Variation and innovation in modern English: Corpus-based studies in the grammaticalization of multiword units

Adam Michael Richard Smith

April 2016

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Linguistics
Macquarie University

Declaration

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirement for a degree to any other university of institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any help and assistance that I have received in my research work and the preparation of the thesis itself have been appropriately acknowledged. All information sources and literature used are indicated in the thesis.

Adam Smith (31538711)

4 April 2016

Abstract

This dissertation is an empirical inquiry into the lexicalization and grammaticalization of several types of multiword units, whose status as fixed lexical units has not been established, and whose grammatical structure and roles are still open to question. They remain on the fringes of codification and classification in current dictionaries and grammars.

The set of four published papers embodied in this dissertation investigate light verbs (e.g. *have a look*), non-numerical quantifiers (e.g. *a lot of*), complex prepositions (e.g. *in spite of)* and complex subordinators (e.g. *the moment*). In their structure, each of these includes a noun phrase, but as units they constitute different grammatical functions, those of the verb, determiner, preposition and subordinator respectively. These four types of multiword unit have been examined to assess how well they meet the standard criteria for grammaticalization, such as fixity, decategorialization and syntactic reanalysis.

A range of standard corpora were used for this study, allowing investigation into the synchronic variation of the items under discussion across different English language regions and registers, along with some research into recent diachronic developments. Corpora of different sizes were selected to provide sufficient data on high- and low-frequency items. For higher frequency items, the Australian, British and New Zealand components of the 1million-word International Corpus of English (ICE), as well as ICE-US (written only), complemented by the spoken Santa Barbara Corpus were used. These smaller corpora also allowed the individual linguistic contexts of examples to be more closely examined. For lower frequency items the British National Corpus and Corpus of Contemporary American English were used, as well as the Corpus of Historical American English, which provided some diachronic data. Selective examples of linguistic contexts were elicited from these larger corpora (100million-word and over) and non-relevant usages were excluded from the frequency counts by the use of search strings adapted to each item.

For each data set, the frequency of fixed and variable forms of the multiword units were compared, and the wider context also examined to find examples of indeterminate grammatical use, manifested by factors such as clause position and inconsistent patterns of concord. Data was also gathered from comprehensive and learner grammars, and dictionaries for first- and second-language users, to gauge the degree of recognition of these marginal/emergent items.

The body of research finds that, while each of the multiword units investigated is lexicalized to some extent, there is also syntagmatic evidence of grammaticalization in two cases. The grammatical status of the unit was indicated in the case of non-numerical quantifiers by whether the singular or plural quantifying noun agrees with the following verb; and for complex subordinators by the absence of a preceding preposition and following relative pronoun, and especially its position at the start of a clause.

The thesis demonstrates that several criteria are necessary to demonstrate the grammatical status of a multiword unit, and that some criteria, such as decategorialization, may be less indicative than others. The study proposes a systematic, corpus-based approach towards identifying and classifying emerging multiword units, so as to improve coverage of their contemporary lexicogrammatical functions within grammars and dictionaries.

Overview of thesis

Chapter 1 Preliminary discussion

Chapter 2 Light verbs in Australian, New Zealand and British English (Paper 1)

Smith, Adam 2009. Light verbs in Australian, New Zealand and British English. In P. Peters, P. Collins and A. Smith (eds.) *Comparative studies in Australian and New Zealand English: Grammar and beyond,* 139-56. Amsterdam/Philadelphia: John Benjamins.

Section 2a Light verbs and grammaticalization

Chapter 3 Non-numerical quantifiers (Paper 2)

Smith, Adam 2009. Non-numerical quantifiers. In P. Peters, P. Collins and A. Smith (eds.) *Comparative studies in Australian and New Zealand English: Grammar and beyond,* 157-180. Amsterdam/Philadelphia: John Benjamins.

Section 3a NNQs and grammaticalization

Chapter 4 Complex prepositions and variation within the PNP construction

(Paper 3)

Smith, Adam 2013. Emergent complex prepositions: variability within the PNP construction. In H. Hasselgård, J. Ebeling, and S.O. Ebeling (eds.) *Corpus perspectives on patterns of lexis,* 153-174. Amsterdam/Philadelphia: John Benjamins.

Section 4a CPs and grammaticalization

Chapter 5 Newly emerging subordinators in spoken/written English (Paper 4)

Smith, Adam 2014. Newly emerging subordinators in spoken/written English. *Australian Journal of Linguistics* 34(1), 118-138.

