

Bell, A., Brenier, J., Gregory, M., Girand, C. and Jurafsky, D. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language 60(1)*. 92-111.

Berg, T. 2009. *Structure in Language. A Dynamic Perspective*. New York: Routledge.

Biber, D. 1991. *Variation across speech and writing*. Cambridge: Cambridge University Press.

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

Biber, D., Conrad, S. and Leech, G. 2002. *Longman Student Grammar of Spoken and Written English*. London: Longman.

Biber, D. and Reppen, R. (eds). 2015. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press.

Bisang, W. 2004. Grammaticalization without coevolution of form and meaning: The case of tense-aspect-modality in East and mainland Southeast Asia. In: Bisang, W., Himmelman, N.P. and Wiemer, B. (eds). *What makes Grammaticalization? A Look from its Fringes and its Components.* Berlin: Mouton de Gruyter. 109-138.

Bisang, W. 2009. On the evolution of complexity – sometimes less is more in East and mainland Southeast Asia. In: Sampson, G., Gil, D. and Trudgill, O. (eds). *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press. 34-49.

Bisang, W. 2011. Grammaticalization and Linguistic Typology. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. Oxford: Oxford University Press. 105-117.

Bisang, W., Himmelman, N.P. and Wiemer, B. (eds). 2004. *What makes Grammaticalization? A Look from its Fringes and its Components.* Berlin: Mouton de Gruyter.

Booij, G.E. 1977. *Dutch morphology: A study of word formation in generative grammar*. Dordrecht: Foris.

Booij, G.E. and Van Marle, J. (eds). 1992. *Yearbook of Morphology 1991*. Dordrecht/Boston/London: Kluwer.

Börjars, K. and Vincent, N. 2011. Grammaticalization and Directionality. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. Oxford: Oxford University Press. 163-176.

Boume, G., Kraemer, I. and Zwarts, J. (eds). 2007. *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science.

Boye, K. and Harder, P. 2012. A Usage-based Theory of Grammatical Status and Grammaticalization. *Language 88(1)*. 1-44.

Breban, T. 2009. Structural persistence: A case based on the grammaticalization of English adjectives of difference. *English Language and Linguistics 13(1)*. 77-96.

Breban, T. 2011. Is there a postdeterminer in the English noun phrase? *Transactions of the Philological Society 108(3)*. 248-264.

Breban, T. 2014. What is secondary grammaticalization? In: Cuyckens, H. (ed). *Refining Grammaticalization*. Folia Linguistica: Acta Societatis Linguisticae Europae 48(2). Berlin/Munich/Boston: Walter de Gruyter. 449-502.

Brems, L. 2007. The Grammaticalization of Small Size Nouns. Reconsidering Frequency and Analogy. *Journal of English linguistics 35(4)*. 293-324.

Brinton, L.J. 2001. Historical Discourse Analysis. In: Schiffrin, D., Tannen, D. and Hamilton, H.E. (eds). *The Handbook of Discourse analysis*. Massachusetts: Blackwell. 138-160.

Brinton, L.J. and Traugott, E.C. 2005. *Lexicalization and Language Change.* Cambridge/New York: Cambridge University Press.

Brisard, F. 2001. Be going to: an exercise in grounding. *Journal of Linguistics 37(2)*.

Budts, S. and Petré, P. 2016. Reading the intentions of be going to. On the subjectification of future markers. *Folia Linguistica 37(1)*.

Burnard, L. (ed). 2007. Reference guide for the British National Corpus (XML edition). Available at: http://www.natcorp.ox.ac.uk/docs/URG/ (last accessed March 2018).

Bybee, J. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. 2002. Sequentiality as the basis of constituent structure. In: Givón, T. and Malle, B.F. (eds). *The Evolution of Language out of Pre-language*. Amsterdam/Philadelphia: John Benjamins Publishing. 109-134.

Bybee, J. 2003. Mechanisms of change in grammaticization: The role of frequency. In: Joseph, B.D. and Janda, J. (eds). *The Handbook of Historical Linguistics*. Oxford: Blackwell Publishing. 602-623.

Bybee, J. 2011. Usage-based Theory and Grammaticalization. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 69-78.

Bybee, J. 2013. Usage-based Theory and Exemplar Representations of Constructions. In: Hoffmann, T. and Trousdale G. (eds). *The Oxford Handbook of Construction Grammar*. New York: Oxford University Press. 49-69.

Bybee, J., Perkins, R. and Pagliuca, W. 1994. *The evolution of grammar. Tense, aspect and modality in the languages of the world.* Chicago/London: The University of Chicago Press.

Bybee, J. and Thompson, S. 1997. Three frequency effect in syntax. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society 23*. 65-85.

Bybee, J. and Hopper P.J. (eds). 2001. *Frequency and the emergence of linguistic structure*. Amsterdam/Philadelphia: John Benjamins Publishing.

Campbell, L. (ed). 2001a. *Grammaticalization: A Critical Assessment*. Special issue of *Language Sciences 23(2-3)*.

Campbell, L. 2001b. What's wrong with grammaticalization?. In: Campbell, L. (ed). *Grammaticalization: A Critical Assessment*. Special issue of *Language Sciences 23(2-3)*. 113-161.

Campbell, L. and Janda, R. 2001. Introduction: conception of grammaticalization and their problems. In: Campbell, L. (ed). *Grammaticalization: A Critical Assessment*. Special issue of *Language Sciences 23(2-3)*. 93-112.

Cech, R., Garabik, R. and Altmann, G. 2015. Testing the thematic concentration of text. *Journal of Quantitative Linguistics 22(3)*. 215-232.

Chomsky, N. 1995. *The Minimalist Program*. Cambridge, Massachusetts: MIT Press.

Company, C.C. 2002. Grammaticalization and category weakness. In: Wischer, I. and Diewald, G. (eds). *New Reflections on Grammaticalization*. Typological Studies in Language 49. Amsterdam/Philadelphia: John Benjamins. 201-216.

Covington, M.A. and McFall, J.D. 2010. Cutting the Gorian Knot: The Moving Average Type-Token Ratio (MATTR). *Journal of Quantitative Linguistics 17(2)*. 94-100.

Cox, D. R. and Snell, E. J. 1989. *The Analysis of Binary Data*. London: Chapman and Hall.

Croft, W.A. 2001. *Radical Construction Grammar: Syntactic theory in typological perspective*. New York: Oxford University Press.

Cuyckens, H. (ed). 2014. *Refining Grammaticalization*. Folia Linguistica: Acta Societatis Linguisticae Europae 48(2). Berlin/Munich/Boston: Walter de Gruyter.

Dahl, Ö. (ed). 2000. *Tense and Aspect in the Language of Europe*. Berlin/New York: Mouton de Gruyter.

Dahl, Ö. 2000. The tense-aspect systems of European languages in a typological perspective. In: Dahl, Ö. (ed). *Tense and Aspect in the Language of Europe*. Berlin/New York: Mouton de Gruyter.

Davidse, K., Breban, T., Brems, L. and Mortelmans, T. (eds). 2012. *Grammaticalization and Language Change: New reflections*. Amsterdam: John Benjamins Publishing.

Davies, M. 2004. BYU-BNC. (Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc/.

Davies, M. 2012. Expanding Horizons in Historical Linguistics with the 400 million word Corpus of Historical American English. *Corpora 7*. 121-57.

Deutscher, G. 2011. The grammaticalization of quotatives. In: Narrog, H. and Heine, B. (eds). 2011. *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 646-655.

Diessel, H. 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology 25(2)*. 108-127.

Diessel, H. and Hilpert, M. 2016. Frequency effects in grammar. In: Aronoff, M. (ed). *Oxford Research Encyclopedia of Linguistics*. New York: Oxford University Press.
Diewald, G. and Smirnova, E. 2010. *Evidentiality in German: Linguistic realization and regularities in grammaticalization*. Berlin/New York: Mouton de Gruyter.

Eckardt, R. Jäger, G. and Veenstra, T. (eds). 2008. *Variation, Selection, Development: Probing the evolutionary model of language change*. Berlin/New York: Mouton de Gruyter.

Evans, N. and Watanabe, H. 2016. *Insubordination*. Typological Studies in Language 115. Amsterdam/Philadelphia: John Benjamins Publishing.

Evans, V. and Pourcel, S. (eds). 2009. *New Directions in Cognitive Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing.

Evert, S. and Lüdeling A. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? *Proceedings of Corpus Linguistics 2001*. 167-175.

Facchinetti, R., Krug, M. and Palmer, F. (eds). 2003. *Modality in Contemporary English*. Berlin/New York: Mouton de Gruyter.

Fernández-Domínguez, J., Díaz-Negrillo, A. and Štekauer, P. 2007. How is low morphological productivity measured? *Atlantis 29(1)*. 29-54.

Fischer, O. 2007. *Morphosyntactic change: Functional and formal perspectives*. Oxford: Oxford University Press.

Fischer, O., Norde, M. and Perridon, H. (eds). 2004. *Up and down the Cline – the Nature of Grammaticalization. Typological Studies in Language 59*. Amsterdam: John Benjamins Publishing.

Flach, S. 2015. Let's go look at usage. A constructional approach to formal constraints on go-VERB. *Yearbook of the German Cognitive Linguistics Association 3(1)*. Berlin: Mouton de Gruyter. 231-252.

Fligelstone, S., Pacey, M. and Rayson, P. 1997. How to generalize the task of annotation. In: Garside, R., Leech, G., and McEnery, A. (eds). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman. 122-136.

Forster, K., Guerrera, C. and Elliot, L. 2009. The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods 41(1)*. 163-71.

Fraser, B. 1999. What are discourse markers? *Journal of Pragmatics 31*. 931-952.

Fried, M. 2013. Principles of constructional change. In: Hoffmann, T. and Trousdale G. (eds). 2013. *The Oxford Handbook of Construction Grammar*. New York: Oxford University Press. 419-437.

Fries, C.C. 1940. On the development of the structural use of word-order in Modern English. *Language 16*. 199-208.

Gallant, S.I. 1991. A Practical Approach for Representing Context and for Performing Word Sense Disambiguation Using Neural Networks. *Neural Computation 3(3)*.

Garside, R. 1987. The CLAWS Word-tagging System. In: Garside, R., Leech, G. and Sampson, G. (eds). *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Garside, R., Leech, G. and Sampson, G. (eds). 1987. *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Garside, R., Leech, G., and McEnery, A. (eds). 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London.

van Gelderen, E. 2004. *Grammaticalization as Economy*. Amsterdam: John Benjamins Publishing.

Giacalone Ramat, A. and Hopper, P.J. (eds). 1998. *The limits of Grammaticalization*. Typological Studies in Language 37. Amsterdam/Philadelphia: John Benjamins Publishing.

Gisborne, N. and Patten, A. 2011. Construction Grammar and Grammaticalization. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 92-104.

Givón, T. 1971. Historical Syntax and Synchronic morphology: An archeologist's field trip. *Chicago Linguistic Society Proceedings 7*. 394-415.

Givón, T. and Malle, B.F. (eds). 2002. *The Evolution of Language out of Pre-language*. Amsterdam/Philadelphia: John Benjamins Publishing.

van Goethem, K., Norde, M., Coussé, E. and Vanderbauwhede, G. (eds). 2018. *Category Change from a Constructional Perspective*. Constructional Approaches to Language 20. Amsterdam/Philadelphia: John Benjamins Publishing.

Goldberg, A.E. (ed.) 1996. *Conceptual structure, discourse and language*. Stanford: CSLI Publications.

Goldberg, A.E. 2006. *Constructions at Work: The nature of generalization in language*. New York: Oxford University Press.

Goldberg, A.E. and Jackendoff, R. 2004. The English Resultative as a Family of Constructions. *Language 80(3)*. 532-568.

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics 27(2).* 153-198.

Gries, S.Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics 13(4)*. 403-437.

Gries, S.Th. 2013. *Statistics for Linguistics with R. A practical introduction*. Berlin/Boston: Mouton de Gruyter.

Gries, S.Th. and Hilpert, M. 2010. Modeling diachronic change in the third person singular: a multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics 14(3)*. 293-320.

Gries, S.Th. and Ellis, N.C. 2015. Statistical Measures for Usage-Based Linguistics. *Language Learning 65(S1)*. 228-255.

Haspelmath, M. 1999. Why is grammaticalization irreversible? *Linguistics 37(6)*. 1043-1068.

Haspelmath, M. 2004. On directionality in language change with particular reference to grammaticalization. In: Fischer, O., Norde, M. and Perridon, H. (eds). *Up and down the Cline – the Nature of Grammaticalization.* Typological Studies in Language 59. Amsterdam: John Benjamins. 17-44.

Hay, J.B. 2002. From speech perception to morphology: Affix ordering revisited. *Language 78*. 527-555.

Hay, J.B. and Plag, I. 2004. What constrains possible suffix combinations? On the interaction of grammatical and processing restrictions in derivational morphology. *Natural language and linguistic theory 22*. 565-596.

Heiman, J. 1994. Ritualization and the development of language. In: Pagliuca, W. (ed). *Perspectives on grammaticalization*. Amsterdam: John Benjamins Publishing. 3-28.

Heiman, J. and Thompson, S.A. (eds). 1988. *Clause Combining in Grammar and Discourse*. Amsterdam: John Benjamins.

Heine, B. 1993. *Auxiliaries: cognitive forces and grammaticalization*. New York: Oxford University Press.

Heine, B. 2002. On the role of context in grammaticalization. In: Wischer, I. and Diewald, G. (eds). *New Reflections on Grammaticalization*. Typological Studies in Language 49. Amsterdam/Philadelphia: John Benjamins. 83-101.

Heine, B. 2003. Grammaticalization. In: Joseph, B.D. and Janda, J. (eds). *The Handbook of Historical Linguistics*. Oxford: Blackwell Publishing. 575-601.

Heine, B. 2013. On discourse markers: Grammaticalization, pragmaticalization, or something else? *Linguistics 51(6)*. 1205-1247.

Heine, B., Claudi, H. and Hünnemeyer F. 1991. *Grammaticalization. A Conceptual Framework*. Chicago: University of Chicago Press.

Heine, B. and Kuteva, T. 2002. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Hengeveld, K. 2011. The Grammaticalization of Tense and Aspect. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 580-594.

Hewings, A., Painter, C., Polias, J., Dare, B. and Rhys, M. 2005. *Getting Started: Describing the Grammar of Speech and Writing*. Milton Keynes: Open University Press.

Hickey, R. and Puppel, S. (eds). 1997. *Language History and Linguistic Modelling. Volume II: Linguistic Modelling. Trends in Linguistics. Studies and Monographs 101*. Berlin/New York: Mouton de Gruyter.

Hilpert, M. 2008. *Germanic Future Constructions. A usage-based approach to language change*. Amsterdam: John Benjamins Publishing.

Hilpert, M. 2010. What can synchronic gradience tell us about reanalysis? Verb-first conditionals in written German and Swedish. In: Traugott, E.C. and Trousdale, G. (eds). *Gradience, Gradualness and Grammaticalization*. Typological Studies in Language 90. Amsterdam: John Benjamins Publishing. 181-202.

Hilpert, M. 2011. Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics 16(4)*, 435-461.

Hilpert, M. 2013. *Constructional Change in English: Developments in Allomorphy, Word Formation and Syntax*. Cambridge: Cambridge University Press.

Hilpert, M. 2014. *Construction Grammar and its Application to English*. Edinburgh: Edinburgh University Press.

Hilpert, M. and Mair, C. 2015. Grammatical change. In: Biber, D. and Reppen, R. (eds). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press. 180-200.

Hilpert, M. and Correia Saavedra, D. 2018. The unidirectionality of semantic changes in grammaticalization: An experimental approach to the asymmetric priming hypothesis. *English Language and Linguistics 22(3)*. 357-380.

Hilpert, M. and Correia Saavedra, D. 2017a. Why are grammatical elements more evenly dispersed than lexical elements? Assessing the roles of frequency and semantic generality. *Corpora 12(3)*. 369-392.

Hilpert, M. and Correia Saavedra, D. 2017b. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory*. (online).

Himmelman, N.P. 2004. Lexicalization and Grammaticization: Opposite or orthogonal? In: Bisang, W., Himmelman, N.P. and Björn, W. (eds). *What makes grammaticalization? A Look from its Fringes and its Components.* Berlin: Mouton de Gruyter. 21-42.

Hoffmann, S. 2004. Are low-frequency complex prepositions grammaticalized? On the limits of corpus data – and the importance of intuition. In: Lindquist, H. and Mair, C. (eds). *Corpus Approaches to Grammaticalization in English*. 171-210.

Hoffmann, S. 2005. *Grammaticalization and English Complex Prepositions. A corpus-based study*. London: Routledge.

Hoffmann, T. and Trousdale G. (eds). 2013. *The Oxford Handbook of Construction Grammar*. New York: Oxford University Press.

Hopper, P.J. 1991. On Some Principles of Grammaticalization. In: Traugott, E.C. and Heine, B. (eds). *Approaches to Grammaticalization*. Amsterdam/Philadelphia: John Benjamins Publishing. 17-35.

Hopper, P.J. and Traugott, E.C. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.

Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. 2013. *Applied Logistic Regression*. Hoboken, New Jersey: John Wiley & Sons.

Huddleston, R. and Pullum, G.K. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hundt, M. 2001. What corpora tell us about the grammaticalisation of voice in get-constructions. *Studies in Language 25(1)*. 49-87.

Israel, M. 1996. The way constructions grow. In: Goldberg, A.E. (ed.) *Conceptual structure, discourse and language*. Stanford: CSLI Publications. 217-230.

Jäger, G. and Rosenbach, A. 2008. Priming and Unidirectional Language Change. *Theoretical Linguistics 34(2)*. 85-113.

Jäger, T.F. 2008. Categorial data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language 59*. 434-446.

Janda, R.D. 2001. Beyond 'pathways' and 'unidirectionality': On the discontinuity of language transmission and the counterability of grammaticalization. *Language Sciences 23*. 265-340.

Jonas, D., Whitman, J. and Garrett, A. (eds). *Grammatical Change: Origins, Nature, Outcomes.* New York: Oxford University Press.

Joseph, B.D. 2001. Is there such a thing as grammaticalization? *Language Sciences 22*. 163-186.

Joseph, B.D. 2011. Grammaticalization: a general critique. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 193-205.

Joseph, B.D. and Janda, J. (eds). 2003. *The Handbook of Historical Linguistics*. Oxford: Blackwell Publishing.

Keller, R. 1994. *Language Change: The Invisible Hand in Language*. London: Routledge.

Kiparsky, P. 2012. Grammaticalization as Optimisation. In: Jonas, D., Whitman, J. and Garrett, A. (eds). *Grammatical Change: Origins, Nature, Outcomes.* New York: Oxford University Press.

Köhler, R. (ed). 2009. *Issues in quantitative linguistics*. Lüdenscheid: RAM.

König, C. 2011. The grammaticalization of adpositions and case marking. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 511-521.

Koops, C. and Lohman, A. 2015. A quantitative approach to the grammaticalization of discourse markers: Evidence from their sequencing behavior. *International Journal of Corpus Linguistics 20(2)*. 232-259.

Krug, M. 2000. *Emerging English Modals. A Corpus-Based Study of Grammaticalization*. Berlin: Mouton de Gruyter.

Krug, M. 2001. Frequency, iconicity, categorization: evidence from emerging modals. In: Bybee, J. and Hopper P.J. (eds). *Frequency and the emergence of linguistic structure*. Amsterdam/Philadelphia: John Benjamins Publishing. 309-336.

Krug, M. 2011. Auxiliaries and grammaticalization. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 547-558.

Kuryłowicz, J. 1965. The evolution of grammatical categories. *Diogenes 51*. 55-71.

Kuteva, T. 2001. *Auxiliation. An Enquiry into the Nature of Grammaticalization*. New York: Oxford University Press.

Lamiroy, B. and de Mulder, W. 2011. Degrees of Grammaticalization Across Languages. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 302-317.

Langacker, R.W. 1987. *Foundations of Cognitive Grammar, Volume 1: Theoretical Prerequisites*. Stanford: Stanford University Press.

Langacker, R.W. 2005. Construction grammars: Cognitive, radical and less so. In: Ruiz de Mendoza Ibàñez, F.J. and Peña Cervel, M.S. *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction.* Berlin: Mouton de Gruyter. 101-159.

Leech, G. 2003. Modality on the move: The English modal auxiliaries 1961-1992. In: Facchinetti, R., Krug, M. and Palmer, F. (eds). *Modality in Contemporary English*. Berlin/New York: Mouton de Gruyter. 223-240.

Leech, G. 2013. Where have all the modals gone? An essay on the declining frequency of core modal auxiliaries in recent standard English. In: Marín-Arrese, J. I., Carretero, M., Arús Hita, J. and van der Auwera, J. (eds). *English Modality: Core, Periphery and Evidentiality*. Berlin/Boston: Mouton de Gruyter. 95-116.

Lehmann, C. 1982. *Thoughts on grammaticalization: A programmatic sketch, vol 1*. Arbeiten des Kölner Universalien-Projekts 48. Cologne: Universität zu Köln.

Lehmann, C. 2002. *Thoughts on grammaticalization*. Arbeitspapiere des Seminars für Sprachwissenschaft der Universität Erfurt 9.

Lehmann, C. 2002b. New reflections on grammaticalization and lexicalization. In: Wischer, I. and Diewald, G. (eds). 2002. *New Reflections on Grammaticalization*. Typological Studies in Language 49. Amsterdam/Philadelphia: John Benjamins. 1-18.

Lehmann, C. 2015. *Thoughts on grammaticalization*. Berlin: Language Science Press.

Levshina, N. 2015. *How to do Linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins Publishing.

Lijffijt, J. and Gries, S.Th. 2012. Correction to "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics 17(1)*. 147-149.

Lindquist, H. and Mair, C. (eds). 2004. *Corpus Approaches to Grammaticalization in English*. Amsterdam/Philadelphia: John Benjamins.

Lohmann, A. 2011. Help vs help to: a multifactorial, mixed-effects account of infinitive marker omission. *English Language and Linguistics 15(3)*. 499-521.

Mair, C. 1995. Changing patterns of complementation, and concomitant grammaticalization, of the verb *help* in present-day British English. In: Aarts, B. and Meyer, C.F. (eds). *The Verb in contemporary English. Theory and description*. Cambridge: Cambridge University Press.

Mair, C. 1997. The spread of the going-to-future in written English: A corpus-based investigation into language change in progress. In: Hickey, R. and Puppel, S. (eds). *Language History and Linguistic Modelling. Volume II: Linguistic Modelling. Trends in Linguistics. Studies and Monographs 101*. Berlin/New York: Mouton de Gruyter. 1537-1544.

Mair, C. 2002. Three Changing Patterns of Verb Complementation in Late Modern English: A Real-Time Study Based on Matching Text Corpora. *English Language and Linguistics 6(1)*. 105-31.

Marín-Arrese, J. I., Carretero, M., Arús Hita, J. and van der Auwera, J. (eds). 2013. *English Modality: Core, Periphery and Evidentiality*. Berlin/Boston: Mouton de Gruyter.

Matras, Y. 2011. Grammaticalization and Language Contact. In: Narrog, H. and Heine, B. (eds). 2011. *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 279-290.

McWorther, J.H. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology 5(2-3)*. 125-166.

Meillet, A. 1912. L'évolution des formes grammaticales. *Scientia (Rivista di Scienza) 12(6)*. 384-400. Reprinted in Meillet 1958 : 130-148.

Meillet, A. 1958. *Linguistique historique et linguistique générale*. Paris: Champion.

Menard, S. 2000. Coefficient of determination for multiple logistic regression analysis. *The American Statistician 54(1)*. 17-24.

von Mengden, F. and Horst, J.S. 2014. What is it then, this Grammaticalization? In: Cuyckens, H. (ed). *Refining Grammaticalization*. Folia Linguistica: Acta Societatis Linguisticae Europae 48(2). Berlin/Munich/Boston: Walter de Gruyter. 347-360.

Minkova, D. and Stockwell, R. (eds). 2002. *Studies in the History of the English Language: A Millennial Perspective*. Berlin: Mouton de Gruyter.

de Mulder, W. and Carlier, A. 2011. The grammaticalization of definite articles. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 522-534.

Nagelkerke, N.J.D. 1991. A note on the general definition of the coefficient of determination. *Biometrika 78(3)*. 691-692.

Narrog, H. and Heine, B. (eds). 2011. *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press.

Neels, J. 2015. The history of the quasi-auxiliary use(d) to: A usage-based account. *Journal of Historical Linguistics 5(2)*. 177-234.

Newmeyer, F. 1998. *Language Form and Language Function*. Cambridge, MA: MIT Press.

Noël, D. 2007. Diachronic construction grammar and grammaticalization theory. *Functions of Language 14(2)*. 177-202.

Norde, M. 2006. Demarcating degrammaticalization: the Swedish s-genitive revisited. *Nordic Journal of Linguistics 29(2)*. 201-238.

Norde, M. 2009. *Degrammaticalization*. Oxford: Oxford University Press.

Norde, M. 2010. Degrammaticalization: three common controversies. In: Stathi, K., Gehweiler, E. and König, E. (eds). *Grammaticalization. Current views and issues*. Amsterdam/Philadelphia: John Benjamins Publishing. 123-150.

Norde, M. 2012. Lehmann's parameters revisited. In: Davidse, K., Breban, T., Brems, L. and Mortelmans, T. (eds). *Grammaticalization and Language Change: New reflections*. Amsterdam: John Benjamins Publishing. 73-110.

Norde, M. and Beijering, K. 2014. Facing interfaces: A clustering approach to grammaticalization and related changes. In: Cuyckens, H. (ed). *Refining Grammaticalization*. Folia Linguistica: Acta Societatis Linguisticae Europae 48(2). Berlin/Munich/Boston: Walter de Gruyter. 385-424.

Pagliuca, W. (ed). 1994. *Perspectives on grammaticalization*. Amsterdam: John Benjamins Publishing.

Palmer, F.R. 2013. *Modality and the English Modals*. Abingdon/New York: Routledge.

Patten, A.L. 2010. Grammaticalization and the it-cleft construction. In: Traugott, E.C. and Trousdale, G. (eds). *Gradience, Gradualness and Grammaticalization*. Typological Studies in Language 90. Amsterdam: John Benjamin. 221-244.

Perek, F. 2018. Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory 14(1)*. 65-97.

Petré, P. 2016a. Unidirectionality as a cycle of convention and innovation : micro-changes in the grammaticalization of [be going to INF]. *Journal of Belgian Linguistics 30(1).* John Benjamins Publishing. 115-146.

Petré, P. 2016b. Grammaticalization by changing co-text frequencies, or why [BE Ving] became the progressive. *English Language and Linguistics 20(1).* 31-54.

Petré, P. and Van de Velde, F. 2018. The real-time dynamics of the individual and the community in grammaticalization. *Language 94(4)*. 867-901.

Pierrehumbert, J.B. 2012. Burstiness of Verbs and Derived Nouns. In: Santos, D., Linden, K. And Ng'ang'a, W. (eds). *Shall we play the Festschrift Game?: Essays on the Occasion of Lauri Carlson's 60th Brithday*. Berlin/Heidelberg: Springer-Verlag.

Plag, I. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Berlin/New York: Mouton de Gruyter.

Plag, I. and Baayen, R.H. 2009. Suffix ordering and morphological processing. *Language 85*. 109-152.

Popescu, I.-I. 2009. *Word Frequency Studies*. Berlin: Mouton de Gruyter.

Pustylnikov, O. and Schneider-Wiejowski, K. 2009. Measuring Morphological Productivity. In: Köhler, R. (ed). 2009. *Issues in quantitative linguistics*. Lüdenscheid: RAM. 1-21.

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

Ramat, P. 1992. Thoughts on degrammaticalization. *Linguistics 30(3)*. 549-560.

Ramat, P. 2011. Adverbial Grammaticalization. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 502-510.

Reinöhl, U. and Himmelmann, N.P. 2017. Renewal: A figure of speech or a process sui generis? *Language 93(2)*. 381-413.

Rhee, S. 2003. The Parallel Reduction Hypothesis in Grammaticalization Revisited: A Case of English Prepositions. *Discourse and Cognition 10(3)*. 187-205.

Richards, B.J. 1986. Type/Token Ratios: what do they really tell us? *Journal of Child language 14(2)*. 201-209.

Roberts, I. and Roussou, A. 2003. *Syntactic Change. A Minimalist Approach to Grammaticalization*. Cambridge: Cambridge University Press.

Rohdenburg, G. and Mondorf, B. (eds). 2003. *Determinants of grammatical variation in English.* Berlin: Mouton de Gruyter.

Rudanko, J. 2006. Watching English grammar change: A case study on complement selection in British and American English. *English Language and Linguistics 10(1)*. 31-48.

Ruiz de Mendoza Ibàñez, F.J. and Peña Cervel, M.S. 2005. *Cognitive Linguistics: Internal Dynamics and Interdisciplinary Interaction.* Berlin: Walter de Gruyter.

Sampson, G., Gil, D. and Trudgill, O. (eds). 2009. *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press.

Savoy, J. 2017. Analysis of the style and the rhetoric of the 2016 US presidential primaries. *Digital Scholarship in the Humanities 33(1)*. 143-159.

Schiering, R. 2006. *Cliticization and the Evolution of Morphology : a Cross-linguistic Study on Phonology in Grammaticalization*. PhD dissertation. Konstanz: Konstanzer Online-Publikations-System, Bibliothek der Universität Konstanz.

Schiffrin, D. 2001. Discourse Markers: Language, Meaning and Context. In: Schiffrin, D., Tannen, D. and Hamilton, H.E. (eds). *The Handbook of Discourse analysis*. Massachusetts: Blackwell.

Schiffrin, D., Tannen, D. and Hamilton, H.E. (eds). 2001. *The Handbook of Discourse analysis*. Massachusetts: Blackwell.

Schmidtke-Bode, K. 2009. *A Typology of Purpose Clauses*. Typological Studies in Language 88. Amsterdam: John Benjamins Publishing.

Shannon, C.E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal 27*. 379-423, 623-656.

de Smet, H. 2010. Grammatical interference: Subject marker for and the phrasal verb particles out and forth. In: Traugott, E.C. and Trousdale, G. (eds). *Gradience, Gradualness and Grammaticalization*. Typological Studies in Language 90. Amsterdam: John Benjamins Publishing. 75-104.

de Smet, H. 2013. *Spreading patterns: Diffusional change in the English system of complementation*. Oxford: Oxford University Press.

Smith, T.J. and McKenna, C.M. 2013. A Comparison of Logistic Regression Pseudo R^2 Indices. *General Linear Model Journal 39(2).*

Sommaruga, G. (ed). 2009. *Formal Theories of Information*. *Lecture Notes in Computer Science 5363*. Berlin/Heidelberg: Springer.

Sommerer, L. 2018. *Article Emergence in Old English. A Constructionalist Perspective.* Berlin/Boston: Mouton de Gruyter.

Stathi, K., Gehweiler, E. and König, E. (eds). 2010. *Grammaticalization. Current views and issues*. Amsterdam/Philadelphia: John Benjamins Publishing.

Stefanowitsch, A. and Gries, S.Th. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics 8(2)*. 209-43.

Sun, L. and Correia Saavedra, D. submitted. Measuring Grammaticalization in Chinese: A quantitative approach to Chinese grammaticalization through corpus analysis.

Sweetser, E.E. 1988. Grammaticalization and Semantic Bleaching. *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*. 389-405.

Tabor, W. and Traugott, E.C. 1998. Structural scope expansion and grammaticalization. In: Giacalone Ramat, A. and Hopper, P.J. (eds). *The limits of Grammaticalization*. Typological Studies in Language 37. Amsterdam/Philadelphia: John Benjamins Publishing. 229-272.

Tagliamonte, S. 2004. In: Lindquist, H. and Mair, C. (eds). *Corpus Approaches to Grammaticalization in English*. 33-56.

Tagliamonte, S., Durham, M. and Smith, J. 2014. Grammaticalization at an early stage: future be going to in conservative British dialects. *English Language and Linguistics 18(1)*. 75-108.

Tausczik, Y.R. and Pennebaker, J.W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology 29(1)*. 24-54.

Taylor, J.R. 2003. *Linguistic Categorization*. New York: Oxford University Press.

Thompson, S.A. and Mulac, A. 1991. A Quantitative Perspective on the Grammaticalization of Epistemic Parentheticals in English. In: Traugott, E.C and Heine, B. (eds). *Approaches to Grammaticalization*. Amsterdam: John Benjamins Publishing.

Torres Cacoullos, R. and Walker, J.A. 2011. Collocations in Grammaticalization and Variation. In: Narrog, H. and Heine, B. (eds). 2011. *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 225-238.

Traugott, E.C. 1995. The role of the development of discourse markers in a theory of grammaticalization. Paper presented at the 12th International Conference on Historical Linguistics, Manchester.

Traugott, E.C. 2002. From etymology to historical pragmatics. In: Minkova, D. and Stockwell, R. (eds). *Studies in the History of the English Language: A Millennial Perspective*. Berlin: Mouton de Gruyter. 19-49.

Traugott, E.C. 2003. Constructions in Grammaticalization. In: Joseph, B.D. and Janda, J. (eds). *The Handbook of Historical Linguistics*. Oxford: Blackwell Publishing. 624-647.

Traugott, E.C. 2008. Grammaticalization, constructions and the incremental development of language: Suggestions from the development of degree modifiers in English. In: Eckardt, R. Jäger, G. and Veenstra, T. (eds). *Variation, Selection, Development: Probing the evolutionary model of language change*. Berlin/New York: Mouton de Gruyter. 219-250.

Traugott, E.C. and Heine, B. (eds). 1991. *Approaches to Grammaticalization*. Amsterdam/Philadelphia: John Benjamins Publishing.

Traugott, E.C. and König, E. 1991. The Semantics-Pragmatics of Grammaticalization Revisted. In: Traugott, E.C. and Heine, B. (eds). 1991. *Approaches to Grammaticalization*. Amsterdam/Philadelphia: John Benjamins Publishing. 189-218.

Traugott, E.C. and Dasher, R.B. 2001. *Regularity in semantic change*. Cambridge: Cambridge University Press.

Traugott, E.C. and Trousdale, G. (eds). 2010. *Gradience, Gradualness and Grammaticalization*. Typological Studies in Language 90. Amsterdam: John Benjamins Publishing.

Traugott, E.C. and Trousdale, G. 2013. *Constructionalization and constructional changes*. New York: Oxford University Press.

Trousdale, G. 2010. Issues in constructional approaches to grammaticalization in English. In: Stathi, K., Gehweiler, E. and König, E. (eds). *Grammaticalization. Current views and issues*. Amsterdam/Philadelphia: John Benjamins Publishing. 51-72.

Turney, P.D. and Pantel, P. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37. 141-188.

Tweedie, F.J. and Baayen, H.R. 1998. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities 32(5)*. 323-352.

Vegnaduzzo, S. 2009. Morphological Productivity Rankings of Complex Adjectives. *Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity*. 79-86.

Vosberg, U. 2006. *Die Große Komplementverschiebung. Außersemantische Einflüsse auf die Entwicklung satzwertiger Ergänzungen im Neuenglischen*. Tübingen: Gunter Narr.

West, B. T., Welch, K. B. and Galecki, A. T. 2007. *Linear mixed models: A practical guide using statistical software*. New York: Chapman & Hall/CRC.

Wiemer, B. 2011. The grammaticalization of passives. In: Narrog, H. and Heine, B. (eds). *The Oxford Handbook of Grammaticalization*. New York: Oxford University Press. 535-546.

Wiemer, B. and Bisang, W. 2004. What makes grammaticalization? An appraisal of its components and its fringes. In: Bisang, W., Himmelmann, N.P. and Wiemer, B. (eds). *What makes Grammaticalization? A look from its Fringes and its Components*. Berlin: Mouton de Gruyter. 3-20.

Willis, D. 2007. Syntactic lexicalization as a new type of degrammaticalization. *Linguistics 45(2)*. 271-310.

Wischer, I. and Diewald, G. (eds). 2002. *New Reflections on Grammaticalization*. Typological Studies in Language 49. Amsterdam/Philadelphia: John Benjamins Publishing.

Xing, J. Z. 2015. A comparative study of semantic change in grammaticalization and lexicalization in Chinese and Germanic languages. *Studies in Language 39(3)*. 594-634.

Zeschel, A. 2009. What's (in) a construction? Complete inheritance vs. full-entry models. In: Evans, V. and Pourcel, S. (eds). *New Directions in Cognitive Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing. 185-200.

Zwanenburg, W. 1983. *Productivité morphologique et emprunt*. Amsterdam/Philadelphia: John Benjamins Publishing.

Corpora

British National Corpus (BNC). http://www.natcorp.ox.ac.uk/ (Last accessed March 2018).
Section on tagging: http://www.natcorp.ox.ac.uk/docs/URG/posguide.html

BNC-BYU. https://corpus.byu.edu/bnc/ (Last accessed March 2018).

Corpus of Contemporary American English (COCA). http://corpus.byu.edu/coca/ (Last accessed March 2018).

Corpus of Historical American English (COHA). http://corpus.byu.edu/coha/ (Last accessed March 2018).

Lancaster Corpus of Mandarin Chinese (LCMC).
http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/ (Last accessed April 2018).

Dictionaries and databases

Early English Books Online: http://eebo.chadwyck.com/home (Last accessed August 2018)

Oxford English Dictionary Online. http://www.oed.com (Last accessed March 2018)

Softwares and programming languages

IBM Corp. 2016. IBM SPSS Statistics for Windows, version 24.0. Armonk, NY: IBM Corp.

Python, version 2.7. https://www.python.org (Last accessed January 2017)

Appendices

Appendix A – Additional tables (chapter 5)

A1. Lexical and grammatical elements database

Word_Tagged_BNC	LexGramm	Word_Tagged_BNC	LexGramm
<w AT*>A	G	<w AJ*>ABLE	L
<w AT*>AN	G	<w AJ*>AVAILABLE	L
<w AT*>EVERY	G	<w AJ*>BEST	L
<w AT*>NO	G	<w AJ*>BIG	L
<w AT*>THE	G	<w AJ*>BRITISH	L
<w CJ*>AFTER	G	<w AJ*>DIFFERENT	L
<w CJ*>ALBEIT	G	<w AJ*>EARLY	L
<w CJ*>ALTHOUGH	G	<w AJ*>ECONOMIC	L
<w CJ*>AND	G	<w AJ*>FULL	L
<w CJ*>AS	G	<w AJ*>GENERAL	L
<w CJ*>AS FAR AS	G	<w AJ*>GOOD	L
<w CJ*>AS LONG AS	G	<w AJ*>GREAT	L
<w CJ*>AS SOON AS	G	<w AJ*>HIGH	L
<w CJ*>BECAUSE	G	<w AJ*>IMPORTANT	L
<w CJ*>BEFORE	G	<w AJ*>LARGE	L
<w CJ*>BUT	G	<w AJ*>LITTLE	L
<w CJ*>COS	G	<w AJ*>LOCAL	L
<w CJ*>EVEN	G	<w AJ*>LONG	L
<w CJ*>EXCEPT	G	<w AJ*>MAIN	L
<w CJ*>FOR	G	<w AJ*>MAJOR	L
<w CJ*>IF	G	<w AJ*>NATIONAL	L
<w CJ*>IN CASE	G	<w AJ*>NEW	L
<w CJ*>LIKE	G	<w AJ*>OLD	L
<w CJ*>NOR	G	<w AJ*>ONLY	L
<w CJ*>NOW THAT	G	<w AJ*>OTHER	L
<w CJ*>ONCE	G	<w AJ*>POLITICAL	L
<w CJ*>OR	G	<w AJ*>POSSIBLE	L
<w CJ*>PROVIDED	G	<w AJ*>PUBLIC	L
<w CJ*>RATHER THAN	G	<w AJ*>REAL	L
<w CJ*>SINCE	G	<w AJ*>RIGHT	L
<w CJ*>SO	G	<w AJ*>SMALL	L
<w CJ*>SUCH THAT	G	<w AJ*>SOCIAL	L
<w CJ*>THAN	G	<w AJ*>SURE	L
<w CJ*>THAT	G	<w AJ*>YOUNG	L
<w CJ*>THOUGH	G	<w AV*>A	L
<w CJ*>TILL	G	<w AV*>ABOUT	L
<w CJ*>UNLESS	G	<w AV*>ACTUALLY	L
<w CJ*>UNTIL	G	<w AV*>AGAIN	L
<w CJ*>WHEN	G	<w AV*>ALL	L
<w CJ*>WHERE	G	<w AV*>ALMOST	L
<w CJ*>WHEREAS	G	<w AV*>ALREADY	L
<w CJ*>WHETHER	G	<w AV*>ALSO	L
<w CJ*>WHETHER OR NOT	G	<w AV*>ALWAYS	L
<w CJ*>WHILE	G	<w AV*>AS	L
<w CJ*>WHILST	G	<w AV*>AT	L
<w CJ*>AS WELL AS	G	<w AV*>AT LEAST	L
<w DPS>HER	G	<w AV*>AWAY	L
<w DPS>HIS	G	<w AV*>BACK	L
<w DPS>ITS	G	<w AV*>BOTH	L
<w DPS>MY	G	<w AV*>OF COURSE	L
<w DPS>OUR	G	<w AV*>DOWN	L
<w DPS>THEIR	G	<w AV*>ENOUGH	L
<w DPS>YOUR	G	<w AV*>EVEN	L
<w DT*>ALL	G	<w AV*>EVER	L

<w DT*>ANOTHER	G	<w AV*>FAR	L
<w DT*>ANY	G	<w AV*>FOR EXAMPLE	L
<w DT*>BOTH	G	<w AV*>HERE	L
<w DT*>EACH	G	<w AV*>HOW	L
<w DT*>EITHER	G	<w AV*>HOWEVER	L
<w DT*>ENOUGH	G	<w AV*>IN	L
<w DT*>FEW	G	<w AV*>JUST	L
<w DT*>FEWER	G	<w AV*>LATER	L
<w DT*>FORMER	G	<w AV*>LESS	L
<w DT*>HALF	G	<w AV*>MORE	L
<w DT*>LATTER	G	<w AV*>MOST	L
<w DT*>LESS	G	<w AV*>MUCH	L
<w DT*>LITTLE	G	<w AV*>NEVER	L
<w DT*>MANY	G	<w AV*>NOW	L
<w DT*>MORE	G	<w AV*>OF	L
<w DT*>MOST	G	<w AV*>OFF	L
<w DT*>MUCH	G	<w AV*>OFTEN	L
<w DT*>NEITHER	G	<w AV*>ON	L
<w DT*>OWN	G	<w AV*>ONCE	L
<w DT*>SAME	G	<w AV*>ONLY	L
<w DT*>SEVERAL	G	<w AV*>OUT	L
<w DT*>SOME	G	<w AV*>OVER	L
<w DT*>SUCH	G	<w AV*>PERHAPS	L
<w DT*>THAT	G	<w AV*>PROBABLY	L
<w DT*>THESE	G	<w AV*>QUITE	L
<w DT*>THIS	G	<w AV*>REALLY	L
<w DT*>THOSE	G	<w AV*>RIGHT	L
<w DT*>WHAT	G	<w AV*>SO	L
<w DT*>WHATEVER	G	<w AV*>STILL	L
<w DT*>WHICH	G	<w AV*>THEN	L
<w DT*>WHOSE	G	<w AV*>THERE	L
<w P**>ANYBODY	G	<w AV*>THEREFORE	L
<w P**>ANYONE	G	<w AV*>TODAY	L
<w P**>ANYTHING	G	<w AV*>TOGETHER	L
<w P**>EACH OTHER	G	<w AV*>TOO	L
<w P**>EVERYBODY	G	<w AV*>UP	L
<w P**>EVERYONE	G	<w AV*>VERY	L
<w P**>EVERYTHING	G	<w AV*>WELL	L
<w P**>HE	G	<w AV*>WHEN	L
<w P**>HER	G	<w AV*>WHERE	L
<w P**>HERS	G	<w AV*>WHY	L
<w P**>HERSELF	G	<w AV*>YET	L
<w P**>HIM	G	<w CRD>MILLION	L
<w P**>HIMSELF	G	<w N**>AREA	L
<w P**>HIS	G	<w N**>AREAS	L
<w P**>I	G	<w N**>BACK	L
<w P**>IT	G	<w N**>BODY	L
<w P**>ITSELF	G	<w N**>BOOK	L
<w P**>LOTS	G	<w N**>BRITAIN	L
<w P**>ME	G	<w N**>BUSINESS	L
<w P**>MINE	G	<w N**>CAR	L
<w P**>MYSELF	G	<w N**>CASE	L
<w P**>NO	G	<w N**>CENTRE	L
<w P**>NOBODY	G	<w N**>CHILD	L
<w P**>NONE	G	<w N**>CHILDREN	L
<w P**>NOTHING	G	<w N**>CITY	L
<w P**>ONE	G	<w N**>COMMUNITY	L
<w P**>ONE ANOTHER	G	<w N**>COMPANY	L
<w P**>OURS	G	<w N**>CONTROL	L
<w P**>OURSELVES	G	<w N**>COUNCIL	L
<w P**>PLENTY	G	<w N**>COUNTRY	L
<w P**>SHE	G	<w N**>COURT	L
<w P**>SOMEBODY	G	<w N**>DAY	L
<w P**>SOMEONE	G	<w N**>DAYS	L

<w P**>SOMETHING	G	<w N**>DEVELOPMENT	L
<w P**>THEM	G	<w N**>DOOR	L
<w P**>THEMSELVES	G	<w N**>EDUCATION	L
<w P**>THEY	G	<w N**>END	L
<w P**>US	G	<w N**>ENGLAND	L
<w P**>WE	G	<w N**>EYES	L
<w P**>WHO	G	<w N**>FACE	L
<w P**>WHOEVER	G	<w N**>FACT	L
<w P**>WHOM	G	<w N**>FAMILY	L
<w P**>YA	G	<w N**>FATHER	L
<w P**>YE	G	<w N**>FORM	L
<w P**>YOU	G	<w N**>GOD	L
<w P**>YOURS	G	<w N**>GOVERNMENT	L
<w P**>YOURSELF	G	<w N**>GROUP	L
<w PRF>OF	G	<w N**>HAND	L
<w PRP>ABOUT	G	<w N**>HEAD	L
<w PRP>ABOVE	G	<w N**>HEALTH	L
<w PRP>ACCORDING TO	G	<w N**>HOME	L
<w PRP>ACROSS	G	<w N**>HOUSE	L
<w PRP>AFTER	G	<w N**>INFORMATION	L
<w PRP>AGAINST	G	<w N**>INTEREST	L
<w PRP>AGED	G	<w N**>JOHN	L
<w PRP>AHEAD OF	G	<w N**>LAW	L
<w PRP>ALONG	G	<w N**>LEVEL	L
<w PRP>ALONGSIDE	G	<w N**>LIFE	L
<w PRP>AMID	G	<w N**>LONDON	L
<w PRP>AMONG	G	<w N**>MAN	L
<w PRP>AMONGST	G	<w N**>MARKET	L
<w PRP>APART FROM	G	<w N**>MEMBERS	L
<w PRP>AROUND	G	<w N**>MEN	L
<w PRP>AS	G	<w N**>MONEY	L
<w PRP>AS OPPOSED TO	G	<w N**>MONTHS	L
<w PRP>AT	G	<w N**>MOTHER	L
<w PRP>AWAY FROM	G	<w N**>MR	L
<w PRP>BECAUSE OF	G	<w N**>NAME	L
<w PRP>BEFORE	G	<w N**>NIGHT	L
<w PRP>BEHIND	G	<w N**>NUMBER	L
<w PRP>BELOW	G	<w N**>OFFICE	L
<w PRP>BENEATH	G	<w N**>OTHERS	L
<w PRP>BESIDE	G	<w N**>PART	L
<w PRP>BETWEEN	G	<w N**>PARTY	L
<w PRP>BEYOND	G	<w N**>PEOPLE	L
<w PRP>BUT	G	<w N**>PER CENT	L
<w PRP>BY	G	<w N**>PERIOD	L
<w PRP>BY MEANS OF	G	<w N**>PERSON	L
<w PRP>BY WAY OF	G	<w N**>PLACE	L
<w PRP>CONCERNING	G	<w N**>POINT	L
<w PRP>CONTRARY TO	G	<w N**>POLICE	L
<w PRP>DEPENDING ON	G	<w N**>POLICY	L
<w PRP>DESPITE	G	<w N**>POWER	L
<w PRP>DOWN	G	<w N**>PROBLEM	L
<w PRP>DUE TO	G	<w N**>PROBLEMS	L
<w PRP>DURING	G	<w N**>QUESTION	L
<w PRP>EXCEPT FOR	G	<w N**>RESEARCH	L
<w PRP>FOLLOWING	G	<w N**>RIGHT	L
<w PRP>FOR	G	<w N**>ROAD	L
<w PRP>FROM	G	<w N**>ROOM	L
<w PRP>IN	G	<w N**>SCHOOL	L
<w PRP>IN ACCORDANCE WITH	G	<w N**>SERVICE	L
<w PRP>IN ADDITION TO	G	<w N**>SERVICES	L
<w PRP>IN CHARGE OF	G	<w N**>SIDE	L
<w PRP>IN CONJUNCTION WITH	G	<w N**>SOCIETY	L
<w PRP>IN CONNECTION WITH	G	<w N**>SOUTH	L
<w PRP>IN FAVOUR OF	G	<w N**>STATE	L

<w PRP>IN FRONT OF	G	<w N**>SYSTEM	L
<w PRP>IN LINE WITH	G	<w N**>THING	L
<w PRP>IN RELATION TO	G	<w N**>THINGS	L
<w PRP>IN RESPECT OF	G	<w N**>TIME	L
<w PRP>IN RESPONSE TO	G	<w N**>TIMES	L
<w PRP>IN SPITE OF	G	<w N**>USE	L
<w PRP>IN SUPPORT OF	G	<w N**>WAR	L
<w PRP>IN TERMS OF	G	<w N**>WATER	L
<w PRP>IN THE LIGHT OF	G	<w N**>WAY	L
<w PRP>IN TOUCH WITH	G	<w N**>WEEK	L
<w PRP>IN VIEW OF	G	<w N**>WOMEN	L
<w PRP>INCLUDING	G	<w N**>WORDS	L
<w PRP>INSIDE	G	<w N**>WORK	L
<w PRP>INSTEAD OF	G	<w N**>WORLD	L
<w PRP>INTO	G	<w N**>YEAR	L
<w PRP>LIKE	G	<w N**>YEARS	L
<w PRP>MINUS	G	<w V**>'D	L
<w PRP>NEAR	G	<w V**>ASKED	L
<w PRP>NEXT TO	G	<w V**>BECOME	L
<w PRP>OFF	G	<w V**>BEGAN	L
<w PRP>ON	G	<w V**>BEING	L
<w PRP>ON BEHALF OF	G	<w V**>CA	L
<w PRP>ON THE PART OF	G	<w V**>CALLED	L
<w PRP>ON TOP OF	G	<w V**>CAME	L
<w PRP>ONTO	G	<w V**>CAN	L
<w PRP>OTHER THAN	G	<w V**>COME	L
<w PRP>OUT	G	<w V**>COULD	L
<w PRP>OUTSIDE	G	<w V**>DOES	L
<w PRP>OVER	G	<w V**>DOING	L
<w PRP>PAST	G	<w V**>FEEL	L
<w PRP>PER	G	<w V**>FELT	L
<w PRP>PLUS	G	<w V**>FIND	L
<w PRP>PRIOR TO	G	<w V**>FOUND	L
<w PRP>RATHER THAN	G	<w V**>GET	L
<w PRP>REGARDING	G	<w V**>GIVE	L
<w PRP>ROUND	G	<w V**>GIVEN	L
<w PRP>SINCE	G	<w V**>GOT	L
<w PRP>SUBJECT TO	G	<w V**>HAVING	L
<w PRP>SUCH AS	G	<w V**>HELD	L
<w PRP>THROUGH	G	<w V**>HELP	L
<w PRP>THROUGHOUT	G	<w V**>KEEP	L
<w PRP>TILL	G	<w V**>KNEW	L
<w PRP>TO	G	<w V**>KNOW	L
<w PRP>TOWARD	G	<w V**>KNOWN	L
<w PRP>TOWARDS	G	<w V**>LEFT	L
<w PRP>UNDER	G	<w V**>LET	L
<w PRP>UNLIKE	G	<w V**>LIKE	L
<w PRP>UNTIL	G	<w V**>LOOK	L
<w PRP>UP	G	<w V**>LOOKED	L
<w PRP>UPON	G	<w V**>LOOKING	L
<w PRP>VIA	G	<w V**>MADE	L
<w PRP>WITH	G	<w V**>MAKE	L
<w PRP>WITH REGARD TO	G	<w V**>MAKING	L
<w PRP>WITHIN	G	<w V**>MAY	L
<w PRP>WITHOUT	G	<w V**>MEAN	L
<w PRP>WORTH	G	<w V**>MIGHT	L
<w V**>'LL	G	<w V**>MUST	L
<w V**>'M	G	<w V**>NEED	L
<w V**>'RE	G	<w V**>PUT	L
<w V**>'S	G	<w V**>SAID	L
<w V**>'VE	G	<w V**>SAW	L
<w V**>AM	G	<w V**>SAY	L
<w V**>ARE	G	<w V**>SAYS	L
<w V**>BE	G	<w V**>SEE	L

<w V**>BEEN	G	<w V**>SEEN	L
<w V**>DID	G	<w V**>SET	L
<w V**>DO	G	<w V**>SHOULD	L
<w V**>DONE	G	<w V**>TAKE	L
<w V**>GO	G	<w V**>TAKEN	L
<w V**>GOES	G	<w V**>TELL	L
<w V**>GOING	G	<w V**>THINK	L
<w V**>GONE	G	<w V**>THOUGHT	L
<w V**>HAD	G	<w V**>TOLD	L
<w V**>HAS	G	<w V**>TOOK	L
<w V**>HAVE	G	<w V**>TURNED	L
<w V**>IS	G	<w V**>USE	L
<w V**>WAS	G	<w V**>USED	L
<w V**>WERE	G	<w V**>USING	L
<w V**>WILL	G	<w V**>WANT	L
<w XX0>N	G	<w V**>WENT	L
<w XX0>N'T	G	<w V**>WORK	L
<w XX0>NOT	G	<w V**>WOULD	L

A2. Top 100 and bottom 100 items using the written BNC, punctuation included

Top100	L/G	Score	Bottom100	L/G	Score
<w AT*>A	G	1	<w N**>POWER	L	0.087
<w AT*>THE	G	1	<w V**>KEEP	L	0.085
<w AV*>A	L	1	<w N**>MARKET	L	0.084
<w CJ*>AND	G	1	<w N**>ROOM	L	0.084
<w CJ*>COS	G	1	<w N**>SIDE	L	0.084
<w CJ*>THAT	G	1	**<w V**>GONE**	G	0.084
<w DPS>HIS	G	1	**<w V**>'VE**	G	0.081
<w DT*>THIS	G	1	<w N**>YEAR	L	0.079
<w P**>HE	G	1	<w V**>SAY	L	0.079
<w P**>I	G	1	<w AJ*>SOCIAL	L	0.078
<w P**>IT	G	1	<w N**>HEAD	L	0.078
<w P**>YA	G	1	<w N**>NIGHT	L	0.078
<w P**>YOU	G	1	<w N**>NUMBER	L	0.078
<w PRF>OF	G	1	<w N**>CASE	L	0.077
<w PRP>ALONGSIDE	G	1	<w N**>GOVERNMENT	L	0.076
<w PRP>AMID	G	1	<w N**>LAW	L	0.076
<w PRP>AS OPPOSED TO	G	1	<w N**>WEEK	L	0.072
<w PRP>AT	G	1	<w V**>TAKEN	L	0.072
<w PRP>BUT	G	1	<w AJ*>RIGHT	L	0.071
<w PRP>BY	G	1	<w N**>WORDS	L	0.071
<w PRP>BY MEANS OF	G	1	**<w V**>'RE**	G	0.071
<w PRP>BY WAY OF	G	1	<w AJ*>ABLE	L	0.067
<w PRP>DOWN	G	1	<w N**>POINT	L	0.067
<w PRP>EXCEPT FOR	G	1	<w AJ*>SURE	L	0.065
<w PRP>FOLLOWING	G	1	<w N**>LEVEL	L	0.065
<w PRP>FOR	G	1	<w V**>CA	L	0.065
<w PRP>FROM	G	1	<w V**>SAYS	L	0.065
<w PRP>IN	G	1	<w V**>TAKE	L	0.065
<w PRP>IN ACCORDANCE WITH	G	1	<w N**>LIFE	L	0.064
<w PRP>IN CONJUNCTION WITH	G	1	<w N**>MAN	L	0.063
<w PRP>IN CONNECTION WITH	G	1	<w V**>KNOW	L	0.062
<w PRP>IN FAVOUR OF	G	1	<w N**>WORK	L	0.06
<w PRP>IN LINE WITH	G	1	<w V**>COME	L	0.06
<w PRP>IN RESPONSE TO	G	1	<w V**>FEEL	L	0.06
<w PRP>IN SUPPORT OF	G	1	<w V**>MAKE	L	0.059
<w PRP>IN THE LIGHT OF	G	1	<w V**>PUT	L	0.059
<w PRP>IN VIEW OF	G	1	<w N**>FATHER	L	0.058
<w PRP>INSTEAD OF	G	1	<w N**>WAR	L	0.058
<w PRP>MINUS	G	1	<w V**>BECOME	L	0.058
<w PRP>ON	G	1	**<w V**>DONE**	G	0.058

<w PRP>ON BEHALF OF	G	1	<w V**>GET	L	0.058
<w PRP>ON THE PART OF	G	1	<w N**>INTEREST	L	0.057
<w PRP>ON TOP OF	G	1	**<w V**>GO**	G	0.057
<w PRP>REGARDING	G	1	<w N**>NAME	L	0.055
<w PRP>TILL	G	1	<w V**>TELL	L	0.055
<w PRP>TO	G	1	<w N**>PROBLEM	L	0.054
<w PRP>TOWARD	G	1	<w V**>SEEN	L	0.054
<w PRP>UNTIL	G	1	<w N**>ROAD	L	0.053
<w PRP>WITH	G	1	**<w V**>'M**	G	0.053
<w PRP>WITH REGARD TO	G	1	<w V**>GIVEN	L	0.052
<w V**>ARE	G	1	**<w V**>GOING**	G	0.052
<w V**>BE	G	1	<w N**>HOUSE	L	0.05
<w V**>HAD	G	1	<w N**>COUNTRY	L	0.048
<w V**>IS	G	1	<w N**>COURT	L	0.048
<w V**>WAS	G	1	**<w P**>ONE**	G	0.048
<w XX0>NOT	G	1	<w V**>GIVE	L	0.048
<w CJ*>BUT	G	0.999	<w V**>MEAN	L	0.048
<w CJ*>IN CASE	G	0.999	<w N**>MOTHER	L	0.044
<w CJ*>LIKE	G	0.999	<w V**>THINK	L	0.044
<w CJ*>OR	G	0.999	<w N**>WORLD	L	0.042
<w CJ*>RATHER THAN	G	0.999	<w V**>FIND	L	0.042
<w CJ*>SUCH THAT	G	0.999	<w V**>HELD	L	0.042
<w CJ*>TILL	G	0.999	<w V**>USE	L	0.037
<w PRP>AHEAD OF	G	0.999	<w N**>FORM	L	0.035
<w PRP>IN CHARGE OF	G	0.999	**<w V**>'LL**	G	0.033
<w PRP>IN RELATION TO	G	0.999	<w N**>PERIOD	L	0.031
<w PRP>PRIOR TO	G	0.999	<w V**>LEFT	L	0.031
<w PRP>UP	G	0.999	<w V**>NEED	L	0.031
<w PRP>VIA	G	0.999	<w AJ*>LITTLE	L	0.03
<w V**>HAVE	G	0.999	<w N**>BODY	L	0.029
<w AT*>AN	G	0.998	<w V**>LOOK	L	0.028
<w CJ*>ALBEIT	G	0.998	<w N**>PLACE	L	0.027
<w P**>OURS	G	0.998	<w N**>CHILD	L	0.026
<w PRP>AMONGST	G	0.998	<w N**>USE	L	0.026
<w PRP>BENEATH	G	0.998	<w V**>MADE	L	0.025
<w PRP>CONCERNING	G	0.998	<w V**>WANT	L	0.025
<w PRP>IN RESPECT OF	G	0.998	<w V**>HELP	L	0.024
<w CJ*>EVEN	G	0.997	<w V**>LIKE	L	0.023
<w P**>ONE ANOTHER	G	0.997	<w V**>WORK	L	0.022
<w P**>THEY	G	0.997	<w N**>END	L	0.021
<w P**>WHOEVER	G	0.997	<w N**>WAY	L	0.021
<w P**>YE	G	0.997	<w V**>KNEW	L	0.021
<w PRP>NEXT TO	G	0.997	<w N**>DOOR	L	0.02
<w PRP>ONTO	G	0.997	<w N**>QUESTION	L	0.018
<w PRP>OUT	G	0.997	<w N**>HAND	L	0.014
<w DT*>WHICH	G	0.996	<w N**>STATE	L	0.012
<w PRP>IN SPITE OF	G	0.996	<w N**>FACE	L	0.011
<w PRP>OTHER THAN	G	0.996	<w N**>FACT	L	0.011
<w CJ*>NOW THAT	G	0.994	<w V**>'D	L	0.011
<w P**>SHE	G	0.994	<w N**>THING	L	0.009
<w PRP>DEPENDING ON	G	0.994	<w V**>TURNED	L	0.009
<w CJ*>PROVIDED	G	0.993	<w N**>EYES	L	0.008
<w CJ*>WHETHER OR NOT	G	0.993	<w V**>FELT	L	0.007
<w PRP>INSIDE	G	0.993	<w V**>FOUND	L	0.007
<w PRP>CONTRARY TO	G	0.992	<w N**>PERSON	L	0.006
<w PRP>IN ADDITION TO	G	0.992	<w V**>THOUGHT	L	0.004
<w V**>WERE	G	0.991	<w V**>ASKED	L	0.003
<w CJ*>AS LONG AS	G	0.989	<w V**>TOLD	L	0.003
<w CJ*>WHILST	G	0.989	<w V**>GOT	L	0.002
<w DPS>THEIR	G	0.989	<w V**>LOOKED	L	0.002

Appendix B – Random samples for classification data (chapter 6)

B1-A. Keep

0 in between: (total=224)	Correct?
<pphs1> he <vvd> rebuffed <nnu> me. <vv0_nn1> // <ppis1> I <vvd> kept <vvg_nn1> thinking <cst> that <at> the <np1> White	y
<at> the <nn1> moment <nn1_vv0> pass <y> , <pphs1> he <vvd> kept <vvg> trying <cc> and <ddq> trying-What <at> the	y
<jj_nn1@> white <y> , <nn1> fleecy <nn2_vvz%> clouds <cst> that <vvd> kept <vvg> changing <appge> their <nn1> shape <at1> every	y
<ppx122> another <ge> 's <nn2> arms <y> . <np1> Saunders <vvd> kept <vvg> sniffing <ii> at <at> the <nn1> blood	y
<nn1> floor <y> , <ccb> but <rt> then <pphs1> he <vvd> kept <vvg> trying <to> to <vvi> unlock <at> the	y
<vbz> 's <xx> not <vdg> doing <pph1> it <to> to <vvi> keep <vvg> feeling <jj> young <y> . <ppis1> I	y
<nn1> officer <cc> and <md> 2nd <jj> young <nn1> officer <vv0> keep <vvg_jj@> moving <y> . <pphs2> They <vv0> start	y
<csn> than <at1> a <nn1> twosome <y> . <ppis1> I <vvd> kept <vvg> wondering <cs_csw@> if <ppis1> I <vhd> 'd	y
<xx> not <vvi> look <jj> familiar <y> . <ppis1> I <vvd> kept <vvg> wishing <rg> so <da1> much <cst> that	y
<"> " <"> " <rr> Well <y> , <ppis1> I <vv0> keep <vvg_jj@> moving <y> . <ppis1> I <vvd> suspected	y
<y> , <at> the <nn2_np1> Moros <iw> with <ppio1> me <vvd> kept <vvg> turning <rp> around <cc> and <vvg> looking	y
<to> to <y> , <ccb> but <jj> certain <nn2> things <vv0> keep <vvg> bothering <ppio1> me <y> . <ppis1> I	y
<jj_nn1> brain-damaged <np1> Larry <np1> Belleau <y> , <pnqs> who <vvd> kept <vvg> slobbering <ppho1> her <iw> with <nn2> kisses	y
<appge> his <nn2> eyes <vvd_vvn> closed <y> . <np1@_nn1> Lucille <vvd> kept <vvg> holding <rp_ii21> on <ii_ii22> to <np1> Chris	y
<io> of <at> the <jj> battered <nn2> wives <pnqs> who <vv0> keep <vvg> coming <rp> back <if> for <dar_rrr> more	y
<vbdz> was <dd1> this <nn1> Jew <y> ? <pphs1> he <vvd> kept <vvg> asking <ppx1> himself <ii> in <at1> a	y
<nn1> mouth <cc> and <vv0> listen <y> . <ppy> You <vvd> kept <vvg> asking <ppio1> me <ddq> what <vbz> 's	y
<np1> Lucille <vvg> acting <rr> strangely <y> . <pphs1> She <vvd> kept <vvg_nn1> crossing <cc> and <jj> uncrossing <appge> her	y
<pph1> it <db_rr@> all <vvi> mean <y> ? <pphs1> he <vvd> kept <vvg> asking <ppx1> himself <csa> as <pphs1> he	y
<jj> still <cc> or <vv0_nn1> retrogress <y> ; <ppis2> we <vv0> keep <vvg> going <rp@_ii> on <cc> and <rp> up	y
<vv0_nn1> upand <appge> his <nn1> chin <vvg> bobbing <pph1> it <vvd> kept <vvg> coming <rp> out <np1> Je <ge> '	y
<vvg> getting <iw> with <ppio2> us <y> . <ppis1> I <vv0> keep <vvg> waiting <if> for <ppy> you <to> to	y
<y> . <cs> After <ppis1> I <vvd> retired <ppis1> I <vvd> kept <vvg> going <ii> to <at> the <nn2> races	y
<ii> to <appge> its <da> former <nn1> direction <pphs2> they <vvd> kept <vvg_jj@> hawking <cc> and <vvg_jj@_nn1@> spitting <cs21> as	y
<to> to <vvi> shield <ppho1> him <y> , <pphs1> he <vvd> kept <vvg> sticking <appge> his <jj> great <nn1> head	y
Total Correct	25

1 in between: (total=83)	
<ppio1> me <y> , <cs> although <ppis1> I <vvd> kept <ppho2> them <vvg> waiting <dd> some <nnt1> time <cs> while	y
<vvg> biting <appge> his <nn2> lips <to> to <vvi> keep <ii> from <vvg> laughing <y> . <nn1_np1@> Rieber <vvz> looks	from
<rr21> a <rr22> little <y> . <ppis1> I <vvd> kept <rp> on <vvg> telling <ppho1> him <ii> about <np1> Skalla	y
<vdd> did <xx> n't <vvi> want <to> to <vvi> keep <pn1> anybody <vvg> waiting <y> , <nn1> miss <y> .	y
<ppy> you <vvd_vvn> left <-> - <to> to <vvi> keep <nn2> things <vvg> going <y> . <vv0> Know <ddq> what	y
<at1> a <nn1> profit <y> . <nn2> Subsidies <vv0> keep <ppho2> them <vvg> going <y> . <ccb> But <pph1> it	y
<appge> our <nn1> turf <y> , <cc> and <vvi> keep <rp> on <vvg> fighting <if> for <pph1> it <y> .	y
<mc> two <nn2_vvz> sticks <cs> if <ppis2> we <vv0> keep <pph1> it <vvg> waiting <at1> a <nnt1> minute <y> .	y
<nn2> hands <cs21> as <cs22> if <to> to <vvi> keep <ii> from <vvg> falling <y> . <"> " <jj@_vvn> Saved	from
<rrr> later <y> . <"> " <np1> Morris <vvd> kept <rp> on <vvg> talking <y> . <vv0_nn1_np1@> Lagarge <y> ,	y
<ppy> You <vm> 'd <vvi> love <to> to <vvi> keep <ppio1> me <vvg> hustling <if> for <ppy> you <y> ,	y
<vd0> do <xx> n't <vvi> want <to> to <vvi> keep <rp> on <vvg> giving <ppy> you <jj_vv0%> smart <nn2> answers	y
<y> ; <ccb> but <rr> still <ppis1> I <vvd> kept <rp> on <vvg> saving <if> for <pph1> it <y> ,	y
<rr22> little <y> , <ccb> but <ppis1> I <vvd> kept <rp> on <vvg> running <rr_jj> fast <cs21> even <cs22> though	y
<y> , <rr> just <rr_dd> enough <to> to <vvi> keep <ppio1> me <vvg_nn1> reading <if> for <at> the <md> next	y
<mc1> one <nn1> end <y> . <pph1> It <vvd> kept <rp> on <vvg_nn1> ringing <cc> and <at1> a <nn1> horn	y
<nn1> sir <y> ; <vd0> do <xx> not <vvi> keep <ppio1> me <vvg> waiting <y> . <"> " <ppis1> I	y
<y> . <"> " <rr> Well <!> ! <vvd> kept <ppy> you <vvg> waiting <y> . <ppis1> I <vdd> did	y
<vbi> be <jj> noble <y> , <cc> and <vv0> keep <ii> from <vvg> envying <appge> your <nn2> neighbors <y> ,	from
<cc> And <y> , <cs> if <ppis1> I <vvd> kept <ppho1_appge> her <vvg_nn1@> waiting <y> , <rr> even <if> for	y
<np1> Lydia <y> , <jj> sorry <to> to <vvi> keep <ppy> you <vvg> waiting <y> . <"> " <"> "	y
<y> . <"> " <jj> Sorry <to> to <vvi> keep <ppy> you <vvg> waiting <y> , <"> " <np1> MacLoon	y
<mc> ten <nn2> feet <cst> that <pphs1> he <vvd> kept <if> for <vvg_jj@> measuring <nn2> purposes <y> . <ccb> But	n
<appge> His <nn2> legs <vvg> fighting <nn1> desperately to <vv0> keep <ii> from <vvg> sliding <rp> off <y> . <at> The	from
<nn1> temperature <y> ; <pphs1> she <vm> could <vvi> keep <nn2> tomatoes <vvg> growing <db> all <nnt1> winter <y> ,	y
Total Correct	20
2 in between: (total=82)	
<vbr> are <jj_vvg@> healing <pphs2> they <vv0> keep <nn1_rr21%> kind <io_rr22> of <vvg> tickling <rp@_ii> on <rl> there <y> .	y
<rr> just <vm> ca <xx> n't <vvi> keep <at> the <nn1> pressure <vvg> going <jj_rr%> full <nn1_vv0> blast <db> all	y
<vbi> be <y> , <to> to <vvi> keep <ppy> you <ii> from <vvg> becoming <at1> a <nn1> drunkard <cc> and	from

<nn2> taxes <ppis2> we <vm> could <vvi> keep <at> the <nn1> war <vvg> going <if> for <da2> several <nn2> campaigns	y
<ii> in <nnt1> time <to> to <vvi> keep <ppho1> her <ii> from <vvg> opening <pph1> it <iw> with <appge> her	from
<y> , <z> -- <"> " <vv0> keep <ppio1> me <rl> here <vvg> talking <y> , <cc> and <rt> then	y
<jj_rr%> strong <at1> a <nn1> necessity <vvd> kept <ppio1> me <ii> from <vvg_nn1> writing <ii> to <ppy> you <ii> before	from
<rt> again <y> ? <uh_zz1> O <vv0> keep <ppio1> me <ii> from <vvg> sinning <ii> against <ppy> thee <y> ,	from
<nn1> modernization <ccb> but <to> to <vvi> keep <at> the <nn2> Soviets <vvg> moving <ii> at <at> the <nn1> negotiation	y
<cst> that <nnb> Mr. <np1> Field <vvd> kept <dd2> these <"> " <vvg> talking <nn2> books <"> " <ii> on	n
<dd_rr> enough <nn1> deuterium <to> to <vvi> keep <at> the <nn1> reaction <vvg> going <y> . <ccb> But <cs31> as	y
<dd1> this <vbdz> was <to> to <vvi> keep <ppho2> them <ii> from <vvg> forming <dd> any <nn2> attachments <y> ,	from
<ppge> mine <y> , <ddq> which <vvd> kept <ppio1> me <ii> from <vvg> drifting <rr@_rg> quite <ii21_rl> away <ii22_ii> from	from
<jj> only <nn1> way <to> to <vvi> keep <pph1> it <ii> from <vvg> coming <ii> to <ppy> you <ii_rp@> on	from
<vm> will <vvi> arise <to> to <vvi> keep <jj> national <nn2> ships <vvg> moving <rg> so <rr> efficiently <cst> that	y
<pphs1> she <vvd> needed <to> to <vvi> keep <ppho1> her <ii> from <vvg> committing <at> the <jj_nn1> criminal <nn1> blunder	from
<rrq> How <to> to <to> to <vvi> keep <np1> Bally <ii> from <vvg> running <rl> away <y> ? <mc1> 1	from
<bcl21> so <bcl22> as <to> to <vvi> keep <ppho2> them <ii> from <vvg_nn1@> thawing <y> . <dd1> This <rr> also	from
<"> " <pph1> It <vm> will <vvi> keep <ppy> you <ii> from <vvg_jj@> catching <nn2> vapors <ii> in <at> the	from
<nn1> armful <rr> enough <to> to <vvi> keep <appge> their <nn2> moms <vvg_nn1> running <y> , <cs21> even <cs22> if	y
<at> The <nn2> ads <vm> would <vvi> keep <rr_nn1_jj> right <ii> on <vvg> appearing <y> . <at> The <jj> blind	y
<y> , <cc> and <to> to <vvi> keep <nn1> papa <ii> from <vvg> getting <jj> nervous <cc> and <vvd_vvn> worked	from
<vhd> had <vvn> struggled <to> to <vvi> keep <at> the <nn1> theatre <vvg> going <y> , <vbdz> was <xx> not	y
<vvi> love <ppio1> me <cc> and <vvi> keep <rr_nn1_jj> right <ii> on <vvg> loving <ppio1> me <y> , <ccb> but	y
<nn2> questions <y> , <to> to <vvi> keep <ppho1> him <ii> from <vvg_jj> growing <jj> angry <!> ! <"> "	from
Total Correct	11
3 in between: (total=47)	
<ppy> you <db_rr@> all <vv0> keep <rr_jj@> still <cc> and <vv0_vvd_vvn> quit <vvg> asking <appge> her <nn2> questions <db_rr@> all	n
<nn2> men <to> to <vvi> keep <"@> ' <fu> em <ii> from <vvg> pondering <rg> too <rr> heavily <ii_vvg> concerning	from
<jj> sufficient <to> to <vvi> keep <at> the <nn2> sails <ii> from <vvg> flapping <ii> against <at> the <nn2> masts	from
<rr> slowly <to> to <vvi> keep <appge> my <nn1> voice <ii> from <vvg> shaking <rr@> too <y> . <"> "	from
<y> , <pphs1> he <vvd> kept <at> the <nn2> lawyers <ii> from <vvg> freeing <appge> their <jj> guilty <nn2> clients	from
<nn1> voice <to> to <vvi> keep <appge> his <nn1> attention <y> . <vvg> Speaking <ii> about	n

<np1> God <vbz> is	
<jj> careful <to> to <vvi> keep <appge> his <nn1> fog <nn2> machines <vvg_jj> rolling <y> , <rr> never <vvg> permitting	y
<nn1> money <vm> would <vvi> keep <at> the <nn1> heiress <ii> from <vvg> running <rl> away <rt> again <y> .	from
<nn1> county <vm> could <vvi> keep <jj> green <nn1> space <cs> while <vvg_nn1> steering <nn1> growth <ii> into <nn2> areas	n
<vdz> does <xx> n't <vvi> keep <ppio1> me <jj> awake <nnt2> nights <vvg@> cabling <ppio1> me <ddq> what <pph1> it	n
<nn1> Failure <to> to <vvi> keep <dd2> these <nn2> countries <ii> from <vvg_nn1@> defaulting <vm> could <vvi> topple <dar> more	from
<cs22> if <to> to <vvi> keep <appge> their <nn1> craft <ii> from <vvg_nn1@> capsizing <y> , <cc> and <appge> their	from
<ddq> which <pphs1> he <vvd> kept <jj> great <nn1> pride <ii> in <vvg> exhibiting <ppho2> them <ii> to <appge> his	n
<y> , <vm> will <vvi> keep <pn1> one <ii> from <nn1_jj> // <vvg> throwing <nn2> clothes <ii> on <nn2> chairs	from
<nn1> talent <to> to <vvi> keep <at> the <nn1> country <ii> from <vvgk> going <to> to <vvi> wreck <y> .	from
<cst_dd1> that <to> to <vvi> keep <at1> a <jj> current <ppis1_zz1%> I <vvg_jj> flowing <ii> through <pph1> it <nn1> resistance	y
<cc> and <rr> thus <vv0> keep <at> the <nn2> vibrations <ii> from <vvg> dying <rl> away <y> . <at> The	from
<ppy> you <to> to <vvi> keep <at> the <nn1> devil <ii> from <vvg> slipping <ii> into <ppy> you <cc> and	from
<ppho2> them <to> to <vvi> keep <at> the <nn1> straw <ii> from <vvg> pricking <appge> their <jj> hardened <nn2> hides	from
<y> , <to> to <vvi> keep <at> the <nn1_jj> light <ii> from <vvg@> dazzling <appge> your <nn2> eyes <y> .	from
<rr> frantically <to> to <vvi> keep <ii> in <nn1> step <y> . <vvg> Moving <ii_rp@> in <at1> a <jj> blind	n
<nn1> stuff <to> to <vvi> keep <at> the <nn1> theater <nn1> company <vvg> going <y> . <rrq> When <vd0> do	y
<rr> only <to_ii> to <vvi> keep <ppx1> himself <ii> from <rr> really <vvg> talking <ii> about <ppho1> her <ii> to	from
<jj_rr@> high <to> to <vvi> keep <at> the <jj> nuclear <nn2> reactions <vvg> going <y> , <at> the <nn1> liberation	y
<cc> and <ppis2> we <vv0> keep <ppx2> ourselves <jj> pure <y> , <vvg> bathing <cs> before <ppis2> we <vv0> eat	n
Total Correct	4
4 in between: (total=18)	
<pphs1> he <vvd> kept <appge> his <jj> sad <nn1_jj> secret <y> , <vvg> recalling <cst> that <np1> Freud <vhd> had	n
<to> to <vvi> keep <dd> some <jj> poor <nn1> soul <ii> from <jj_vvg@> starving <y> ; <rt> sometimes <pph1> it	from
<vh0> 've <vvn> kept <jj> close <nn2> tabs <ii> on <pn1> everyone <vvg> leaving <cc> or <vvg> entering <np1> hughes	n
<y> . <vv0> keep <rr> still <z> some <mc1> one <vbz> is <vvg_nn1> knocking <y> , <"> " <pphs1> she	n
<rr> ever <vvi> keep <jj> Strange <nn1> pastime <y> , <rr> ever <vvg> striving <to> to <vvi> unfold <appge> Their	n
<cst> that <vvd> kept <np1> Delilah <ge> 's <nn2> men <ii> from <vvg> falling <rl> overboard <y> . <mc> 10	from
<to> to <vvi> keep <at> the <nn1> water <ii> from <nn1_jj> // <vvg_nn1@> breaking <ii> into <at> the <jj> intermediate	from
<to> to <vvi> keep <jjr> Balder <nn1> company <y> . <ii> After <vvg> standing <at1> a <da2>	n

few <nn2> moments	
<to> to <vvi> keep <rp> up <y> . <nn2> Couples <vbr> are <vvg> kissing <ii> in <at> the <nn2> corners	n
<to> to <vvi> keep <at> the <nn1> sewer <nn1> gas <ii> from <vvg> coming <ii> into <at> the <nn1> room	from
<pphs1> he <vvd> kept <np1> Brett <ge> 's <nn1> hand <ii> from <vvg> touching <at1> a <nn1> length <io> of	from
<xx> not <vvi> keep <rp> on <nn1@_jj> growing <ge> ' <cc> and <jj_vvg_nn1@> growing <y> . <ii21> Except <ii22> for	y
<vbdz> was <vvn> kept <vvn_jj@_vvd> informed <io> of <ddq> what <vbdz> was <vvg> going <rp@> on <y> , <ccb> but	n
<to> to <vvi> keep <appge> her <nn1> brother <vvn> occupied <ii> by <vvg> asking <ppho1> him <nn2> questions <ii> about	n
<!> ! <vv0> keep <ppho1> her <if> for <ppx1> yourself <y> . <vvg> going <ppis1> i <vv0> wish <ppis1> i	n
<to> to <vvi> keep <at> the <jj> ugly <nn2> words <ii> from <vbg> being <vvn> spoken <y> . <rr> once	from
<pphs1> He <vvd> kept <appge> his <nn1> aim <jj_rr@> steady <y> , <vvg> holding <at> the <nn1> firearm <ii> on	n
<to> to <vvi> keep <at> the <jj_nn1> conservative <jj> old <nn1> line <vvg> growing <iw> with <at> the <nnt2> times	y
Total Correct	2

B1-B. Keep

0 in between: (total=224)	**Correct?**
<rr> weakly <y> . <np1> Rita <y> , <pphs1> she <vvd> kept <vvg> telling <ppx1> herself <y> , <ppy> you	y
<y> ? <np1> JAY <rrq> Why <vd0> do <ppy> you <vvi> keep <vvg> asking <if> for <nn1> permission <y> ?	y
<io> of <at> the <jj> unpleasant <nn1> feeling <cst> that <vvd> kept <vvg> pawing <ii> at <appge> his <nn2> insides	y
<nn2> labs <y> , <cc> and <jj> skilled <nn2> engineers <vv0> keep <vvg> converting <ppho2> them <ii> into <jj> practical	y
<vvd> said <np1> Nannie <y> . <ccb> But <np1> Betty <vvd> kept <vvg_nn1> dancing <ii_rp> around <np1> Jack <y> ,	y
<at> the <nn1> moment <nn1_vv0> pass <y> , <pphs1> he <vvd> kept <vvg> trying <cc> and <ddq> trying-What <at> the	y
<nn1> officer <cc> and <md> 2nd <jj> young <nn1> officer <vv0> keep <vvg_jj@> moving <y> . <pphs2> They <vv0> start	y
<rt> then <y> . <"> " <"> " <ppy> You <vv0> keep <vvg> saying <ppy> you <vbr> 're <at1> a	y
<ii> on <at> the <nn1> porch <y> . <np1> Kenny <vvd> kept <vvg> looking <ii> at <ppho1> her <y> .	y
<at> the <nn1> forecastle <nn1> door <y> , <pphs1> he <vvd> kept <vvg> swinging <appge> his <jj_nn1> right <nn1> fist	y
<pn1> Everything <!> ! <cc> And <vd0> do <xx> n't <vvi> keep <vvg> saying <ddq> what <!> ! <"> "	y
<vvg> tickling <rp@_ii> on <rl> there <y> . <ppis1> I <vv0> keep <vvg_nn1> feeling <rr> just <vv0@_cs%_ii> like <ppis1_zz1%> I	y
<io> of <nn1> wire <nn2> sculptures <cst> that <pphs1> he <vvd> kept <vvg> adding <ii> to <if> for <nnt2> decades	y
<at> the <nn1> house <y> , <cs> if <ppis1> I <vv0> keep <vvg> coming <rp> back <y> . <"> "	y
<jj> rhetorical <jj> theatrical <nn2> phrases <y> . <pphs1> She <vvd> kept <vvg> saying <ii> to <ppx1> herself <rp> over	y
<nn1_vv0> chase <y> , <ccb> but <ppis1> I <vm> must <vvi> keep <vvg_jj@_nn1@> moving <y> . <ppy> You <vm> will	y

<rl> together <cc> and <nn1> packing <ge_"@> ' <to> to <vvi> keep <vvg> running <rp> up <cc> and <ii> down	y
<npm1> June <mc> 30. <vv0_jj> // <"> " <ppis1> I <vvd> kept <vvg> going <rl> home <at1> every <nn1> chance	y
<vvi> smile <ii> at <ppho1> him <csa> as <pphs1> he <vvd> kept <vvg> looking <rp> up <ii> at <ppho1> her	y
<vhz> has <to> to <vvi> borrow <nn1> money <to> to <vvi> keep <nn1_vvg> operating <y> . <vv0> PUSH <vbz> is	y
<nn2> things <cs> if <ppy> you <vhd> had <to> to <vvi> keep <vvg> coming <rl_nn1> home <to> to <vvi> blank	y
<xx> not <vvi> cover <y> , <pn1> something <cst> that <vvd> kept <vvg> flaming <rp> up <to> to <vvi> demand	y
<vvi> stay <ii_rp@> on <appge> their <nn2> feet <cc> and <vvi> keep <vvg> exercising <appge> their <nn2> limbs <z> --	y
<xx> not <vvi> look <jj> familiar <y> . <ppis1> I <vvd> kept <vvg> wishing <rg> so <da1> much <cst> that	y
<vm> could <vm> would <vbi> be <jk> able <to> to <vvi> keep <vvg> going <ii21> up <ii22> to <at> the	y
Total Correct	25
1 in between: (total=67)	
<vbdz> was <xx> n't <vvgk> going <to> to <vvi> keep <ppho2> them <vvg> working <rg> very <jj_rr> late <y> .	y
<rr> consequently <uh_at_rr%> no <vv0_nn1@> pull <to> to <vvi> keep <pph1> it <vvg> turning <y> . <rex21> For <rex22> example	y
<vd0> do <xx> n't <vvi> want <to> to <vvi> keep <rp> on <vvg> giving <ppy> you <jj_vv0%> smart <nn2> answers	y
<xx> not <y> , <cs> because <pphs1> he <vvd> kept <rp> on <vvg> swimming <if> for <dar> more <nn2> mines	y
<mc1> one <nn1> end <y> . <pph1> It <vvd> kept <rp> on <vvg_nn1> ringing <cc> and <at1> a <nn1> horn	y
<ppio1> me <y> , <cs> although <ppis1> I <vvd> kept <ppho2> them <vvg> waiting <dd> some <nnt1> time <cs> while	y
<nn1> abode <y> ; <dd2> these <ppis1> I <vvd> kept <rr> wearily <vvg> peopling <iw> with <jj_nn1@> white <y> ,	y
<"> " <mc> three <"> " <rr> still <vv0> keep <rp> on <vvg> praying <if> for <pph1> it <y> .	y
<y> , <cs> if <at> the <np1> U.S. <vvd> kept <rp> on <vvg> bqilding <cc> and <vvg_nn1> operating <nn2> reactors	y
<vdd> did <xx> n't <vvi> want <to> to <vvi> keep <pn1> anybody <vvg> waiting <y> , <nn1> miss <y> .	y
<y> , <rr> just <rr_dd> enough <to> to <vvi> keep <ppio1> me <vvg_nn1> reading <if> for <at> the <md> next	y
<nn1> man <vvi> want <ppio1> me <cc> and <vv0> keep <rp> on <vvg> wanting <ppio1> me <y> . <rr> Well	y
<nn1> sir <y> ; <vd0> do <xx> not <vvi> keep <ppio1> me <vvg> waiting <y> . <"> " <ppis1> I	y
<nn2> stairs <pphs2> they <vm> 'd <rr> just <vvi> keep <ppio1> me <vvg> trotting <ii_rp@> about <vvg> waiting <ii> on	y
<cc> And <y> , <cs> if <ppis1> I <vvd> kept <ppho1_appge> her <vvg_nn1@> waiting <y> , <rr> even <if> for	y
<at1> a <nn1> profit <y> . <nn2> Subsidies <vv0> keep <ppho2> them <vvg> going <y> . <ccb> But <pph1> it	y
<np1> Lydia <y> , <jj> sorry <to> to <vvi> keep <ppy> you <vvg> waiting <y> . <"> " <"> "	y
<rr21_rg> as <rr22> well <y> . <pph1> It <vvd> kept <rp> on <vvg> barking <cs> while <pphs1> he <vvd> cleaned	y
<ppy> you <vvd_vvn> left <-> - <to> to <vvi> keep <nn2> things <vvg> going <y> . <vv0> Know <ddq> what	y

<y> . <"> " <rr> Well <!> ! <vvd> kept <ppy> you <vvg> waiting <y> . <ppis1> I <vdd> did	y
<y> . <"> " <jj> Sorry <to> to <vvi> keep <ppy> you <vvg> waiting <y> , <"> " <np1> MacLoon	y
<nn2> creatures <vv0_nn1> divide <y> . <pphs2> They <vv0> keep <rr> always <vvg> trying <to> to <vvi> make <appge> your	y
<nn1> temperature <y> ; <pphs1> she <vm> could <vvi> keep <nn2> tomatoes <vvg> growing <db> all <nnt1> winter <y> ,	y
<pphs1> she <vhz> 's <vvn> got <to> to <vvi> keep <rp> on <vvg> wearing <jj_nn1> dark <nn2> glasses <y> ,	y
<if> for <at1> a <nnt1@> while <cc> and <vv0> keep <ppho1> him <vvg> guessing <y> . <cs_rrq> When <pphs1> he	y
Total Correct	25
2 in between: (total=31)	
<pph1> It <vm> ca <xx> n't <vvi> keep <at> the <nn2> railroads <vvg_nn1> running <y> , <cc> or <vv0> unsnarl	y
<nn1> solution <vvn_vvd> designed <to> to <vvi> keep <at> the <nn1> program <vvg> going <ii@_cs> until <mc> 1961 <nn1> Congress	y
<rr> just <vm> ca <xx> n't <vvi> keep <at> the <nn1> pressure <vvg> going <jj_rr%> full <nn1_vv0> blast <db> all	y
<vm> will <vvi> arise <to> to <vvi> keep <jj> national <nn2> ships <vvg> moving <rg> so <rr> efficiently <cst> that	y
<vvi> love <ppio1> me <cc> and <vvi> keep <rr_nn1_jj> right <ii> on <vvg> loving <ppio1> me <y> , <ccb> but	y
<jj> other <nn2> women <cc> and <vv0> keep <rr_nn1_jj> right <ii> on <vvg> loving <ppho2> them <db> all <io> of	y
<dd_rr> enough <nn1> deuterium <to> to <vvi> keep <at> the <nn1> reaction <vvg> going <y> . <ccb> But <cs31> as	y
<vm> will <vvi> flow <to> to <vvi> keep <at> the <nn1> motor <vvg> going <y> , <cs@_rr> so <at> the	y
<vhd> had <vvn> struggled <to> to <vvi> keep <at> the <nn1> theatre <vvg> going <y> , <vbdz> was <xx> not	y
<vm> would <vhi> have <to> to <vvi> keep <jj> considerable <nn2> forces <vvg> facing <at> the <nd1> east <cc> and	n
<cst> that <nnb> Mr. <np1> Field <vvd> kept <dd2> these <"> " <vvg> talking <nn2> books <"> " <ii> on	n
<nn1> armful <rr> enough <to> to <vvi> keep <appge> their <nn2> moms <vvg_nn1> running <y> , <cs21> even <cs22> if	y
<y> , <z> -- <"> " <vv0> keep <ppio1> me <rl> here <vvg> talking <y> , <cc> and <rt> then	y
<at> the <nn1> market <cc> and <vv0> keep <at> the <np1> Dow <vvg_jj> rolling <rr> inexorably <rl> forward <y> .	y
<y> ; <cc> and <ppis1> I <vvd> kept <rp@> on <y> , <vvg> laughing <rr> grimly <ii> at <ppx1> myself	y
<nn1> labor <nn2_vvz> costs <cc> and <vvd> kept <at> the <nn1> place <vvg> humming <iw> with <nn2> films <ii21> such	y
<at> The <nn2> ads <vm> would <vvi> keep <rr_nn1_jj> right <ii> on <vvg> appearing <y> . <at> The <jj> blind	y
<ii> by <np1> Adams <to> to <vvi> keep <db> all <nn1> information <ii_vvg> regarding <at> the <nn1> utility <nn2> appraisals	n
<ppis1> I <vvd> wanted <to> to <vvi> keep <at> the <nn1> conversation <vvg> going <y> , <dd> any <nn1> conversation	y
<vvd> sat <y> , <ccb> but <vvd> kept <appge> my <nn2> feet <vvg> going <y> . <ppis1> I <vvd> blew	y
<cs_csw@> if <np1> God <vm> could <vvi> keep <db> all <ppho2> them <vvg> hanging <rl> there <ii> on <pn1> nothing	y
<nn1> modernization <ccb> but <to> to <vvi> keep <at> the <nn2> Soviets <vvg> moving <ii> at <at> the <nn1> negotiation	y

<vbr> are <jj_vvg@> healing <pphs2> they <vv0> keep <nn1_rr21%> kind <io_rr22> of <vvg> tickling <rp@_ii> on <rl> there <y> .	y
<vv0> learn <rrq> how <to> to <vvi> keep <at> the <nn2> thrills <vvg> coming <y> . <vv0> GET <jj> TEMPORARY	y
<appge> our <nn2> servants <to> to <vvi> keep <at> the <nn1> air <vvg> moving <ii> in <at> the <nn1> room	y
Total Correct	22
Lexical: (total=1598)	
<np1@_nn1> dubin <vvd> replaced <at> the <nn1> pane <cc> and <vvd> kept <at> the <nn1> barn <nn2> doors <vvd_vv0_vvn> shut	y
<jj%_vvg_nn1%> regulating <nn1> power <y> , <dd1> that <ddq> which <vvz> keeps <at1> every <nn1> part <io> of <at> the	y
<rr21> at <rr22> first <ppis1> i <vm> could <rr> hardly <vvi> keep <appge> my <nn1> voice <jj_rr@> steady <y> ,	y
<at1> a <nn1> fiddle <y> , <vvn_jj_vvd@> determined <to> to <vvi> keep <rg21> up <rg22> to <at> the <md> last	y
<y> . <pphs1> she <vvd> said <y> , <xx> n't <vvi> keep <ppho1> her <rl> there <rr21> any <rr22> longer	y
<pph1> it <ii> to <np1> tom <y>) <ppy> you <vv0> keep <pph1> it <if> for <ppio1> me <y> ,	y
<cc> and <nn1> petroleum <nn2> derivatives <vvn_vvd> needed <to> to <vvi> keep <appge> its <nn2> industries <jj> alive <ii21> because	y
<cc> and <nnb> mr. <np1> winthrop <y> . <ppis1> i <vvd> kept <jj> silent <at> the <rgr> more <rr> keenly	y
<csn> than <at1> a <m> thousand <nn2> shares <to> to <vvi> keep <appge> their <nn2> hands <rp@_ii> in <y> ;	y
<nn1> case <y> , <rr> however <y> , <cc> and <vv0> keep <ii_rp@> in <nn1> mind <cst> that <ppis2> we	y
<nn1> food <y> , <rr> also <rrq> how <pphs2> they <vvd> kept <nn2> fruits <cc> and <nn2> vegetables <ii> in	y
<cst_dd1> that <vm> must <vbi> be <vvn@> periodicallyflooded <to> to <vvi> keep <pph1> it <rg_rr> so <jj> green <y> .	y
<ii@> to <ppho1> her <ii> in <nnt1> time <to> to <vvi> keep <ppho1> her <ii> from <vvg> opening <pph1> it	y
<ppio2> us <y> ; <vv0> see <rrq> how <pphs1> he <vvz> keeps <rp> off <y> ; <pphs1> he <vm> will	y
<pnqs> who <vvz> drives <at> the <nn1_np1> studebaker <cc> and <vvz> keeps <at> the <nn1> house <y> . <md> last	y
<cc> and <vvi> boil <ppho2> them <y> . <vv0_nn1> // <vv0> keep <at1> an <jj> old <nn1> pan <rr> exclusively	y
<vm> would <rt> then <vbi> be <jj> necessary <to> to <vvi> keep <at> the <nn1> water <io> of <at> the	y
<cc> and <pphs1> he <vhd> had <vvn> agreed <to> to <vvi> keep <np1> claudius <cc> and <np1> margaret <rl> apart	y
<jj> white <nn1> cord <y> , <ddq> which <pphs1> he <vvd> kept <vvn_vvd> scrubbed <rr> decently <jj> clean <y> .	y
<ppx221> each <ppx222> other <y> . <"> " <pphs1> he <vvd> kept <at> the <nn1> smile <vvn> fixed <ii> on	y
<rr> just <nnu> $3 <nn1> change <y> . <"> " <vv0> keep <pph1> it <y> , <"> " <vvd> said	y
<cst> that <ii> by <rr> so <vdg> doing <pphs1> he <vvd> kept <ii21> out <ii22> of <appge> his <da> own	y
<nn1> benefit <y> ? <at> the <nn1> snow <nn1> cover <vvz> keeps <at> the <nn1> ground <jjr> warmer <cc> and	y
<nn1> water <cc> and <nn2> blankets <y> , <cc> and <vvd> kept <at> the <nn1> tv <vvd_vvn> turned <ii> to	y
<vvd_jj_vvn> added <jj_nn1@_np1@> green <y> . <"> " <ppis1> i <vv0> keep <at1> a <nn1> memorandum <nn1> book <ii> on	y

<vv0> want <nn1> vengeance <!> ! <"> " <"> " <vv0> keep <ppx1> yourself <jj> easy <z> -- <ppy> you	y
<y> . <ppis1> i <vbm> 'm <vvgk> going <to> to <vvi> keep <jj_nn1@> quiet <if> for <at1> a <da2> few	y
<null> <p> <np1> montgomery <nn1> county <vvz> wants <to> to <vvi> keep <appge> its <jj> vast <nn1> countryside <nn1> --	y
<nn1> baggage <z> -- <nn1> air <np1> france <vm> could <vvi> keep <pph1> it <if> for <ppho2> them <ii> in	y
<jj> other <nn2> lenders <y> . <nn1> failure <to> to <vvi> keep <dd2> these <nn2> countries <ii> from <vvg_nn1@> defaulting	y
<ppis1> i <vm> 'll <vvi> help <ppy> you <to> to <vvi> keep <pph1> it <y> . <vv0_nn1> exit <np1_nn1@> grace	y
<y> . <y> (<"> " <pph1> it <vm> will <vvi> keep <ppy> you <ii> from <vvg_jj@> catching <nn2> vapors	y
<cc> and <at> the <np1> serbs <vvd> wanted <to> to <vvi> keep <pph1> it <dd1> that <nn1> way <y> .	y
<cc> and <vm> would <vbi> be <vvn> developed <to> to <vvi> keep <nn1> pace <iw> with <at> the <nn1> threat	y
<vvd> sat <ii> in <at> the <nn1> window <cc> and <vvd> kept <appge> her <nn1> eye <ii> on <ppho2> them	y
<io> of <at> the <nn2> natives <y> , <to> to <vvi> keep <at> the <nn1> church <nn1@_jj_rr@> cool <y> .	y
<rr21_at1> a <rr22_jj> little <pn1> something <rr> else <to> to <vvi> keep <ppx1> myself <ii> in <nn1> health <y> ,	y
<if> for <at> the <md> third <nn1> rep <y> , <vv0> keep <appge> your <nn1> torso <rp> up <if> for	y
<vvi> hurdle <at> the <nn1> obstacle <y> . <to> to <vvi> keep <appge> his <nn1> seat <vv0_jj> secure <y> ,	y
<np1> vecchi <y> , <at1> a <nn1> dancer <pnqs> who <vvd> kept <appge> his <nn1> money <ii> in <appge> his	y
<pphs1> he <vbz> 's <jj> tough <cc> and <pphs1> he <vvz> keeps <pn1> everything <jj> low <cc> and <pphs1> he	y
<vvn> improved <rr> wonderfully <rr> lately <y> ; <vd0> do <vvi> keep <ii21> up <ii22> to <dd1> that <nn1> reputation	y
<at1> an <jj@> illustrated <nn1> catalogue <ddq> which <ppis1> i <vvd> kept <ii_rp@> in <appge> my <jj> little <nn1> coin	y
<np1> peter <ge> 's <nn1> face <y> . <"> " <vv0> keep <ii_rp@> on <at> the <nn1> way <ppy> you	y
<ppx121> one <ppx122> another <rp> along <y> , <nn1> fee <vvz> keeps <appge> his <jjt> deepest <nn2> feelings <ii> to	y
<io> of <ppho1> her <cst> that <pphs1> he <vm> must <vvi> keep <jj> intact <cc> and <jj> operative <cs_csw@> if	y
<io> of <nn1> chemistry <vvz> enables <ppio2> us <to> to <vvi> keep <ii> in <nn1> health <cc> and <nn1> productiveness	y
<rg> too <rr> well <cc> and <vvi> try <to> to <vvi> keep <"@> ' <ppio1> me <ii> in <np1> indiana	y
<jj> five-day <nn1> ban <rt> yesterday <cst_dd1> that <vm> will <vvi> keep <at> the <jj@_nn1@> slugging <nn1> star <ii21> out	y
<jj> necessary <if> for <at> the <nn1> nutritionist <to> to <vvi> keep <at> the <nn1_jj> level <ii_rg@> under <jj> constant	y
Total Correct	50

B2. Quit

0 in between: (total=43)	**Correct?**
<ccb> but <ppy> you <vh0> 've <vvn> got <to> to <vvi> quit <vbg> being <jj> ashamed <y> . <"> "	y
<nn1> brother <y> . <"> " <vm21> let <vm22> 's <vvi> quit <vvg> searching <rt> now <y> , <cc> and	y
<vvd> said <y> , <"@_ge> ' <vm_nn1@> will <ppy> you <vvi> quit <vvg> playing <iw> with <dd1> that <nn1> telephone	y
<vv0> pull <at> the <nn1> strongback <nn1> pin <!> ! <vv0> quit <vvg> jockeying <dd1_cst> that <nn1_vv0> davit <nn1> winch	y
<ge> 's <nn1> sake <y> , <np1_vv0> maude <y> . <vv0> quit <vvg> acting <ii> like <at1> a <nn1> baby	y
<y> , <pphs1> he <vvd> encouraged <ppio1> me <to> to <vvi> quit <nn1_jj_vvg> smoking <y> , <ddq> which <ppis1> i	y
<y> , <ppis1> i <vv0> tell <ppy> you <y> . <vv0> quit <vvg_nn1> spying <ii> on <ppio1> me <!> !	y
<ii> to <np1> canada <y> . <nn2> hogs <vm> 'll <vvi> quit <nn1_vvg> eating <nn1> corn <if> for <nn2> corpses	y
<y> . <ppy> you <vh0> 've <vvn> got <to> to <vvi> quit <vvg_jj> trailing <ii> after <ppio2> us <y> ,	y
<rr> well <y> ? <cc> and <ppy> you <vm> can <vvi> quit <vvg> looking <if> for <nn1> trouble <rl> ahead	y
<rr> much <vvn_vvd_jj> obliged <cs_csw@> if <ppy> you <vm> 'd <vvi> quit <vvg> popping <uh> yo <ge> ' <nn1> finger	y
<y> , <pn1> everyone <vm> might <vhi> have <to> to <vvi> quit <vvg_jj_nn1> smoking <rr> temporarily <y> , <ccb> but	y
<nn2> men <ii> in <dd1> this <nn1> country <vm> will <vvi> quit <vvg> making <pph1> it <jj> other <nn2> men	y
<vvn_vvd> looked <cs> like <ppis2> we <rr> never <vm> would <vvi> quit <vvg> going <ii> through <ppho2> them <db> all	y
<y> . <ppis1> i <vbm> 'm <vvgk> gon <to> na <vvi> quit <vbg> being <rg_rr> so <rg_jj> damned <jj> easy-going	y
<y> . <"> " <ppis2> we <vh0> have <to> to <vvi> quit <vvg_nn1_jj> running <ii> on <nn1> labor <ge> 's	y
<rrr> more <y> , <rt> then <ppis2> we <vm> 'll <vvi> quit <vvg> talking <nn1> shop <cc> and <vvi> proceed	y
<xx> n't <np1> godzilla <rr_dd1> either <y> . <np1> foreman <vvd> quit <vvg_nn1> fighting <if> for <mc> 10 <nnt2> years	y
<rr21> at <rr22> first <ppis1> i <vvd> tried <to> to <vvi> quit <vvg_nn1> drinking <ii_rp@> on <appge> my <da> own	y
<y> , <cc> and <np1> ken <vhd> had <to> to <vvi> quit <vvg> sending <np1> molly <nn2> clothes <cs> because	y
<!> ! <ppis1> i <vv0> guess <ppis1> i <vm> 'll <vvi> quit <vvg> tryin <"@_ge> ' <to> to <vvi> shoot	y
<vm22> 's <vvi> talk <ii> about <pph1> it <y> . <vv0> quit <vvg_nn1@> disputing <y> . <"> " <np1> flint	y
<rt> then <ppis1> i <vm> 'd <vhi> have <to> to <vvi> quit <vvg> getting <vvn> married <y> . <cs> if	y
<null> <jj_nn2_nn1> timi <nn1_vv0> wish <nn1> mother <vm> would <vvi> quit <vvg> acting <rr> all <ii> over <at> the	y
<nn1> wife <pnqs> who <vvd> nagged <ppio1> me <to> to <vvi> quit <nn1_jj_vvg> smoking <y> . <y> . <y> .	y
Total Correct	25
1 in between: (total=4)	
<"@> ' <rrq> when <vdd> did <ppy> you <vvi> quit <appge> your <nn1@_vvg_jj@> peddling	n

<ii> in <np1> georgia <y> ?	
<np1> pheelie <np1> enos <y> , <ppy> you <vv0> quit <dd1> dat <nn1@_vvg> making <nn1> fun <y> . <nn1> boy	n
<nn1_vv0> look <y> , <vm21> let <vm22> 's <vvi> quit <dd1> this <nn1@_vvg_jj@> stalling <y> , <ppis1> i <vm> 'll	n
<vhd_vhn@> had <at1> a <nn1> notice <to> to <vvi> quit <y> , <vvg> expiring <rt> to-morrow <y> . <"> "	n
Total Correct	0
2 in between: (total=3)	
<ccb> but <vm> can <xx> not <vvi> quit <pph1> it <iw> without <vvg_nn1@_jj@> remarking <y> , <dd1_cst> that <csa> as	n
<ppis1> i <vhd> had <to> to <vvi> quit <y> . <ii> in <vvg> meeting <dd> any <nn1> crisis <y> ,	n
<np1> edward <vm> can <xx> not <vvi> quit <np1> america <iw> without <vvg> sacrificing <appge> his <nn2> prospects <y> ,	n
Total Correct	0
Lexical: (total=225)	
<np1> cox <vbdz> was <vvn> fired <cc> and <np1> richardson <vvd> quit <y> . <db2> both <nn2> men <vvd> appeared	y
<jj_vv0_nn1> /i000 <cs21> so <cs22> that <pphs1> he <vm> could <vvi> quit <appge> his <nn1> job <cc> and <vvi> design	y
<rr> just <rpk> about <to> to <vvi> call <pph1> it <vvz> quits <rrq_cs> when <at> the <jj> attentive <np1> henry	y
<ccb> but <ppis1> i <vvd> advised <ppho1> him <to> to <vvi> quit <cc> and <vvi> concentrate <ii_rp@> on <appge> his	y
<if> for <ppy> you <rr_cs@> so <pphs2> they <vm> 'll <vvi> quit <y> ? <vdd> did <ppy> you <vvi> make	y
<rr> really <vvn_vvd> got <vvg> going <y> . <pphs1> he <vvd> quit <to> to <vvi> go <ii> to <np1> hollywood	y
<np1_nn1> clancy <vm> will <rr> only <vvi> promise <to> to <vvi> quit <rr> entirely <!> ! <"> " <vvd> interjected	y
<io> of <jj> civil <nn1> cook <vvd> telephoned <to> to <vvi> quit <y> , <cc> or <y> , <csa> as	y
<if> for <pph1> it <y> , <ppis1> i <vm> 'll <vvi> quit <at> the <nn1> memorial <nn1> church <rr_nn1> right	y
<y> , <vvg> panting <rr> furiously <y> . <"> " <vv0> quit <pph1> it <y> , <vv0_nn1> y'fool <!> !	y
<vbdr> were <rr21> in <rr22> vain <y> . <pphs1> she <vvd> quit <ii> at <rg> around <mc> six <ra> o'clock	y
<ppho1> him <ii> by <appge> his <nn1> physician <to> to <vvi> quit <at> the <nn1> chamber <y> , <cs_rrq> where	y
<nn1> operation <y> . <cs> if <at> the <nn1> radar <vvz> quits <y> , <cc> or <at1> a <nn1> radio	y
<y> . <"> " <ppis1> i <vd0> do <xx> n't <vvi> quit <nn2> things <y> , <ccb> but <ppis1> i	y
<rg_rr> so <jj_vvn_vvd> disposed <y> , <pphs2> they <vm> can <vvi> quit <appge> their <nn1> loom <if> for <at> the	y
<at> the <nn1> disease <y> . <vvn> forced <to> to <vvi> quit <appge> her <nn1> job <ii_csa> as <at1> a	y
<vvg> assuming <at> the <nn1> premiership <y> . <np1> castro <vvd> quit <csa_ii@> as <nn1> commander <io> of <at> the	y
<rr@> too <y> . <"> " <"> " <pphs1> he <vvd> quit <y> , <"> " <vvd> yelled <at> the	y
<y> , <pphs1> he <vbdz> was <vvn> compelled <to> to <vvi> quit <at> the <nn1> board <io> of <at> the	y
<md> next <nnt1> morning <cs> before <ppis1> i <vm> would <vvi> quit <y> , <ppy> you <vv0> know <y> ,	y

<pph1> it <y> . <ccb> but <cs> when <pphs1> he <vvd> quit <y> , <pphs1> he <rr> soon <vvd_vvn@> discovered	y
<ii> on <nnb> mister <np1> moore <vd0> do <xx> n't <vvi> quit <rt> now <y> . <y> . <y> .	y
<nn1> sir <y> , <ppis2> we <vm> must <rr> either <vvi> quit <at> the <nn1> country <cc> or <vv0_nn1> attack	y
<mc1> one <nn1> recourse <y> ? <pphs1> he <vm> can <vvi> quit <y> . <ddq> what <vbz> is <vvn> demanded	y
<nn1> part <io> of <appge> their <nn1> intention <to> to <vvi> quit <appge> their <nn2> farms <y> , <vvg> owing	y
<ii> at <at> the <nnt1> time <y> . <pphs1> he <vvd> quit <np1> martin <y> , <ddq> which <vbdz> was	y
<ppho1> him <y> ; <pphs1> he <vm> could <xx> n't <vvi> quit <y> . <rr> silently <pphs1> he <vvd> hung	y
<y> . <uh> bah <!> ! <"> " <"> " <vv0> quit <jj> worrying <y> . <mc1> one <io> of	y
<nn2> guts <y> . <pphs1> he <vdz> does <xx> n't <vvi> quit <y> . <ii> in <np1_nn1> hill <nn1> business	y
<ppis1> i <vbm> am <xx> not <vvgk> going <to> to <vvi> quit <nn1> work <y> . <ppis1> i <vv0> like	y
<nn1> outfit <y> , <at1> a <nn1> beret <y> . <vvi> quit <rt> now <y> , <vv0> stay <rl> home	y
<nn1> house <nn> sales <nn1> manager <cs> before <pphs1> he <vvd> quit <appge> his <nn1> office <npd1> friday <vbdz> was	y
<fu> ai <xx> n't <at1> a <nn1> fight <y> . <vv0> quit <pph1> it <y> , <np1> haney <!> !	y
<iw> with <np1_nn1> park <y> . <"> " <np1> tommy <vvd> quit <at> the <nn1> board <cs_rrq> when <pphs1> he	y
<io> of <rr> wisely <vvd_jj@_vvn@> counselled <np1_nn1_vv0> overbeck <to> to <vvi> quit <at> the <nn2> regions <io> of <jj> creative	y
<y> . <rg> too <jj_rr%> bad <pphs1> she <vm> 'd <vvi> quit <y> . <md> second <nnt1> time <dd1> this	y
<da> own <nn2> hands <y> , <cc> and <vm> can <vvi> quit <at> the <jj> old <nn1> man <cs_rrq> when	y
<at> the <jj> old <nn1> man <nn1_rr_jj> right <to> to <vvi> quit <ppho1> him <rr21> at <rr22> once <y> ,	y
<appge> your <nn1> head <vvn> blown <rp> off <!> ! <vv0> quit <pph1> it <z> -- <z> -- <"> "	y
<jj> grim <nn1> voice <vvg_nn1@> saying <y> : <"> " <vv0> quit <pph1> it <cc> or <ppy> you <vv0> get	y
<nnt2> months <ra> ago <y> , <cs_rrq> when <pphs1> he <vvd> quit <ii21> because <ii22> of <y> , <at> the	y
<vvi> put <nn1> pressure <ii> on <np1> bush <to> to <vvi> quit <y> . <vvz> says <np1> reagan <y> :	y
<y> , <cst_dd1> that <np1> ed <vvd> refused <to> to <vvi> quit <cc> and <cst_dd1> that <np1_nn1> keenan <vvd> came	y
<null> <p> <ppis1> i <vbm> 'm <vvg> trying <to> to <vvi> quit <y> , <ppy> you <vv0> see <y> .	y
<y> , <cst> that <pph1> it <vm> could <xx> not <vvi> quit <rr21> for <rr22> good <iw> without <nn1> action	y
<vm> would <vvi> make <pph1> it <jj> painful <to> to <vvi> quit <dd1> this <nn1> neighborhood <y> . <np1> jo	y
<y> . <pphs1> he <vbdz> was <vvn> forced <to> to <vvi> quit <nn1_vv0> work <ii_csa> as <at1> a <nn1> construction	y
<nnt1> time <pphs1> he <vvd> saw <jj_vv0> fit <to> to <vvi> quit <appge> his <nn1> pulpit <y> (<nnt1> time	y
<rl_nn1> home <y> , <cs_rrq> when <at> the <nn1> baboon <vvz> quits <appge> his <nn1> concealment <y> , <cc> and	y
<"> " <ppy> you <vbdr> were <jj> smart <to> to <vvi> quit <y> , <"> " <pphs1> he <vvd> said	y
Total Correct	50

B3. Have

0 in between: (total=30255)	**Correct?**
<da> same <jj%_vvn> fallen <nn1> tree <y> . <pphs1> he <vhz> has <vvn> forgotten <appge> his <jj> tender <nn1_jj> rear	y
<nn1> footnote <fu> * <at> the <nn1> reader <vm> may <vhi> have <vvn> noticed <cst> that <at> the <md> first	y
<y> , <ppy> you <vv0> know <y> . <pphs2> they <vh0> have <vvn> arranged <db> all <appge> your <nn2> affairs	y
<at> the <jj> blind <jj_rr> little <nn1_vv0> fool <pphs1> she <vhd> had <vbn> been <!> ! <db> all <dd2> those	y
<nn2> officers <vvd_vvn> assured <np1_vv0@> krantz <at> the <nn1> money <vhd> had <vbn> been <vvn> sent <ii> by <appge> his	y
<at> the <da> same <jj> mysterious <nn1> way <cst> that <vhd> had <vvn> worked <iw> with <ppio2> us <y> ,	y
<vbdz> was <at> the <md> last <nn1> thing <pphs1> he <vhd> had <vvn> given <at> the <jj> american <nn1> credit	y
<cc> and <np1> j.c <y> . <nn1_np1@> penney <y> , <vh0> have <vvn> accepted <jj> illegal <jj_nn1> japanese <nn2> kickbacks	y
<y> . <ii> in <nn1> fact <y> , <pph1> it <vhz> has <vbn> been <vvn> calculated <cst_dd1> that <rr> only	y
<io> of <db2> both <nn2> sexes <y> . <ppis1> i <vh0> have <vbn> been <vvn> forced <to> to <vvi> limit	y
<at> the <nn2> horrors <io> of <np1> siberia <vm> could <vhi> have <vbn> been <y> , <ccb> but <pph1> it	y
<ii> in <at> the <np1> u.s. <y> , <nn2> researchers <vh0> have <vvn> shown <cst_dd1> that <jj> mechanical <nn1> tillage	y
<cs_rrq> when <appge> her <nn1> house <vmk> ought <to> to <vhi> have <vbn> been <at_rr%> no <dar_rrr> more <csn> than	y
<ii21> out <ii22> of <at> the <nn1> mess <pphs2> they <vhd> had <vvn> got <ppx2> themselves <ii> into <y> .	y
<y> , <rr> still <jj> unimproved <y> . <pphs1> he <vhd> had <vvn> said <ii> to <nn1_np1> elfride <cst> that	y
<at1> a <jj_nn1@> blank <y> . <rt> now <ppis1> i <vhd> had <vvn> got <rl> here <y> , <ddq> what	y
<np1> sawyer <ge> 's <nn1> experience <iw> with <nn2> thump-wallopers <vhd> had <vbn> been <vvn> limited <y> . <ex> there	y
<nn2> details <ii> about <npd1> wednesday <y> . <pphs1> he <vhd> had <vbn> been <rr@_db> all <vvn> ginned <rp> up	y
<nn2> stresses <ii> on <jj> jewish <nn1> family <nn1> solidarity <vh0> have <vvn> appeared <y> . <jj> other <nn1> group	y
<rp_ii@> up <at> the <nn2> wrappers <at> the <nn2> boys <vhd> had <vvn> tossed <ii> on <at> the <jj> pale	y
<at> the <nn1> city <ge> 's <jj> tight <nn1> budget <vhz> has <vvn> tied <rp_ii@> up <dd> some <io> of	y
<at1> a <da2> few <nnt2> days <rrr> later <pphs1> he <vhd> had <vvn> made <rp_ii@> up <appge> his <nn1> mind	y
<dd> some <nn> means <cc> or <jj> other <to> to <vhi> have <vvn> persuaded <ppho1> her <to> to <vvi> stay	y
<nn1> hotel <y> , <appge> his <jj> new <nn1> secretary <vhz> has <vvn> booked <ppho1> him <ii> into <at1> a	y
<"> " <np1> harold <vvd> interrupted <y> . <pphs1> he <vhd> had <vhn> had <appge> his <nn1@> fill <io> of	y
Total correct	25
1 in between: (total=6132)	
<y> ? <"> " <"> " <ppis1> i <vh0> have <xx> not <vvn> thought <pn1> anything <io> of <nn1> marriage	y

<at> the <jj> real <nn1> tenant <pnqs> who <vhz> has <xx> n't <vvn> returned <ii21> up <ii22> to <dd1> this	y
<dd1> this <md> first <nn1> mirror <ppis1> i <vh0> have <rr> always <vvn> considered <csa_ii@> as <mc1> one <io> of	y
<nn2> orders <y> . <ppis1> i <rr> still <vh0> have <xx> n't <vvn> received <at> the <nn2> guns <cc> or	y
<xx> not <vvi> destroy <y> , <cc> and <vh0> have <dd1> each <vvn> thought <vvn_vvd> expressed <ii> in <nn2> tears	y
<appge> his <nn2> poems <y> . <"> " <vh0> have <ppy> you <vvn> read <pph1> it <y> ? <"> "	y
<vvd> parted <vbdz> was <cst> that <ppis1> i <vhd> had <xx> n't <vvn> asked <ppho1> her <ii> about <np1> tom	y
<io> of <jj%> fallen <np1_nn2@> states <nn1> -- <vhd> had <rt> now <vvn> come <y> . <at> the <nn1_jj@> republican	y
<ii> in <at> the <nnt1> morning <pphs1> she <vhd> had <xx> not <vvn> returned <y> . <dd1> this <nn1> agent	y
<to> to <vvi> ebb <y> . <pphs1> he <vhd> had <xx> n't <vvn> lied <ii> about <dd1> that <y> .	y
<at> the <jjt> worst <nn1_vv0> waitress <pphs1> he <vhd> 'd <rr> ever <vhn> had <y> . <ii_cs> after <mc> two	y
<jj> aloof <jj> major <nn1> general <pnqs> who <vhd> had <rr> never <vvn> held <at1> an <jj> independent <nn1> command	y
<jj> good <cc> and <rg_rr> so <jj_nn1> clear-headed <vhd> had <xx> not <vvn> flared <rp> up <rr21> at <rr22> once	y
<y> . <nn1_np1@> eugenia <ccb> but <ppis1> i <vh0> have <xx> not <vvn> complained <io> of <nn1> injury <y> .	y
<nn1> enjoyment <y> . <cs> before <pphs1> he <vhd> had <rr@_db> half <vvn> finished <appge> his <nn1> pipe <pphs1> he	y
<cc> nor <dd> any <jj> other <nn1> man <vhd> had <rr> ever <vvn> seen <rl> there <y> . <"> "	y
<io> of <nn2> memories <y> . <ddq> what <vhd> had <pphs1> he <vvn> forgotten <y> ? <rt> then <y> ,	y
<at> the <nn2> students <cc> and <nn2> counselors <vh0> have <rr> already <vdn> done <rr31@> a <rr32> good <rr33> deal	y
<at> the <nn1> belief <cst> that <pphs1> he <vhd> had <rr> really <vvn> rendered <appge> his <nn1> country <at1> a	y
<nnb> mr. <np1> winthrop <ge> 's <nn1> displeasure <vhd> had <rr> probably <vbn> been <at1> a <md> second <nnt1> time	y
<nn1> thing <io> of <np1> jesus <y> ? <vh0> have <ppy> you <vbn> been <rl> here <rt@_ii> before <y> ?	y
<nn1> grievance <cst> that <appge> my <nn1> father <vhd> had <xx> not <vbn> been <jj> content <to> to <vvi> take	y
<cc> nor <vv0_nn1> smoke <y> ; <ppis2> we <vh0> 've <rr> never <vvn> spoken <ii> about <pph1> it <ii> to	y
<y> , <ii> by <ppx2> themselves <y> , <vh0> have <xx> not <vvn> stabilized <nn1> employment <y> . <at> the	y
<vv0_nn1> lead <y> , <cc> and <nn1_vv0> tungstate <vh0> have <rr@_db> all <vbn> been <vvn> reported <to> to <vvi> affect	y
Total correct	25
2 in between: (total=1391)	
<cc> nor <appge> her <nn1> government <vhd> had <rg> so <rr> far <vvn> lost <appge> their <nn1> attachment <ii> to	y
<cst> that <at> the <nn1> core <vhz> has <to> to <vbi> be <vvn@_jj> laminated <y> . <y> (<jj> loth	n
<y> , <cs> although <ppis1> i <vhd> had <rr> never <rt> before <vhn> had <rg_rr> so <jj> favorable <at1> an	y
<io> of <nn1> sow <nn1> yard <vh0> have <jj@> feeding <nn2> stalls <vvn_vvd@> incorporated <y> . <at1_zz1> a <jj> recent	n
<ii22> of <nnt1> season <vm> will <vhi> have <to> to <vbi> be <vvn> relied <ii21_rp> on <ii22_to_ii> to <vv0> achieve	n

<nn1> school <y> . <ppis1> i <vh0> have <rr> ever <cs_rr@> since <vvn> entertained <at1> a <nn1> hatred <ii> against	y
<jj_nn1> fifty-foot <nn2> women <y>) <vh0> have <rr21> at <rr22> last <vbn> been <vvn> defeated <cc> and <appge> our	y
<jj> consecutive <nn2> congresses <y> . <vhd> had <nnb> president <np1> roosevelt <vvn> said <mc1> one <nn1> word <ii> in	y
<io> of <y> , <cc> nor <vhz> has <nn1> eternity <rr> yet <vvn> heard <pph1> it <y> ; <rr_cs@> so	y
<at1> a <nn1> group <cst_dd1> that <vhz> has <rg_rr> so <rr> long <vvn> defined <ppx1> itself <ii> through <nn1> resistance	y
<at> the <nn1> pain <pphs1> she <vhd> had <rg_rr> so <rr> cheerfully <vvn> borne <y> . <ddq> what <vm> could	y
<at1> an <nn1> ollieer <pnqs> who <vhd> had <xx> not <rr> yet <vvn> spoken <rt> now <vvd_vvn> made <ppio1> me	y
<nd1> south <y> , <np1> nashville <vhz> has <at1> a <nn1_jj> future <vvn> promised <to_ii> to <ccb> but <da2> few	n
<y> . <dd1> that <nn1> town <vhd> had <nnt2> years <rt@> before <vbn> been <at> the <nn1> centre <io> of	y
<nn2> appliances <y> , <pphs2> they <vhd> had <xx> not <rr> yet <vvn> discovered <at> the <nn> means <io> of	y
<io> of <ddq> which <np1> taine <vhz> has <rg_rr> so <rr> skillfully <vvn> made <at1> a <nn1_jj> mosaic <ii> from	y
<cs> if <pn121> some <pn122> one <vhd> had <rr@_db> all <ii_rp> along <vbn> been <jj> close <ii> beside <ppho1> him	y
<if> for <np1> india <y> . <vhd> had <rg_rr> so <rr> long <vvn> anticipated <at> the <jj@> trying <nn1> scene	y
<rt> then <pphs1> she <vm> would <vhi> have <to> to <vbi> be <vvn> brought <ii> into <appge> his <nn1> circle	n
<dd2> those <jj> fir-built <nn2> frigates <vhd> had <rg_rr> so <rr> often <vvn> redeemed <appge> their <nn1> character <ii> in	y
<nn1_jj_vv0> // <nn2> services <y> . <vhd> had <pph1> it <rr> also <vbn> been <appge> our <nn1> policy <ii> from	y
<cst> that <ppy> you <rr> only <vh0> have <mc> four <nnt2> years <vvn> left <to> to <vvi> live <y> ?	n
<ccb> but <ppy> you <vm> might <vhi> have <appge> your <nn2> boots <vvn> blacked <y> . <ex> there <vbz> is	n
<nn1> effect <io> of <ddq> which <vhd> had <xx> not <rr> yet <vvn> worn <rp> off <y> , <cs> although	y
<ddq> which <nnt1> time <pphs1> he <vhd> had <rg_rr> so <rr> completely <vvn> entwined <ppx1> himself <ii_rp> around <appge> his	y
Total correct	18
3 in between: (total=405)	
<y> , <cc> and <vh0> have <appge> his <nn1> whole <vbg> being <vvn> ennobled <ii> by <at> the <nn1> grandeur	?
<jj> lawless <nn1> fury <vhz> has <at> the <jj@> predestined <nn1> victory <vbn> been <vvg> drawing <rl> near <!> !	y
<rr> somewhat <np1> delia <vhd> had <y> , <rr> unwittingly <y> , <vvn> exercised <at1> an <nn1> influence <ii> over	y
<at> the <nn1> village <vhd> had <ii> by <dd1> this <nnt1> time <vvd_vvn> assembled <to> to <vvi> see <ppio2> us	y
<nn1> existence <ppis1> i <vhd> had <if> for <at1> a <nnt1> time <vvn> forgotten <y> . <at> the <nn1> calf	y
<nn1> hair <pphs1> he <vhd> had <y> , <fo> |p154flat <cc> and <jj_vvn> mashed <y> , <vbdz> was <nn1_jj@> silver	n
<y> , <cc> and <vhd> had <rr_jj> long <cc> and <rr> honorably <vvn_vvd> filled <jj> various <jj> important <cc> and	y
<np1> whitney <rr> still <vhz> has <appge> his <nn2> role-players <rr> still <vv0_vvn@> come	n

<ii> to <ppho1> him <y> ,	
<y> , <pphs1> he <vhd> had <dd> some <nn1> x-ray <nn2> plates <vvn> brought <rp> up <ii> from <at> the	y
<cc> and <np1> barker <vhd> had <dar> more <csn> than <rr> once <vvd_vvn@> confessed <ii> to <at> the <nn1> countess	y
<jj> sure <to> to <vhi> have <appge> their <nn2> morals <rr> well <vvn_vvd> attended <ii_to> to <y> , <cc> and	n
<ppho2> them <y> , <vhd> had <pph1> it <xx> not <rr> unluckily <vvn> happened <cst> that <pphs2> they <vvd> became	y
<y> . <rrq> when <vhd> had <at> the <nn1> family <xx> not <vvn> noticed <pn1> anything <cst> that <jj@_vvd@> affected	y
<cs22> if <ppis2> we <vhd> had <rg> as <da1> much <csa> as <vvn> killed <ppy> you <"> " <y> (	n
<y> , <ddq> which <vhd> had <ccb> but <pn1> one <rr> fully <jj_vvn> licensed <jj_nn1> latin <nn1> teacher <ii> during	n
<io> of <vvg> knowing <vhd> had <ppis1_mc1%> i <xx> not <rr> fortunately <vbn> been <vvn> rescued <ii> by <at1> a	y
<nn1> complaint <!> ! <vhd> had <at> the <nn1> countess <rr> ever <vvn> thought <io> of <pph1> it <y> ?	y
<y> , <pph1> it <vhd> had <dd1> this <nnt1> time <rr> only <vvn> brushed <ppho2> them <y> . <dd1> that	y
<nnt1> evening <y> , <vhz> has <if> for <dd> some <nnt1> time <vbn> been <vvn> conducted <ii> in <np1_nn1> burman	y
<csa> as <np1> peter <vhd> had <rr> gleefully <ii> re <y> ; <vvn_jj> marked <ii> upon <nn1> confrontation <iw> with	y
<nn> people <cc> and <vhz> has <at1> a <jj> large <nn1> plant <vvd_vvn_vv0> spread <rp_ii> over <m> hundreds <io> of	n
<nn1> position <pphs1> he <vhd> had <to> to <vbi> be <rr> forcibly <vvn_vvd> retired <dar> more <csn> than <mc> fifty	y
<y> , <nn2> railways <vh0> have <ii> in <dat> most <nn2> countries <vbn> been <vvn> selected <ii> as <at1> a	y
<vvd> concluded <to> to <vhi> have <appge> my <nn1> question <rr> carefully <vvn> written <ppio1> me <nn1> hand <pph1> it	y
<jj> white <nn1> dwarf <vhz> has <xx> not <rr> yet <rr@_rg> quite <vvd_vvn> reached <at> the <jj_nn1> final <nn1> stage	y
Total correct	18
4 in between: (total=315)	
<pnqs> who <vhz> has <xx> not <at> the <nn1> fear <rr> sufficiently <vvn_vvd_jj> commended <y> . <"> " <ppis2> we	n
<nn2> steamships <vhd> had <jj_nn1> walking-beam <nn2> engines <y> , <rr> still <vvn> seen <ii> on <nn1> ferry <nn2> boats	?
<rr22> least <vhi> have <at> the <nn1> honor <io> of <vbg> being <vvn> consigned <ii> to <at> the <nn2> flames	n
<xx> n't <vhi> have <to> to <y> . <at> the <nn1> vehicle <vvd_vvn> rolled <rp> over <rr_cs@_rg> so <appge> its	n
<ddq> which <vhd> had <rr> yet <rr_vv0%> even <to> to <vbi> be <vvn> addressed <y> . <if_cs%> for <pph1> it	n
<to> to <vhi> have <appge> his <nn1> application <cc> or <nn1> patent <vvn_vvd> declared <ii> in <"> " <nn1> interference	n
<ppis1> i <vhd> had <at1> a <jj> vague <y> , <db_rr@> half <vvd_vvn> formed <nn1> notion <io> of <vvg> seeking	n
<np1> garner <vhd> had <iw> with <at1> an <jj> historic <nn1> gesture <vvd_vv0_vvn> put <at> the <jj> modified <nn1> court	y
<nn2> republicans <vh0> have <at1> a <nn1> right <to> to <vbi> be <vvn@_jj> gratified <y> , <cc> and <pph1> it	n
<y> , <vhd> had <xx> not <appge> my <jj_nn1> present <nn2> circumstances <vbn> been <io> of <da> such <at1> a	y

<ppis1> i <vhd> had <cc> and <at> the <nn2> books <vhd> had <vvn> arrived <cc> and <vbn> been <vvn> paid	n
<nn1> table <vh0> have <dd> any <jj> extra <nn2> plates <cc> or <jj_vvn> prepared <jj_nn1> cold <nn2> dishes <cst_dd1> that	n
<pphs1> he <vhd> had <db> all <appge> his <jj_nn1> convertible <nn2> funds <vvn_vvd@> transferred <y> . <rr_cs%> only <appge> his	n
<ppis1> i <vh0> have <xx> not <y> , <rr> however <y> , <vvn_vvd> offended <ii> against <ppy> you <y> ,	y
<y> . <vhd> had <at1> a <nn1> man <ge> 's <nn1> glove <vbn> been <vvn> dashed <ii> in <appge> his	y
<jjr> longer <vh0> have <dd> any <jj> subatomic <nn1> energy <nn2> sources <vvd_vvn> left <y> . <jj_vvn> deprived <io> of	n
<pnqs> who <vhd> had <at> the <nn2> arrangements <ii> in <nn1> hand <vvd_vvn> believed <cst_dd1> that <da> such <at1> a	n
<pphs1> she <vhd> had <dd1> another <nn1> engagement <y> , <cc> and <vvn> dressed <iw> with <jj> great <nn1> care	n
<rrq> why <vhd> had <pphs1> he <xx> not <rr21> at <rr22> least <vvn> tried <cst_dd1> that <jj> ordinary <rr> experiencebefore	y
<y> . <vhd> had <pn1> everything <y> , <rt> then <y> , <vbn> been <at1> a <jj> gargantuan <nn1> mistake	y
<nn1> control <vh0> have <to> to <vbi> be <rg> very <rr> carefully <vvn_vvd> studied <ii> before <vvg> embarking <ii_rp@> on	n
<ppis1> i <vhd> had <nn2> heaps <io> of <jj> little <nn2> fortifications <vvn_vvd@> scattered <db_rr@> all <ii_rp> around <np1> paris	n
<to> to <vhi> have <nn2> industries <pnqs> who <vbdr> were <xx> not <vvn> entitled <to> to <vhi> have <ppho2> them	n
<"> " <vhz> has <pn1> anyone <ppy> you <vv0> know <rr> ever <vbn> been <rr> ritually <jj_vvn@> mutilated <cs> because	y
<rr> ever <vhd> had <if> for <nn1> poetry <vhz> has <rr> nearly <vvn> forsaken <ppio1> me <y> ; <rr> besides	n
Total correct	7
Lexical: (total=13377)	
<at> the <jj> black <nn1> country <y> . <ppis1> i <vhd> had <at> no <nn2> letters <io> of <nn1> introduction	y
<nn1> culture <ii> in <vvg> molding <at> the <nn1> individual <vhz> has <at1> a <nn1@> bearing <ii> on <nn1> psychology	y
<cc> and <vv0> jump <rp> out <y> : <ppis1> i <vhd> had <to> to <vvi> sit <rp> up <rg> very	y
<...> ... <nn2> rolly-coasters <rg_rr@> too <...> ... <pphs2> they <vh0> have <da> same <nn1> effect <y> . <rt> now	y
<jj> weak <y> , <at> the <jj> immortal <nn1> mind <vhd> had <da2> many <nn2> tumbles <y> ; <ccb> but	y
<jj> very <nn1> brute <y> , <nn1> gentleman <pnqs> who <vhz> has <nn1> charge <io> of <at> the <nn2> dogs	y
<ppis1> i <vd0> do <xx> n't <vvi> think <ppy> you <vhd> had <da1> much <nn1> trouble <iw> with <np1_nn1> thoreau	y
<vdd> did <pphs2> they <vvi> know <cst_dd1> that <np1> kanika <vhd> had <appge> her <nn1> mind <ii> on <jjr> bigger	y
<cc> and <vv0> wear <at> the <nn2> laurels <ppis2> we <vh0> have <np1> thoulouse <ge> 's <nn1> train <y> ,	y
<ii> at <at> the <nn1> lake <y> . <pphs1> she <vhd> had <ii> on <at1> an <jj> uncolored <nn1> linen	y
<rr21> at <rr22> first <cs> because <db> all <nn2> ideas <vh0> have <nn2> bugs <y> , <"> " <vvz> recalls	y
<nn1> sieve <y> . <"> " <ppis1> i <vm> 'll <vhi> have <at> the <nn2> boys <vv0> drive <ppy> you	y
<ddq> what <np1> russell <vvz> seems <rt> sometimes <to> to <vhi> have <ii> in <nn1> mind <y>) <cst> that	y

<iw> with <mc> seven <jj> first-rate <nn2> linebackers <y> , <vh0> have <at1> an <nn1> embarrassment <io> of <nn1> talent	y
<y> : <"> " <at> the <np1_jj> scarlet <np1_nn1@> pimpernel <vhz> has <xx> n't <at1> a <jj> damned <nn1> thing	y
<ppy> you <vbr> 're <vvg> goin <"@_ge> ' <to> to <vhi> have <dd> some <nn2> police <nn1> protection <cs31> as	y
<ppho2> them <nn2> sons <io> of <nn2> chiefs <y> , <vhd> had <nn1_jj> been <vvg> receiving <at1> a <jj> christian	n
<cs> because <da2> many <io> of <at> the <nn2> carriers <vh0> have <nn2> tonnages <io> of <nn2> tracks <jj> available	y
<vbn> been <vvn> born <rt> again <y> , <ppis1> i <vh0> have <nn1> reason <to> to <vvi> think <y> ,	y
<cc> and <ppis2> we <vvd> started <y> . <pphs1> he <vhd> had <at1> a <jj> nice <nn1> face <y> ,	y
<at> the <dar> more <nn1> luck <ppy> you <vm> will <vhi> have <y> . <at> the <nn1> rice <vvz> signifies	y
<y> . <rr> so <rt> again <y> , <ppis2> we <vh0> have <nn1> animal <nn2> mechanics <ii> in <at> the	y
<appge> my <jj@> lost <nn1> wife <y> . <ppis2> we <vhd> had <xx> not <ii_rp@> in <dar_rrr> less <csn> than	y
<ppy> you <vv0> act <cs21> as <cs22> though <ppy> you <vhd> had <at> no <nn1> sense <y> , <rr> thus	y
<nn1> money <y> ? <"> " <"> " <ppis1> i <vh0> have <pn> plenty <y> . <ppis1> i <vbdz> was	y
<y> , <nnb> miss <np1> james <y> , <ppis1> i <vh0> have <dd> some <nn2> things <to> to <vdi> do	y
<io> of <jj> appreciable <nn1> magnitude <cc> and <vm> would <vhi> have <at> no <nn1@_vvg> tuning <nn2> properties <ddqv> whatever	y
<vbz> is <vvn> begun <y> , <nnb> mr. <np1> howells <vhz> has <at1> a <nn1> farce <y> , <nnb> mr.	y
<at1> an <nn1> experiment <y> , <ccb> but <ppis1> i <vhd> had <at_rr%> no <rrr_dar> more <y> , <to> to	y
<rr> therefore <y> , <cst> that <pphs1> he <vm> should <vhi> have <nn1> faith <ii> in <ppx1> himself <y> ,	y
<nn2> scarves <y> , <cst> that <at> the <nn2> police <vhd> had <to> to <vvi> turn <rp> back <rr@> too	y
<ge> 's <y> . <np1> robert <y> . <np1> robert <vhd> had <xx> n't <ppis1> i <rrr> better <np1> fleming	y
<mc> three <nn2> circuits <y> . <dd1> each <nn1> circuit <vhd> had <appge> its <da> own <nn1> knife <nn1> switch	y
<y> , <cc> and <dd1> that <nnt1> afternoon <ppis2> we <vhd> had <at1> a <jj> gorgeous <nn1> sunset <y> ;	y
<nn2> recipes <-> - <ppy> you <vm> wo <xx> n't <vhi> have <to> to <vvi> wait <dd1> another <nnt1> minute	y
<ppho1> him <vbdr> were <nn2> women <ccb> but <pphs1> he <vhd> had <pn> none <io> of <ppho2> them <y> .	y
<nn1> foretaste <io> of <at> the <nn1> life <ppis1> i <vh0> have <at1> a <nn1> right <to> to <vvi> live	y
<at> the <jj> very <nn1> name <io> of <nn1> home <vhz> has <dar> more <csn> than <at1> a <jj> common	y
<vbi> be <rr> strictly <jj> identical <cs> since <pphs2> they <vh0> have <jj> different <jj_nn1> past <nn2> histories <y> .	y
<y> , <vbi> be <y> ? <vvz%> sides <ppis1> i <vhd> had <ii> on <at1> a <jj> heavy <nn1> veil	y
<vvg_nn1@> acting <if> for <ppx1> itself <y> : <dd1> each <vhg> having <at1> a <nn1> power <io> of <jj> absolute	y
<ccb> but <pphs1> she <vvd> understood <cst> that <ppis1> i <vhd> had <to> to <vvi> follow <nn2> orders <y> .	y
<vvg_jj%> looking <nn2> objects <y> , <csa> as <np1> kitty <vhd> had <nn1> cause <to> to <vvi> remember <y> .	y

<nn1> body <ppy> you <vv0> mean <pph1> it <to> to <vhi> have <y> , <jj_nn1_vv0> // <iw> with <at1> a	y
<nn1> capt <y> . <rr_np1@> thieleunfortunately <y> , <ppis1> i <vh0> have <to> to <vbi> be <jj> answerable <xx> not	y
<y> , <vvg> planning <at> the <nn1> surprise <pphs1> he <vhd> had <to> to <vvi> propose <z> -- <cst> that	y
<rt> then <y> ? <"> " <"> " <ppis1> i <vh0> have <at1> an <nn1> education <y> ; <iw> with	y
<mc> eighty-two <y> . <"> " <ccb> but <pphs1> he <vhz> has <at1> a <nn1> caller <iw> with <ppho1> him	y
<y> . <dd1> each <nn1> class <io> of <nn1> service <vhz> has <appge> its <da> own <jj> peculiar <nn2> rates	y
<jj> handsome <y> . <at1> a <nn1> woman <vm> would <vhi> have <to> to <vbi> be <jj> blind <xx> not	y
Total correct	49

B4. Go

0 in between: (total=1085)	**Correct?**
<ii> in <at> the <nn1> world <vbr> are <ppy> you <vvgk> going <to> to <vvi> ask <pn1> somebody <to> to	y
<y> , <cc> and <np1> collins <vbz> is <xx> not <vvgk> going <to> to <vvi> stop <ppho1> him <y> ,	y
<ppho1> him <y> . <cc> and <ppis1> i <vbm> 'm <vvgk> going <to> to <vvi> get <ppho1> him <y> .	y
<vv0_nn1_np1@> freytagwell <y> , <ppis1> i <vbm> 'm <xx> not <vvgk> going <to> to <vvi> take <pph1> it <y> .	y
<"> " <"> " <ddq> what <vbr> are <ppy> you <vvg> going <to> to <vdi> do <iw> with <ppx1> yourself	y
<vvg> going <y> ? <null> <nn1_np1> amparoi <vbm> 'm <vvgk> going <to> to <vvi> wash <rp> up <y> .	y
<y> : <dd1> this <nn1> phone <nn1> call <vbz> is <vvgk> going <to> to <vbi> be <jj> important <y> ,	y
<vvd> asked <ppio1> me <cs_csw@> if <ppis1> i <vbdz> was <vvgk> going <to> to <vvi> surrender <rr@> too <y> .	y
<jj_nn1> right <nn1> thing <cst> that <ppis1> i <vbm> 'm <vvgk> going <to> to <vdi> do <y> ; <vv0> suppose	y
<y> , <cc> and <rt> now <pphs1> he <vbdz> was <vvgk> going <to> to <vvi> pitch <ii> into <ppho1> him	y
<nn1> position <rt> again <y> , <pphs1> he <vbdz> was <vvgk> going <to> to <vvi> put <ppio2> us <rl> somewhere	y
<at1> a <nnu1> cent <y> . <ppis1> i <vbdz> was <vvgk> going <to> to <vvi> charge <ppy> you <nnu> $10,000	y
<rr> well <y> , <vv0> thank <np1> god <ppy> you <vvgk> gon <to> na <vbi> be <mc> eighteen <rr> soon	y
<ii_rp@> on <nn2> hikes <y> , <cc> and <nn2> kids <vvgk> going <to> to <vvi> see <appge> their <nn1> grandmother	y
<rp> off <y> . <pph1> it <fu> ai <xx> n't <vvgk> gon <to> na <vbi> be <rr%_jj@_cs%_ii> like <ppis1> i	y
<nn1@_jj> red <vvd_vvn> told <ppio1> me <pphs1> he <vbz> 's <vvgk> going <to> to <vvi> borrow <at> the <nn1> car	y
<jj> gusty <nn1> wind <y> , <ppis1> i <vbdz> was <vvgk> going <to> to <vvi> risk <vvg_jj> continuing <ii> to	y
<nn1> mouth <cs21> as <cs22> if <pphs1> she <vbdz> was <vvgk> going <to> to <vbi> be <jj> sick <y> .	y
<nn1> migrant <nn1> worker <y> . <pnqs> who <vbz> 's <vvgk> going <to> to <vvi> listen <ii> to <ppio1> me	y

<np1_nn1_jj> bremershaven <y> . <ddq> what <vbm> am <ppis1> i <vvgk> going <to> to <vvi> tell <ppho1> her <y> ?	y
<"> " <at> the <nn1> president <vbz> 's <xx> not <vvgk> going <to> to <vvi> run <y> ? <"> "	y
<"> " <uh> oh <y> , <ppis2> we <vbr> 're <vvgk> going <to> to <vhi> have <at1> a <nn1> storm	y
<y> . <rr> maybe <ppy> you <vbr> 're <rr> never <vvgk> going <to> to <vvi> get <rp> out <y> .	y
<vvd_vvn> passed <rp@> in <ii21> out <ii22> of <vbdr> were <vvgk> going <to> to <vvi> blow <rp> out <at> the	y
<rr32> the <rr33> same <y> , <ppis1> i <vbm> 'm <vvgk> going <to> to <vvi> talk <ii> to <ppho1> him	y
Total Correct	25
1 in between: (total=4)	
<y> , <ppis1> i <vbm> am <rr> only <vvg> going <to> to <"> " <vvi> elope <"> " <rex> viz	y
<jj_nn1> italian <jj@_vvd> motorized <nn1> division <vbdz> was <vvg> going <to> to <"> " <vvi> dash <"> " <at> the	y
<vvi> come <y> , <pphs2> they <vbr> 're <vvg> going <to> to <rr> really <vvi> throw <rl> away <at> the	y
<"> " <ccb> but <ddq> what <ppis2> we <vvgk> gon <to> na <to> to <vvi> fix <at> no <jj_nn1> fancy	?
Total Correct	3
Lexical: (total=158)	
<y> . <ccb> but <cs> if <ppis2> we <vbr> are <vvg> going <ii> about <at> the <nn1> business <io> of	y
<ii> to <appge> her <nn1> place <y> , <cc> and <vvg> going <ii> to <at> the <jj> other <nn1> side	y
<vvd> wrote <y> : <"> " <ppis2> we <vbr> are <vvg> going <rp_ii> over <at1> a <nn1> threshold <ii> in	y
<y> , <iw> with <at> the <nn1> pretext <io> of <vvg> going <ii> to <at1> a <nn1> lavatory <y> ,	y
<y> , <vbdz> was <da2> several <nnt2> days <ii> in <vvg> going <ii> down <at> the <nn1> river <y> ;	y
<y> . <ppis1> i <vd0> do <xx> n't <vvi> mind <vvg> going <ii> in <at> the <nn1> procession <y> ,	y
<cst> that <pphs1> he <vvd> thought <pphs2> they <vbdr> were <vvg> going <ii> about <at> the <nn1_jj> whole <vvg_nn1@> dismantling	y
<nn1> offer <y> . <pphs2> they <vbdr> were <rr> already <vvg> going <rp_ii@> down <at> the <nn1> ladder <ii_cs> before	y
<nn2> bills <ii_cs> after <nn1> marriage <y> , <cc> and <vvg> going <ii> to <at> the <nn2> movies <ii_cs> before	y
<dd1_cst> that <nn1> wher <ge_"@> ' <pphs1> he <vbdz> was <vvg> going <ii> to <at1> a <nn1> body <vm> would	y
<pph1> it <vbz> is <mc1%_ppis1> i <pnqs> who <vbm> am <vvg> going <ii> on <at> the <nn1> errand <y> .	y
<nn1> minister <y> , <cc> and <rt> then <y> , <vvg> going <ii> to <appge> his <nn1> grandfather <y> ,	y
<ii> by <np1> jack <y> . <ii21> instead <ii22> of <vvg> going <ii> in <at> the <nn1> country <rr_jj> early	y
<nn1_np1@> gusman <pnqs> who <vvz> calls <y> ? <y> (<vvg> going <ii> toward <at> the <nn1> cavern <y>)	y
<appge> my <nn1> hat <ii_rp@> on <cc> and <vbdz> was <vvg> going <ii> through <at> the <nn1> door <cs_rrq> when	y
<vbn> been <iw> with <np1> judd <np1> steiner <y> . <vvg> going <ii_rp> over <at> the <nn1> story <at1> a	?
<y> , <csa_ii> as <at> the <nn1> vessel <vbdz> was <vvg> going <ii> at <at1> a <jj> rapid <nn1> rate	y

<nn2> anecdotes <ii> about <np1_nn2> evarts <y> . <ii> on <vvg> going <ii> to <at> the <nn1_jj> polling-place <ii> in	?
<vvd> heard <at> the <jj> friendly <nn1> coal <nn1_vv0> heaver <vvg> going <ii> down <at> the <nn1> corridor <ii> to	y
<cc> and <vvd> made <appge> his <nn1> breakfast <y> . <vvg> going <ii> to <at> the <nn1> edge <io> of	y
<pphs1> she <vvz> feels <at1> a <jj_nn1@> slight <vvi> shudder <vvg> going <ii_rp@> through <appge> his <nn1> body <y> .	y
<rrr> better <at1> every <nnt1> day <y> , <ii> after <vvg> going <ii> to <at1> an <nn1> extra-hole <to> to	y
<dd1_cst@> that <cs_nnt1@> while <at> the <nn1> ship <vbdz> was <vvg> going <ii_rp@> through <appge> its <jj> extensive <nn1> repair	y
<cc> and <cs> while <dd1> this <nn1> thought <vbdz> was <vvg> going <ii_rp@> through <appge> my <nn1> brain <y> ,	y
<nn1> value <to> to <vbi> be <vvn> gained <ii> by <vvg> going <ii> into <at> the <jj> old <nn1> investigation	y
<at> the <nd1> north <jj_np1> western <nn1> station <ii> before <vvg> going <ii> to <at> the <nn1_np1> amphitheater <y> .	y
<vvd> nodded <y> . <"> " <ppis1> i <vbm> 'm <vvg> going <ii_rp@> on <at1> an <nn1> adventure <y> .	y
<dar_rgr> more <nn1_jj@> sound <csn> than <at1> a <nn1> finger <vvg> going <ii> into <at1> a <nn1> glove <y> .	y
<at> the <nn2> fortunes <io> of <nn1> war <vvd> seemed <vvg> going <ii> against <at> the <nn1_np1> union <y> ,	y
<nn1> refusal <to> to <vvi> permit <appge> his <nn1> name <vvg> going <ii> before <at> the <jj> democratic <nn1> convention	y
<pph1> it <rr> just <cs> because <jj> terrible <ii> about <vvg> going <ii> to <at> the <np1> folies <np1_vv0> berg'ore	?
<nnb> captain <np1> dial <cc> and <ppis1> i <vbdr> were <vvg> going <ii> to <at> the <nn1> war <rl> together	y
<nn2> tourists <vvg> visiting <np1> egypt <vv0> outnumber <nn2> egyptians <vvg> going <ii> in <at> the <jj> other <nn1> direction	y
<vvg> rising <y> . <"> " <ppis1> i <vbm> am <vvg> going <ii> to <appge> my <nn1> cabin <to> to	y
<ccb> but <pphs1> he <vvd> talked <ppho1> her <ii> into <vvg> going <ii> to <at1> a <nn1> movie <cst> that	y
<ppis1> i <vh0> have <dd> some <nn2> thoughts <io> of <vvg> going <ii> as <at1> a <nn1> teacher <ii> to	y
<y> . <pphs1> he <vbz> is <fo> thinkin* <io> of <vvg> going <ii> into <at> the <nn1> church <y> ,	y
<vvi> welcome <at> the <nn1_jj> reprobate <y> , <rr> even <vvg> going <ii> to <at> the <jj_nn1> front <nn1> door	y
<jj> western <nn2> allies <ge> ' <nn1> purpose <ii> in <vvg> going <ii> before <at> the <jj> united <nn2> nations	y
<rr> well <y> , <cs@_rr> so <ppy> you <vbr> 're <vvg> going <ii> to <at> the <nn2> dogs <y> .	y
<y> . <at> the <nn1> war <vhd> had <vbn> been <vvg> going <ii_rp@> on <at1> a <jj> long <nnt1> time	n
<nn2> brushes <cc> and <nn2> sponges <vvn_vvd@> attached <y> , <vvg> going <ii_rp> over <at> the <nn1> floor <ii> at	y
<nn2> places <io> of <at> the <nn1> earth <y> , <vvg> going <ii> from <at> the <nn2> wigwams <io> of	y
<jj> poor <nn1> father <cc> and <at> the <nn2> children <vvg> going <ii> on <at> the <nn1> parish <y> .	y
<vvd> sounded <ii> like <at1> a <nn1_jj> circular <vvd_nn1@> saw <vvg> going <ii_rp@> through <at1> a <nn1> pine <nn1> knot	y
<cs> after <ppis1> i <vvd> retired <ppis1> i <vvd> kept <vvg> going <ii> to <at> the <nn2> races <cc> and	y
<y> . <"> " <ppis1> i <vbdz> was <rr> just <vvg> going <ii> to <at> the <nn1> bathroom <y> .	y

<nn1> catcher <vvd_vvn> saved <at> the <nnt1_md> second <ii> from <vvg> going <ii> into <at> the <nn2> stands <y> .	y
<vvn> asked <ppho1> him <rrq> why <pphs1> he <vbdz> was <vvg> going <ii> to <at> the <nn1> city <y> .	y
<nn1> enthusiasm <y>) <fu> ai <xx> n't <ppy> you <vvg> going <ii> to <at> the <nn1> church <jj> social	y
Total Correct	46

B5. Die

Grammatical instances: (total=48)	**Correct?**
<y> , <np1> lilian <z> -- <ppis1> i <vbm> am <vvg_jj> dying <to> to <vvi> see <ppho1> him <y> .	y
<y> , <rr> just <cs_rrq> when <pn1> everybody <vbz_ge> 's <vvg_jj> dying <to> to <vvi> meet <ppho1> him <y> .	y
<vm> can <vvi> imagine <y> , <ppis1> i <vbm> 'm <vvg_jj> dying <to> to <vvi> ask <at1> a <m> thousand	y
<vv0> come <rp> along <!> ! <ppis1> i <vbm> 'm <vvg_jj> dying <to> to <vvi> see <nn1_np1> ma <y> .	y
<y> , <np1> buddy <y> . <ppis1> i <vbm> 'm <vvg_jj> dying <to> to <vhi> have <at1> a <jj> riotous	y
<y> , <ccb> but <cs> although <pphs1> she <vbdz> was <vvg_jj> dying <to> to <vhi> have <ppho1> him <vv0> tell	y
<jj_nn1> descendingmy <nn2> legs <y> . <ppis1> i <vbdz> was <vvg_jj> dying <to> to <vvi> get <ii> to <ppho1> her	y
<at> the <nn1> truth <y> , <ppis1> i <vbm> 'm <vvg_jj> dying <to> to <vvi> go <rp_ii> over <appge> your	y
<rl> away <y> ; <pphs2> they <vbr> are <ii> in <vvg_jj> dying <to> to <vvi> see <at> the <nn1> wedding	y
<at> the <nnt1> time <y> . <ppis1> i <vbdz> was <vvg_jj> dying <to> to <vvi> show <np1> tom <ii> around	y
<nn1> question <ppis1> i <vv0> know <ppy> you <vbr> 're <vvg_jj> dying <to> to <vvi> ask <y> : <vm> will	y
<appge> his <nn2> sins <cs> whilst <np1> christ <vbdz> was <vvg_jj> dying <to> to <vvi> redeem <nn2> sinners <y> .	y
<ccb> but <ppy> you <vm> must <vvi> think <ppho1_appge> her <vvg_jj> dying <to> to <vvi> come <rg> so <rr> soon	y
<nn1> soda <y> ? <ppis1> i <vbm> 'm <rr> just <vvg_jj> dying <to> to <vhi> have <dd> some <y> .	y
<rp> up <cc> and <rt> now <pphs1> she <vbz> 's <vvg_jj> dying <to> to <vvi> meet <ppy> you <y> .	y
<"> " <"> " <y> . <pphs2> they <vbr> 're <vvg_jj> dying <to> to <vbi> be <iw> with <ppy> you	y
<uh> no <y> , <ccb> but <ppy> you <vbr> 're <vvg_jj> dying <to> to <vvi> know <ii> about <at> the	y
<rr> quickly <y> , <cs> for <ppis1> i <vbm> 'm <vvg_jj> dying <to> to <vvi> see <pph1> it <y> .	y
<mc2> 1980s <y> . <"> " <ppis1> i <vbdz> was <vvg_jj> dying <to> to <vvi> know <y> , <"> "	y
<ii> over <np1> new <np1> york <pnqs> who <vbdr> were <vvg_jj> dying <to> to <vvi> find <at1> an <nn1> apartment	y
<nn2> eyes <vv0> brighten <y> . <ppy> you <vbr> 're <vvg_jj> dying <to> to <vvi> show <ppho2> them <ii> to	y
<appge> my <nn2> adventures <y> , <ppis1> i <vbm> am <vvg_jj> dying <to> to <vvi> tell <ppy> you <pn1> something	y
<ppis1> i <vvd> saw <cst> that <pphs1> he <vbdz> was <vvg_jj> dying <to> to <vbi> be <vvn@> asked <to> to	y
<rl> home <y> , <cs> because <pphs1> he <vbdz> was <vvg_jj> dying <to> to <vvi> see <ppho1>	y

her <y> .	
<nn2> saints <y> , <cc> and <ppis1> i <vbdz> was <vvg_jj> dying <to> to <vhi> have <mc1> one <io> of	y
Total Correct	25
Lexical instances: (total=735)	
<at> the <nn> small-pox <y> , <cc> and <vbdz> was <vvg> dying <if> for <at1> a <jj> little <nn1> food	y
<vvi> tell <ppio1> me <cst> that <ppis1> i <vbm> am <vvg_jj> dying <rr_jj> far <ii> over <at> the <nn1> ocean	y
<cs> if <pphs2> they <vv0> think <pphs2> they <vbr> are <vvg_jj> dying <pphs2> they <vm> may <vvi> come <cc> and	y
<vm> would <vbi> be <vhg> having <ppy> you <rl> there <vvg_jj> dying <io> of <nn1> cancer <y> . <"> "	y
<nn1> scene <ii> like <dd1> this <ii> by <z> -- <vvg_jj> dying <if> for <nn1> joy <!> ! <at> the	y
<y> , <ppis1> i <vv0> see <at> the <nn1> sun <vvg> dying <dd1> each <nnt1> night <ii> in <at> the	y
<jj> naked <y> , <cc> and <vvd_jj@> thumbed <cc> and <vvg_jj> dying <y> : <z> -- <rr@_db> all <vvn_vvd%> destined	y
<jj> ill <cst> that <ppis1> i <vvd> thought <ppho1_appge> her <vvg_jj> dying <y> . <pphs1> she <vhd> had <rr> scarcely	y
<rt> again <cc> and <vv0> know <ppis1> i <vbdz> was <vvg_jj> dying <y> . <cc> and <cst_dd1> that <pn1> everything	y
<cs_rrq> when <pphs1> she <vvd> learned <pphs1> she <vbdz> was <vvg_jj> dying <y> , <jj> great <nn1> grandmother <np1> lisa	y
<y> . <dd> some <io> of <ppho2> them <vbdr> were <vvg_jj> dying <y> , <cc> and <rrr> more <vhd> had	y
<y> . <at> the <nn1> problem <vbz> is <cst_dd1> that <vvg_jj> dying <ii_cs> after <nn1> illness <vm> can <vvi> make	y
<da> own <jj_nn1@> private <nn2> lives <ii21> instead <ii22> of <vvg> dying <ii_rp@> on <nn2> barricades <y> , <cs> if	y
<md> last <io> of <appge> his <nn2> dogs <vbz> is <vvg_jj> dying <jj> dying <jj> dying <y> , <rrq> how	y
<nn1> artist <y> , <at> the <jj> young <ccb> but <vvg_jj> dying <np1_nn1@> goupil <y> , <ddqge> whose <nn2> sketches	n
<nn1_jj> patient <y> . <pphs1> she <vbdz> was <xx> not <vvg_jj> dying <y> ; <pphs1> she <vm> might <vvi> linger	y
<ppho2> them <rr> just <csa> as <pphs1> she <vbdz> was <vvg_jj> dying <y> . <np1_nn1> ra <y> . <nn1_np1@> ramona	y
<db> all <appge> his <nn2> soldiers <pnqs> who <vbdr> were <vvg_jj> dying <io> of <at> the <nn1> scurvy <y> ,	y
<nn1> train <y> . <cs> while <np1> murphy <vbdz> was <vvg> dying <appge> his <nn2> comrades <y> , <dd1> each	y
<y> , <rr> perhaps <ii> at <dd1> this <nn1> moment <vvg_jj> dying <y> , <vv0@> aid <ppio1> me <!> !	y
<vvd> answered <y> , <"> " <ppis1> i <vbm> am <vvg_jj> dying <np1> alfred <y> , <cc> and <cs_ii> before	y
<vvg_jj> smiling <y> , <cs> because <pphs1> he <vbz> 's <vvg_jj> dying <y> . <pphs1> he <vvz> thinks <rg> so	y
<vbg> being <vvn> poisoned <y> . <nn2_np1> birds <vbr> are <vvg> dying <ii> from <vvg> eating <jj> toxic <nn> fish	y
<y> , <dd> some <nn1> blood <nn1> relation <vbdz> was <vvg_jj> dying <y> . <at> the <nn2_np1> turcopoles <y> ,	y
<at1> a <nn1> provision <io> of <nn1> nature <ii> against <vvg_jj> dying <iw> with <nn2> things <jj> unexplained <y> ,	y
Total Correct	24

B6. Help

0 in between: (total=65)	**Correct?**
<nnt1> time <y> , <ppis1> i <vm> could <xx> not <vvi> help <vvg> observing <y> . <"> " <ddq> what	y
<y> ; <ccb> but <ppis1> i <vm> could <xx> n't <vvi> help <vvg> telling <ppy> you <y> . <"> "	y
<at> the <nn1> argument <y> , <vm> can <xx> not <vvi> help <vbg> being <vvn> lifted <ii> from <appge> his	y
<nn1> position <y> , <ppis1> i <vm> could <xx> not <vvi> help <vvg> laughing <ii> at <at> the <jj> ludicrous	y
<"> " <"> " <rrq> how <vm> could <pphs2> they <vvi> help <vvg> knowing <cst_dd1> that <y> , <z> for	y
<vbz> is <jj> controversial <cc> and <vm> can <xx> not <vvi> help <vbg> being <rr> so <cs_ii@> until <nn2> men	y
<pn1_mc1> one <ge> 's <nn2> olfactories <vm> could <xx> not <vvi> help <vvg> noticing <y> , <rr> either <y> ,	y
<vvd> looked <y> . <ppis1> i <vm> could <xx> n't <vvi> help <vvg> saying <pn1> something <ii> about <pph1> it	y
<nn1_vvg> blessing <y> . <pphs1> she <vm> can <xx> not <vvi> help <vvg_jj> loving <ddq> what <pphs1> he <vvz> loves	y
<nn1> doubt <y> . <pn1> one <vm> can <xx> not <vvi> help <vvg> asking <ppx1> himself <cs> if <at> the	y
<y> , <cc> and <ppis1> i <vm> could <xx> not <vvi> help <vvg> touching <appge> her <nn1> hand <ii> in	y
<nn1> englishman <y> . <pphs1> he <vm> could <xx> not <vvi> help <vvg> suspecting <at1> a <nn1> pose <y> .	y
<rr> long <jj> secluded <ppis1> i <vm> could <xx> not <vvi> help <vvg> remembering <cst> that <db> all <appge> my	y
<cs_rr@> though <jj_nn1_vv0> sharp-set <y> , <vm> could <xx> not <vvi> help <vvg> perceiving <at1> a <jj> woful <nn1> inspidity	y
<vbr> are <nn2> things <ppy> you <vm> ca <xx> n't <vvi> help <vvg_jj@> knowing <y> . <null> <p> <ii> at	y
<y> , <cs@_rr> so <pn1> one <vm> could <xx> n't <vvi> help <nn1_vvg> hearing <ddq> what <pphs1> he <vvd> said	y
<dd1> that <y> . <ppis1> i <vm> could <xx> not <vvi> help <vvg_nn1> feeling <rr> also <at1> a <jj> positive	y
<nn1> haste <y> , <np1> judd <vm> could <xx> not <vvi> help <vvg> noticing <at1> a <jj> subtle <jj> fleeting	y
<rr> altogether <y> , <ppis1> i <vm> can <xx> not <vvi> help <vvg_nn1> thinking <cst_dd1> that <ii> in <nn1> point	y
<"> " <dd2> those <ii@_rrt_jjt> nearest <vm> could <xx> not <vvi> help <nn1_jj_vvg> listening <y> ; <pph1> it <vbdz> was	y
<rl> away <y> . <ppis1> i <vm> could <xx> n't <vvi> help <vvg> wishing <cst> that <at> the <np1_nn2> ervengs	y
<np1_nn1> tono <z> -- <ppis1> i <vm> ca <xx> n't <vvi> help <vvg> calling <ppy> you <nn1_np1> tono <rr> anyway	y
<ppx122> another <y> , <ppy> you <vm> can <xx> not <vvi> help <vvg_nn1> thinking <ii> about <jj> big <nn2> questions	y
<nn1> path <y> , <np1> sophie <vm> could <xx> not <vvi> help <vvg_nn1> feeling <cst> that <pphs1> she <vhd> had	y
<y> . <"> " <ppis1> i <vm> could <xx> not <vvi> help <vvg_jj> smiling <ii> at <at> the <jj> veiled	y
Total Correct	25
1 in between: (total=4)	
<vvi> enlighten <at> the <nn> people <cc> and <vvi> help <vv0> accelerate <vvg_nn1> thinking <nn1> anc <nn2> events <ii> toward	n
<vm> can <vbi> be <vvn> taken <to> to <vvi> help <jj_nn1> non-english <jj@_nn1@_vvg>	n

speaking <nn2> citizens <vv0> learn <ii_rp@> about	
<y> . <jj@_vvn> organized <nn1> religion <vm> can <vvi> help <ii> by <vvg> giving <nn1> leadership <cc> and <nn1_vv0@> voice	n
<cc> and <at> the <np1> us <vm> can <vvi> help <ii> by <vvg> making <pph1> it <jj_rr@> plain <cst> that	n
Total Correct	0
2 in between: (total=6)	
<appge> his <nn2> symptoms <vm> will <vvi> help <ppio1> me <ii> in <vvg> treating <appge> your <nn1> brother <ge> 's	n
<if> for <ppho1> him <to> to <vvi> help <ppho1> him <y> , <vvg_cs_ii> considering <appge> his <nn1> environment <y> .	n
<bcl21> so <bcl22> as <to> to <vvi> help <rp> out <ii> in <vvg> starting <at> the <jj> new <nn1> work	n
<"> " <ppy> you <vm> can <vvi> help <ppho1> him <ii> by <vvg> telling <ppho1> him <ppy> you <vv0> know	n
<y> . <"> " <vm> may <vvi> help <dd> some <ii> in <vvg> dilating <at> the <nn2> veins <y> .	n
<ppis1> i <vm> could <xx> not <vvi> help <at1> a <jj%> decided <vvg_jj@> surging <ii21> out <ii22> of <nn1> gratitude	n
Total Correct	0
Lexical: (total=648)	
<appge> her <nn1> pocket <pphs1> she <vvd> offered <to> to <vvi> help <ppho1> him <vvi> carry <rp_ii> off <at> the	y
<y> . <ppy> you <vm> 'll <vhi> have <to> to <vvi> help <ppio1> me <vvi> get <ppho1> him <ii> outside	y
<cs31> as <cs32> long <cs33> as <ppis1> i <vm> can <vvi> help <ppho1> him <y> . <"> " <"> "	y
<jj> new <jj> transgenic <nn1> rice <nn1> plant <vm> can <vvi> help <to> to <vvi> cure <nn1> blindness <ii> in	y
<vvg_nn1_jj> living <ii> in <at> the <jj> main <to> to <vvi> help <nn2> others <ii> in <at> the <nn1> way	y
<ppy> you <vv0> describe <y> . <vm_vv0%> can <ppy> you <vvi> help <ppio2> us <y> ? <"> " <nn1_vv0> //	y
<nn1@_vvg_jj@> calling <ii> in <at1> an <nn1> attempt <to> to <vvi> help <vvi> preserve <np1> kenya <ge> 's <nn1> wildlife	y
<vvi> see <rrq> how <at> the <nd1> west <vm> could <vvi> help <iw> with <at> the <jj_nn1> mind-boggling <nn1> task	y
<jj@_rr> just <cc> and <nn1_jj@> kind <nn1> action <vm> will <vvi> help <ppy> you <rrr> farther <ii> on <at> the	y
<csa_rg@> as <jj> non-disabled <nn2> sportsmen <vv0> want <to> to <vvi> help <nn2> hunters <cc> and <nn2> fishermen <iw> with	y
<np1> i. <ppis1> i <vm_vv0%> can <cc> and <vm> will <vvi> help <ppy> you <y> , <cc> and <rr> so	y
<dd1> each <io> of <dd2> these <nn2> trends <vm> will <vvi> help <ppy> you <vvi> get <appge> your <jj> new	y
<nn2> eyes <rr> wide <jj> open <y> , <to> to <vvi> help <ppio1> me <vvi> beg <appge> thy <nn1> forgiveness	y
<cc> and <ppis1> i <vbm> 'm <vvgk> going <to> to <vvi> help <pph1> it <rp> along <y> . <pphs2> they	y
<y> . <"> " <"> " <vv0> need <rr%> any <vvi> help <y> ? <"> " <"> " <uh> no	y
<cst> that <vvg> lengthening <appge> your <nn1> stride <vm> would <vvi> help <ppy> you <vvi> run <rrr_jjr> faster <y> .	y
<y> . <"> " <"> " <ppy> you <vd0> do <vvi> help <ppio1> me <y> , <"> " <pphs1> she	y
<y> . <cc> and <cs_csw@> if <pph1> it <vm> will <vvi> help <vvi> fix <at> the <nn1> fact <ii> in	y
<ppy> you <zz> -- <cc> and <vvd> tried <to> to <vvi> help <ppy> you <y> . <cc> and <vvg> making	y

<nn> people <vm> would <vbi> be <jj> willing <to> to <vvi> help <...> .... <pphs1> he <vbz> 's <vvg> staring	y
<y> . <nn1> nor-den <vvd_vvn> moved <rl> forward <to> to <vvi> help <ppho1> him <y> , <cc> and <dd1> this	y
<y> , <iw> with <jj> only <ppio1> me <to> to <vvi> help <y> , <ppy> you <vbdr> were <rr> already	y
<rp> up <appge> your <nn1> pleasure <rt> tonight <cc> and <vvi> help <ppio1> me <vvi> find <appge> my <nn1> wife	y
<ii> in <at1> a <nn1> way <cst_dd1> that <vm> would <vvi> help <dd> some <to> to <vvi> improve <appge> their	y
<y> . <ccb> but <ppis1> i <vm> could <xx> not <vvi> help <pph1> it <y> . <"> " <"> "	y
<np1> van <np1> buren <cc> nor <nn1> clay <vm> can <vvi> help <at> the <pn1_mc1> one <y> , <cc> nor	y
<y> , <ccb> but <pphs1> he <vm> ca <xx> n't <vvi> help <pph1> it <y> . <pphs1> he <rr> just	y
<y> , <vvn_vvd@> authorized <at> the <nn1> agency <to> to <vvi> help <rr> poorly <vvn_vvd> housed <jj_nn1> low-income <nn2> families	y
<rp> down <rr@_jj> close <csa> as <vm> ca <xx> n't <vvi> help <pph1> it <y> , <"> " <vvd_vvn> concluded	y
<ppy> you <vvd> came <rp> back <rl> here <to> to <vvi> help <ppio1> me <rp> out <y> , <"> "	y
<vm> can <ppis1> i <vdi> do <pn1> anything <to> to <vvi> help <y> ? <vm> shall <ppis1> i <vvi> run	y
<pn1> anything <ppis2> we <vm> can <vdi> do <to> to <vvi> help <appge> your <nn1> thinking <y> . <"@_ge> '	y
<nnt1> time <y> , <vv0> stand <to> to <to> to <vvi> help <ppho2> them <vvi> bring <rp> in <at1> a	y
<jj> basic <nn2> necessities <y> , <vm> would <rr> also <vvi> help <y> . <zz1_np1@> h <rl_cs> ere <vbz> is	y
<nn1> shoulder <y> . <"> " <ppis2> we <vm> can <vvi> help <to> to <vvi> set <ppy> you <rp> up	y
<at1> a <nn1> simile <cst_dd1> that <vm> may <rrr> further <vvi> help <at> the <nn1> realization <io> of <ddq> what	y
<ppho1> him <y> . <pphs1> he <vm> ca <xx> n't <vvi> help <pph1> it <y> . <"> " <"> "	y
<cc> and <vvd_vvn> got <at1> a <nn1> man <to> to <vvi> help <ii> on <at> the <jj> homeward <nn1> trip	y
<vvi> declare <cst> that <pphs1> he <vm> could <xx> n't <vvi> help <pph1> it <y> , <cc> and <cst> that	y
<cst> that <pphs1> she <vm> would <vvi> try <to> to <vvi> help <ppio1> me <vvi> get <at1> an <nn1> audition	y
<appge> your <da> own <y> , <ppy> you <vm> might <vvi> help <ppho1> her <ppis1> i <vv0> object <y> ,	y
<rr> even <rgr> more <rr> lavishly <y> . <to> to <vvi> help <vvi> publicize <at> the <nn1> opening <io> of	y
<nn1> plan <vm> must <vvi> aim <rr> chiefly <to> to <vvi> help <at> the <jj> unemployed <y> . <at> the	y
<vvz> comes <ii21> out <ii22> of <nn1> prison <y> . <vvi> help <to> to <vvi> repay <ppho1> him <if> for	y
<csa_rg@_ii@> as <jj> president- <nn1> elect <y> . <to> to <vvi> help <vvi> strengthen <nd1> south <np1> korea <ge> 's	y
<vbdz> was <pn1> something <pphs1> he <vm> could <xx> not <vvi> help <y> , <ccb> but <at> the <nn1_vv0> viewingit	y
<at1_zz1> a <jj> great <nn1> artist <vm> could <xx> n't <vvi> help <pph1> it <cs> if <pphs1> he <cst> that	y
<nn1> study <y> , <rr> always <vvn> prepared <to> to <vvi> help <ppho2> them <cs> if <pphs1> he <vm> could	y
<rr> else <ppis1> i <vm> can <vdi> do <to> to <vvi> help <ppy> you <cc> and <at> the <nn1> committee	y
<at1> a <nn1> psi <nn1> unit <cst_dd1> that <vm> 'd <vvi> help <ppy> you <rg> quite <rr21> a <rr22> bit	y
Total Correct	50

B7. Use

Grammatical instances: (total=318)	**Correct?**
<appge> my <nn1> father <cc> and <appge> his <nn2> companions <vmk> used <to> to <vvi> go <ii> to <at> the	y
<vvi> change <appge> your <nn1> mind <y> ? <ppy> you <vmk> used <to> to <vvi> condemn <ppio1> me <if> for	y
<"> " <pphs1> he <vvz> explains <y> , <"> " <vmk> used <to> to <vvi> say <pph1> it <vbz> 's	y
<nn1> mother <y> , <rgq> how <rr> often <pphs1> he <vmk> used <to> to <vvi> play <dd1> this <ii> to	y
<rrr_jjr> better <csn> than <at> the <nn1> work <ppis1> i <vmk> used <to> to <vdi> do <y> . <"> "	y
<rr21_at> no <rr22_nn1> doubt <y> . <appge> my <nn1> brother <vmk> used <to> to <vbi> be <jk> able <to> to	y
<ii> by <at1> an <nn1> actress <pnqs> who <ppis2> we <vmk> used <to> to <vbi> be <jj> good <nn2> friends	y
<y> , <cc> and <cs_rrq> where <pph1> it <rr> just <vmk> used <to> to <vbi> be <at1> a <nn1> blur	y
<iw> with <appge> my <nn1> father <cs> because <pphs1> he <vmk> used <to> to <vvi> sing <ppio1> me <dd1> that	y
<ddq> what <nn1> kind <io> of <nnt2> winters <ex> there <vmk> used <to> to <vbi> be <cs_rrq> when <pphs2> they	y
<nn1> name <vbz> is <np1> schnitzel <y> . <pphs1> he <vmk> used <to> to <vvi> work <if> for <at> the	y
<ppis1> i <vhd> had <vvn> raised <y> . <ppis1> i <vmk> used <to> to <vvi> go <cc> and <vvi> return	y
<nn1> contrast <ii> to <at> the <nn1> depression <ddq> which <vmk> used <to> to <vvi> haunt <ppge@_appge> his <z> boarding-house -- vaguely	y
<io> of <ppho2> them <y> . <"> " <ppis1> i <vmk> used <to> to <vvi> believe <at> the <nn1> world	y
<io> of <ddq> what <at> the <jj> southern <nn2> masters <vmk> used <to> to <vdi> do <ii> to <appge> their	y
<at> no <jjr@> further <nn2> questions <y> . <pphs1> he <vmk> used <to> to <vvi> jump <ppio1> me <ii> from	y
<y> . <csa_ii@> as <appge> my <nnb> aunt <np1> sarah <vmk> used <to> to <vvi> say <z> -- <z> --	y
<vvi> sing <at> the <nn1> psalm <cst> that <ppis2> we <vmk> used <to> to <vvi> sing <ii> at <nnt1> night	y
<ppis1> i <md> first <vvd> knew <ppho1> him <pphs1> he <vmk> used <to> to <vvi> wear <at1> a <nn1> balustrade	y
<ppho2> them <rr> regularly <y> . <ppis1> i <rr> often <vmk> used <to> to <vvi> weep <y> , <cs> when	y
<appge> your <nn1> master <cst> that <at> the <nn> means <vmk> used <to> to <vvi> vindicate <at> the <nn1> freedom	y
<y> . <"> " <dd2> these <"> " <nn2> methods <vmk> used <to> to <vvi> assess <at> the <nn2> interests	y
<vvg> collecting <at> the <nn2> coins <y> , <ppis1> i <vmk> used <to> to <vvi> make <jj> careful <nn2> drawings	y
<ii> in <at> the <nn1> world <y> . <ppis1> i <vmk> used <to> to <vvi> suppose <db> all <nn2> sorts	y
<ii_cs> after <nn1_vvg> swimming <y> . <appge> my <nn1> mother <vmk> used <to> to <vvi> say <cst> that <appge> his	y
Total Correct	25
Lexical instances: (total:142)	
<cc> and <vvd> quoted <at> the <nn1> verse <pphs1> he <vvd> used <ii> at <at> the <nnt1> time <to> to	y

<at> the <nn1> subway <nn1> strike <y> . <pphs1> she <vvd> used <"> " <ppis2> we <"> " <rrr> more	y
<ii> from <at> the <nn1> kitchen <nn1> tap <cc> and <vvd> used <pph1> it <to> to <vvi> wash <rp> down	y
<at1> a <nnt1> christmas <nn1> present <y> . <pphs1> she <vvd> used <pph1> it <if> for <at1> a <nn1> couple	y
<pphs1> he <vvd> engaged <np1> fabian <y> , <pphs1> he <vvd> used <at1> an <jj> occasional <nn1> polo <vvg> playing	y
<"> " <uh> yeah <y> . <"> " <pphs1> she <vvd> used <appge_ppho1> her <md> second <nn1> drink <y> .	y
<rt> then <ii> to <at> the <nn2> sportsmen <pnqs> who <vvd> used <ppho2> them <y> , <cc> and <rt> then	y
<vvg> requiring <da> such <nn2> tanks <ii> to <pphs1> he <vvd> used <rl> outdoors <rr> only <y> . <at> the	adj
<at1> a <jj> dramatic <nn1> age <y> . <pphs1> he <vvd> used <at> the <jj> accumulated <nn2> materials <io> of	y
<vdg> doing <rr_rg> so <y> , <ppis1> i <md> first <vvd> used <at> the <nn1> example <io> of <ppx1> myself	y
<at> the <nn1> subject <y> , <cc> and <pphs1> he <vvd> used <at> the <jjt> best <jj> possible <nn1> bait	y
<"> " <ppis1> i <vbm> 'm <jj> afraid <ppis1> i <vvd> used <appge> your <nn1> name <y> . <cc> and	y
<y> , <at> the <nn1> confidence <nn1> man <pnqs> who <vvd> used <at> no <jj> concealed <nn1> weapon <ii21> other	y
<y> . <jj_np1@_nn1> lary <nn1_np1> damron <y> , <pnqs> who <vvd> used <at> the <nn1> alias <"> " <np1_nn1> lary	y
<jj_nn1> gen <jj_nn1> eral <nn1> government <y> , <pphs2> they <vvd> used <appge> their <nn1> power <if> for <vvg> sending	y
<vvi> go <rr@_rg> too <y> , <ccb> but <ppis2> we <vvd> used <np1> bebb <ge> 's <nn1> attendance <ii> at	y
<nn1> life <y> . <dd2> those <nn2> men <pnqs> who <vvd> used <nn1> liquor <ii> to <nn1> condition <at1> a	y
<y> . <np1> josephine <np1> baker <y> , <pnqs> who <vvd> used <nn2> pearls <ii> in <da2> many <io> of	y
<dd1> that <io> of <at1> a <nn1> man <pnqs> who <vvd> used <nn2> drugs <y> . <np1> calvin <nnl1> hill	y
<nn1> part <io> of <ddq> what <np1> lucille <np1> bebb <vvd> used <rrq_cs> when <pphs1> she <vbdz> was <vvg> telling	y
<y> ; <ccb> but <at> the <nn1> tone <pphs1> he <vvd> used <vbdz> was <at1> a <nn1> magnet <to> to	y
<csa> as <pphs1> he <vm> could <y> , <pphs1> he <vvd> used <at> the <nn2> words <io> of <np1> washington	y
<nn1> post <vvd_vvn> reported <cst> that <at> the <nn1> operation <vvd> used <np1> u.s. <nn2> submarines <vvg> operating <ii> in	y
<jj> official <nn1> technicality <y> . <cs> although <pphs1> he <vvd> used <at> the <nn1_vv0> pronoun <"> " <appge> my	y
<pph1> it <vbdz> was <jj> clear <cst> that <pphs1> he <vvd> used <nn1> machinery <ii21> rather <ii22> than <at> the	y
Total Correct	24

B8. Long

Grammatical instances: (total=142)	**Correct?**
<da> own <nn1> project <y> , <cc> and <cs31> as <cs32> long <cs33> as <pphs1> he <vbz> is <rl> there	y
<if> for <nn1> bulk <nn1> traffic <y> . <cs31> as <cs32> long <cs33> as <ex> there <vbr> are	y

<nn2> vessels	
<nn1> run <io> of <at> the <nn1> house <cs31> as <cs32> long <cs33> as <ppy> you <vvd> desired <pph1> it	y
<to> to <vvi> agree <to> to <vvi> come <cs31> as <cs32> long <cs33> as <pn121> no <pn122> one <rr> else	y
<cc> and <nn1> effort <y> . <if_cs%> for <cs31> as <cs32> long <cs33> as <ppis1> i <vm> could <vvi> remember	n
<vm> can <vvi> see <ppy> you <y> , <cs31> as <cs32> long <cs33> as <ppis1> i <vh0> have <at> the	y
<vbdz> was <at1> an <jj> easy <nn1> concept <cs31> as <cs32> long <cs33> as <pph1> it <vvd> took <nn1> place	y
<at> the <jj> evident <nn2> differences <if> for <cs31> as <cs32> long <cs33> as <dd1> this <vvd> lasted <y> ;	n
<ddq> what <pphs1> he <vm> could <vvi> expect <cs31> as <cs32> long <cs33> as <pphs1> he <vvd> lived <ii_rp@> under	y
<vvi> scold <rg@_csa> as <jj_rr> loud <cc> and <cs31> as <cs32> long <cs33> as <dd> any <y> , <ppis1> i	n
<jj> effective <nn1> enforcement <io> of <nn1> prohibition <cs31> as <cs32> long <cs33> as <pph1> it <vbz> is <at> the	y
<"> " <ppis1> i <vm> 'll <vvi> stay <cs31> as <cs32> long <cs33> as <ppy> you <vv0> need <ppio1> me	n
<vvnk> bound <to> to <vvi> find <ppy> you <cs31> as <cs32> long <cs33> as <ppis2> we <vv0> keep <ppy> you	y
<jj> dainty <z> shoe -- it <vbdz> was <rp@> about <cs31> as <cs32> long <cs33> as <at1> a <nn1> baby <ge> 's	n
<ii> in <"> " <jj> deep <nn2> backgrounders <cs31> as <cs32> long <cs33> as <ppis2> we <vvd> agreed <xx> not	y
<to> to <vbi> be <vvn> undervalued <y> . <cs31> as <cs32> long <cs33> as <at> the <nn2> companies <vv0> follow	y
<nn1_jj> fun <ppis2> we <vvd> wanted <y> , <cs31> as <cs32> long <cs33> as <ppis2> we <vvd> behaved <ppx2> ourselves	y
<nn1> home <iw> with <ppio2> us <y> , <cs31> as <cs32> long <cs33> as <vm> would <vvi> suit <appge> his	y
<ii> in <appge> our <nn2> hands <y> . <cs31> as <cs32> long <cs33> as <ppis2> we <vv0> let <at> the	y
<vbdz> was <vvg> carrying <at1> a <nn1> flashlight <cs31> as <cs32> long <cs33> as <at1> a <nn1> nightstickand <vvd_vvn> arrived	n
<vbz> is <vvn> considered <jj_rr@> likely <cst> that <cs31> as <cs32> long <cs33> as <at> the <nn2> soviets <vh0> have	y
<cc> and <jj> other <nn2> supplies <y> ? <cs31> as <cs32> long <cs33> as <np1> germany <vhd> had <nn1_jj@> gold	y
<vbdz> was <jj> complete <y> . <null> <p> <cs31> as <cs32> long <cs33> as <pph1> it <vbdz> was <ii> like	y
<rp@_ii> on <rr> doggedly <y> , <"> " <cs31> as <cs32> long <cs33> as <ppis1> i <vbm> am <rr@> all	y
<if> for <jj> real <nn2> courtesies <y> . <cs31> as <cs32> long <cs33> as <nn> people <vv0> live <rr> simply	y
Total Correct	19
Lexical instances: (total=2848)	
<vvd> cried <ii_rp@> on <dd1> that <nn1> bed <rg_rr> so <rr> long <ra> ago <y> . <rr> only <xx> not	y
<pphs1> he <vhz> has <vbn> been <rp_rg@_ii> around <rg> so <jj_rr> long <cst> that <jjr> younger <nn> people <rt> sometimes	y
<nn1> battle <nn1> cry <null> <p> <ii> in <appge> its <jj> long <nn1> fight <if> for <dar> more <nn1> security	y
<pphs1> she <vvd> wearied <io> of <at> the <z> day -- so <rr_jj> long <y> , <rg_rr> so <jj> hot <y> ,	y

<jj> int <y> . <nn1> shelf <nn1> room <-> - <rr> long <vvd_vvn> shot <np1> tom <vvz> appears <iw> with	y
<appge> her <nn1> grief <cc> and <rr_jj> long <y> , <rr> long <vvg> waiting <dd1_cst> that <jj> only <nn1> death	y
<y> , <pph1> it <vm> would <vbi> be <at1> a <jj> long <nnt1> night <y> . <at> the <nn1> phone	y
<rp> out <appge> my <nn2> ongoings <if> for <at1> a <jj> long <nn1@> stretch <io> of <at> the <nn1_jj> future	y
<nn1> beauty <y> . <at> the <nnt1> time <vvd> seemed <jj> long <to> to <np1> claudius <y> , <cs> though	y
<y> . <pphs1> he <vvd> felt <y> , <ii> before <rr> long <y> , <at> the <nn1> need <io> of	y
<jj> unjealous <ii> among <appge> his <nn2> confreres <vhd> had <rr> long <ra> ago <vvn> decided <nn1_jj> -- <ddq> which	y
<"> " <dd2> those <vbr> are <xx> n't <rr> really <jj> long <y> , <"> " <pphs1> he <vvd> said	y
<y> , <np1_jj> sleepless <ii> through <db> all <at> the <jj> long <npm1> december <nnt1> night <y> , <vvz> hears	y
<nn1> face <y> . <pph1> it <vvz> lasts <at1> a <jj> long <nnt1> time <y> , <cc> and <rt> then	y
<vhd> had <vvn> rehabilitated <ppx1> himself <ii> during <appge_ppge@> his <jj> long <nn1@_vv0> stay <ii> in <nn1> prison <y> .	y
<nn1> sound <pphs1> he <vvd> remembered <ii> from <at1> a <jj> long <nnt1> time <ra> ago <y> . <null> <p>	y
<vm> would <vhi> have <vbn> been <vvn> passed <y> , <rr> long <ra> ago <y> , <cs> if <at1> a	y
<vvn_jj_vvd> revived <io> of <rr> late <y> , <cc> and <rr> long <dar_rrr> more <csn> than <rr> ever <to> to	y
<ii> upon <at> the <nn2> stairs <y> . <ii> before <rr> long <pphs1> he <vvd> heard <at1> an <jj> ominous	y
<rrr> more <pphs1> she <vvz> boasts <y> . <xx> not <rr> long <ra> ago <y> , <ex> there <vbdz> was	y
<y> , <jj> motionless <y> , <vvg> waiting <iw> with <jj_rr> long <y> , <jj> dry <z> jaws -- for <ddq> what	y
<at1> a <jj> common <nn1> impulse <ii> to <at> the <jj> long <jj> neglected <nn1> field <io> of <jj> private	y
<vvn> enacted <y> . <ppis1> i <vbdz> was <xx> not <jj> long <ii> in <nn1> doubt <z> -- <ii> from	y
<ii> by <vvg_nn1@> asking <y> , <fu> * <rgq@> how <rr> long <at1> a <nnt1> time <vm> will <pph1> it	y
<at> the <np1> appleton <md> fifth <nn1> reader <!> ! <rr> long <vm> may <pphs1> she <vvi> read <!> !	y
Total Correct	25

B9. Well

Grammatical instances: (total=331)	**Correct?**
<ppis1> i <vbdz> was <rr> naturally <jj> courageous <ii31> as <ii32> well <ii33> as <jj> unconventional <y> , <cc> and	y
<nn1@> chewing <io> of <at> the <nn1> feed <ii31> as <ii32> well <ii33> as <jj> occasional <jj> biting <cc> and	y
<at> the <np1> united <np1> states <y> , <ii31> as <ii32> well <ii33> as <at> the <nn1> person <vvd> intrusted	y
<nn1> room <rr> enough <if> for <ppx1> myself <ii31> as <ii32> well <ii33> as <nnb> mr. <np1> c. <ii> at	y
<jj> strong <nn1> pressure <y> , <jj> governmental <ii31> as <ii32> well <ii33> as <jj> financial <y> , <vm> will	y
<mc1> one <io> of <ppho2> them <y> , <ii31> as <ii32> well <ii33> as <at1> a <nn> pair <io> of	y
<vbdz> was <xx> not <vvn> invited <y> , <ii31> as <ii32> well <ii33> as <at> the <nn2>	y

particulars <io> of	
<ii> on <dd1> this <nn1> topic <y> , <ii31> as <ii32> well <ii33> as <da2> many <nn2> others <y> .	y
<vbr> are <rr> easily <jj_vvn@_vvd@> distinguished <y> , <ii31> as <ii32> well <ii33> as <appge> his <nn2> moons <y> ,	y
<y> , <cc> and <np1_nn1> bedminster <y> , <ii31> as <ii32> well <ii33> as <np1> hilltown <y> , <ii_rl%> between	y
<nn1> feed <cc> and <vv0> clothe <y> , <ii31> as <ii32> well <ii33> as <nn1> shelter <y> , <appge> your	y
<y> , <cc> or <nn2> eggs <y> , <ii31> as <ii32> well <ii33> as <at1> a <jj> starchy <nn1> vegetable	y
<iw> with <at> the <np1> united <np1> states <ii31> as <ii32> well <ii33> as <np1> england <y>) <vh0> have	y
<nn1> weather <nn1_jj_vv0> // <nn2> conditions <y> , <ii31> as <ii32> well <ii33> as <at> the <nn1> provision <cc> and	y
<if> for <dd1> this <nn1> opinion <y> , <ii31> as <ii32> well <ii33> as <at> the <nn1> evidence <io> of	y
<dar_rrr> more <ii> at <nn1> hand <y> , <ii31> as <ii32> well <ii33> as <vvg> giving <at> the <nn1> commander-in-chief	y
<rr21> at <rr22> once <jj> inevitable <jj> political <ii31> as <ii32> well <ii33> as <jj> physical <nn1> death <ii> in	y
<at> the <jj_nn1> official <nn1> world <rl> here <ii31> as <ii32> well <ii33> as <at> the <nn2> critics <io> of	y
<nn1> principle <io> of <nn1> selection <y> , <ii31> as <ii32> well <ii33> as <at1> a <jj> cheerful <nn1> compliance	y
<nn1> understanding <io> of <jj_nn1> past <nn1> history <ii31> as <ii32> well <ii33> as <at> the <nn1> solution <io> of	y
<nn1> pride <cc> and <nn1> resentment <y> , <ii31> as <ii32> well <ii33> as <at> the <nn1> hope <io> of	y
<appge> its <da> own <jj> peculiar <nn1> beauty <ii31> as <ii32> well <ii33> as <nn1> usefulness <y> . <appge> their	y
<io> of <at> the <nn1> amino <nn1> acid <ii31> as <ii32> well <ii33> as <nn2> diets <io> of <jj> different	y
<at> the <jj> olden <nnt1> time <y> , <ii31> as <ii32> well <ii33> as <at> the <nn1> engineering <nn2> establishments	y
<jj> unknown <ii> to <ppho2> them <y> , <ii31> as <ii32> well <ii33> as <nn1> disease <y> . <at> the	y
Total Correct	25
Lexical instances: (total=3061)	
<nn1> snow <nn1> storm <y> ; <ccb> but <rg_rr> so <rr> well <vvn> prepared <vbdr> were <pphs2> they <cst> that	y
<z> -- <nn1> fare <z> -- <vv0_nn1> fare <ppy> ye <rr> well <!> ! <rrq> how <dd1> this <y> ,	y
<null> <p> <to> to <vvi> make <ppho2> them <rg> very <rr> well <y> . <null> <p> <vbr> are <pphs2> they	y
<to> to <vbi> be <vvn> expected <y> . <"> " <rr> well <y> , <pphs1> he <vbz> is <y> ,	y
<ii> by <nn1> name <y> , <vvd> knew <rr> perfectly <rr> well <pnqs> who <fo> shep137was <y> , <ccb> but	y
<y> , <cs> though <ppis1> i <vv0> know <rg> very <rr> well <dd1_cst> that <y> , <rg> so <rr> far	y
<cc> and <nn2> anecdotes <ddq> which <vvd> served <rg> very <rr> well <to> to <vvi> pass <at> the <nnt1> time	y
<vvg> beginning <rr> already <y> . <"> " <"> " <rr> well <y> , <jj> good <nn1> luck <ii> to	y
<at> the <nn1> language <io> of <ppge> yours <rg> pretty <rr> well <ii> on <dd2> those <nn2> occasions <y> ;	y
<y> . <rt> now <vvd> came <at1> a <nnt1> season <rr> well <vvn> known <ii> to <da2> many <at1> a	y

<y> . <"> " <ppis1> i <vv0> know <rg> damned <rr> well <pph1> it <vbz> is <xx> n't <appge> your	y
<y> ; <vvd> said <appge> his <nn2> relations <vvd_vvn> meant <rr> well <y> , <rr> perfectly <rr> well <y> ,	y
<dd1> this <nnt1> day <iw> without <nn1> discovery <y> . <rr> well <y> , <at_rr%> no <rrr> sooner <vd0> do	y
<cc> and <pphs1> he <vvd> said <pph1> it <vbdz> was <jj> well <y> , <cc> and <cst> that <appge> my	y
<vbn> been <rg> quite <jj> high <cc> and <vvg> holding <rr> well <rr> heretofore <y> , <jj> other <nn2> sections	y
<vv0> go <y> , <cc> and <rr> just <vv0> get <rr> well <cc> and <jj> strong <rt> again <y> ;	y
<vv0> know <pphs1> he <vbz> 's <jj> rich <y> . <rr> well <y> , <pphs1> he <vdz> does <at1> a	y
<ppho2> them <if> for <vvg_jj@> following <nn2> orders <rg_rr> so <rr> well <y> . <null> <p> <at> the <md> third	y
<y> ; <ccb> but <y> , <cs_rr@> though <rg_rr> so <rr> well <vvn> incorporated <y> , <at_uh> no <"> "	y
<jj> educational <cc> and <nn1> adjustment <nn2> problems <vbz> is <jj> well <cst_dd1> that <jj> responsible <nn2> adults <vbr> are	y
<np1> lindsey <ii> at <at> the <nn1> bar <y> . <rr> well <y> , <to> to <vvi> tell <at> the	y
<ppho2> them <db2> both <y> . <pphs2> they <vvd> knew <rr> well <ddq> what <pphs2> they <vm> must <vdi> do	y
<nn> people <y> , <at> the <nn> pair <vm> might <rr> well <vvi> reflect <y> , <vdd> did <xx> not	y
<jj> fervent <y> . <appge> his <nn2> illustrations <vbr> are <rr> well <vvn> chosen <y> , <cc> and <ddq> what	y
<y> . <at> the <jj> friendly <nn1> creature <vvd_vvn> meant <rr> well <y> , <ccb> but <csa_ii> as <at> the	y
Total Correct	25

B10. Soon

Grammatical instances: (total=281)	**Correct?**
<cc> and <vvd> took <nn1> leave <y> , <cs31> as <cs32> soon <cs33> as <pphs1> he <vm> could <rr> decently	y
<ii> to <nn1> brownie <rr> immediately <y> . <cs31> as <cs32> soon <cs33> as <pphs1> he <vvd> answered <y> ,	y
<rr> firmly <vvn_vvd> intended <to> to <vdi> do <cs31> as <cs32> soon <cs33> as <at> the <nnt1> hour <nn1> hand	y
<dd1> that <at> the <nn1> gospel <y> , <cs31> as <cs32> soon <cs33> as <ppis1> i <vm> could <vvi> speak	y
<ppio2> us <to> to <vvi> issue <y> , <cs31> as <cs32> soon <cs33> as <nn2> preparations <vm> could <rr> possibly	y
<y> , <nnb> mr. <np1> gordon <y> ? <cs31> as <cs32> soon <cs33> as <pphs1> he <vvd> said <np1> johnny	y
<nn1> fire <cc> and <vvd_vvn> burned <rp> up <cs31> as <cs32> soon <cs33> as <pph1> it <vbdz> was <jj> dry	y
<rp@> on <y> ; <cc> and <rr> just <cs31> as <cs32> soon <cs33> as <pph1> it <vvd> began <to> to	n
<mc> 3 <np1> dall. <mc> 91 <y> . <cs31> as <cs32> soon <cs33> as <at> the <nn2> provinces <vvd> took	y
<np1> beacon <nnl1> hill <nn1_np1@> bank <y> . <cs31> as <cs32> soon <cs33> as <pphs1> he <vbdz> was <rl> there	y
<ppis1> i <vm> will <vdi> do <dd1> that <cs31> as <cs32> soon <cs33> as <ppis1> i <vbm> am <jj> sure	y
<vhi> have <at> no <nn1> pause <y> . <cs31> as <cs32> soon <cs33> as <at1> a <nn1> president	y

<vbdz> was	
<ppis1> i <vm> 'll <vvi> join <ppho1> him <cs31> as <cs32> soon <cs33> as <ppis1> i <vh0> 've <vhn> had	y
<"> " <pphs1> she <vvd> said <y> , <cs31> as <cs32> soon <cs33> as <pphs1> she <vvd> saw <ppho1> him	y
<y> , <pph1> it <vbdz> was <jj> evident <cs31> as <cs32> soon <cs33> as <appge> her <nn1> machinery <vvd> gave	y
<nn1_jj> five-count <y> . <cc> and <vv0_nn1> stop <cs31> as <cs32> soon <cs33> as <appge> your <nn2> hands <vv0> reach	y
<ge> 's <nn1> chest <y> , <ccb> but <cs31> as <cs32> soon <cs33> as <pphs1> he <vvd> felt <pph1> it	y
<nn1> mechanism <ii> by <rt> then <y> . <cs31> as <cs32> soon <cs33> as <appge> my <nn2> thoughts <vvd> gave	y
<y> . <rr21> of <rr22> course <y> , <cs31> as <cs32> soon <cs33> as <ppis1> i <vvd> heard <cst_dd1> that	y
<rp> up <iw> with <ppho1> her <y> . <cs31> as <cs32> soon <cs33> as <pphs1> she <vvd> hoisted <at> the	y
<rrq> how <nn2> kids <vbr> are <y> . <cs31> as <cs32> soon <cs33> as <nn1> masturbation <vbz> is <vvn> brought	y
<jj> sure <to> to <vvi> tell <ppho1> him <cs31> as <cs32> soon <cs33> as <pphs1> she <vvd> knew <y> ,	y
<nn1> brother <vm> can <vvi> tell <ppio2> us <cs31> as <cs32> soon <cs33> as <pphs1> he <vvz> comes <rp> out	y
<at> the <nn1> contraction <vm> must <vvi> stop <cs31> as <cs32> soon <cs33> as <at> the <nn2> distances <ii> between	y
<rr_jj@> just <mc1_pn1> one <y> . <ccb> but <cs31> as <cs32> soon <cs33> as <pphs1> he <vvd> died <y> ,	y
Total Correct	24
Lexical instances: (total=1067)	
<y> , <cc> and <vvi> leave <pph1> it <rg> as <rr> soon <csa> as <jj> possible <y> . <vv0> bring	y
<y> ? <at1> an <nn1> assumption <ddq> which <vbdz> was <rr> soon <vvn_vvd> proved <rr> empirically <ii> by <nn1> observation	y
<nn1> expense <y> , <if> for <at> the <nn1> glass <rr> soon <vvd> went <ii> to <nn2> bits <y> ,	y
<"> " <pphs1> she <vm> 'll <vbi> be <rl> here <rr> soon <y> . <rr> so <ddq> what <vbz> is	y
<nn1> earth <ge> 's <nn2> lips <y> , <cc> and <rr> soon <ex> there <vm> will <vbi> be <pn> none	y
<ii_cs> after <at> the <nn1> battle <ppis1> i <vm> will <rr> soon <vvi> return <z> -- <jj> victorious <z> --	y
<y> , <cs> because <at1> a <jj> polished <nn1> table <rr> soon <vvz> becomes <jj@_vvn@> marred <y> . <at1> a	y
<pphs2> they <vhd> had <rrr> better <vvi> eat <rg> as <rr> soon <csa> as <jj> possible <cs> because <pphs1> he	y
<if_cs%> for <xx> not <vvg> surrendering <ii> to <np1> washington <rr> soon <rr_dd> enough <y> . <appge> her <nn1> composition	y
<vbi> be <at1> a <jj> social <nn1> success <y> . <rr> soon <pphs1> he <vvd> published <at1> a <nn1> book	y
<nn1> sumfin <ge> ' <to> to <vvi> eat <rg> pretty <rr> soon <y> , <cc> or <np1> ise <ge> '	y
<to> to <vvi> accompany <at> the <nn1> party <vbdz> was <rr> soon <vvn> bruited <ii> throughout <at> the <nn1> village	y
<nn1> bureaucracy <y> . <dd1> that <nn1> war <vvd> began <rr> soon <ii_cs> after <at> the <jj> 35-year-old <np1_nn1@> hastie	y
<xx> n't <vbn> been <vvgk> going <to> to <vvi> quit <rr> soon <y> , <ppis1> i <vv0> think <pphs1> he	y
<dd2> these <nn2> changes <ii> into <nn1> effect <rg> as <rr> soon <csa> as <jj> possible <y> , <np1> ross	y

<nn1> law <y> , <pphs1> he <vvd> commenced <y> , <rr> soon <ii> after <vvg> leaving <nn1> college <y> ,	y
<y> , <ccb> but <rg_rr> so <jj> strong <pph1> it <rr> soon <vvd> gave <ppio1> me <at1> a <nn1> headache	y
<appge> my <nn2> goods <y> , <ppis1> i <vbdz> was <rr> soon <vv0_nn1> fotch <vm_vhd> 'd <rp> up <ii> in	y
<at> the <nn1> mining <nn2> machineries <nn2> cases <y> , <rr> soon <to> to <vbi> be <vvn> heard <y> ,	y
<at> the <nn1> draught <y> , <pph1> it <vm> will <rr> soon <vbi> be <vvn> gained <y> . <nn1> music	y
<cs> for <ppis1> i <vbm> 'm <vvg> coming <rl> home <rr> soon <cc> and <ppis1> i <vm> 'll <vvi> write	y
<cc> and <ppis1> i <vv0> believe <pphs1> she <vm> will <rr> soon <vbi> be <jjr> better <y> . <at> the	y
<vbi> be <ii> in <at> the <nn1> city <rg> very <rr> soon <y> , <"> " <cc> and <rt> then	y
<dar> more <csn> than <mc> three <nnt2> years <y> . <rr> soon <ii_cs> after <appge> his <nn1> resignation <y> ,	y
<y> ? <rt> to-morrow <vm> will <vbi> be <rg> too <rr> soon <y> , <ppis1> i <vbm> am <jj> afraid	y
Total Correct	25

B11. Regard

Grammatical instances: (total=153)	**Correct?**
<nn2> administrations <if> for <mc> eight <nnt2> years <ii31> in <ii32> regard <ii33> to <ppho2> them <vhz> has <nnu> been.inevitable	y
<rr> apparently <vvn> received <dd> some <nn1> instruction <ii31> in <ii32> regard <ii33> to <ppio1> me <y> , <vvd> led	y
<io> of <np1> scandinavia <ge> 's <nn1> position <ii31> in <ii32> regard <ii33> to <db2> both <nn2> wars <vbz> is	y
<nn1> course <to> to <vbi> be <vvn> pursued <ii31> in <ii32> regard <ii33> to <pph1> it <y> . <"> "	y
<vbn> been <vvn> attributed <appge> his <nn1> conduct <ii31> in <ii32> regard <ii33> to <at> the <nn1> politics <io> of	y
<nn1> premise <y> . <appge> his <nn1> action <ii31> in <ii32> regard <ii33> to <np1> mexico <y> ; <nnb> mr.	y
<at> the <nn2> laws <cc> and <nn1> proclamation <ii31> in <ii32> regard <ii33> to <nn1> slavery <y> ? <dd2> those	y
<vvz> gives <at> no <nn1> order <y> . <ii31> in <ii32> regard <ii33> to <appge> your <jj> sacred <nn2> books	y
<vvz> touches <at1> an <jj> important <nn1> principle <ii31> in <ii32> regard <ii33> to <at> the <nn1> stimulation <io> of	y
<vhd> had <vvn> issued <dd> some <nn1> order <ii31> in <ii32> regard <ii33> to <at> the <nn1> observance <io> of	y
<nnb> president <np1> roosevelt <ge> 's <nn1> suggestion <ii31> in <ii32> regard <ii33> to <nn1> prohibition <io> of <jj> interstate	y
<da> such <vbr> are <at> the <nn2> facts <ii31> in <ii32> regard <ii33> to <at> the <nn1> withdrawal <io> of	y
<io> of <at> the <nn1> railway <nn1> company <ii31> in <ii32> regard <ii33> to <nn1> freight <nn2> charges <vbz> is	y
<nn1> president <io> of <jj> provincial <nn1> blindness <ii31> in <ii32> regard <ii33> to <at> the <nn1> necessity <if> for	y
<vvd> said <at> the <jj> other <nnt1> day <ii31> in <ii32> regard <ii33> to <at> the <jj> asiatic <nn1> aspect	y
<nn2> proclamations <y> , <iw> with <nn2> provisions <ii31> in <ii32> regard <ii33> to <at> the <nn1> liberation <io> of	y

<vvg> sitting <rr> always <y> . <"> " <ii31> in <ii32> regard <ii33> to <at> the <nn1> danger <io> of	y
<vhn@> had <at1> a <jj> susceptible <nn1> imagination <ii31> in <ii32> regard <ii33> to <nn2> women <y> ? <at> the	y
<nnt2> years <ra> ago <jj> similar <nn2> comments <ii31> with <ii32> regard <ii33> to <at> the <jj_nn1@> commercial <csa> as	y
<rr> indeed <ii> on <jj_nn1> classic <nn1> ground <ii31> in <ii32> regard <ii33> to <pn1> everything <vvn_jj> connected <iw> with	y
<ii> in <db> all <nn2> speculations <y> , <ii31> in <ii32> regard <ii33> to <db> all <jj> other <nn2> subjects	y
<vvi> call <ii> on <nnb> mr. <np1> erveng <ii31> in <ii32> regard <ii33> to <at> the <nn1_np1> fetich <y> .	y
<vvg_jj@> mounting <nn1> fiftyfour <nn2> guns <y> . <ii31> with <ii32> regard <ii33> to <at> the <nn1> absence <io> of	y
<xx> not <y> ; <rr> also <y> , <ii31> with <ii32> regard <ii33> to <vvg> affording <nn1> refuge <ii> to	y
<y> . <"> " <dd1> this <nn1> action <ii31> in <ii32> regard <ii33> to <np1> louisiana <vbdz> was <vvn> accompanied	y
Total Correct	25
Lexical instances: (total=90)	
<ccb> but <cs> if <y> , <ii> from <at1> a <nn1> regard <y> , <pph1> it <vvz> becomes <jj> implicit	y
<at1> a <jj> military <nn1> force <y> , <iw> without <nn1> regard <ii> to <nn2> numbers <y> , <vbz> is	y
<fo> couldp228be <vvn_vvd> played <y> . <ii21> out <ii22> of <nn1> regard <if> for <at> the <jj@_np1_nn1@> aachen <nn2> comrades	y
<y> , <cc> and <iw> with <rr> almost <jj> reverential <nn1> regard <if> for <appge_ppho1> her <jj> strong <cc> and	y
<nn1> character <csa_ii@> as <dd1> this <y> . <appge> his <nn1> regard <if> for <appge> your <nn1> father <vbz> is	y
<nn1> inquiry <cst_dd1> that <vvz> prevails <rg> so <rr> extensively <ii32> regard <ii33> to <at> the <nn2> incidents <io> of	n
<y> , <csa> as <pph1> it <vvz> affects <at> the <nn1> regard <ii> in <ddq> which <pphs1> he <vbz> is	y
<ddq> which <db> all <nn2> differences <ii> in <dd1> that <nn1> regard <vm> must <vbi> be <vvn> determined <y> .	y
<y> , <cc> and <vvd_vvn> forced <appge> his <jj> unwilling <nn1> regard <y> . <ppis1> i <vvd> shrank <iw> with	y
<at> the <np1> atlantic <y> , <ddqv> whatever <jj> personal <nn1> regard <cc> or <nn1> affection <if> for <ppho1> him	y
<jj> thin <nn1> disguise <io> of <at1> a <jj@> pretended <nn1> regard <if> for <jj_nn1> public <nn1> health <y> ,	y
<nn1> ordinance <y> , <rr> purely <ii21> out <ii22> of <nn1> regard <nn1> to'his <nn1_vv0> command <y> , <cc> and	y
<nn1> use <y> , <xx> not <rr> only <ii> in <nn1> regard <ii> to <appge> their <jj> great <nn1> power	n
<nn_vv0@> fish <nn1> meal <y> , <vv0_nn1> yeast <y> , <nn1> regard <ii> to <at> the <jj> available <nn1_jj@> nutrient	n
<io> of <ii> by <ppx1> itself <y> , <iw> without <nn1> regard <ii> to <at> the <jj_nn1@> other <y> .	y
<io> of <nn1> danger <y> . <io> of <appge> my <nn1> regard <if> for <dd2> these <y> , <dd1> this	y
<nn1> house <io> of <nn2> representatives <ii> in <dd1> that <nn1> regard <y> , <cc> and <vvn_vvd> re-echoed <ii> on	y
<vbn> been <vvn> taught <ii> from <nn1> infancy <ii> to <nn1> regard <ppx1> myself <ii> as <at1> a <jj> superior	n
<dd2> these <vbdr> were <nn2> articles <io> of <jj> superstitious <nn1> regard <y> , <cc> and	y

<vvn_vvd> held <ii> in	
<ii> re <y> ? <nn1> verse <io> of <jj> due <nn1> regard <if> for <at> the <nn2> interests <io> of	y
<y> ; <pph1> it <vvz> cherishes <at1> a <jj> tender <nn1> regard <if> for <at> the <"> " <nn1_jj> pocket-nerve	y
<rr> even <vvg> preferring <ppho1> him <ii> in <dd1> this <nn1> regard <ii_to> to <rg_rr> so <jj_vvg@> promising <at1> a	y
<ii> in <jj_nn1@> earnest <ii> in <appge> their <jj> mutual <nn1> regard <to> to <vvi> accept <at> the <nn1> test	y
<io> of <nn1> opinion <cst_dd1> that <y> , <vhg> having <nn1> regard <ii> to <at> the <jj> necessary <jj_nn1> official	y
<np1> united <np1> states <y> , <cc> and <vhg> having <nn1> regard <ii> to <at> the <nn1> responsibility <io> of	y
Total Correct	21

B12. Respect

Grammatical instances: (total=110)	**Correct?**
<rr> certainly <vdz> does <xx> not <vvi> hold <ii31> with <ii32> respect <ii33> to <dat> most <nn2> sortals <y> .	y
<jj> inner <nn1@> half <y> . <cc> and <ii31> with <ii32> respect <ii33> to <fo> | <at1> every <jj> other	y
<vvz> holds <at> the <da> same <nn1> language <ii> in <nn1> respect <ii> to <at> the <nn1> constitution <y> .	y
<io> of <jj> certain <nn1> railroad <nn2> companies <ii31> with <ii32> respect <ii33> to <appge> their <jj> enviable <nn2> records	y
<nn1> theory <io> of <nn1> motion <y> , <ii> in <nn1> respect <ii> to <ddq> which <at> the <jj> classical	y
<ii> by <vvg> stabilizing <at> the <nn1> object <ii31> with <ii32> respect <ii33> to <dd> some <jj_nn1@> particular <mc1> one	y
<nn1> chapter <y> , <ii> upon <cc> and <ii31> with <ii32> respect <ii33> to <nn1> income <vvn> derived <ii> from	y
<jj> relative <nn1> motion <io> of <pn1_mc1> one <ii31> with <ii32> respect <ii33> to <at> the <jj_nn1@> other <y> .	y
<cst_dd1> that <vdd> did <vvi> exist <y> . <ii> in <nn1> respect <ii> to <dd2> these <y> , <at> the	y
<csa_ii@_rg@> as <xx> not <rg> very <jj> pressing <ii31> with <ii32> respect <ii33> to <at> the <nn1> case <io> of	y
<jj> costly <nn1> source <io> of <nn1> power <ii> in <nn1> respect <ii> to <nn1> steam <y> . <np1> watt	y
<jj> independent <io> of <ppx221> each <ppx222> other <ii> in <nn1> respect <ii> to <appge> their <jj> domestic <nn2> concerns	y
<vm> would <rr> evidently <xx> not <vvi> arise <ii31> with <ii32> respect <ii33> to <dd1> this <nn1> sense <y> (	y
<vvn> designed <to> to <vvi> demonstrate <nn1> motion <ii31> with <ii32> respect <ii33> to <dd1> this <nn1> ether <y> .	y
<vbz> is <ii> at <at> no <nn1> disadvantage <ii31> with <ii32> respect <ii33> to <appge> his <nn1_jj@> patent <nn2> rights	y
<jj> long <nn1> interview <iw> with <nn1_np1> maudslay <ii31> with <ii32> respect <ii33> to <at> the <nn1> planning <cc> and	y
<vbdr> were <xx> not <jj> revolutionary <cs_ii> except <ii> in <nn1> respect <ii> to <nn1> geography <cc> and <at> the	y
<y> , <"> " <zz1> t <cs> because <ii> in <nn1> respect <ii> to <at> the <nn2> powers <vvn_vvd> confided	y
<"> " <jj> improper <nn2> procedures <"> " <ii31> with <ii32> respect <ii33> to <nn1> expense <nn2> accounts <y> .	y

<io> of <at> the <nd1> south <y> , <ii> in <nn1> respect <ii> to <at> the <nd1> north <y> ,	y
<jj> professional <nn2> men <io> of <nn2> letters <ii> in <nn1> respect <ii> to <at> the <jj> so-called <nn2> corruptions	y
<nn1_np1> price <nn1> control <y> . <"> " <ii31> with <ii32> respect <ii33> to <dd> some <io> of <at> the	y
<np1> russia <y> , <rr> so <vvd_vvn@> enlighted <ii> in <nn1> respect <ii> to <appge> her <nn2> interests <y> ,	y
<vvi> help <at> the <nn2> students <"> " <ii31> with <ii32> respect <ii33> to <nn2> deadlines <y> , <nn1> application	y
<ii21> out <ii22> of <appge> her <nn2> commitments <ii31> with <ii32> respect <ii33> to <np1> nato <y> , <at> the	y
Total Correct	25
Lexical instances: (total=317)	
<nn1> family <y> , <vhd> had <dar_rrr> more <io> of <nn1> respect <cc> and <nn1> gratitude <csn> than <rr> commonly	y
<y> ; <pph1> it <vvd> differed <ii> in <at> no <nn1> respect <ii> from <at> the <nn2> others <cs_vv0%_ii> except	y
<y> , <ppis1> i <vvd> noticed <cst> that <appge> her <nn1> respect <vhd> had <rr> considerably <vvn> increased <y> ;	y
<ppis1> i <vbdz> was <jj> fortunate <ii> in <dd1> another <nn1> respect <y> , <ii> in <vbg> being <vvn> admitted	y
<csa_rg@> as <rr> passively <y> , <mc> iii <dd1> this <nn1> respect <y> , <csa_ii> as <at> the <nn1> post	y
<jj> old <nn1> age <y> , <ii> in <dd> some <nn2> respects <y> , <vhd> had <at1> a <jj> similar	y
<rr> hardly <vvd_vvn> surpassed <ppho1> him <ii> in <dd1> this <nn1> respect <y> . <appge> his <nn1> integrity <vbdz> was	y
<vbm> am <xx> not <jj> unique <ii> in <dd1> this <nn1> respect <ii21> according <ii22> to <at1> a <jj> reliable	y
<at1> a <jj> mediocre <nn1> performance <y> , <ii> in <nn2> respects <ddq> which <nnb> mr. <np1> downes <vm> would	y
<y> , <ppis1> i <vbm> am <y> , <iw> with <nn1> respect <y> , <ppge> yours <y> , <rr> truly	y
<rr> certainly <vbdz> was <y> ; <ii> in <dd1> this <nn1> respect <vvg> corresponding <iw> with <pn1> everything <ppis1> i	y
<vbdz> was <xx> not <rr> only <ii21> out <ii22> of <nn1> respect <io> of <ppho1> him <cc> and <appge> his	y
<at> the <nd1> south <vbr> are <ii> in <dd> some <nn2> respects <jj> peculiar <y> , <cc> and <to> to	y
<nn1> man <y> , <cc> and <ii> in <da2> many <nn2> respects <at> the <nnt2> times <y> , <at> the	y
<ddq> which <y> , <cs> while <ii> in <da2> many <nn2> respects <jj> similar <ii> to <jj> ordinary <nn2> protons	y
<vbr> are <jj> sure <to> to <vvi> lose <jj_nn1> public <nn1> respect <y> , <cc> and <ii> in <at> the	y
<nnb> mrs. <np1> tomasic <vhd> had <vvn> paid <vvg_jj@> moving <nn1> respect <ii> to <dd1> that <jj> peerless <nn1> seaman	y
<cc> or <nn1> misery <ii> in <dd> any <jj> other <nn1> respect <y> ; <cs> whereas <y> , <ii> in	y
<nn1> station <ge> 's <nn1> record <ii> in <dd1> this <nn1> respect <ii> in <vvg_nn1> passing <ii_rp@> on <appge> its	y
<if> for <pnqo> whom <pphs1> he <vhd> had <at> no <nn1> respect <rr21_ii> at <rr22_db> all <y> , <cc> and	y
<vvz> minimizes <nn1> change <ii> in <at1> a <jj> certain <nn1> respect <y> (<rex> i.e. <y> , <ii31> in	y
<vbi> be <vvn> regarded <iw> with <at> the <da> same <nn1> respect <cc> and <nn1> confidence <ii> by <db> all	y
<vm> should <vhi> have <vvn> withstood <ii> in <dd1> this <nn1> respect <at> the <nn1> clamor <io> of <jj> political	y

<y> , <vbr> are <y> , <ii> in <mc1> one <nn1> respect <cc> or <dd1> another <y> , <csn> than	y
<jj> earnest <nn1> desire <to> to <vvi> show <at1> every <nn1> respect <ii> to <at> the <jj> royal <nn1> visitor	y
Total Correct	25

B13-A. Spite

Grammatical instances: (total=156)	**Correct?**
<to> to <vvi> marry <ppho1> him <y> ? <ii31> in <ii32> spite <ii33> of <appge> her <nn1> excitement <y> ,	y
<nn1> science <cc> and <nn1> art <y> . <ii31> in <ii32> spite <ii33> of <at> the <jj_nn1> whole <jj> modern	y
<vvi> moan <y> . <np1> lowenthal <y> , <ii31> in <ii32> spite <ii33> of <ppx1> himself <y> , <vvz> turns	y
<nn1_jj> future <!> ! <if_cs%> for <y> , <ii31> in <ii32> spite <ii33> of <mc> four <nnt2> centuries <cc> and	y
<ppis1> i <vvd> protested <rr> faintly <y> ? <ii31> in <ii32> spite <ii33> of <at> the <nn1> chill <ppis1> i	y
<np1> david <y> (<rr> tremulously <y> ; <ii31> in <ii32> spite <ii33> of <ppx1> himself <y>) <np1> jenny	y
<jjr> larger <nn2> earnings <ii> per <nn1> share <ii31> in <ii32> spite <ii33> of <da2> many <nn2> decreases <ii> in	y
<nnt1> afternoon <vbdz> was <jj> hot <y> , <ii31> in <ii32> spite <ii33> of <at> the <jj> steady <nn1> breeze	y
<vvn> quoted <csa_ii@> as <vvg> saying <cst_dd1> that <ii31> in <ii32> spite <ii33> of <nn2> denials <y> , <dd> some	y
<nn1_jj> future <y> . <rr> so <y> , <ii31> in <ii32> spite <ii33> of <at> the <nn1_jj@> kind <nn1> assurance	y
<vm> 'll <vvi> beat <np1> tom <y> , <ii31> in <ii32> spite <ii33> of <appge> his <nn1@_jj@> bragging <y> .	y
<vvg> exceeding <jj> glad <cc> and <nn1_jj> sunshiny <ii31> in <ii32> spite <ii33> of <appge> her <nn1> refusal <to> to	y
<io> of <appge> her <nn1> veil <y> . <ii31> in <ii32> spite <ii33> of <dd> some <jj> undefined <nn2> apprehensions	y
<np1> schumann <y> (<vvz_nn2> smiles <y> , <ii31> in <ii32> spite <ii33> of <ppx1> himself <y>) <ppis1> i	y
<rr> regularly <y> , <ppx1> myself <y> , <ii31> in <ii32> spite <ii33> of <jj> furious <nn2> storms <cc> and	y
<at> the <nn1> end <y> , <vvn> detained <ii31> in <ii32> spite <ii33> of <ppx1> himself <ii> by <at> the	y
<rr22> yet <y> ; <if> for <y> , <ii31> in <ii32> spite <ii33> of <appge> her <nn1> annoyance <y> ,	y
<nnt1> night <iw> with <ppho1> him <y> , <ii31> in <ii32> spite <ii33> of <appge> her <nn1> shyness <cc> and	y
<vvd> stuck <ii> to <dd1> this <nn1> model <ii31> in <ii32> spite <ii33> of <dd1> this <nn1> contradiction <y> ,	y
<vvg_jj> rasping <ii> towards <ppho1> him <y> , <ii31> in <ii32> spite <ii33> of <at> the <nn1> laziness <io> of	y
<vbr> are <rr> absolutely <jj> different <nn2> things <ii31> in <ii32> spite <ii33> of <appge> their <jj> similar <nn2> names	y
<y> , <cc> and <nn1> movement <y> . <ii31> in <ii32> spite <ii33> of <nn1> smallness <y> , <nn1> bareness	y
<at1> a <jj> bewildered <nn1> expression <ii> on <appge> his <ii32> spite <ii33> of <at> the <nn2> tears <ii> on	y
<nn1> world <rr21> at <rr22> large <y> . <ii31> in <ii32> spite <ii33> of <appge_ppge@> his <rg> rather <jj> defective	y
<at> the <jj> outside <nn1> world <y> , <ii31> in <ii32> spite <ii33> of <nnb> professor <np1> sulkowski <ge> 's	y

Total Correct	25
Lexical instances: (total=19)	
<jj> dead <nn1> weight <cst> that <vvd> steadied <ppho1> him <nn1> spite <io> of <appge> his <jjt> best <nn2> efforts	n
<nn2> souls <to> to <vvi> glut <at1> a <jj> meaningless <nn1> spite <y> ? <ppy> you <vh0> have <vbn> been	y
<np1> margaret <vhd> had <vvn@> become <jj> interested <ii> in <nn1> spite <io> of <ppx1> herself <y> , <cs> though	n
<cst> that <np1> marian <y> , <if> for <dd> some <nn1> spite <y> , <vm> would <xx> not <vvi> awaken	y
<cs21> as <cs22> though <at1> a <vhd_vhn@> had <at1> a <nn1> spite <ii> against <nnu> th <ge> ' <jj@> very	y
<vm> must <nn2> needs <vvi> open <pph1> it <y> , <nn1> spite <zz1> o <ge> ' <db> all <appge> my	n
<jj> wondrous <jj> fair <nn1> woman <y> , <ii> in <nn1> spite <zz1> o <ge> ' <appge> her <da> own	n
<nn1_vv0> ride <ii> to <at> the <nn2> wars <ii> in <nn1> spite <zz1> o <ge> ' <db_rr@> all <rr> forth	n
<rrq_cs> where <nn1_jj> ill-nature <y> , <nn1> malice <y> , <nn1> spite <y> , <vv0_nn1> envy <y> , <vv0_nn1> deceit	y
<vvi> carry <ppy> you <rp@> in <y> , <ge> ' <nn1> spite <io> of <nn1> hell <"> " <"> "	n
<to> to <vvi> carry <ppy> you <rp@> in <y> , <nn1> spite <io> of <nn1> hell <"> " <pphs2> they	n
<vvi> carry <ppy> you <rp@> in <y> , <ge> ' <nn1> spite <io> of <nn1> hell <y> . <"> "	n
<nn2> origins <y> , <pphs1> he <vvd> vented <appge> his <nn1> spite <ii> in <nn1> nihilism <y> . <appge> his	y
<np1> mcclintock'seyes <vvd> shone <iw> with <nn1> whiskey <cc> and <nn1> spite <y> . <pphs1> he <jj> ap/proached <nn1> vridar	y
<xx> n't <vdi> do <pph1> it <rr> just <if> for <nn1> spite <y> . <pphs1> she <vm> 'd <vhi> have	y
<jj> bad <nn2> words <rr> just <if> for <at> the <nn1> spite <io> of <pph1> it <y> . <pphs1> he	y
<pph1> it <y> , <vvg> sitting <jj> close <nn1> together in <nn1> spite <io> of <at> the <nn1> heat <y> ,	n
<at1> a <jj> mighty <cc> and <jj> triumphant <nn1> faction-in <nn1> spite <io> of <ppho2> them <pphs1> he <vvd> carried	n
<nn1> ticket <ii> into <at> the <nn1> field <if> for <nn1> spite <y> . <cc> and <vvd_vvn> declared <cst> that	y
Total Correct	9

B13-B. Spite

Grammatical instances: (total=889)	**Correct?**
<vvn_vvd> looked <nn2> volumes <y> , <cc> and <ii31> in <ii32> spite <ii33> of <at> the <nn2> glasses <y> ,	y
<jj> handsome <nn2> shops <y> . <ccb> but <ii31> in <ii32> spite <ii33> of <db> all <dd1> this <y> ,	y
<ii> from <appge> her <nn1> office <y> , <ii31> in <ii32> spite <ii33> of <at> the <jj@> borrowed <nn1> hat	y
<csa_ii@> as <nn1> confederate <nn2> men-of-war <y> , <ii31> in <ii32> spite <ii33> of <nn1> neutrality <nn2> treaties <y> ,	y
<at1> a <jj> true <nn1> woman <y> , <ii31> in <ii32> spite <ii33> of <appge> his <nn1> shame <cc> and	y
<cc> and <nn1_vvg%_jj> ruling <at> the <nn1> scene <ii31> in <ii32> spite <ii33> of <appge> her <nn1> absence <ii> from	y

<jj> consistent <"> " <nn1> hustler <"> " <ii31> in <ii32> spite <ii33> of <at> the <jj_nn1@> friendly <cc> and	y	
<y> , <npm1> july <y> , <vv0_nn1_jj> // <ii31> in <ii32> spite <ii33> of <jj> external <nn2> difficulties <cc> and	y	
<vvi> change <db> all <dd1> this <y> , <ii31> in <ii32> spite <ii33> of <at> the <jj> abominable <nn2> stretches	y	
<vvz> sees <rr> clearly <dd1_cst> that <y> , <ii31> in <ii32> spite <ii33> of <appge> her <nn1> littleness <cc> and	y	
<nnt1> day <y> . <rr> however <y> , <ii31> in <ii32> spite <ii33> of <jj> certain <nn2> assurances <y> ,	y	
<vvd> decided <ii> against <pph1> it <y> , <ii31> in <ii32> spite <ii33> of <at> the <nn1> money <nn1> place	y	
<nn1> rectitude <y> . <rr> nevertheless <y> , <ii31> in <ii32> spite <ii33> of <dd1> this <nn1> consciousness <y> ,	y	
<nn1> movie <nn1> piece <y> , <if_cs%> for <ii31> in <ii32> spite <ii33> of <at> the <nn1> care <if> for	y	
<ddq> what <pph1> it <vbz> is <y> . <ii31> in <ii32> spite <ii33> of <db> all <at> the <nn2> efforts	y	
<ii> in <nn1> place <y> . <vvg_nn1_jj> sweating <ii31> in <ii32> spite <ii33> of <at> the <jj_nn1> fifty-five-degree <nn1> weather	y	
<vh0> have <rr> greatly <vvn> increased <y> , <ii31> in <ii32> spite <ii33> of <at> the <nn1> reduction <ii> in	y	
<nn1> life <ii> in <np1> india <y> , <ii31> in <ii32> spite <ii33> of <appge> her <nn1> mother <ge> 's	y	
<cst> that <ppis1> i <vdd> did <rr> so <ii31> in <ii32> spite <ii33> of <jj> certain <nn2> provisions <ddq> which	y	
<vvd> suited <ppio2> us <rg> very <rr> well <ii31> in <ii32> spite <ii33> of <at> the <nn1> fact <cst> that	y	
<nn1> immobility <vvd_vvn> made <ppho1> him <jj> nervous <ii31> in <ii32> spite <ii33> of <ppx1> himself <y> . <pphs1> he	y	
<nn1> prosecution <io> of <appge> my <nn2> studies <ii31> in <ii32> spite <ii33> of <at> the <nn2> difficulties <ii> by	y	
<vbdz> was <vvn_jj> ousted <y> . <"> " <ii31> in <ii32> spite <ii33> of <db> all <nn1> repression <y> ,	y	
<cc> or <nn1> spine <vvd> occurred <y> , <ii31> in <ii32> spite <ii33> of <at> the <jj> cheerful <nn1> hopefulness	y	
<at> the <nn1> village <y> , <cc> and <ii31> in <ii32> spite <ii33> of <at> the <nn1> rain <at1> a	y	
Total Correct	25	
Lexical instances: (total=64)		
<io> of <nn1> favoritism <y> , <nn1> jealousy <y> , <nn1> spite <at> the <nn1> play <io> of <at> the	y	
<nn1> family <y> , <ccb> but <rr> merely <ii> to <nn1> spite <appge> his <jj_nn1> english <nn2> cousins <y> ,	y	(v)
<rr> pertly <cc> and <appge> her <nn1> eye <vvd> flashed <nn1> spite <y> . <"> " <rr> all <ii> over	y	
<vbz> 's <vdg> doing <pph1> it <ii21> out <ii22> of <nn1> spite <y> . <ccb> but <pphs1> he <vhz> has	y	
<nn1> figure <y> , <cs> if <rr> only <ii> to <nn1> spite <at> the <np2> pinkertons <y> . <cs> before	y	(v)
<cc> and <at1> every <nn1> impulse <io> of <jj> low <nn1> spite <vm> would <vvi> combine <to> to <vvi> urge	y	
<iw> with <nn1> violence <y> , <cc> and <jj> holy <nn1> spite <y> , <cc> and <nn1> malice <y> ,	y	
<cs_rrq> when <pphs1> he <vvz> refuses <y> , <ii> from <nn1> spite <ppio2> us <cst> that <pphs1> he <vbz> is	y	
<iw> with <nn1> jealousy <cc> and <nn1> rage <cc> and <nn1> spite <y> , <appge> her <nn1> father <vvd@_vvn@_jj> married	y	

<ii32> front <ii33> of <appge> his <nn2> eyes <if> for <nn1> spite <y> . <nn1_np1@> lyndell <vm> would <vvi> drift	y	
<nn1_np1@> lenore <vvz> regards <ppho1> him <iw> with <jj> covert <nn1> spite <y> . <y> (<vv0_nn1> note <y> ;	y	
<y> , <cc> and <at> the <jj> poor <ii> from <nn1> spite <y> . <pphs1> he <vvz@> dresses <rg> very	y	
<nn1> curiosity <iw> with <nn1> scorn <y> , <cc> and <nn1> spite <iw> with <nn2> imprecations <y> . <at1> a	y	
<vvd> said <y> . <"> " <ii21> out <ii22> of <nn1> spite <y> . <ii21> out <ii22> of <jj> pure	y	
<vbm> 'm <vvgk> gon <to> na <vvi> switch <at> the <nn1> spite <ii21> out <ii22> of <dd1> that <fu> dee	y	
<rp> over <ii> to <jj> low <nn1> anger <cc> and <nn1> spite <y> , <vvg> drinking <rl> away <appge> your	y	
<y> . <"> " <pph1> it <vbz> 's <xx> not <nn1> spite <y> , <"> " <pphs1> she <vvd> said	y	
<at> the <nn1> downfall <io> of <vd0> do <appge> thy <nn1> spite <!> ! <"> " <cs_rrq> when <jj> humble	y	
<ppy> you <vbr> 're <vdg> doing <pph1> it <ii> to <nn1> spite <nn1> rue <y> . <ppy> you <vv0> want	y	(v)
<rp_ii> off <pn1_mc1> one <ge> 's <nn1> nose <ii> to <nn1> spite <pn1_mc1> one <ge> 's <nn1> face <y> ?	y	(v)
<y> , <rr> simply <to> to <vvi> gratify <jj> personal <nn1> spite <y> , <rr> designedly <cc> and <rr> maliciously	y	
<ii> at <appge> your <nn1_jj@> return <z> -- <ii> through <nn1> spite <cc> and <nn1> scorn <y> , <appge> your	y	
<y> , <at1> a <nn1> spectacle <io> of <jj> fiendish <nn1> spite <y> . <"@_ge> ' <dd2> those <nn2> indians	y	
<"> " <pphs1> she <vvd> added <y> , <iw> with <nn1> spite <cc> and <vv0> somethingnot <jj> unlike <nn1> jealousy	y	
<ii> into <nn2> hearts <rr> already <vvg> festering <iw> with <nn1> spite <y> . <at> the <nn1> house <vvd_vvn> raised	y	
Total Correct	25	
Probably grammatical instances: (total=33)		
<y> , <ii> on <dd1> that <nnt1> day <y> . <nn1> spite <io> of <at> the <nn1_jj> cold <y> ,	y	arch
<dd1> this <pph1> it <vdd> did <y> , <ii> in <nn1> spite <y> , <ppis1> i <vbm> am <jj> convinced	y	
<jj> other <jj> neutral <y> , <cc> and <ii> in <nn1> spite <rr@_rg> too <io> of <at> the <jj> reactionary	y	
<vbr> are <rr> necessarily <jj> great <y> , <ii> in <nn1> spite <y> , <db> all <dd1_cst> that <vm> can	y	
<nn1> man <to@_ii> to <vvi_nn1> work <if> for <y> , <nn1> spite <io> of <appge> his <jj> queer <nn2> ways	y	arch
<nn2> meannesses <y> , <cc> and <ii> in <jj_nn1> // <nn1> spite <io> of <db> all <appge> its <nn2> meannesses	y	
<ii> in <appge> my <nn1> breast <y> , <cc> and <nn1> spite <io> of <nn1> nature <ge> 's <jj> feeble	y	arch
<appge> its <jj_nn1> manifold <nn2> teachings <y> , <ii> in <nn1> spite <rr@_rg> too <io> of <at> the <nn2> sacraments	y	
<cc> and <vm> will <vvi> maintain <pph1> it <y> , <nn1> spite <io> of <db> all <at> the <nn1> world	y	arch
<rl> apart <y> , <vh0> have <vvn> concentrated <ii> in <nn1> spite <io> of <at> the <nn1> river <cc> and	y	
<jj> great <nn1> city <ii> in <mc> 6 <nnt2> years <nn1> spite <io> of <nn2> overflows <cc> and <nn1> fever	y	arch
<mc> 10 <ccb> but <pnqv> whoever <y> , <ii> in <nn1> spite <y> , <ii> at <jj> american <nn1_jj> right	y	
<pphs1> he <vvi> roll <ii> in <nn1> wealth <y> ; <nn1> spite <io> of <jj> scrupulous <nn1>	y	arch

economy <y> ,		
<jj> conceivable <y> ; <cc> and <dd1_cst> that <y> , <nn1> spite <io> of <at> the <nn2> sores <cc> and	y	arch
<pphs1> she <rt> now <vdz> does <nn1_jj> -- <ii> in <nn1> spite <y> , <rr@_rg> too <y> , <io> of	y	
<appge> his <jj> round <nn1> face <y> . <ii> in <nn1> spite <y> , <rr> however <y> , <io> of	y	
<y> , <cs> if <vvn> carried <rp> on <ii> in <nn1> spite <io> of <appge> their <nn1> opposition <y> ,	y	
<vvz> belongs <ii> to <pph1> it <xx> not <ii> in <nn1> spite <io> of <ccb> but <ii21> because <ii22> of	y	
<nn1> voice <vvg> trembling <rr21_at1> a <rr22_jj> little <ii> in <nn1> spite <io> of <appge> her <nn2> efforts <to> to	y	
<ge> 's <nn1> garden <y> . <rr> so <ii> in <nn1> spite <io> of <ddq> what <vm> might <vvi> seem	y	
<ii> about <ppho1> him <y> , <cc> and <nn1> shiveredin <nn1> spite <io> of <at> the <nn1> flame <y> .	y	arch
<y> , <vvg> blushing <rr21_at1> a <rr22_nn1> bit <ii> in <nn1> spite <io> of <ppx1> herself <y> . <"> "	y	
<io> of <appge> his <nn1> existence <y> ; <cc> and <nn1> spite <io> of <ppx1> himself <y> , <at1> a	y	arch
<da1> much <nn1> nobleness <ii> in <ppho1> him <y> , <nn1> spite <io> of <appge> his <nn2> vagaries <y> .	y	arch
<vvi> hang <rp_ii21> on <ii_ii22> to <pph1> it <ii> in <nn1> spite <at1> a <nn1> hell <cc> and <jj> high	y	
Total Correct	25	

B14. Terms

Grammatical instances: (total=127)	**Correct?**
<vm> can <xx> not <vbi> be <vvn@> analyzed <ii31> in <ii32> terms <ii33> of <jj> warm <nn1> weather <y> ,	y
<rr> publicly <vvn_jj@_vvd> traded <nn1_vv0> brokerage <nn1_jj> firm <ii31> in <ii32> terms <ii33> of <nn1> revenue <ii> in <mc> 1976	y
<at> the <jj> compositional <nn1> criterion <y> , <ii31> in <ii32> terms <ii33> of <at> the <nn1> persistence <io> of	y
<rr> soberly <vvi> visualize <da> such <nn2> particles <ii31> in <ii32> terms <ii33> of <jj> ordinary <nn2> qualities <ii> like	y
<nn2> criteria <if> for <jj> ordinary <nn2> objects <ii31> in <ii32> terms <ii33> of <at1> a <jj> prior <nn1> notion	y
<rr> easily <vvi> define <jj_nn1> spatiotemporal <nn1> continuity <ii31> in <ii32> terms <ii33> of <pph1> it <y> . <rt> now	y
<dd1> this <vm> may <vbi> be <vvn> described <ii31> in <ii32> terms <ii33> of <jj> digestible <cc> and/or <jj> metabolisable	y
<nn1_jj> level <iw> with <np1> bob <np1> dylan <ii31> in <ii32> terms <ii33> of <vbg> being <at> the <nn1> voice	y
<nn1> politics <y> . <pph1> it <vbz> is <ii31> in <ii32> terms <ii33> of <at> the <nn1> risk <io> of	y
<np1_nn2> fparts <y> (<rex> e.g. <y> , <ii31> in <ii32> terms <ii33> of <at> the <nn1> overlap <nn2> relations	y
<vvg> drinking <rr> moderately <y> . <ge_"@> ' <ii31> in <ii32> terms <ii33> of <at> the <nn1> health <io> of	y
<vbz> is <rr> thus <rr> ultimately <vvn> analyzed <ii31> in <ii32> terms <ii33> of <at> the <jj_nn1> sortal-covered <nn1> persistence	y
<nn1> tree <cc> and <at> the <nn1> trunk <ii31> in <ii32> terms <ii33> of <at> the <nn2> successions <cc> or	y

<io> of <appge> our <nn1> identity <nn2> concepts <ii31> in <ii32> terms <ii33> of <at> the <jj> persistence-free <nn1> level	y
<y>) <vm> can <vbi> be <vvn@> analyzed <ii31> in <ii32> terms <ii33> of <at> the <jj> following <jj> sortal-neutral	y
<at1> a <rr> probably <vvn_jj> misplaced <y> . <ii31> in <ii32> terms <ii33> of <at> the <jj> dutch <nn1> class	y
<vbi> be <rgt> most <rr> fully <vvn_vvd> understood <ii31> in <ii32> terms <ii33> of <dd2> these <nn2> manifestations <y> ;	y
<vvg> interpreting <dd1_cst> that <nn1_vv0@> phrase <rr> liberally <ii31> in <ii32> terms <ii33> of <nn1> thrift <y> , <nn1> economy	y
<y> . <nn2> luminosities <vbr> are <vvn> given <ii31> in <ii32> terms <ii33> of <at> the <nn1> luminosity <io> of	y
<to> to <vbi> be <jj> definable <rr> purely <ii31> in <ii32> terms <ii33> of <nn1> continuity <nn2> considerations <y> ,	y
<xx> not <vhi> have <jj> observational <nn2> criteria <ii31> in <ii32> terms <ii33> of <ddq> which <to> to <vvi> analyze	y
<nn1> degree <rr> causally <vvn_jj> connected <y> (<ii31> in <ii32> terms <ii33> of <ddq> which <y> , <csa> as	y
<nn2> obligations <y> , <rr21> at <rr22> least <ii31> in <ii32> terms <ii33> of <jj> civil <nn1> law <y> .	y
<ge> 's <nn1> unity <ii> through <nn1> space <ii31> in <ii32> terms <ii33> of <rrq> how <appge> its <nn2> parts	y
<rrq> when <ppis2> we <vv0> consider <dd1> this <ii31> in <ii32> terms <ii33> of <nn2> successions <io> of <nn2> object-stages	y
Total Correct	25
Lexical instances: (total=360)	
<to> to <vvi> arrange <ii21> as <ii22> to <at> the <nn2> terms <y> . <ppis1> i <vbdz> was <vvn> offered	y
<np1> abm <nn1> treaty <y> , <cc> and <at> the <nn2> terms <io> of <at> the <nn1> treaty <ppx1> itself	y
<vvn@> said <y> , <ii_rp@> in <rr> crudely <jj> unanalyzed <nn2> terms <y> , <to> to <vvi> preserve <at> the	y
<vm> would <xx> not <vvi> submit <ii> to <appge> his <nn2> terms <y> , <vbdr> were <vvn> left <ii> on	y
<vvi> accept <at> the <nn1> challenge <ii> on <dd2> those <nn2> terms <y> . <cc> and <nn1> maugre <dd2> these	y
<vhd> had <vbn> been <vvn> described <ii> in <da> such <nn2> terms <ii> by <at1> a <jj_nn1> french <nn1_jj> official	y
<io> of <at> the <nn1> use <io> of <at> the <nn2> terms <y>) <vh0> have <rr> always <vbn> been	y
<y> , <vvg> looking <ii> at <at> the <jj_nn1> general <nn2> terms <io> of <at> the <nn2> commissions <y> ,	y
<vbr> are <xx> n't <rr> exactly <ii> on <vvg_jj@> speaking <nn2> terms <ccb> but <dd1> that <vbz> 's <vvg> playing	y
<xx> not <vvgk> going <to> to <vvi> quote <at> the <nn2> terms <io> of <nn1> abuse <rg_rr> so <rr> lavishly	y
<nn1> patient <y> . <jj> obstructive <nn1> sleep <np1_nn1@> apnea <nn2> terms <null> <p> <nn1_np1@> apnea <vbz> is <at> the	y
<y> , <ccb> but <rr> also <ii> upon <jj> intimate <nn2> terms <io> of <nn1> friendship <y> . <np1> great	y
<vbdz> was <xx> n't <jj> reducible <ii> to <jj> ordinary <nn2> terms <io> of <nn1> expression <y> . <at> the	y
<io> of <appge> his <nn2> clients <vbdz> was <at> the <nn2> terms <vbdr> were <jj> simple <z> -- <ppis1> i	y
<mc> 1965 <y>) <y> . <ii> in <jj> general <nn2> terms <pph1> it <vvz> seems <jj> unnecessary <to> to	y
<to> to <vbi> be <vvn> made <ii> beyond <appge> its <nn2> terms <y> , <pphs2> they <vbr> are <nn2> implications	y

<vhi> have <vvn> enabled <ppho1> him <to> to <vvi> dictate <nn2> terms <io> of <nn1> reconstruction <rr> strictly <ii> on	y
<ii> between <jj> different <nn2> towns <y> . <at> the <nn2> terms <io> of <at> the <nn1> settlement <vm> will	y
<y> ; <dd2> these <cc> and <jj> other <jj@> contrasted <nn2> terms <vv0> oppose <ppx221> each <ppx222> other <ii> in	y
<nn1> restoration <io> of <at> the <nn1_np1> union <ii> on <nn2> terms <jj> consistent <iw> with <rt_jj%> then <jj> existing	y
<nn1_jj> **25;985;toolong <y>) <y> . <ii> in <dd2> these <nn2> terms <y> , <at> the <nn1> figure <io> of	y
<cc> or <rr> even <ii21> as <ii22> to <ddq> which <nn2> terms <vbr> are <jj> subordinate <ii> to <ddq> which	y
<xx> n't <rr21> in <rr22> general <ii> on <jj> friendly <nn2> terms <iw> with <at> the <nn1> quality <y> ,	y
<vm> will <rr> please <vvi> send <rp> in <appge> their <nn2> terms <y> , <jj_vvn_vvd> sealed <cc> and <vvd_jj@_vvn> endorsed	y
<ii> in <np1> china <at1> a <nnt1> decade <rrr> later <ii32> terms <ii33> of <at> the <nn1> empire <cc> and	y
Total Correct	25

B15. View

Grammatical instances: (total=76)	**Correct?**
<at> the <jj_nn1> future <nn2_vvz> needs <y> . <ii31> in <ii32> view <ii33> of <at> the <jj> financial <nn2> difficulties	y
<io> of <nn1> state-craft <y> , <cc> and <ii31> in <ii32> view <ii33> of <at> the <nn2> conditions <io> of	y
<io> of <at> the <nn1> war <y> . <ii31> in <ii32> view <ii33> of <appge> its <nn1> success <y> ,	y
<y> . <pph1> it <vbz> is <vvn> justified <ii31> in <ii32> view <ii33> of <at> the <jj> small <nn1> wave-length	y
<jj_nn1%> good <ii> in <ppho2> them <y> . <ii31> in <ii32> view <ii33> of <dd1> that <y> , <at> the	y
<at> the <nn1> act <vbdz> was <vvn> framed <ii31> in <ii32> view <ii33> of <at> the <nn1> existence <io> of	y
<cc> and <jjr> happier <rr> still <y> , <ii31> in <ii32> view <ii33> of <at> the <jj> eternal <nn1> world	y
<if> for <at> the <nn1> spark <y> . <ii31> in <ii32> view <ii33> of <at> the <jj> high <nn1> speed	y
<fo> madep283during <at> the <nnt1> night <y> . <ii31> in <ii32> view <ii33> of <dd2> those <y> , <pph1> it	y
<csa> as <vbdz> was <xx> not <jj> unnatural <ii31> in <ii32> view <ii33> of <at> the <nn2> developments <y> ,	y
<nnt1> night <ppis2> we <vvd> arrived <y> , <ii31> in <ii32> view <ii33> of <appge> our <nn2> prospects <y> ;	y
<to> to <vbi> be <dd1_cst> that <y> , <ii31> in <ii32> view <ii33> of <at> the <nn1> example <vvn> set	y
<vbz> is <xx> not <jj> unavoidable <y> . <ii31> in <ii32> view <ii33> of <at> the <jj_nn1> manifold <nn2> abuses	y
<vbz> is <rr> sufficiently <jj> comprehensive <y> . <ii31> in <ii32> view <ii33> of <at> the <nn1> fact <cst> that	y
<pph1> it <vbdz> was <jj_nn1> inexpedient <y> , <ii31> in <ii32> view <ii33> of <jjr@_rrr> further <jj> possible <nn2> complications	y
<rr> slightly <dar_rrr> more <csn> than <dd1> this <ii31> in <ii32> view <ii33> of <at> the <jj> limited <nn1> evidence	y
<rp> down <ii> in <nn1> despondency <y> , <ii31> in <ii32> view <ii33> of <at> the <nn1_jj> dark <cc> and	y

<nn1> rest <nn1> attention <y> , <rr> especially <ii31> in <ii32> view <ii33> of <at> the <jj> basic <nn1> similarity	y
<nnt1_md> second <cst> that <vv0_nn1> demand <y> . <ii31> in <ii32> view <ii33> of <dd1> this <nn1> intensity <io> of	y
<rr> logically <pphs2> they <vm> should <vbi> be <ii31> in <ii32> view <ii33> of <at> the <jj> increased <nn2> costs	y
<cs_ii> before <fo> ns <y> , <cc> and <ii31> in <ii32> view <ii33> of <at> the <rg> very <jj> clear	y
<nn2> materials <y> . <rt> now <y> , <ii31> in <ii32> view <ii33> of <at> the <jj> rough <nn1@_vvg> shaking	y
<nn1> energy <nn1> efficiency <y> , <rr> particularly <ii31> in <ii32> view <ii33> of <at> the <nn1> fact <cst_dd1> that	y
<ii> in <nn1> popularity <y> , <rr> particularly <ii31> in <ii32> view <ii33> of <appge> its <jj> low <nn1> capital	y
<pphs2> they <vmk> ought <to> to <y> , <ii31> in <ii32> view <ii33> of <np1> ahmed <y> , <np1> sophie	y
Total Correct	25
Lexical instances: (total=940)	
<vhd> had <rr> thus <rr> modestly <vvn> expressed <appge> their <nn2> views <y> , <vmk> ought <xx> not <ii> to	y
<mc> 47 <z> -- <mc> 115 <jj> medium <nn1> shot <nn1> view <nn1> page <nn1> image <null> <np1> tom	n
<iw> with <at> no <jj@> declared <nn1> object <ii> in <nn1> view <y> , <jjr> balder <vvd_vvn> grasped <at> the	y
<nn1_vv0> witness <nn1_vv0> stand <-> - <jj_rr> late <nnt1> afternoon <nn1> view <nn1> page <nn1> image <np1_nn1> injun <np1> joe	n
<y> ; <ccb> but <ii> to <appge> our <jj> imperfect <nn1> view <y> , <pph1> it <vdz> does <xx> not	y
<y> , <ii> in <ddq> which <at> the <jj> new <nn1> view <vvz> destines <at> the <nn2> creeds <io> of	y
<y> , <cs> for <ppis2> we <vv0> think <appge> his <nn2> views <vbr> are <rr> just <rr31> in <rr32> the	y
<jj_nn1> front <nn1> yard <io> of <np1> sawyer <nn1> house <nn1> view <nn1> page <nn1> image <nnb> aunt <np1> polly	n
<vbz> is <mc1> one <io> of <appge> my <jj_nn1> favorite <nn2> views <y> . <ppis1> i <vbm> am <jj> glad	y
<io> of <jj_nn1> female <nn1> loveliness <vvd_vvn> met <appge> my <nn1> view <ii21> such <ii22> as <ppis1> i <vhd> had	y
<mc> 94 <z> -- <mc> 270 <jj> medium <nn1> shot <nn1> view <nn1> page <nn1> image <ii_csa> as <at> the	n
<ii> in <nn1> corroboration <io> of <at> the <da> same <nn2> views <y> . <"> " <at> the <nn1> union	y
<mc> 254 <jj> close <nn1> shot <-> - <np1> sid <nn1> view <nn1> page <nn1> image <pphs1> he <vbz> is	n
<nn1> flood <y> , <cc> and <appge> my <md> last <nn1> view <io> of <ppho1> him <vbz> is <csa> as	y
<ccb> but <rr> merely <to> to <vvi> present <at> the <nn2> views <io> of <mc1> one <io> of <at> the	y
<rr> quietly <vvz> goes <z> -- <mc> 79 <z> -- <nn1> view <nn1> page <nn1> imageto <at> the <jj> other	n
<jj> anxious <to> to <vvi> carry <rp> out <appge> his <nn2> views <y> . <pphs1> he <vvd> sought <if> for	y
<pphs1> he <rr> cordially <vvn_vvd> entered <ii> into <appge> my <nn2> views <y> . <pphs1> he <vvd> left <appge> his	y
<nn1> ext <y> . <np1_nn2> woods <-> - <nnt1> night <nn1> view <nn1> page <nn1> image <zz1_at1@> a <nn1> dog	n
<mc> 200 <nn2> photographs <vv0> provide <at1> a <jj_nn1> close-up <nn1> view <io> of <appge> his <nn1> vision <y> ,	y

<ppis1> i <vvd> thought <pph1> it <vbdz> was <ii> in <nn1> view <y> . <rr%_nn1%_uh> no <y> : <"@> '	y
<mc> fourteen <nn2> guineas <dd1> each <y> (<jj> small <nn2> views <ii> in <np1> hampshire <y>) <y> ,	y
<nn1> responsibility <io> of <nn1> action <nn1> -- <appge> his <nn2> views <vvd> became <jj_vvn> modified <y> , <cs> till	y
<nn2> channels <y> . <ii> in <appge> my <da> own <nn1> view <y> , <cst_dd1> that <vvd> appeared <to> to	y
<y> . <vh0> have <vvn> gained <rr_da1> much <jjr> clearer <nn2> views <io> of <at> the <jj> christian <nn1> religion	y
Total Correct	18

B16. Event

Grammatical instances: (total=20)	**Correct?**
<io> of <nn1> priesthood <y> . <ii> in <at> the <nn1> event <io> of <appge> his <nn1> escape <y> ,	y
<vbi> be <vvn> extinguished <y> . <ii> in <at> the <nn1> event <io> of <appge> its <vvg_nn1%> ceasing <to> to	y
<rp> up <ii> around <pph1> it <ii> in <at> the <nn1> event <io> of <at1> an <nn1> attack <y> .	y
<pn1> one <pphs2> they <vvd> emptied <ii> in <at> the <nn1> event <io> of <at1> a <nn1> stickup <y> .	y
<vvd> awaited <ppho2> them <y> , <ii> in <at> the <nn1> event <io> of <np1> jefferson <ge> 's <nn1> success	y
<y> , <vvg> asking <y> , <ii> in <at> the <nn1> event <io> of <appge> your <xx> not <vvg> playing	y
<vm> will <vhi> have <ppho2> them <ii> in <at> the <nn1> event <io> of <nn1> war <y> . <ccb> but	y
<vbi> be <vvn> considered <rr> only <ii> in <at> the <nn1> event <io> of <nn1> failure <io> of <at> the	y
<at1> a <nnt1> year <y> . <ii> in <at> the <nn1> event <io> of <jj> heavy <nn2> layoffs <y> ,	y
<vvi> expedite <nn1> troop <nn2> movements <ii> in <at> the <nn1> event <io> of <nn1> war <y> , <cc> and	y
<to> to <vbi> be <vvn> destroyed <ii> in <at> the <nn1> event <io> of <at1> an <nn1> attack <ii> on	y
<nn1> matter <io> of <nnt2> months <ii> in <at> the <nn1> event <io> of <at1> a <jj> national <nn1> emergency	y
<vm> would <vbi> be <vvn> met <ii> in <at> the <nn1> event <io> of <nn1> unemployment <y> ; <at1> a	y
<cs21> so <cs22> that <y> , <ii> in <at> the <nn1> event <io> of <at> the <nn1> treaty <vbg> being	y
<nn1> law <y> , <"> " <ii> in <at> the <nn1> event <io> of <at1> a <nn1> failure <to> to	y
<if> for <at> the <nn2> americans <ii> in <at> the <nn1> event <io> of <at1> a <nn1> break <y> .	y
<jj_nn1> most-discussed <nn1_jj@> republican <nn1> ticket <ii> in <at> the <nn1> event <io> of <at1> a <nn1> convention <nn1_vv0> deadlock	y
<mc> two <nn2> eventualities <y> : <ii> in <at> the <nn1> event <io> of <np1> cyprus <vvg> falling <ii> into	y
<at> the <jj> presidential <nn1> office <ii> in <at> the <nn1> event <io> of <jj> final <nn1> failure <to> to	y
<at> the <nn1> president <cst> that <ii> in <at> the <nn1> event <io> of <nn1_jj> non-concurrence <ii> by <at> the	y
Total Correct	20
Lexical instances: (total=558)	
<jj> political <nn2> institutions <y> , <ii> to <at> the <nn2> events <io> of <appge> our <nn1> history <y> ,	y

<jj_vv0@> moderate <y> (<nn1_np1> aht <io> of <mcmc> 15-30 <nn2> events <ii> per <nnt1> hour <y>) <y> ;	y
<y> . <csa_ii@_rg@> as <db> all <dd2> these <jj> undesirable <nn2> events <vm> will <vhi> have <vvn> taken <nn1> place	y
<to> to <vbi> be <vvn> supported <ii> at <db> all <nn2> events <y> , <cs21> even <cs22> though <rr> somewhat	y
<jj> probable <csn> than <da2> many <jj> other <jj> unpleasant <nn2> events <cst_dd1> that <vm> can <vvi> happen <ii@> to	y
<cc> and <vvi> help <vv0> accelerate <vvg_nn1> thinking <nn1> anc <nn2> events <ii> toward <dd2> these <nn2> objectives <ii21> instead	y
<dd> any <nn2> difficulties <vvn> encountered <ii> during <dd1> this <nn1> event <y> . <null> <p> <ii> in <jj> obese	y
<vm> will <vbi> be <at> the <rgt> most <jj> scrumptious <nn1> event <ii> in <appge> my <nn1> life <y> .	y
<at> the <nn1> country <y> , <ii> to <at> the <nn2> events <ii> in <at> the <jj> far <nd1> west	y
<ii> at <at> the <nnt1> time <rrq> when <at> the <nn2> events <pphs2> they <vv0_nn1> recount <vbdr> were <vvg> taking	y
<vh0> have <vvn> mentioned <y> . <ii> at <db> all <nn2> events <y> . <iw> without <nn1> compliment <y> ,	y
<ge> 's <nn1> hypothesis <ii_vvg> concerning <at> the <jj> catastrophic <nn2> events <vvn_vvd> observed <ii> in <at> the <nn2> supernovae	y
<jj_nn1> present <nn1> course <nn1@_vm> will <ii> at <db> all <nn2> events <vv0> improve <appge> my <nn1> health <y> ,	y
<at> the <jj_nn1> human <nn1> voice <cc> and <jj_nn1> human <nn2> events <ppx2> themselves <ii> into <nn1> motion <nn1> picture	y
<nn1> regularity <ii> in <at> the <nn1> happening <io> of <nn2> events <y> . <ppis2> we <vv0> observe <cst_dd1> that	y
<vvn> represented <at1> a <rr> much <rgr> more <jj> common <nn1> event <y> , <cc> and <dd1> each <nn1> star	y
<ii> in <np1> america <y> . <"> " <at1> an <nn1> event <vvd> occurred <y> , <ii> at <dd1> this	y
<nn1> space <jj> spectacular <y> , <at1> a <jj> special <nn1> event <vvg> combining <jj> special <nn1> lighting <nn2> effects	y
<cst> that <at1> a <nn1> miracle <vbz> is <at> no <nn1> event <cst_dd1> that <vvz> transpires <rr> singly <cc> or	y
<pph1> it <vvz> pleases <to> to <vvi> govern <vvg_jj@> coming <nn2> events <cc> and <at> the <jj_nn1> future <nn1> conduct	y
<jj> white <nn1_jj> captive <y> ; <ii> in <dd1> this <nn1> event <pph1> it <vbz> is <vvn> referred <ii> to	y
<at> the <nn1> ocean <y> . <ii> in <dd2> these <nn2> events <y> , <np1> i. <vm> would <vvi> acknowledge	y
<at> the <nn1> state <nn1@_vm> will <ii> at <db> all <nn2> events <vv0> adhere <ii> to <at> the <nn1_np1> union	y
<jj> little <nn1> daughter <nn1> talk <io> of <dd1> this <nn1> event <y> , <vvz> goes <rr> understandably <nn2> bananas	y
<vhd> had <vvn> lost <y> . <ii> at <db> all <nn2> events <pphs1> he <vvd> thought <rr> so <y> ,	y
Total Correct	25

B17. Matter

Grammatical instances: (total=47)	**Correct?**
<vvg_nn1> flowering <ii> from <appge> its <nn1> seed <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <cs_rrq> when <jj> golden <nnt2> days	y
<ii> in <appge> his <jj> black <nn1> robe <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <pph1> it <vvz> gets <rr> really	y

<vvnk> bound <to> to <vvi> respect <y> . <cs> if <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <rl_ex> there <vhz> has <vbn> been <nn1> unfairness	y
<nn1> signalman <ii> on <dd1> each <nn1> boat <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <pphs1> he <vhd> had <vvn> told	y
<ccb> but <pphs2> they <vdd> did <xx> n't <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <ppis1> i <vv0> think <pph1> it	y
<io> of <at1> a <nn1> railroad <nn1> journey <y> ? <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <at> the <nn2> risks <vbr> are	y
<pphs1> she <vbdz> was <vvg@> warning <ppho1> him <y> ? <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <np1_nn1_jj> cutty <rr> never <vvd_vvn> heard	y
<ii> to <jj> timid <nn2> travelers <y> , <ccb> but <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <at1> a <nn1> steam-ship <vvz> goes <ii> to	y
<"> " <pphs1> he <vvd> said <y> . <"> " <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <ppis1> i <rr> only <vvd_vvn> met	y
<appge> your <nn1> excellency <y> . <uh> no <y> , <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> . <ppis1> i <vbm> 'm <vvg> calling	y
<jj> smooth <nn1> curve <y> , <cc> and <y> , <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <da2> several <da> such <nn2> deviations	y
<jj> mysterious <cc> and <jj> secret <nn1> mission <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <vhd> had <ppis1_mc1%> i <vvn> gone	y
<vvd> said <y> , <rr> quietly <y> , <"> " <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <ppy> you <vvd> knew <pphs2> they	y
<nn1> report <y> . <pphs1> he <vbdz> was <y> , <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <rr21> a <rr22> little <jj_vvn@_vvd@> annoyed	y
<io> of <ddq> which <pphs1> he <vbdz> was <y> , <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <at1> a <nn1_jj> self-respecting <cc> and	y
<io> of <nn1> helium <y> . <nn1> helium <y> , <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <vm> can <rr> always <vbi> be	y
<cs> if <ppy> you <vd0> do <z> -- <np1> hastings <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <ppis1> i <vd0> do <vvi> know	y
<vvn> affected <ii> by <at> the <nn1> law <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> ; <pph1> it <vbz> is <rr> well	y
<io> of <nn1> business <vvz> remains <jj_nn1@> constant <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <at> the <jj> continual <nn1> crop	y
<vbz> is <xx> not <rr> so <y> , <cc> and <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <pph1> it <vbz> is <rr_da1> much <jjr> easier	y
<pphs1> she <vbdz> was <vbg> being <vvn> followed <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <pphs1> she <vbdz> was <xx> not	y
<vvn_jj> marked <fo> + <cc> and <y> ? <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <pph1> it <vvz> makes <at> no	y
<nn1> credit <if> for <at> the <nn1> job <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <at> the <nn1> brain <nn1> trust <vhd> had	y
<to> to <vvi> hit <at1> a <nn1> nucleus <y> . <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <at> the <nn2> a-particles <vvn_vvd> used	y
<nn1@> making <io> of <appge> its <nn2> rates <y> ; <ii> as <at1> a <jj31_nn131> matter <jj32_nn132> of <jj33_nn133> fact <y> , <rr> however <y> , <vvg_jj> competing	y
	25
Lexical instances: (total=1384)	
<xx> not <vvi> know <pn1> anything <ii> about <dd2> these <nn2> matters <y> . <pphs2> they <vbr> are <if> for	y
<nn1> profit <y> , <cc> and <nn1> fear <io> of <nn1> matter <io> of <nn1> life <cc> or <nn1> death	y
<vvi> starve <ii> to <nn1> death <ii> in <at1> a <nn1> matter <io> of <nnt2> hours <cs> if <pph1> it	y
<ii> throughout <at> the <nd1> south <if> for <jj> incendiary <nn1> matter <y> , <iw> with <at1> a <jj_nn1> //	y
<nn1> item <rr> thus <vvn_vvd> traced <cc> and <at> the <nn1> matter <ddq> which <y> , <ii> at <dd> any	y

<cst> that <ex> there <vbdr> were <da2> many <jj> little <nn2> matters <to> to <vbi> be <vvn> attended <ii_to> to	y
<at> the <jj> usual <nn1> way <io> of <vvg_nn1> hearing <nn2> matters <vvn> submitted <ii> by <at> the <da2> several	y
<pphs2> they <vvd> continued <to> to <vvi> discuss <at> the <nn1> matter <dd> some <nnt1> time <y> , <y> ,	y
<nn2> views <io> of <at> the <da> same <jj> interesting <nn1> matter <y> . <ppis1> i <vv0> feel <ppx1> myself	y
<y> . <"> " <ddq> what <vbz> 's <at> the <nn1> matter <iw> with <appge> your <nn1> voice <y> ?	y
<at1> every <nnt1_jj@> minute <nn1> part <io> of <at> the <nn1> matter <vv0> remain <at> the <da> same <y> .	y
<vbdz> was <vvg> talking <iw> with <ppho1> him <ii_rp@> on <nn2> matters <io> of <nn1> importance <ii> at <at> the	y
<vvz> covers <dd1> this <nn1> trial <vbz> is <at1> a <nn1> matter <if> for <at> the <nn1> press <to> to	y
<to> to <vhi> have <vvn> ignored <at> the <jj> whole <nn1> matter <cc> and <vvn> taken <np1_nn1> garry <ge> 's	y
<rl> here <y> , <ddq> what <vbz> 's <at> the <nn1> matter <iw> with <np1> felix <y> ? <"> "	y
<y> . <"> " <ddq> what <vbz> 's <at> the <nn1> matter <y> ? <"> " <np1@_nn1> cutty <vvd> whispered	y
<at1> a <jjr> larger <pn1> one <rt> then <at> the <nn1> matter <ddq> which <vvz> makes <rp_ii@> up <at> the	y
<vbi> be <vvn> noted <y> , <vbz> is <at1> a <nn1> matter <io> of <nn1> degree <y> , <csa> as	y
<jj> cold <y> ; <cc> and <ii> in <db> all <nn2> matters <ii> in <ddq> which <pphs2> they <vhd> had	y
<pph1> it <y> . <mc> 2 <y> . <dd1> that <nn1> matter <ppx1> itself <y> , <cc> and <db> all	y
<iw> with <at> the <nn1> fact <cst> that <dd2> these <nn2> matters <vd0> do <vvi> involve <nn2> conversations <iw> with	y
<appge> his <nn1> cheek <cc> and <to> to <vvi> make <nn2> matters <rrr@> worse <pphs1> he <vbz> 's <vvg_jj> grinning	y
<pnqo> whom <ppis1> i <vvd> talked <io> of <da> such <nn2> matters <y> , <"> " <vvz> sets <rp> around	y
<fo> | <"> " <ddq> what <vbz> 's <at> the <nn1> matter <y> , <np1> lydia <y> , <ppy> you	y
<at> the <nn2> gentlemen <ii31> in <ii32> regard <ii33> to <nn2> matters <io> of <da1_rr> much <jjr> higher <nn1> concernment	y
Total Correct	25

B18. Means

Grammatical instances: (total=118)	**Correct?**
<vbz> is <vvn> made <to> to <vvi> flow <ii31> by <ii32> means <ii33> of <at1> a <nn1> generator <y> ,	y
<rr> both <nn2> employers <cc> and <vvz_nn2> employes <ii31> by <ii32> means <ii33> of <jj> fake <nn1> labor <nn2> unions	y
<vvg_nn1_jj@> setting <nn2> things <ii> to <nn2> rights <ii31> by <ii32> means <ii33> of <jj> physical <nn1> force <y> .	y
<dd1> this <nn1> effect <vbz> is <vvn> produced <ii31> by <ii32> means <ii33> of <at1> a <jj@_nn1> travelling <jj> electric	y
<nn2> wave-lengths <vvd_vvn> used <ddq> which <vbz> is <ii31> by <ii32> means <ii33> of <at> the <nn1> electron <nn1> microscope	y
<at1> a <nn1> wave <vbz> is <vvn> calculated <ii31> by <ii32> means <ii33> of <at> the <np1> de <np1> broglie	y
<nn1> merit <y> , <cc> or <vvn_vvd> expected <ii31> by <ii32> means <ii33> of <pph1> it <to> to <vvi> establish	y

<io> of <at> the <jj> magnetic <nn1> field <ii31> by <ii32> means <ii33> of <at> the <nn1> field <nn1@_jj> current	y
<vm> would <xx> not <vvi> slip <rp@> in <ii31> by <ii32> means <ii33> of <at1> a <nn1> crowd <y> .	y
<nn1> transformation <ii> to <jj> other <nn2> aspects <ii31> by <ii32> means <ii33> of <at1> a <nn1> probability <nn1> statement	y
<nn1> paddle <nn2_vvz> wheels <y> , <rrr_jjr> later <ii31> by <ii32> means <ii33> of <nn1> screw <nn2> propellers <y> ;	y
<nn1> potential <vbz> is <vvn_jj> applied <y> , <ii31> by <ii32> means <ii33> of <at1> a <jj> strong <jj_nn1> short-wave	y
<nn1> signal <vm> can <vbi> be <vvn> measured <ii31> by <ii32> means <ii33> of <at1> a <jj_nn1> cathode-ray <nn1> tube	y
<vbi> be <rgt> most <rr> reliably <vvn_vvd> measured <ii31> by <ii32> means <ii33> of <jj> standardized <nn2> tests <ii> at	y
<nn1> survey <io> of <at> the <nn2> skies <ii31> by <ii32> means <ii33> of <nn1> photography <vvz> indicates <cst> that	y
<cc> and <rr> perhaps <nn2> murderers <nn1_jj> -- <ii31> by <ii32> means <ii33> of <ppho2> them <y> . <pph1> it	y
<nn1> wireless <y> , <vbz> is <vdn> done <ii31> by <ii32> means <ii33> of <at1> a <jj> relay-like <nn1> feedback	y
<y> , <vm> can <vbi> be <vvn> counted <ii31> by <ii32> means <ii33> of <at1> a <nn1> counter <vvn> surrounded	y
<cc> and <at> the <jj> mechanical <nn2> devices <ii31> by <ii32> means <ii33> of <ddq> which <at> the <nn1> method	y
<pphs1> he <vvd@> determined <to> to <vvi> undertake <ii31> by <ii32> means <ii33> of <appge> his <da> own <nn2> funds	y
<pph1> it <vbdz> was <vvn> conducted <y> , <ii31> by <ii32> means <ii33> of <nn1> glass <nn2> pipes <y> ,	y
<y> , <vvg> gaining <at1> an <nn1> entrance <ii31> by <ii32> means <ii33> of <nn2> stairs <vvg> leading <rp> down	y
<at> the <nn1@_jj> negative <vbg> being <vvn> rectified <ii31> by <ii32> means <ii33> of <at> the <nn1> commutator <y> .	y
<nn1> machine <vm> should <vbi> be <vvn> invented <ii31> by <ii32> means <ii33> of <ddq> which <db> all <nn2> kinds	y
<rrt> best <vvn> described <cc> and <vvn_vvd> addressed <ii31> by <ii32> means <ii33> of <nn1> poetry <y> . <nn1> poetry	y
Total Correct	25
Lexical instances: (total=400)	
<nn1> government <vvz> gives <appge> its <nn2> possessors <at> the <nn> means <io> of <ii> per <y> ? <vvg_nn1@> petuating	y
<cc> and <pphs1> she <vhz> has <xx> not <jj> sufficient <nn> means <to> to <vvi> support <rr_db2> both <rr> comfortably	y
<ppis2> we <vm> shall <vhi> have <vvn@> become <at> the <nn> means <io> of <vvg> introducing <ii> to <at> the	y
<nn1> order <ddq> which <vvz> requires <cst_dd1> that <jj> appropriate <nn> means <vm> should <vbi> be <vvn> used <to> to	y
<nn2> signals <y> . <zz1> a <vhd> had <nn1> odor <nn> means <cs> if <ppy> you <vv0> smell <nn1> gas	y
<nn2> converts <vv0> see <ii> in <pph1> it <at> the <nn> means <io> of <vvg_nn1_jj@> turning <nn1> depression <ii> into	y
<ppho2> them <y> , <cc> and <jj_vvd> contrived <jj> various <nn> means <ii> to <appge> their <nn1> mitigation <y> ,	y
<vbz> is <rr> clearly <jj> unattainable <ii> by <dd> any <nn> means <ii> at <appge> our <nn1> disposal <y> .	y
<np1> peterson <nn1_vv0> sacrifice <y> ? <ii> by <db> all <nn> means <y> , <vv0> let <ppio2> us <vvi> get	y
<vbz> is <jj> bountiful <y> ; <ccb> but <dd2> these <nn> means <vbr> are <rr_rg> so <vvn_jj@>	y

distributed <y> ,	
<rt> before <y> , <cc> and <cs_rrq> when <at> the <nn> means <io> of <jj> agricultural <nn1> production <y> ,	y
<nn1> device <y> , <nn1> contrivance <y> , <cc> or <nn> means <vbr> are <vvn> imputed <ii> to <nn1> president	y
<io> of <appge> his <nn2> connections <ii> by <at1> every <nn> means <appge> your <nn2> affairs <ii> to <at> the	y
<to> to <vvi> supply <ppio1> me <iw> with <at> the <nn> means <io> of <vvg> completing <appge> my <jj> proper	y
<ppis1> i <vvd> concluded <cs_csw@> if <ppis1> i <vhd> had <nn> means <io> of <appge> my <da> own <y> ,	y
<vvn@_jj@> misdirected <y> , <cc> or <cst> that <at> the <nn> means <ii> at <nn1> hand <vbdr> were <rr> generally	y
<ppis1> i <vm> would <rr> somehow <vvi> find <at> the <nn> means <to> to <vvi> prove <pph1> it <y> .	y
<y> . <uh> yes <y> , <ii> by <db> all <nn> means <vv0> get <at> the <nn1> watch <y> .	y
<y> ; <at1> a <nn1> neglect <io> of <at> the <nn> means <io> of <vvg> creating <nn2> seamen <if> for	y
<if> for <vvg_jj@> laboring <nn2> men <io> of <jj> limited <nn> means <to> to <vvi> acquire <nn1> property <ii> in	y
<nn1> son <y> . <ii21> as <ii22> to <appge> my <nn> means <io> of <nn1> knowledge <y> , <cs_rrq> when	y
<nn1> vat <vdz> does <"> " <vvz> elopes <"> " <nn> means <y> ? <nn1_vv0@> bunk <y> . <nnb> major	n
<ii> by <ddq> which <to> to <vvi> earn <jj> sufficient <nn> means <to> to <vvi> enable <ppio2> us <to> to	y
<nd1> south <y> , <io> of <at> the <jj> ordinary <nn> means <ii> at <appge> its <nn1> command <y> .	y
<rr> neither <at> the <nn1@> will <cc> nor <at> the <nn> means <to> to <vvi> enforce <at> the <jj@_nn1@> antidumping	y
Total Correct	24

B19. Order

Grammatical instances: (total=340)	**Correct?**
<y> , <ddq> which <vbdz> was <vdn> done <bcl21> in <bcl22> order <to> to <vvi> make <at> the <nn1> ground	y
<jj> intelligent <nn1> use <io> of <nn1> a.i. <bcl21> in <bcl22> order <to> to <vvi> supply <jj> large-scale <nn2> fatteners	y
<nn1> way <io> of <nn1> business <y> , <bcl21> in <bcl22> order <to> to <vvi> supply <ppio1> me <iw> with	y
<xx> not <rt> now <np1> expedient. <y> ? <bcl21> in <bcl22> order <to> to <vvi> decide <csw> whether <at> the	y
<vvd> called <ii> upon <appge> my <nn1> father <bcl21> in <bcl22> order <to> to <vvi> consult <ppho1> him <ii31> with	y
<vvg> exercising <appge> its <vvg_jj%> exempting <nn2> powers <bcl21> in <bcl22> order <to> to <vvi> give <rr> otherwise <jj> unlisted	y
<nn2> companies <vvn_vvd> needed <to> to <vvi> combine <bcl21> in <bcl22> order <to> to <vvi> compete <cs> after <pphs2> they	y
<at1> a <nn1> separation <vvd_vvn> meant <y> , <bcl21> in <bcl22> order <to> to <vvi> serve <appge> her <nn2> interests	y
<nn1> heart <ii> against <pph1> it <"> " <bcl21> in <bcl22> order <to> to <vvi> emphasize <at> the <nn1> damage	y
<jj> virile <nn2> neighbors <y> . <ccb> but <bcl21> in <bcl22> order <to> to <vvi> see <ddq> what <jj_nn1> corrective	y

<at> the <jj%> produced <cc> and <y> , <bcl21> in <bcl22> order <to> to <vvi> obtain <at1> a <jj> progressive	y
<np1> philadelphia <ge> 's <np1_nn1@> temple <nn1> university <bcl21> in <bcl22> order <to> to <vvi> conduct <at> the <zz1_jj> u	y
<ccb> but <cs> when <y> , <"@> ' <bcl21> in <bcl22> order <to> to <vvi> form <at1> a <rgr> more	y
<jj> electrical <cc> or <nn1_vvg@> plumbing <nn2> businesses <bcl21> in <bcl22> order <to> to <vvi> moonlight <y> , <cs21> rather	y
<y> , <cc> and <nn1_vv0@> trade <nn2> practices <bcl21> in <bcl22> order <to> to <vvi> safeguard <nn1> solvency <cc> and	y
<appge> your <nn2> pets <ii> to <nn2> pieces <bcl21> in <bcl22> order <to> to <vvi> find <rp> out <appge> their	y
<y> ; <cs@_rr> so <pph1> it <vbdz> was <bcl21> in <bcl22> order <to> to <vvi> search <appge> her <nn2> rooms	y
<ii> inside <at> the <jj_nn1> red <nn2> giants <bcl21> in <bcl22> order <to> to <vvi> ascertain <at> the <jj> physical	y
<jj> necessary <ii> to <db> all <y> , <bcl21> in <bcl22> order <to> to <vvi> secure <at1> a <jj@> confirmed	y
<cc> and <nn2> crops <vvn_vvd> swept <rl> away <bcl21> in <bcl22> order <to> to <vvi> save <nn2> buildings <cc> and	y
<vbdz> was <vvg> talking <ii> to <y> , <bcl21> in <bcl22> order <to> to <vvi> judge <rr> fairly <y> .	y
<pph1> it <vvz> follows <cst_dd1> that <y> , <bcl21> in <bcl22> order <to> to <vvi> become <jj_nn1@> stable <y> ,	y
<nn1> son <y> , <rt> then <y> , <bcl21> in <bcl22> order <to> to <vvi> bring <ppho1> him <rp> back	y
<pn1> something <ii> in <mc> 1.787 <y> . <bcl21> in <bcl22> order <to> to <vvi> prevent <jj> rational <nn1> thinking	y
<cc> and <vvg_jj@> receiving <nn1> baptism <y> , <bcl21> in <bcl22> order <to> to <vvi> become <appge> his <nn1> disciple	y
Total Correct	25
Lexical instances: (total=960)	
<vm> would <vbi> be <vvn> stopped <ii> by <at> the <nn1> order <y> .	y
<pph1> it <y> : <zz1_at1@> a <jj> spatial <at1> an <nn1> order <io> of <nn1_vvg> reading <y> . <ppis1> i	y
<io> of <at> the <nn2> natives <ii_cs> after <jj> good <nn1> order <cc> and <nn2> morals <y> ; <cc> and	y
<"> " <"> " <ppis1> i <vv0> give <ppy> you <nn2> orders <y> , <nn1_jj> general <y> . <ppy> you	y
<vvi> put <appge> its <jj> financial <nn1> house <ii> in <nn1> order <cs_ii@> since <at> the <jj> polish <nn1> debt	y
<jj> basic <nn1> conformity <iw> with <at> the <jj> existing <nn1> order <y> , <vvn> accentuated <ii> by <at1> an	y
<y> . <"> " <pphs1> he <vvd> gave <appge> their <nn1> order <y> , <rt> then <vvd_vvn@> recounted <if> for	y
<at> the <nn1> manner <ii> in <ddq> which <at1> an <nn1> order <vbdz> was <fo> | <vvn_jj@> given <cc> and	y
<y> , <cc> and <vvd> said <cst> that <at> the <nn2> orders <io> of <at> the <nn1> commander-inchief <vbdr> were	y
<at> the <nn1> viceroy <vvd> gave <at> no <jj> decisive <nn1> order <y> ; <ccb> but <rr> merely <vvd_vvn@> said	y
<vhd> had <xx> not <vvn> given <ppio1> me <at1> an <nn1> order <y> . <ccb> but <at> the <nn2> strangers	y
<nn1> existence <io> of <at> the <da2> many <jj> distinct <nn2> orders <cc> and <nn2> establishments <ii> in <at> the	y
<vm> can <vvi> relish <y> . <ii> from <at> the <nn1> order <cc> and <nn1> regularity <io> of <nn2> things	y

<vvi> put <appge_ppho1> her <jj> naval <nn1> house <ii> in <nn1> order <y> ; <cc> and <rl> here <vbdz> was	y
<ppho2> them <nn2> rookies <vm> wo <xx> n't <vvi> obey <nn2> orders <y> . <ppy> you <vv0> see <y> ,	y
<vvz_nn2> states <y> , <cst> that <"> " <at> the <nn1> order <io> of <nn1> battle <vvn_vvd@> required <ppho2> them	y
<y> , <at> the <nn1> effect <io> of <appge> his <nn1> order <y> , <vvn> given <ii> in <nn1> violation	y
<at> the <jjt> best <nn2> types <io> of <jj> social <nn1> order <y> , <ii21> such <ii22> as <at> the	y
<ii> toward <at1> a <nn1> novelty <io> of <dd1> this <nn1> order <y> . <jj> famous <nn2> names <vbdr> were	y
<at> the <nn1> senate <y> ; <cc> and <dd> any <nn2> orders <cc> or <nn2> instructions <vvg> relating <ii> to	y
<cst> that <pphs1> he <vhd> had <vvn> issued <at1> an <nn1> order <if> for <at> the <nn1> removal <io> of	y
<y> . <ccb> but <nn1> competition <io> of <dd1> this <nn1> order <vbdz> was <jj> foreign <ii> to <np1> fabian	y
<vvd> asked <ppio1> me <y> . <"> " <ddq> what <nn1> order <vd0> do <ppy> you <vvi> refer <ii> to	y
<ppis1> i <vbdz> was <ii> under <at> no <jj> specific <nn2> orders <y> . <ccb> but <nn1> lieutenant <nn1> dial	y
<jj_nn1> general <jj> commanding <to> to <vvi> change <at> the <nn1> order <io> of <nn1> attack <ii> in <dd> some	y
Total Correct	25

B20-A. Back

Grammatical: (total=2662)	**Correct?**
<y> . <at> the <nn1> news <vvd_vv0_vvn> set <ppio2> us <rp> back <at1> a <nn1> pace <cc> or <mc> two	y
<ge> 's <nn1> stomach <csa> as <pphs1> he <vvd> walked <rp> back <ii> to <appge> his <pphs1> he <vvn> heard	y
<y> . <"> " <pphs1> he <vm> 'll <vbi> be <rp> back <rr> soon <y> . <at> the <nn1> smell	y
<vvn> passed <y> . <rr> surely <np1> judd <vbdz> was <rp> back <ii> from <rl_jj> downtown <y> . <at> the	y
<y> , <vvg> permitting <np1_nn1@_vv0@> mendel <to> to <vvi> slip <rp> back <ii> into <at> the <jj> muddy <nn1> water	y
<y> , <ccb> but <ppis1> i <vbm> 'm <vvg> driving <rp> back <rt> tomorrow <y> . <np1> roger <vvd> invited	y
<rl> here <!> ! <"> " <pphs1> he <vvd> pushed <rp> back <appge> his <np1> panama <nn1> hat <y> .	y
<y> . <"> " <vv0> send <cst_dd1> that <vv0@_nn1> dish <rp> back <rl> here <y> , <np1_nn1> breen <y> ,	y
<nn1> emperor <y> . <null> <np1> rahl <vvd> glanced <rp> back <ii_csa> as <at> the <nn1> frigate <vvd_vvn> pitched	y
<y> . <to> to <vvi> bring <at> the <nn1> argument <rp> back <ii> to <nn2> houses <y> : <ii> in	y
<vv0> suppose <ppy> you <vv0> mean <to> to <vvi> come <rp> back <rr> soon <y> ? <"> " <"> "	y
<io> of <at> the <nn1> hill <cc> and <vvd_vvn> looked <rp> back <ii@_rp> down <at> the <nn1> slope <y> ,	y
<np1_nn1@> kathie <!> ! <uh> o-o-o-h <!> ! <vv0> bring <rp> back <np1> kath-i-e <!> ! <"> " <vvi> lay	y
<vvn> begun <to> to <vvi> bring <at> the <nn1> line <rp> back <ii> into <jjr> clearer <nn1> sailing <nn1> weather	y
<rp> back <vv0_nn1> stand <rp_nn1> back <!> ! <vv0> stand <rp> back <!> ! <np1@_nn1@> will <y> . <np1@> will	y

<nn1> bailiff <y> , <vv0> take <nnb> mr. <np1> choo <rp> back <ii> to <vvg_nn1> booking <rr_cs@> so <pphs1> he	y
<jj> storm-bound <cc> and <vm> could <xx> n't <vvi> get <rp> back <ii> to <at> the <nn1> cabin <cst_dd1> that	y
<io> of <nn1> reminiscence <y> . <pphs1> he <vvd> sat <rp> back <ii> in <appge> his <jj> low <nn1> arm-chair	y
<at> the <nn1> street <y> , <cc> and <vvd> walked <rp> back <y> . <at> the <nn1> bell <vdd> did	y
<rp> off <ii> like <at> the <nn2> swallows <vvg> going <rp> back <ii> to <np1_nn1> capistrano <y> . <ppis1> i	y
<vvd_vvn> found <ppio1> me <cc> and <vvd> brought <ppio1> me <rp> back <ii> to <appge> his <nn1> piazza <y> ,	y
<nnl1> street <y> . <csa> as <pphs1> he <vvd> settled <rp> back <cc> and <vvd> took <ii> in <at> the	y
<nn1> baby <y> , <cc> and <vvg> keeping <ppio2> us <rp> back <dd1> that <nn1> way <y> . <md> first	y
<jj> flat-out <nn1> gallop <cc> and <y> , <vvg> looking <rp> back <y> , <vvd_vvn_jj@> caught <nn1> sight <io> of	y
<vvd> seemed <jj_rr@> likely <to> to <vvi> sweep <ppio1> me <rp> back <rr> dangerously <ii_vv0@_rl> near <dd1_cst> that <nn1_vv0@> shore	y
<at1> a <ppy> you <vd0> do <xx> n't <vvi> go <rp> back <y> ? <"> " <"> " <at> the	y
<vvi> fool <rp> around <cc> and <rt> then <vv0> come <rp> back <y> , <vm> will <ppy> you <y> ?	y
<y> . <appge> his <nn1> mind <vbdz> was <rr> always <rp> back <ii> at <at> the <nn1> house <cs_rrq> where	y
<rp> off <pphs1> he <vvz> goes <y> , <vvg> waving <rp> back <rr> bravely <y> . <rr> regardless <io> of	y
<io> of <nn1> pain <y> , <pphs1> she <vvz> comes <rp> back <ii> to <ppio1> me <ii> in <dd1> this	y
<at> the <rr> already <vvn> forgotten <y> . <vv0> look <rp> back <y> , <vv0> hold <rp@_ii> on <y> :	y
<pphs1> he <vbz> is <vvg> thinking <ii> about <vvg> going <rp> back <ii> to <nn1> school <to> to <vvi> get	y
<vm> could <xx> not <vvi> reconsider <cc> and <vvi> get <rp> back <ii> in <at> the <nn1> race <y> .	y
<rr> just <rg> as <jj> bright <z> -- <vv0> take <rp> back <at> the <jj> ugly <nn2> things <!> !	y
<rrq> why <vd0> do <xx> n't <ppy> you <vvi> go <rp> back <ii> to <pph1> it <y> ? <"> "	y
<ii> in <np1> peter <ge> 's <cc> and <vvd> threw <rp> back <appge> her <jj> heavy <nn1> crepe <nn1_vv0> veil	y
<io> of <at> the <nn1> band <vhd> had <vvn> come <rp> back <iw> with <at> the <nn2> leaders <y> ,	y
<vvg> lifting <appge> her <nn1> head <cc> and <vvg> leaning <rp> back <cs21> so <cs22> that <pphs1> she <vm> could	y
<at> the <nn1> light <rp_ii> off <cc> and <vvd> went <rp> back <ii> into <at> the <nn1> sitting <nn1> room	y
<nn2> men <vdd> did <pn1> nothing <cs> but <vvi> shout <rp> back <rr> uncertainly <cc> and <nn1> mill <ii> towards	y
<y> , <cc> and <pphs2> they <vvd> pulled <ppho1> him <rp> back <ii> by <at> the <nn2> heels <y> .	y
<y> . <"> " <np1@_nn1> padua <vvd> gave <ppho1> him <rp> back <at> the <nn1> smile <y> . <"> "	y
<vvd_vvn> learned <cst> that <at> the <nn1> princess <vbdz> was <rp> back <ii> in <nn1> town <cs_rrq> when <at> the	y
<pn1> anyone <jj> special <y> ? <np1@_nn1> artie <vvd> held <rp> back <at> the <nn1> news <ii> about <np1> judd	y
<rr> shortly <rt> thereafter <y> , <pphs1> she <vvd@> called <rp> back <iw> with <at1> a <nn1> bit <io> of	y
<ppho1> him <rr> half-way <y> . <ccb> but <vvg> glancing <rp> back <rt> again <ii> before <vvg> passing <rp> out	y

<y> . <"> " <ppis1> i <vh0> have <vvn> come <rp> back <ii> to <ppy> thee <y> , <"> "	y
<nn1> window <pphs1> he <vvd> watched <ppho1> her <vvg> going <rp> back <ii> in <at> the <nn1> snow <cs_ii@> until	y
<np1> beijing <y> , <np1_nn1> huang <vbdz> was <vvn> sent <rp> back <ii> to <np1_nn1_jj> beihai <cc> and <vvn> detained	y
<nn1> plane <y> . <cs_rrq> when <pphs1> he <vvd> ran <rp> back <to> to <vvi> jump <pphs1> he <vvd> found	y
Total Correct	50
Lexical: (total=574)	
<y> . <at> the <nn1> flesh <io> of <appge> her <nn1> back <vbdz> was <jj> warm <cc> and <jj> alive	y
<dd> some <ppis1> i <vv0> want <appge> my <nn1> money <nn1> back <"> " <vdd> did <xx> not <vvi> bother	n
<jj> dull <jj_nn1> flat <nn1> sound <ii> at <at> the <nn1> back <io> of <at> the <nn1> place <y> ,	y
<vvd> watched <at> the <nn1> pertormance <ii> from <at> the <nn1> back <md> last <nn1> row <io> of <nn2> seats	y
<vvn> assisted <ii> to <appge> my <nn1> horse <ge> 's <nn1> back <y> , <cc> and <vvd_vvn> recommended <to> to	y
<vvi> show <rl_ex> there <vbz> is <at1> a <nn1> passageway <nn1> back <io> of <at> the <nn2> falls <y> ,	y
<vbn> been <at1> a <nn1> question <ii> in <at> the <nn1> back <io> of <appge> my <nn1> mind <y> ,	y
<y> . <np1@_nn1> pacomo <y> (<vvg> rubbing <appge> his <nn1> back <y>) <rr> perfectly <y> , <nn1> sir	y
<jj> startling <jj> black <nn2> seams <ii@_rp> up <at> the <nn1> back <y> . <np1> dick <np2> meyers <vbdz> was	y
<iw> with <appge> his <nn1> hat <ii> on <at> the <nn1> back <io> of <appge> his <nn1> head <y> ,	y
<cc> and <vvd> threw <pph1> it <rp_ii@> down <at> the <nn1> back <io> of <appge> his <nn1> throat <y> .	y
<appge> her <nn1> hair <ge> 's <nn1> growin <ge> ' <nn1> back <y> . <"> " <"> " <nn1_np1@> atta	n
<appge> his <jj%> clasped <nn2> hands <ii> behind <appge> his <nn1> back <y> , <cc> and <vvd> rocked <ppx1> himself	y
<jj@> crowding <nn2> beasts <y> , <ii> over <ddqge> whose <nn2> backs <pphs2> they <vbdr> were <vvn> obliged <to> to	y
<vm> must <vhi> have <nn2> eyes <ii> in <at> the <nn1> back <io> of <appge> her <nn1> head <y> .	y
<nn1_vvg> turning <cc> and <nn1> hair <ii> on <at> the <nn1> back <fo> ofp173it <cc> and <jj> soft <nn1@_rl> inside	y
<vm> should <vbi> be <vvn> shot <ii> in <at> the <nn1> back <cc> and <vv0> die <ii_rp@> in <nn1> disgrace	y
<at> the <nn1> counter <rl@_ii> beyond <y> , <appge> their <nn2> backs <ii> to <at> the <nn1> door <y> ,	y
<pphs1> he <vvd> knew <cst_dd1> that <ii> behind <appge> his <nn1> back <pphs2> they <vhd> had <xx> n't <at1> a	y
<at1> a <jj> brown <nn1> hat <ii> on <at> the <nn1> back <io> of <appge> his <nn1_jj@> head <nn1> reading a	y
<nn1> achievement <rr> merely <vvz> rides <ii> on <at> the <nn1> back <io> of <nn1> intelligence <y> : <"> "	y
<io> of <nn2> pains <jj> rheumatic <ii> in <appge> his <nn1> back <y> ; <cc> and <md> last <nnt1> year	y
<nn2> pajamas <cc> and <vvn> lay <ii_rp@> on <appge> his <nn1> back <ii> in <appge> his <nn1> underwear <y> ,	y
<vvn_vvd> arched <ii> like <at1> a <nn1> dolphin <ge> 's <nn1> back <y> , <vbdr> were <vvn@_jj@> sandalled <y> ;	y

<nn1> noise <pphs1> she <vvd> made <ii> in <at> the <nn1> back <io> of <appge> her <nn1> throat <cs_rrq> when	y
<pphs1> she <vvd> patted <ppho1> him <ii> on <at> the <nn1> back <y> . <"> " <vv0> bless <appge> your	y
<y> , <"> " <ccb> but <ii> in <at> the <nn1> back <io> of <appge> our <nn2> minds <ppis2> we	y
<ii> in <at> the <nn1> niche <io> of <appge> its <nn1> back <y> , <at> no <nn1> choice <vv0_jj_rr@> open	y
<ii> up <at> the <nn1> hill <ii> at <at> the <nn1> back <io> of <at> the <nn1> fort <y> .	y
<rp> out <cc> and <vvn_vvd> laid <ii> on <at> the <nn1> back <io> of <nn1> mouth <cs_rrq> when <at> the	y
<jj> draughty <nn1_jj> country-house <y> , <cs_rrq> where <appge> your <nn1> back <vm> would <vvi> freeze <cs> while <appge> your	y
<db2> both <y> . <vvn_jj> seated <iw> with <appge> his <nn1> back <ii> to <at> the <nn1> wall <vbdz> was	y
<vvg> hiding <at1> a <nn1> cigarette <ii> behind <appge> his <nn1> back <y> , <cc> and <nn1> grandma <vbdz> was	y
<pphs2> they <nn1_vv0@> don <ge_"@> ' <zz1> t <vv0> come <nn1> back <ge> ' <nn1_vv0_jj> // <dd1_rr> either <nn1> figure	n
<y> . <pphs1> he <vvd> sat <iw> with <appge> his <nn1> back <ii> to <at> the <jj> broad <nn1_jj@> spruce	y
<nnb> lieutenant <np1> trevor <y> , <vvg> turning <appge> his <nn1> back <ii> on <at> the <nn2> belles <io> of	y
<jj> short <jj_nn1> crisp <nn1> hair <ii> along <at> the <nn1> back <io> of <appge> his <nn1> head <y> ,	y
<y> , <vvg> prodding <ppho1> him <ii> in <at> the <nn1> back <iw> with <at1> a <nn1> stick <vvn_vvd> fashioned	y
<nn2> stairs <pphs2> they <vvd> went <ii> behind <appge> his <nn1> back <cc> and <vvn_vvd> returned <ii> to <at> the	y
<rp> out <appge> her <nn1> hand <if> for <at> the <nn1> back <io> of <nn1> fee <ge> 's <nn1> chair	y
<cst> that <vvd> led <ii> to <at> the <nn1> bedroom <nn1> back <io> of <ppio1> me <y> . <"> "	?
<ge> ' <at> the <nn1> hall <at1> an <ge> ' <nn1> back <y> . <vv0> step <jj> lively <y> .	?
<appge> your <nn2> britches <vbz> is <vvg> comin <ge> ' <nn1> back <y> . <pphs1> he <vvz> shrieks <iw> with	n
<ii> in <at> the <jj_nn1%> small <io> of <appge> his <nn1> back <cs> until <pphs1> he <vhd> had <vvn> convoyed	y
<np1> tom <cc> and <np1_nn1_vv0> huck <y> , <iw> with <nn2> backs <ii> to <nn1> camera <y> , <vvg> standing	y
<ii> on <at> the <nn2> shoes <ii> in <at> the <nn1> back <io> of <appge> my <nn2> parents <ge> '	y
<ppis1> i <vvd_vv0> put <ppho1> him <ii_rp@> on <appge> his <nn1> back <ii> on <at> the <nn1> highway <y> .	y
<vv0> carry <nn2> babies <vvn_vvd> bound <ii> to <appge> their <nn2> backs <iw> with <nn1> cloth <y> . <cc> and	y
<ii> in <jj_nn1@> white <ii> on <at> the <nn1> mirror <nn1> back <io> of <at> the <nn1> bar <y> .	y
<np2> mis <ge> ' <np1> abel <y> , <appge> my <nn1> back <vhz> 's <vbn> been <nn1> behavin <ge> '	y
Total Correct	44

B20-B. Back

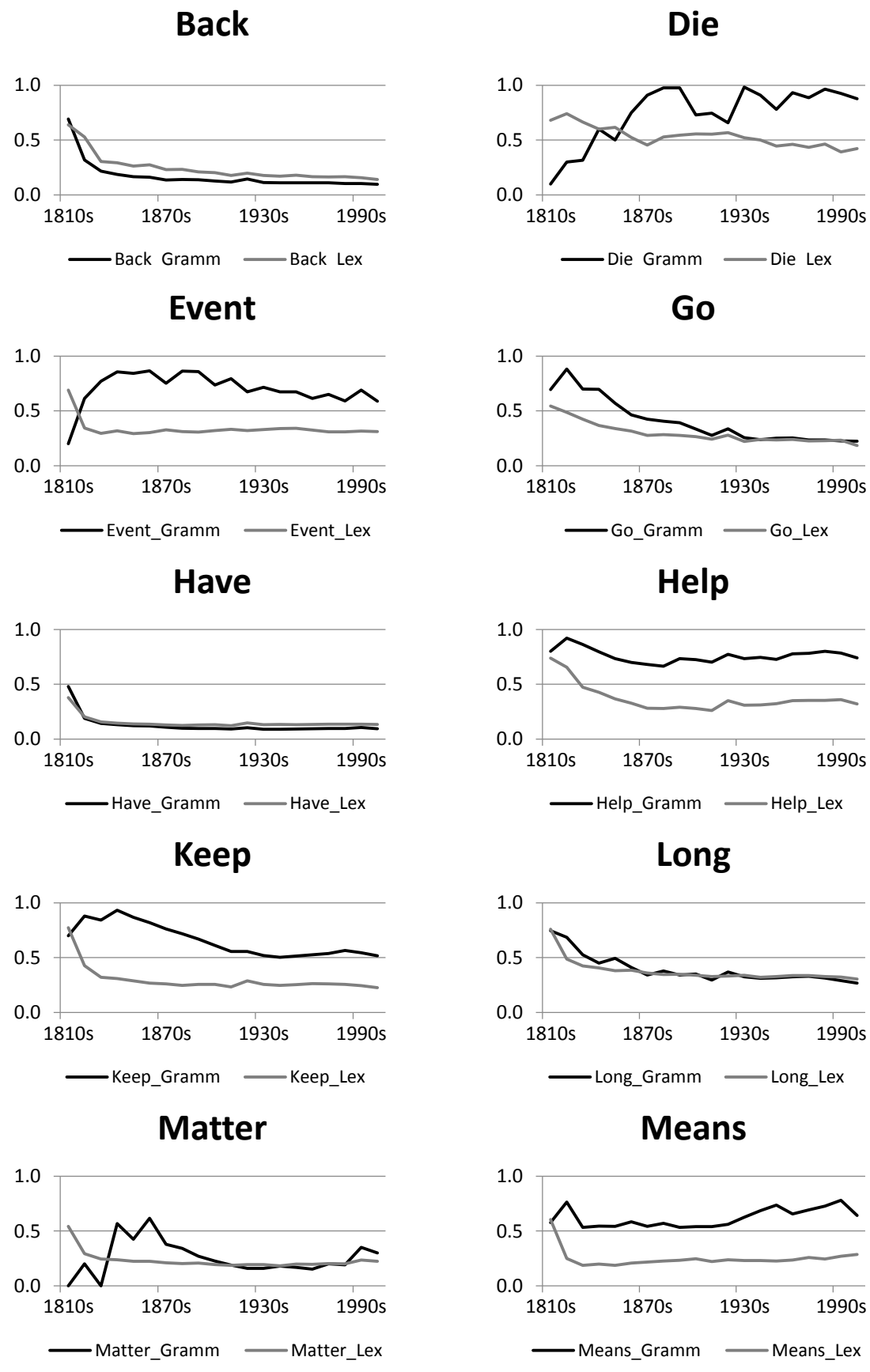
Ambiguous NN: (total=928)	**Correct?**	**Tag**
<cc> and <ppis1> i <vv0> want <dd2> those <nn2> beads <rp_nn1_jj@> back <nn1> beforecourt <vvz> gets <jj> wise <y> .	y	rp

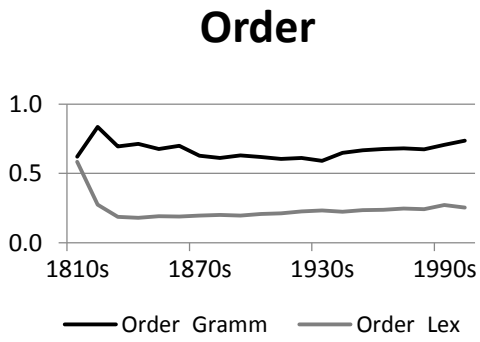
<vvd_vvn> stopped <y> , <rr> only <to_ii> to <vvi> crawl <rp_nn1> back <cc> and <vbi> be <vvn> swallowed <ii> by	y	rp
<cc> and <at1> a <nn1> jump <np1> bruce <vbz> is <rp_nn1> back <ii> to <appge> his <nn1> car <y> .	y	rp
<y> , <iw> with <appge> her <nn1> tail <cc> and <nn1_rp_jj@> back <vvg_nn1@> bristling <y> . <pphs1> she <vvd> made	y	nn1
<csa> as <ppis2> we <vv0> fade <rp> out <cc> and <nn1_rp_vv0%> back <rp@_ii> in <ii> to <y> : <dd1> another	n	rp
<to> to <vvi> throw <at> the <nn1> incident <nn1> particle <nn1_rp> back <y> , <vvg> producing <at1> an <jj> ordinary	n	rp
<csa> as <pphs1> she <vvd> faced <at> the <"> " <nn1_jj@> back <nn1> door <y> , <"> " <vvn_vvd> stretched	y	jj
<ii> from <nn1> dark <ii> to <nn1> light <cc> and <rp_nn1> back <rt> again <y> . <ex> there <vbdr> were	y	rp
<y> , <vvg> drawing <at> the <jj_nn1> red <nn2> curtains <rp_nn1> back <y> ; <vvg_nn1> opening <cc> and <vvg> shutting	y	rp
<cc> and <nn2> associations <y> . <ppis1> i <vvd> came <rp_nn1> back <vvd_jj@> refreshed <if> for <nn1> work <y> ,	y	rp
<nn2> nerves <y> . <ppy> you <vv0> go <ii_rp@> on <nn1_rp> back <y> . <"> " <ppis1> i <vvd> offered	n	rp
<vvi> step <ii21> out <ii22> of <ppge> ours <y> . <nn1_rp> back <y> , <nn2> sirs <y> , <rp_nn1> back	n	rp
<nn1@> laugh <y> . <"> " <np1> claudius <vvg> coming <rp_nn1> back <y> ? <"> " <pphs1> he <"> "	y	rp
<rr31> so <rr32> much <rr33> as <vv0> turn <ppho1_appge> her <rp_nn1> back <y> , <cc> and <xx> not <vvg_jj> sleeping	n	nn1
<cc> and <pphs1> he <vbz> 's <vvg> hugging <ppho1_appge> her <rp_nn1> back <rt> now <-> - <at1> a <jj> real	y	rp
<y> . <pphs1> he <vvd> pulled <at> the <nn1> chair <rp_nn1> back <rr21> a <rr22> little <iw> with <at1> a	y	rp
<ii> from <ddq> what <ppis1> i <vv0> hear <y> , <rp_nn1> back <ii_to@> to <nn1_vvi> work <cc> and <nn1> feelin	y	rp
<np1> johnny <nn2> two-hawks <y> ? <"> " <"> " <nn1_rp> back <ii> to <nn1> earth <y> . <"> "	n	rp
<vvn> attached <y> . <rl> here <y> , <cc> and <nn1_rp> back <pphs1> he <vvd> swung <ii_rp@> in <at1> a	n	rp
<appge> my <jj> old <nn2> friends <ii> in <at> the <nn1_jj@> back <nn2> districts <z> -- <cst_dd1> that <vbz> is	y	jj
<y> . <cs> while <ii> on <appge> their <nn1> way <rp_nn1> back <ii> to <np1> canarsie <at> the <jj_nn1> auxiliary	y	rp
<vvd> bayed <mc1> one <nn1> lion <ii> at <at> the <nn1_jj@> back <nn1> end <io> of <at1> a <jj_nn1> 40-foot	y	jj
<y> , <pphs1> he <jj_nn1> - made <appge> his <nn1> way <rp_nn1> back <ii> into <at> the <jj> mountainous <nn1> country	y	rp
<nn1> effort <ddq> which <vvd> drew <ppho1_appge> her <vv0_nn1> face <rp_nn1> back <ii> into <appge> its <jj> accustomed <nn2> lines	y	rp
<ii> from <at> the <nn1> vegetable <nn1> garden <rp> out <rp_nn1> back <y> , <ppy> you <vm> can <vvi> look	y	rp
<mc> two <nn2> fires <y> . <at> the <nn1> way <rp_nn1> back <ii> to <at> the <nn1> lake <vbdz> was	y	rp
<vv0_nn1> form <at1> a <nn1> crescent <ii> in <at> the <nn1_jj@> back <nn1> ground <y> , <cc> and <vv0@_nn1> level	y	jj
<y> . <pphs1> he <vvd> groped <appge> his <nn1> way <rp_nn1> back <ii> to <at> the <nn1> taffrail <y> .	y	rp
<io> of <nn1_jj@> cut <nn2_vvz@> flowers <ii> in <at> the <nn1_jj@> back <z> -- <jj> white <nn1> iris <y> ,	y	nn1
<nn1> father <vvi> walk <rl> ahead <ii> toward <at> the <nn1_jj@> back <nn1> door <cc> and <rt> then <vvd> ran	y	jj
<nn2> borders <vbr> are <vvn> pressed <rr> still <rrr_jjr@> further <rp_nn1> back <y> ,	y	rp

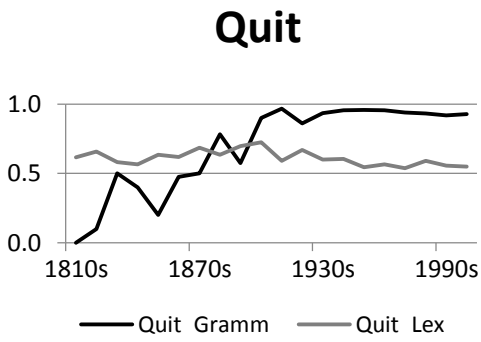
<cc> and <ii> by <at> the		
<to> to <vvi> kill <ii> before <vvg> meeting <np1> quinn <nn1_rp> back <ii> at <at> the <nn1> office <y> ,	n	rp
<jj> imaginary <nn1> conversation <ii> on <np1> joe <ge> 's <nn1_jj@> back <nn1> lawn <y> , <pphs1> he <vvz> feels	y	jj
<np1> jersey <nn2> primaries <at1> a <da2> few <nnt2> months <rp_nn1> back <cc> and <vvd_vvn> resigned <ii> as <nn1> head	y	rp
<vbdz> was <vvn> gone <nn1_jj> -- <rp_ii%> out <at> the <nn1_jj@> back <nn1> door <y> . <"> " <vv0> look	y	jj
<vbz> is <vvg> looking <rl> ahead <y> , <xx> not <rp_nn1_jj@> back <y> . <np1> mason <ge> 's <nn1> foot	y	rp
<vvg> naming <at> the <nn2> weeds <ii> in <at> the <nn1_jj@> back <nn1> yard <cc> and <rt> then <vvg> saying	y	jj
<nn2> teeth <y> , <vvg> sliding <at> the <nn1_np1> leatherman <rp_nn1> back <ii> onto <appge> his <nn1> belt <y> .	y	rp
<vvd> remembered <ii> from <at1> a <jj> long <nnt1> time <rp_nn1> back <y> . <pphs1> she <vhd> had <vvn> believed	y	rp
<at> the <nnb> senator <np1> caldwell <nn1> murder <y> . <nn1_rp> back <iw> with <nn2> details <ii_cs> after <dd1> this	n	rp
<y> , <to> to <vvi> throw <pph1> it <rr_jj> far <rp_nn1> back <ii> into <at> the <jj> poignant <nn1> nothingness	y	rp
<nn1> table <cc> and <vvd> loosened <at> the <nn1> hair <rp_nn1> back <ii> from <appge> her <nn1> face <y> .	y	rp
<vvd> went <rp> out <y> . <ii> through <at> the <nn1_jj@> back <nn1> door <y> . <dd1> that <vbz> 's	y	jj
<y> , <np1> gamaliel <vvd> made <appge> his <nn1> way <rp_nn1> back <ii> to <at> the <nn1> road <cc> and	y	rp
<nn2> nails <ii> into <appge> his <nn2> sides <cc> and <nn1_rp> back <y> , <ii> at <at> the <da> same	y	nn1
<y> . <np1@> strange <vvd> dropped <at> the <nn1> newspaper <rp_nn1> back <ii> on <at> the <nn1> bench <nn1> seat	y	rp
<null> <p> <at> the <jj> good <y> : <vvg_jj> running <rp_nn1> back <np1> barry <np1> sanders <vbz> is <rr> simply	y	rp
<at1> a <nn1> lot <io> of <jj> bombastic <nn2> statements <rp_nn1> back <cc> and <md> fourth <ii> in <at> the	y	rp
<to> to <vvi> live <y> , <vvd_vvn> bent <rr_jj> far <rp_nn1> back <cs21> so <cs22> that <at> the <nn1> eye	y	rp
<if> for <at> the <nn1> elevator <z> -- <cc> and <rp_nn1> back <ii> to <at> the <nn1> doorway <ii> in	y	rp
Total Correct (i.e. first tag is the correct one, note that <nn1_jj@> is considered correct for adjectives as it is consistently used like this)	41	

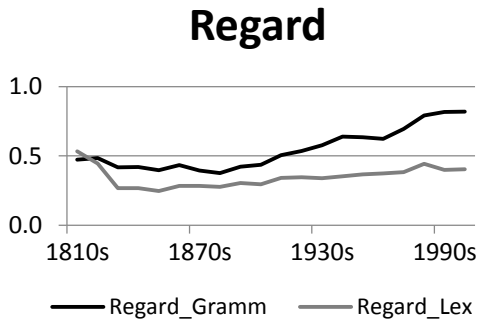
Appendix C – Additional figures (chapter 6)

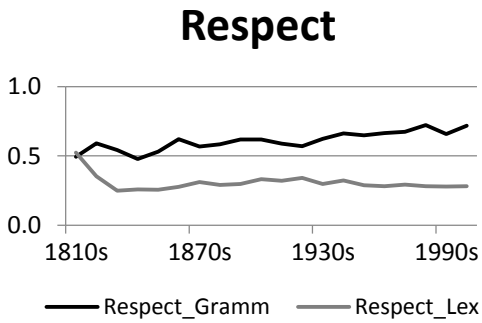
C1. Collocate diversity (R1)

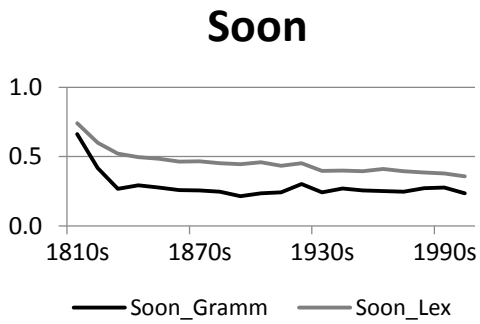


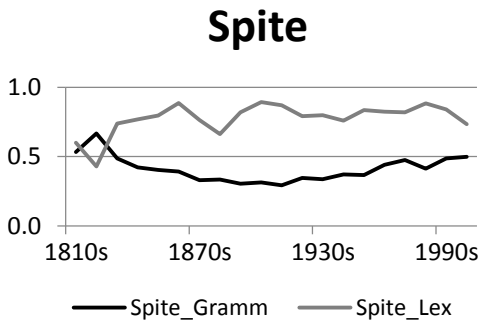

Order
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Order_Gramm
Order_Lex

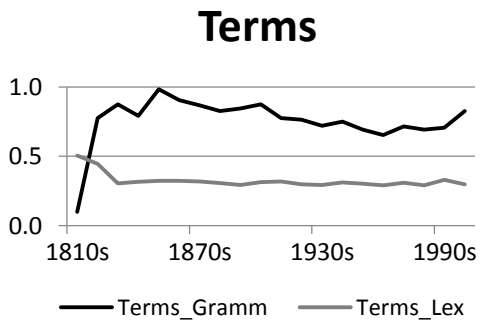

Quit
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Quit_Gramm
Quit_Lex

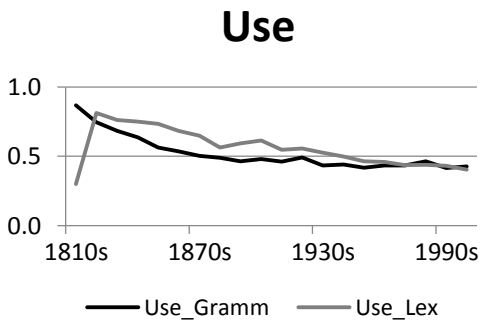

Regard
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Regard_Gramm
Regard_Lex

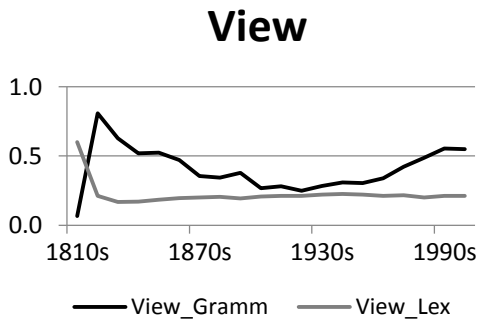

Respect
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Respect_Gramm
Respect_Lex

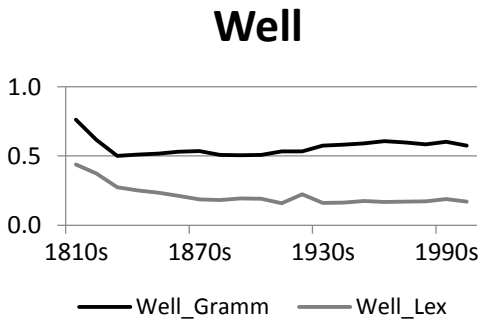

Soon
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Soon_Gramm
Soon_Lex


Spite
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Spite_Gramm
Spite_Lex

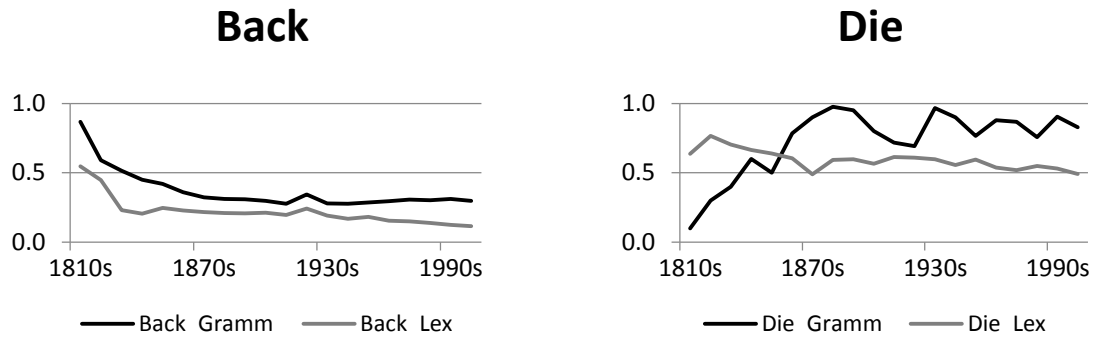

Terms
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Terms_Gramm
Terms_Lex

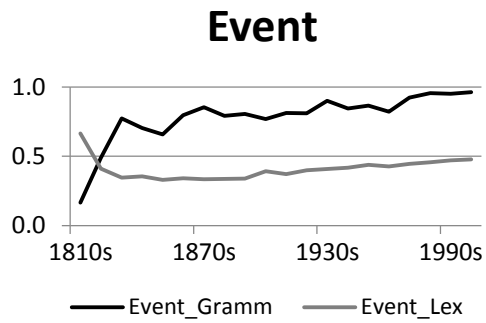

Use
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Use_Gramm
Use_Lex

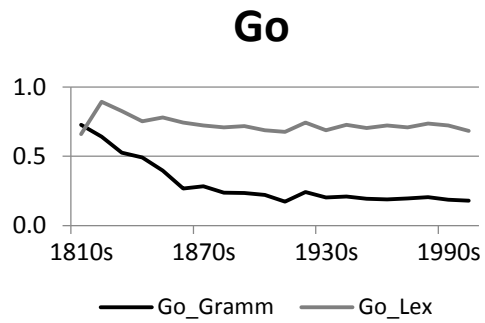

View
1.0
0.5
0.0
1810s
1870s
1930s
1990s
View_Gramm
View_Lex

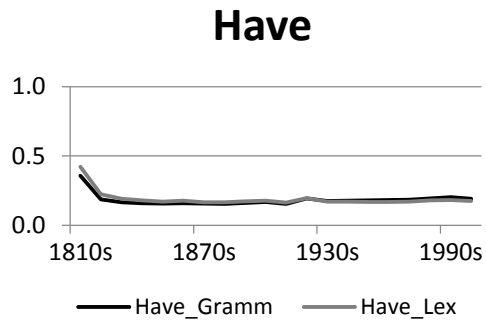

Well
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Well_Gramm
Well_Lex

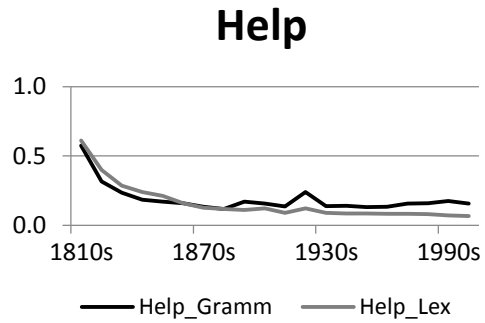
C2. Collocate diversity (L1)

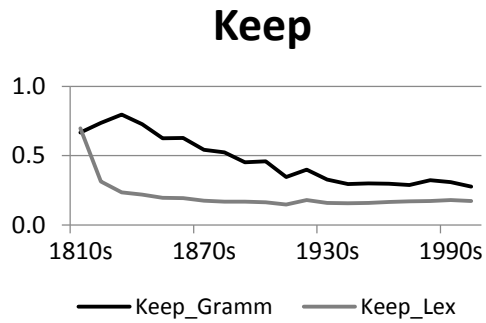


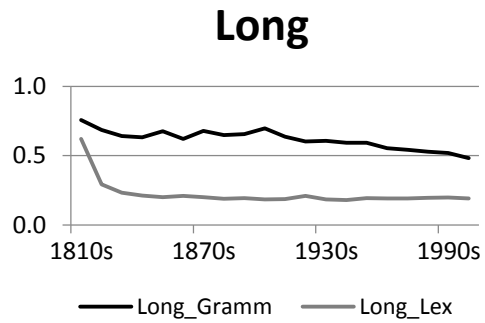


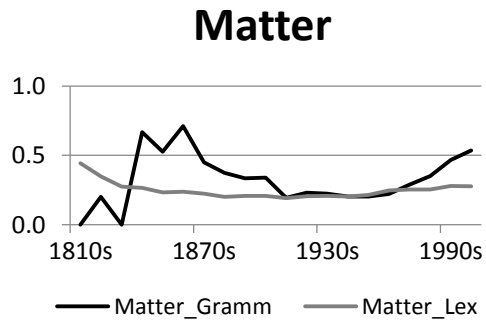


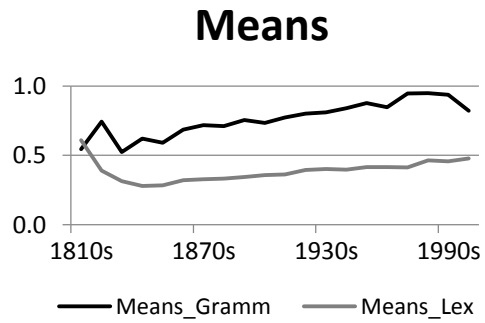


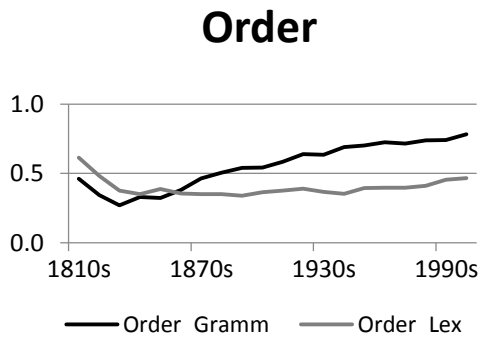


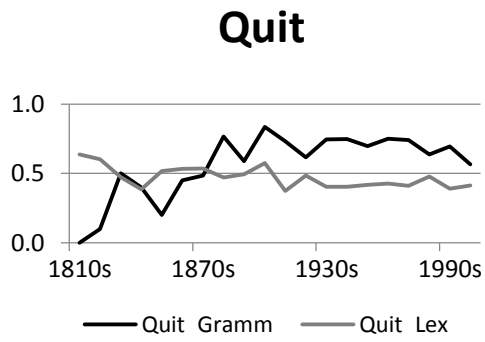


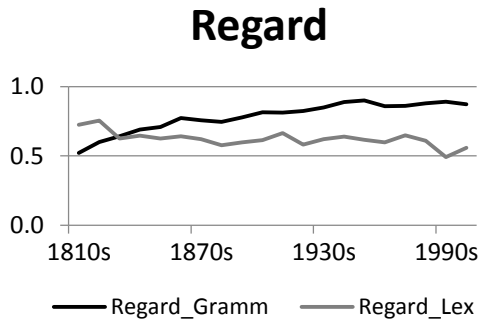


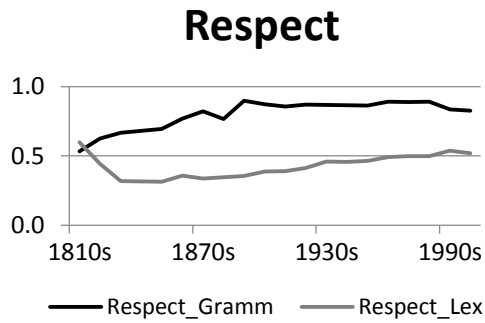


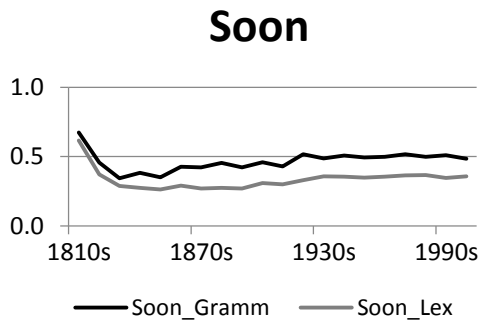


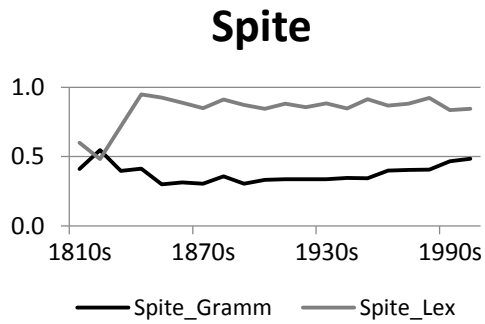

Order
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Order_Gramm
Order_Lex

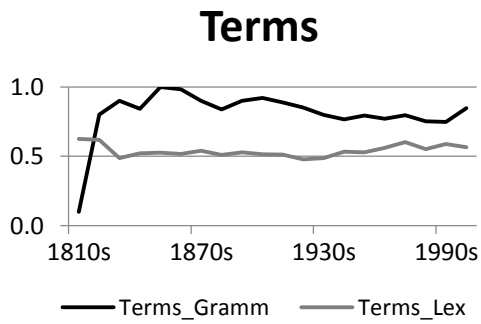

Quit
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Quit_Gramm
Quit_Lex

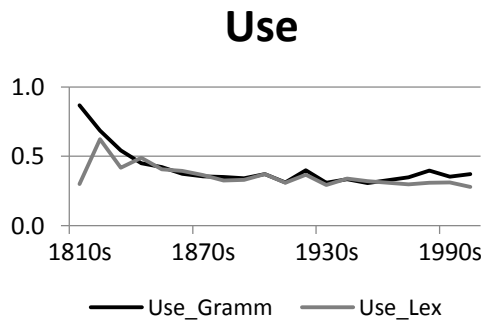

Regard
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Regard_Gramm
Regard_Lex

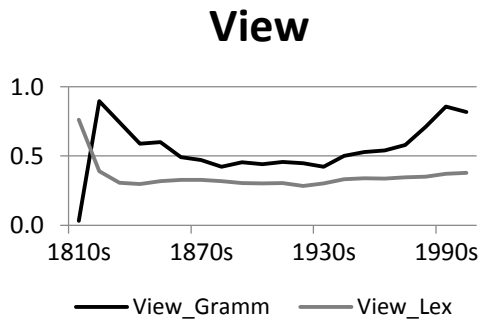

Respect
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Respect_Gramm
Respect_Lex

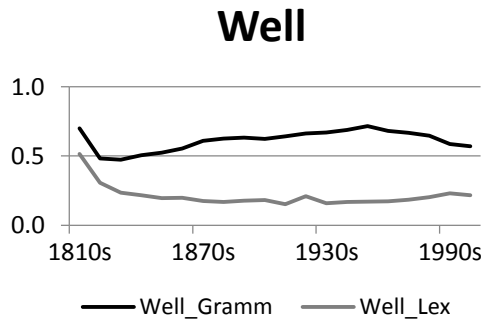

Soon
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Soon_Gramm
Soon_Lex


Spite
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Spite_Gramm
Spite_Lex

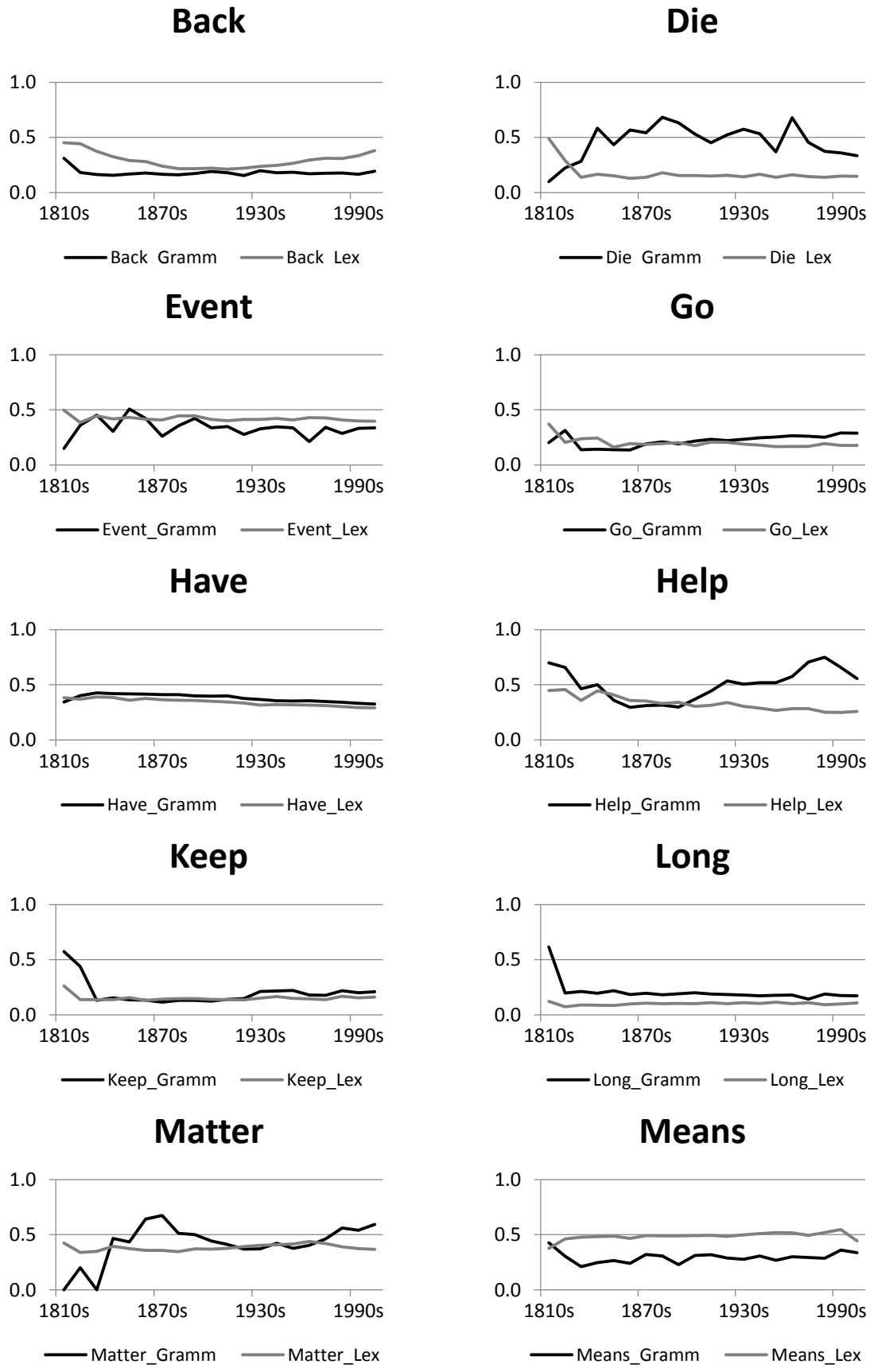

Terms
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Terms_Gramm
Terms_Lex

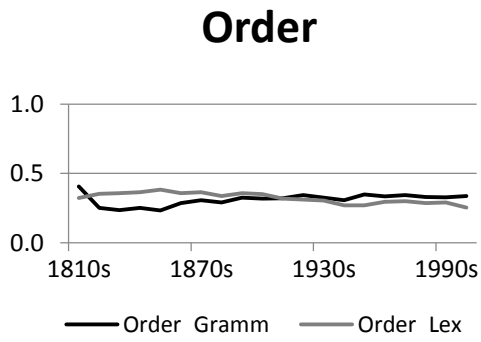

Use
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Use_Gramm
Use_Lex

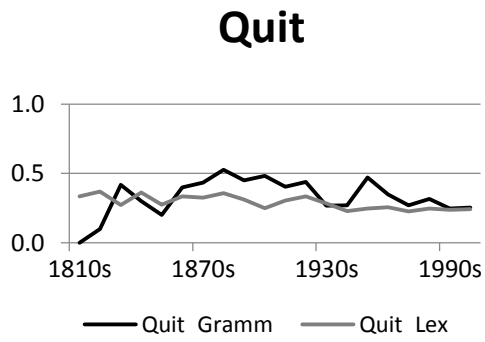

View
1.0
0.5
0.0
1810s
1870s
1930s
1990s
View_Gramm
View_Lex

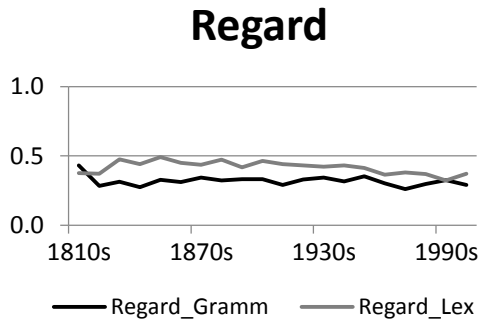

Well
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Well_Gramm
Well_Lex

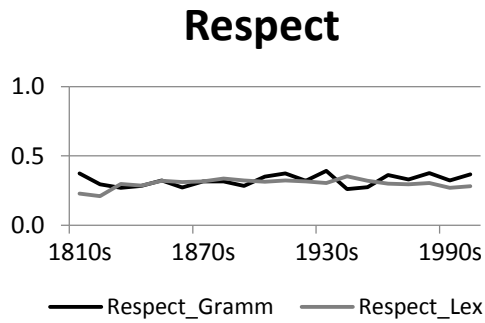
C3. Colligate diversity (L1)

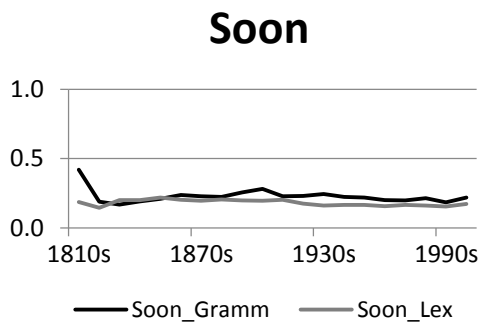


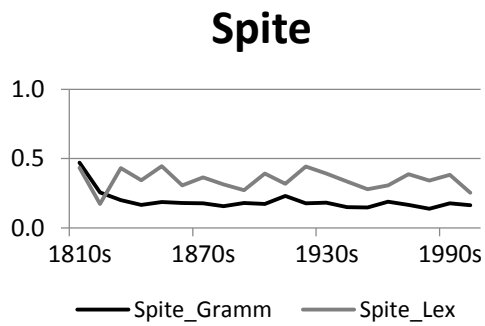

Order
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Order_Gramm
Order_Lex

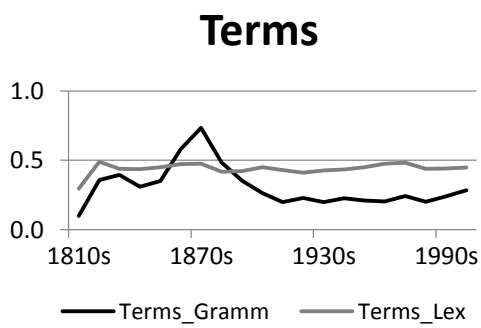

Quit
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Quit_Gramm
Quit_Lex

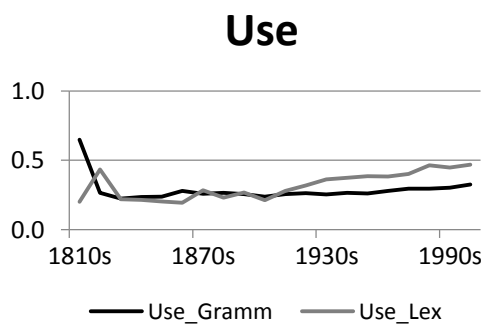

Regard
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Regard_Gramm
Regard_Lex

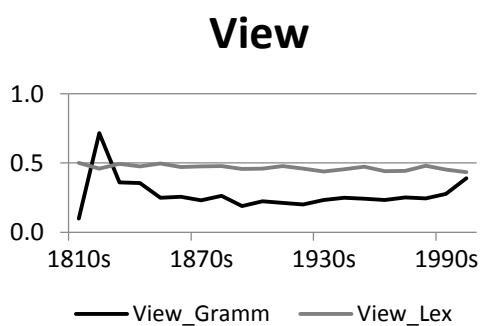

Respect
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Respect_Gramm
Respect_Lex

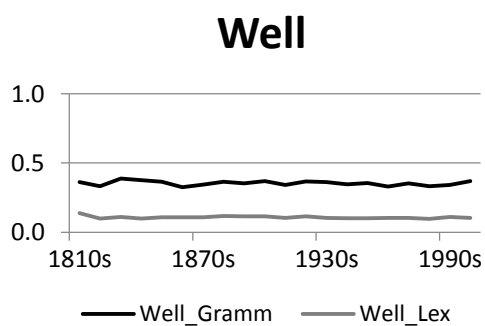

Soon
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Soon_Gramm
Soon_Lex


Spite
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Spite_Gramm
Spite_Lex


Terms
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Terms_Gramm
Terms_Lex


Use
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Use_Gramm
Use_Lex


View
1.0
0.5
0.0
1810s
1870s
1930s
1990s
View_Gramm
View_Lex


Well
1.0
0.5
0.0
1810s
1870s
1930s
1990s
Well_Gramm
Well_Lex

Appendix D – Scripts

D1. BNC written script (section 5.4)

```
import os, re, operator, time

#This function checks if 2 BNC tags are the same, but with jokers
#For example <w DT*> matches <w DT0>
def poScheck(word1, word2):
    #The XXX part in <w XXX> is checked
    for i in range(3,6):
        #If the letter matches or if it's a joker, we go to the next iteration
        if word1[i] == word2[i] or (word1[i]=="*" or word2[i]=="*"):
            next
        else:
            #If wrong match is found
            return False
    return True

#This function calculates the standard deviation on the basis of a list of numbers.
def stdev(my_list):
    average = float(sum(my_list)) / len(my_list)
    variance = 0
    for i in my_list:
        variance += (average - i) ** 2
    return (variance / len(my_list)) ** 0.5

#This function creates the output for the collocate-based values (stdev included)
#It requires a 2 dimensions collocate dictionary as first argument, and a file name for the output as second argument
#The dictionary has the words from the database as first key and each collocate of each word as second key
#(and contains the frequency as value)
#For example, dictionary["the"]["cat"] contains the value of the number of occurrences of "the" and "cat" together
def outputCollocates(colloc_dictionary, fileName):
    with open(fileName, "w") as output:
        for i in sorted(colloc_dictionary):
            count = 0
            valueList = []
            for j in colloc_dictionary[i]:
                #Check if hapax (count if so)
                if colloc_dictionary[i][j] == 1:
                    count += 1
                else:
                    next
                valueList.append(colloc_dictionary[i][j])
            if len(valueList) != 0:
                standard_deviation = stdev(valueList)
            else:
                standard_deviation = "n/a"
            output.write(i + "\t" + str(len(colloc_dictionary[i])) + "\t" + str(count)+ "\t" + str(standard_deviation)+"\n")

#This function creates the output for the colligate-based values
#It requires a 2 dimensions collocate dictionary as first argument, and a file name for the output as second argument
def colligateOutput(colloc_dictionary, fileName):
    with open(fileName, "w") as output:
        #Header
        output.write("Word\tMostFreqColligate\tColliMaxValue\tColliTotal\tMaxColliPercent\n")
        #Colligates output
        for i in sorted(colloc_dictionary):
            dicoTag = {}
            for j in colloc_dictionary[i]:
                #j[0:7] corresponds to <W XXX> tags in the BNC
                try:
                    dicoTag[j[0:7]] += 1
                except KeyError:
                    dicoTag[j[0:7]] = 1

            try:
                tagMax = max(dicoTag.iteritems(), key=operator.itemgetter(1))[0]
                valueMax = dicoTag[tagMax]
            except ValueError:
                tagMax = "n/a"
                valueMax = 0
```

```
        #count total
        total = 0
        for j in dicoTag:
            total += dicoTag[j]

        if total != 0:
            output.write(i + "\t" + str(tagMax) + "\t" + str(valueMax) + "\t" + str(total)+ "\t" + str(round(float(valueMax)/total, 3))+ "\n")
        else:
            output.write(i + "\t" + str(tagMax) + "\t" + str(valueMax) + "\t" + str(total)+ "\t" + "0" + "\n")


###################################
### Settings and Initialization ###
###################################

#Opens the database and get the word list from it (everything is put in upper case, line feeds are removed)
#Words in the database should be tagged as such: <w XXX> word
with open("DBv4_Tagged.txt", "r") as f:
    wordList = []
    for i in f:
        wordList.append(i.upper().strip())

#Corpus directory
corpusDirectory = "./written"

#Get the files from the corpus
#Note: All the files are in the same folder. If they are not, use the Python walk function instead of listdir
files = os.listdir(corpusDirectory)

#Only relevant for Mac (used to remove a useless file that may cause errors in the script)
if ".DS_Store" in files:
    files.remove(".DS_Store")

#Timer (information regarding execution time)
start_time = time.time()

#Used to show progression of the script
fileCount = 0

#Frequency dictionary
frequency = {}

#Collocates parameters
window = 4
windowR = 1
windowL = 1

#Dictionary of collocates
collocatesFinal = {}
collocatesFinal_R = {}
collocatesFinal_L = {}
#Fill the dictionaries with sub-dictionaries (to get the frequencies of each collocate)
for i in wordList:
    collocatesFinal[i] = {}
    collocatesFinal_R[i] = {}
    collocatesFinal_L[i] = {}

#DevProp Settings
chunkSize = 1000

#Deviation of proportions counts
currentChunkCount = 0
chunkNB = 0

#Dictionary, each entry contains a list with the number of words per chunk
#The dictionary starts with its keys (taken from WordList) only and an empty list
wordsInChunks = {}
for i in wordList:
    wordsInChunks[i] = [0]

###################
### Main Script ###
###################

#Corpus is browsed file by file
```

```
for i in range(len(files)):
    with open(corpusDirectory + "/" + files[i]) as f:
        content = f.read()
        #RE that finds anything that begins in <w ...> and contains only letters and/or blank space (or apostrophe)
        check = re.compile("<w \w+>[a-zA-Z' ]+")
        allMatches = check.findall(content)

        #Cleaning weird stuff like line feeds and so on and put everything in upper case
        allMatchesClean = []
        for clean in allMatches:
            allMatchesClean.append(clean.upper().strip())

        #Go through the words of the current file
        for j in range(len(allMatchesClean)):
            #Check every word from the list
            #It's not possible to check if k in list because the tags use jokers
            for k in range(len(wordList)):
                #A word is found in the list: Word must be the same, and PoSTag must match as well
                if wordList[k][7:] == allMatchesClean[j][7:] and poScheck(allMatchesClean[j], wordList[k]):
                    #Frequency count
                    try:
                        frequency[wordList[k]] += 1
                    except KeyError:
                        frequency[wordList[k]] = 1

                    #Collocates counts, symmetrical window
                    collocates =  allMatchesClean[j-window:j] + allMatchesClean[j+1:j+window+1]
                    for collocate in collocates:
                        if collocate not in collocatesFinal[wordList[k]]:
                            collocatesFinal[wordList[k]][collocate] = 1
                        else:
                            collocatesFinal[wordList[k]][collocate] += 1

                    #Note: the left and right widows are computed separately
                    #The symmetrical window could be re-used to make the script more synthetic
                    #However, this might cause unexpected errors due to window size variations at the beginning and end of files
                    #Also, it's possible to use any window sizes (even though, well, no one is probably going to do that).

                    #Collocates counts, right window
                    collocates = allMatchesClean[j+1:j+windowR+1]
                    for collocate in collocates:
                        if collocate not in collocatesFinal_R[wordList[k]]:
                            collocatesFinal_R[wordList[k]][collocate] = 1
                        else:
                            collocatesFinal_R[wordList[k]][collocate] += 1

                    #Collocates counts, left window
                    collocates = allMatchesClean[j-windowL:j]
                    for collocate in collocates:
                        if collocate not in collocatesFinal_L[wordList[k]]:
                            collocatesFinal_L[wordList[k]][collocate] = 1
                        else:
                            collocatesFinal_L[wordList[k]][collocate] += 1

                    #Deviation of proportions count
                    wordsInChunks[wordList[k]][chunkNB] += 1

            #Update chunk count
            if currentChunkCount < chunkSize:
                currentChunkCount += 1
                next
            #If chunkSize is reached
            else:
                #Start a new chunk in the chunk dictionary for each element
                for element in wordsInChunks:
                    wordsInChunks[element].append(0)
                #Reset count and start next chunk
                currentChunkCount=0
                chunkNB += 1



    fileCount += 1
    if fileCount % 10 == 0:
        print str(fileCount) + " done out of " + str(len(files))
```

```
###############
### Outputs ###
###############

#Frequency Output
with open("Output_freqWritten.txt", "w") as output:
    for i in sorted(frequency):
        output.write(i + "\t" + str(frequency[i])+"\n")

#Collocates Output Symmetrical Window: number of different collocates, number of hapaxes, stdev
outputCollocates(collocatesFinal, "Output_collocates_written_4-4.txt")

#Collocates Output Right Window: number of different collocates, number of hapaxes, stdev
outputCollocates(collocatesFinal_R, "Output_collocates_written_1R.txt")

#Collocates Output Left Window: number of different collocates, number of hapaxes, stdev
outputCollocates(collocatesFinal_L, "Output_collocates_written_1L.txt")

#Colligates Output - Right Window
colligateOutput(collocatesFinal_R, "Output_colligates_written_1R.txt")

#Colligates Output - Left Window
colligateOutput(collocatesFinal_L, "Output_colligates_written_1L.txt")

#Deviation of proportions Output
with open("Output_DevProp_written.txt", "w") as output:
    for i in sorted(wordsInChunks):
        total = 0
        frequency = sum(wordsInChunks[i])
        #Usually, words should not have a frequency of 0, but this can happen when only working with a small part of the corpus
        if frequency > 0:
            for j in wordsInChunks[i]:
                total +=  abs ( ((float(j)/frequency)) - (1/float(chunkNB)) )

            devProp = total / 2
            output.write(i + "\t" + str(devProp) + "\t" +str(frequency) +"\n")
        else:
            output.write(i + "\t" + "n/a" + "\t" +str(frequency) +"\n")

print "Script done in", time.time() - start_time, "seconds"
```

D2. Selection of lexical and grammatical counterparts (section 6.2)

D2-A. Auxiliary example (keep)

```
#This script only uses the sample version of the COHA (which is enough for keep)

import os, re, time, random

#Writes the output in a file for a given window
def windowOutput(window, nb, output):
    if len(window)>=nb:
        for i in random.sample(window,nb):
            output.write(" ".join(i)+"\n")
    else:
        for i in window:
            output.write(" ".join(i)+"\n")

corpusDirectory = "COHA-SAMPLE_wlp"
files = os.listdir(corpusDirectory)
fileCount = 0

#Mac only
if ".DS_Store" in files:
    files.remove(".DS_Store")

#Settings and initialization
count_ing = 0
count = 0
```

```
window0 = []
window1 = []
window2 = []
window3 = []
window4 = []
lexical = []

#Regular expression that includes anything that contains vvg somewhere between < and > (thus includes vvgk)
#It also includes vbn (be), vhn (have), and vdn(do).
find_vvg = re.compile("<.*v(v|b|h|d)g.*>")

#Regular expression that looks for "from"
find_from = re.compile("<.*> from")

start_time = time.time()

for i in range(len(files)):
   with open(corpusDirectory + "/" + files[i]) as file:

      allMatches = [] #List of all the words in the file

      #File is read line by line
      lines = file.readlines()
      for line in lines:
         line = line.strip()
         line = line.split("\t")
         if "@" in line[0] or len(line) < 3:
            next
         else:
            allMatches.append("<"+line[2].lower()+">"+" "+line[0].lower())

      year = files[i].split("_")[1]

      #Go through the words
      for position_i in range(len(allMatches)):
         if allMatches[position_i] in ["<vv0> keep", "<vvz> keeps", "<vvi> keep", "<vvd> kept", "<vvn> kept"]:
            #ING right after keep
            if find_vvg.match(allMatches[position_i+1]):
               window0.append(allMatches[position_i-5:position_i+5])
            #ING 1 word after
            elif find_vvg.match(allMatches[position_i+2]) and not find_from.match(allMatches[position_i+1]):
               window1.append(allMatches[position_i-4:position_i+6])
            #ING 2 words after
            elif find_vvg.match(allMatches[position_i+3]) and not find_from.match(allMatches[position_i+1]) and not
find_from.match(allMatches[position_i+2]):
               window2.append(allMatches[position_i-3:position_i+7])
##             #ING 3 words after
##             elif find_vvg.match(allMatches[position_i+4]):
##                window3.append(allMatches[position_i-2:position_i+8])
##             #ING 4 words after
##             elif find_vvg.match(allMatches[position_i+5]):
##                window4.append(allMatches[position_i-1:position_i+9])
            #Lexical
            else:
               lexical.append(allMatches[position_i-5:position_i+5])


   #CHECKING
   if fileCount % 100 == 0:
      print str(fileCount) + " Done out of " + str(len(files))

   fileCount += 1


###############
### Outputs ###
###############

#Number of randomly selected sentences
nb = 25

with open("Output_keep_optimal_window.txt", "w") as output:
   output.write("0 in between: (total="+str(len(window0))+")\n")
   windowOutput(window0, nb, output)
   output.write("-"*50+"\n"+"1 in between: (total="+str(len(window1))+")\n")
```

```
    windowOutput(window1, nb, output)
    output.write("-"*50+"\n"+"2 in between: (total="+str(len(window2))+")\n")
    windowOutput(window2, nb, output)
##    output.write("-"*50+"\n"+"3 in between: (total="+str(len(window3))+")\n")
##    windowOutput(window3, nb, output)
##    output.write("-"*50+"\n"+"4 in between: (total="+str(len(window4))+")\n")
##    windowOutput(window4, nb, output)
    output.write("-"*50+"\n"+"Lexical: (total="+str(len(lexical))+")\n")
    windowOutput(lexical, nb*2, output)

print "Script done in", time.time() - start_time, "seconds"
```

D2-B. Complex preposition example (as well as)

```
import os, re, time, random

#Writes the output for a specified sample size
def selectionOutput(selection, nb, output):
    if len(selection)>=nb:
        for i in random.sample(selection,nb):
            output.write(" ".join(i)+"\n")
    else:
        for i in selection:
            output.write(" ".join(i)+"\n")

corpusDirectory = "./COHA-SAMPLE_wlp"
files = os.listdir(corpusDirectory)
fileCount = 0

#Mac only
if ".DS_Store" in files:
    files.remove(".DS_Store")

#All instances we are looking for
#Grammatical instances are: <cs31> as <cs32> well <cs33> as
instances = ["<ii32> well", "<jj> well", "<rr> well"]

grammList = []
lexList = []

start_time = time.time()

for i in range(len(files)):
    with open(corpusDirectory + "/" + files[i]) as file:

        allMatches = [] #List of all the words in the file

        #File is read line by line
        lines = file.readlines()
        for line in lines:
            line = line.strip()
            line = line.split("\t")
            if "@" in line[0] or len(line) < 3:
                next
            else:
                allMatches.append("<"+line[2].lower()+">"+" "+line[0].lower())

        #In the main script, [3] is used because the date is taken from the file which is itself inside a folder
        year = files[i].split("_")[1]

        #Go through the words
        for position_i in range(len(allMatches)):
            if allMatches[position_i] in instances:
                #Grammatical match
                if allMatches[position_i-1]=="<ii31> as" and allMatches[position_i][0:6] == "<ii32>" and allMatches[position_i+1]=="<ii33> as":
                    grammList.append(allMatches[position_i-5:position_i+5])

                #Lexical match
                else:
                    #This removes ambiguous cases (i.e. cases tagged as <ii32> well but not preceded or followed by <ii31> as / <ii33> as)
                    if allMatches[position_i]=="<ii32> well":
                        next
```

```
                else:
                    lexList.append(allMatches[position_i-5:position_i+5])



    #CHECKING
    if fileCount % 100 == 0:
        print str(fileCount) + " Done out of " + str(len(files))

    fileCount += 1


###############
### Outputs ###
###############

#Number of randomly selected sentences
nb = 25

with open("Output_well_selection.txt", "w") as output:
    output.write("Grammatical instances: (total="+ str(len(grammList))+")\n")
    selectionOutput(grammList, nb, output)
    output.write("\nLexical instances:(total="+ str(len(lexList))+")\n")
    selectionOutput(lexList, nb, output)

print "Script done in", time.time() - start_time, "seconds"
```

D3. COHA script example (sections 6.3 and 6.4) (keep)

```
import os, re, operator, time

#This function calculates the standard deviation on the basis of a list of numbers.
def stdev(my_list):
    average = float(sum(my_list)) / len(my_list)
    variance = 0
    for i in my_list:
        variance += (average - i) ** 2
    return (variance / len(my_list)) ** 0.5

#This function fills an empty 2D dictionary with a key for each year
#Each year itself then points to an empty dictionary
def emptyDico():
    new_dico = {}
    for i in range(1810,2010):
        new_dico[str(i)] = {}
    return new_dico

#This function adds frequency to an output table
def frequencyOutput(freq_dic, output_table, header):
    my_list = ["Frequency"+header]
    for i in sorted(freq_dic):
        my_list.append(freq_dic[i])
    output_table.append(my_list)
    return output_table

#This function takes a list of words and checks if they are in a dictionary of collocates
#The current year is also required as input
def checkCollocates(wordList, collocates, year):
    for word in wordList:
        if word not in collocates[year]:
            collocates[year][word] = 1
        else:
            collocates[year][word] += 1

#This function adds the number of collocates and collocate diversity to the output table
#It takes 4 arguments: a dictionary of collocates, a dictionary of frequencies, an output table and a header (str)
def collocatesOutput(collocates, frequency, output_table, header):
    nb_colloc = ["Collocates"+header]
    colloc_div = ["CollocDiv"+header]
    for i in sorted(collocates):
        nb_colloc.append((len(collocates[i])))
        #Avoid divison by 0
```

```
        if len(collocates[i]) > 0:
            colloc_div.append(float(len(collocates[i])) / frequency[i])
        else:
            colloc_div.append(0)
    output_table.append(nb_colloc)
    output_table.append(colloc_div)
    return output_table


#This function adds the number of hapaxes and productivity to the output table
#It takes 4 arguments: a dictionary of collocates, a dictionary of frequencies, an output table and a header (str)
def hapaxesOutput(collocates, frequency, output_table, header):
    nb_hapaxes = ["Hapaxes"+header]
    productivity = ["p"+header]
    for year in sorted(collocates):
        count_hapax = 0
        for collocate in collocates[year]:
            if collocates[year][collocate] == 1:
                count_hapax += 1
            else:
                next
        nb_hapaxes.append(count_hapax)
        #Avoid division by 0
        if count_hapax > 0:
            productivity.append(float(count_hapax)/frequency[year])
        else:
            productivity.append(0)
    output_table.append(nb_hapaxes)
    output_table.append(productivity)
    return output_table

#This function adds the standard deviation to the output table
#It takes 3 arguments: a dictionary of collocates, an output table and a header (str)
def stdevOutput(collocates, output_table, header):
    stdev_list = ["Stdev"+header]
    for year in sorted(collocates):
        valueList = []
        for collocate in collocates[year]:
            valueList.append(collocates[year][collocate])
        #Avoid division by 0
        if len(valueList) != 0:
            standard_deviation = stdev(valueList)
        else:
            standard_deviation = "n/a"
        stdev_list.append(standard_deviation)
    output_table.append(stdev_list)
    return output_table

#This function is a used by the colligateOutput() function to count the number of colligates in a dictionary of collocates
def colligateOutput(collocates, output_table, header):
    collig_mostFreq = ["MostFreqColligate"+header]
    collig_maxValue = ["MaxColliValue"+header]
    collig_maxPercent = ["MaxColliPercent"+header]
    for i in sorted(collocates):
        tag_dico = {}
        #Retrieve the tags only, so stuff between < ... >
        for j in collocates[i]:
            try:
                tag_dico[j[j.index("<")+1:j.index(">")]]+= 1
            except KeyError:
                tag_dico[j[j.index("<")+1:j.index(">")]] = 1

        #Find out which tag is the most frequent
        try:
            maxTag = max(tag_dico.iteritems(), key=operator.itemgetter(1))[0]
            collig_mostFreq.append(maxTag)
            maxValue = tag_dico[maxTag]
            collig_maxValue.append(maxValue)
        except ValueError:
            collig_mostFreq.append("n/a")
            collig_maxValue.append(0)

        #Count total number of tags (should be equal to frequency, used for verification purposes
        total = 0
        for j in tag_dico:
```

```
            total += tag_dico[j]

        if total != 0:
            collig_maxPercent.append(float(maxValue)/total)
        else:
            collig_maxPercent.append(0)
    output_table.append(collig_mostFreq)
    output_table.append(collig_maxValue)
    output_table.append(collig_maxPercent)
    return output_table

#This function is a used by the devPropOutput function to compute the deviation of proportions score
#It takes 3 arguments: a year chunk dictionary as first argument, an output table, and a header
def devPropOutput(years, output_table, header):
    devProp_list = ["DevProp"+header]
    for i in sorted(years):
        total = 0
        frequency = sum(years[i])
        if frequency > 0:
            for j in years[i]:
                total +=  abs ( ( (float(j)/frequency)) - (1/float(len(years[i])) ) )
            devProp = total / 2
            devProp_list.append(devProp)
        #Avoid division by 0
        else:
            devProp_list.append("n/a")
    output_table.append(devProp_list)
    return output_table

#This function fills the output table using most of the other functions
def outputContent(collocates, collocates_R, collocates_L, frequency, years_chunk, output_table, header):
    #Frequency Output
    output_table = frequencyOutput(frequency, output_table, header)
    #Collocates Output Symmetrical Window
    output_table = collocatesOutput(collocates, frequency, output_table, "4-4"+header)
    #Hapaxes Output - Symmetrical Window
    output_table = hapaxesOutput(collocates, frequency, output_table, "4-4"+header)
    #Stdev Output - Symmetrical Window
    output_table = stdevOutput(collocates, output_table, "4-4"+header)
    #Collocates Output Right Window
    output_table = collocatesOutput(collocates_R, frequency, output_table, "R1"+header)
    #Hapaxes Output - Right Window
    output_table = hapaxesOutput(collocates_R, frequency, output_table, "R1"+header)
    #Stdev Output - Right Window
    output_table = stdevOutput(collocates_R, output_table, "R1"+header)
    #Collocates Output Left Window
    output_table= collocatesOutput(collocates_L, frequency, output_table, "L1"+header)
    #Hapaxes Output - Left Window
    output_table = hapaxesOutput(collocates_L, frequency, output_table, "L1"+header)
    #Stdev Output - Left Window
    output_table = stdevOutput(collocates_L, output_table, "L1"+header)
    #Colligates Output - Right Window
    output_table = colligateOutput(collocates_R, output_table, "R1"+header)
    #Colligates Output - Left Window
    output_table = colligateOutput(collocates_L, output_table, "L1"+header)
    #Deviation of proportions Output
    output_table = devPropOutput(years_chunk, output_table, header)
    return output_table

#This function produces a separate output with the top X most frequent collocates for each year
#It takes two dictionaries as first arguments (containing the counts for each year for each case (gramm/lex) and a file name as last argument
def collocatesOutputDetailed(collocates_1, collocates_2, fileName):
    with open(fileName, "w") as output:
        for year in sorted(collocates_1):
            for collocate in collocates_1[year]:
                output.write(year + "\t" + collocate + "\t" + str(collocates_1[year][collocate]) + "\n")
        output.write("-"*50+"\n")
        for year in sorted(collocates_2):
            for collocate in collocates_2[year]:
                output.write(year + "\t" + collocate + "\t" + str(collocates_2[year][collocate]) + "\n")

#This function writes the final output in a separate file
def finalOutput(output_table, fileName):
    with open(fileName, "w") as output:
```

```
    #Header and creation of the first line (dates)
    lines = ["Date"]+range(1810,2010)

    for i in range(len(output_table)):
        for j in range(len(output_table[i])):
            #Note: using += here is not possible given that I used a range function to fill the first slot of lines
            lines[j] = str(lines[j]) + ("\t"+str(output_table[i][j]))

    for line in lines:
        output.write(line+"\n")


##############################
### Item Specific Section ###
##############################

#Grammatical instances we are looking for
instances = ["<vv0> keep", "<vvz> keeps", "<vvi> keep", "<vvd> kept", "<vvn> kept"]

#Maximum number of words after keep (should be minimum 1)
ing_threshold = 3

#Regular expression that includes anything that contains vvg somewhere between < and > (thus includes vvgk and ambiguity tags)
#It also includes vbn (be), vhn (have), and vdn(do).
find_vvg = re.compile("<.*v(v|b|h|d)g.*>")

#Regular expression that looks for "from"
find_from = re.compile("<.*> from")


###################################
### Settings and Initialization ###
###################################

#The final outputs, which are two large tables
output_table_gramm = []
output_table_lex = []

#Used to show progression of the script
fileCount = 0

#Corpus directory
corpusDirectory = "./COHA"
folders = os.listdir(corpusDirectory)
files = []

#Only relevant for Mac (used to remove a useless file that causes errors in the script)
if ".DS_Store" in folders:
    folders.remove(".DS_Store")

#This section looks for the files in a all the sub-folders
for folder in folders:
    folderList = os.listdir(corpusDirectory+"/"+folder)
    for fichier in folderList:
        files.append(folder + "/" + fichier)

#Only relevant for Mac (used to remove a useless file that causes errors in the script)
if ".DS_Store" in files:
    files.remove(".DS_Store")

#Start a dictionary for each year and its associated frequency counts
frequency_gramm = {}
frequency_lex = {}

#The frequency dictionaries are filled with the list of dates as keys which point to a 0 counter
for i in range(1810,2010):
    #Starting a count for gramm (left) and lex (right)
    frequency_gramm[str(i)] = 0
    frequency_lex[str(i)] = 0

#Initializing collocates dicos
collocatesFinal_gramm = emptyDico()
collocatesFinal_lex = emptyDico()
collocatesFinal_R_gramm = emptyDico()
collocatesFinal_R_lex = emptyDico()
```

```
collocatesFinal_L_gramm = emptyDico()
collocatesFinal_L_lex = emptyDico()

#Window Settings
window = 4
windowR = 1
windowL = 1

#DevProp Settings
chunkSize = 1000

#Dictionary, each entry contains a list with the number of words per chunk
#This is used for deviation of proportions only
#The dictionary starts with its keys only (years) and an empty list
years_gramm = {}
years_lex = {}
for i in range(1810,2010):
    years_gramm[str(i)] = [0]
    years_lex[str(i)] = [0]


###################
### Main Script ###
###################

start_time = time.time()

#Current year is used for deviation of proportions later
current_year = "1810"
#Current count dict is used to continue a count (within a chunk) when a year comes back
#This happens because the folders/files are not classified by year (genre first).
current_year_counts = {}
for i in range(1810,2010):
    current_year_counts[str(i)] = 0

#Initialization chunk number and count
chunkNB = 0
count = 0

for i in range(len(files)):
    with open(corpusDirectory + "/" + files[i]) as file:
        allMatches = [] #List of all the words in the file
        #File is read line by line
        lines = file.readlines()
        for line in lines:
            line = line.strip()
            line = line.split("\t")
            #The @symbols are removed from the data
            if "@" in line[0] or len(line) < 3:
                next
            else:
                allMatches.append("<"+line[2].lower()+">"+" "+line[0].lower())

    #Get current year
    year = files[i].split("_")[3]

    #Reset devProp counts if year change
    if year != current_year:
        #We need to get the chunk count for the year in question (0 if first time)
        #Also, years_gramm and years_lex have the same length and -1 is needed to get to the previous chunk
        #This avoids out of range errors because we use indexes (so if 3 elements, the max index is 2, i.e. 3-1)
        chunkNB = len(years_gramm[year])-1
        #Store count for the year that we are leaving
        current_year_counts[current_year] = count
        #Retreive count for the new year (first time is 0)
        count = current_year_counts[year]
        #Update to new year
        current_year = year

    #Find instances in current file (iterating through the words)
    for position_i in range(len(allMatches)):
        #flag used for -ing case
        flag = 0
        if allMatches[position_i] in instances:
            windowING = allMatches[position_i+1:position_i+1+ing_threshold]
```

```
        #Starting the range with 1 is necessary because I use k to find the correct position with the right collocates
        for k in range(1,len(windowING)+1):
            if find_from.match(windowING[k-1]):
                flag = 2 #We remove "from" cases completely
                break
            else:
                if position_i+k+1 < len(allMatches) and find_vvg.match(windowING[k-1]):
                    #A correct grammatical match is found
                    flag = 1
                    collocates =  allMatches[position_i-window:position_i] + allMatches[position_i+1:position_i+window+1]
                    collocates_R = allMatches[position_i+1:position_i+windowR+1] #Ving included into the collocate window
                    collocates_L = allMatches[position_i-windowL:position_i]
                    #Frequency count
                    frequency_gramm[year] += 1
                    #Collocates count
                    checkCollocates(collocates, collocatesFinal_gramm, year)
                    checkCollocates(collocates_R, collocatesFinal_R_gramm, year)
                    checkCollocates(collocates_L, collocatesFinal_L_gramm, year)
                    #Deviation of proportions count
                    years_gramm[year][chunkNB] += 1
                    break

        #Not a match (other than -ing)
        #flag necessary as "from" cases need to be included here, we can't just get away with a normal if/elif/else
        if flag == 0:
            collocates =  allMatches[position_i-window:position_i] + allMatches[position_i+1:position_i+window+1]
            collocates_R = allMatches[position_i+1:position_i+windowR+1]
            collocates_L = allMatches[position_i-windowL:position_i]
            #Frequency count
            frequency_lex[year] += 1
            #Collocates count
            checkCollocates(collocates, collocatesFinal_lex, year)
            checkCollocates(collocates_R, collocatesFinal_R_lex, year)
            checkCollocates(collocates_L, collocatesFinal_L_lex, year)
            #Deviation of proportions count
            years_lex[year][chunkNB] += 1

    #Chunk update (Deviation of proportions only)
    if count < chunkSize:
        count += 1
    else:
        years_gramm[year].append(0)
        years_lex[year].append(0)
        count = 0
        chunkNB += 1

  if fileCount % 100 == 0:
      print str(fileCount) + " Done out of " + str(len(files))

  fileCount += 1


###############
### Outputs ###
###############

#Compute the main output tables
output_table_gramm = outputContent(collocatesFinal_gramm, collocatesFinal_R_gramm, collocatesFinal_L_gramm, frequency_gramm, years_gramm, output_table_gramm, "_gramm")
output_table_lex = outputContent(collocatesFinal_lex, collocatesFinal_R_lex, collocatesFinal_L_lex, frequency_lex, years_lex, output_table_lex, "_lex")

#Write the output tables in separate files (one gramm, one lex)
finalOutput(output_table_gramm, "Output_keep_gramm.txt")
finalOutput(output_table_lex, "Output_keep_lex.txt")

#Collocates Output Symmetrical Window DETAILED
collocatesOutputDetailed(collocatesFinal_gramm, collocatesFinal_lex, "Output_keep_colloc4-4_DETAILED.txt")

print "Script done in", time.time() - start_time, "seconds"
```