Chapter 6 Conclusion

Acknowledgements

My first and biggest debt of gratitude is owed to Professor Pam Peters, who has been my constant supporter, research collaborator, colleague and friend throughout my 20 years as a member of the Linguistics Department at Macquarie University. It was she who first suggested a possible theme for this thesis, and who, as a supervisor, provided continual encouragement, advice, expert knowledge, and the infusion of her own boundless energy when my own was flagging. This thesis would not exist without her.

Thanks also to my associate supervisor, Professor Jan Tent, whose close reading and positive comments helped me to draw the thesis together. I must also acknowledge the supportive community of the ICAME (International Computer Archive of Modern and Medieval English) conferences, that provided an environment within which some of these papers could be tested, refined and ultimately published. In particular, thanks to Professor Kristin Davidse for her invariably useful ideas on areas where our research intersected.

Finally, I must thank my wife and children, Jenny, Hannah and Ben, who generously gave me the time and space to carry out this work, and the love and belief to complete it.

Table of Contents

Declaration i

Abstract and Overview ii

Acknowledgements iv

Table of contents v

List of tables xi

List of figures xiii

List of appendices xiii

List of abbreviations used xiv

Chapter 1: Preliminary discussion 1

1. Introduction 1

1.1 Lexis and grammar as separate entities 1

1.2 The integration of lexis and grammar 3

1.3 Emergent grammar 5

2. Lexicalization and Grammaticalization 6

2.1 Classification and criteria: Lexicalization 6

2.2 Classification and criteria: Grammaticalization 8

2.3 Linguistic and discourse contexts for category change 11

2.3.1 Linguistic contexts 11

2.3.2 Discourse contexts 12

2.4 Analogy 13

3. Challenges for language description and analysis 14

3.1 Forms of multiword units 14

3.2 Lexical meaning/grammatical structure 15

4. Examples of multiword units to be investigated 16

4.1 Verb phrase – Light Verbs 16

4.1.1 Previous literature 16

4.1.2 Current research 17

4.2 Noun phrase – Non-numerical quantifiers (NNQs) 17

4.2.1 Previous literature 18
4.2.2 Current research 18
4.3 Prepositional phrase – Complex preposition 19
4.3.1 Previous literature 19
4.3.2 Current research 19
4.4 Complex subordinators 20
4.4.1 Previous literature 20
4.4.2 Current research 20
5. Corpus methodology 20
5.1 Using corpus data 20
5.2 Choice of corpora 21
5.3 Corpus data 22
5.3.1 Relative frequency 22
5.3.2 Regional variation 23
5.3.3 Variation across different genres 24
5.3.4 Diachronic changes 25
5.3.5 Evidence of indeterminate contexts from corpus data 25
6. Summary – Research questions 26
Light verbs in Australian, New Zealand and British English 28
1. Definition of "light verb" 28
2. Evidence for regional divergence 30
2.1 Research studies 30
2.2 Dictionary evidence 32
3. Frequency of common light verbs in the ICE corpora 35
3.1 Spoken vs written 36
3.1.1 Informality vs. interpersonal courtesy 36
3.1.2 Hedging and other uses of pre-modification 38
3.1.3 Colloquial productivity 40
3.2 Regional and temporal differences 41
3.2.2 Corpus comparisons 41
3.2.3 Regional choices of have or take 43

4. Conclusions 44

Section 2A: Light verbs (LVs) and grammaticalization 47

1. Issues of lexicalization and grammaticalization 47

2. Grammars 48

3. Dictionaries 49

4. Summary 50

Non-numerical quantifiers 52

1. Introduction 52

2. Classification of NNQs 53

3. Issues that apply to NNQs 55

3.1 Grammar: Verbal agreement and noun complementation 55

3.2 Collocation 56

3.3 Semantic weight 57

3.4 Variation 57

3.4.1 Regional divergence 57

3.4.2 Register 58

4 Previous corpus studies 58

5 Corpus findings 59

5.1 A lot/lots of 59

5.1.1 Regional and register features 60

5.1.2 Grammatical features 60

5.2 ONNQs that have a singular or plural quantifying noun 64

5.2.1 Regional and register features 65

5.2.2 Collocation with heap(s)/load(s)/bunch(es) 68

5.2.3 Grammatical features 69

6 NNQs with singular or plural forms only 72

7 Conclusions 74

References 75

Appendix 3.1 77

Section 3A: Non-numerical quantifiers (NNQs) and grammaticalization 79

1. Issues of lexicalization and grammaticalization 79

2. Grammars 80

3. Dictionaries 80

4. Summary 81

Complex prepositions and variation within the PNP construction 83

1. The status of the CP as a grammaticalised unit 83

2. Types of variation 84

3. Variation and grammaticalisation 85

3.1 Previous studies on the grammaticalisation of PNP sequences 86

4. Representation of CPs in contemporary grammars 88

5. The definite article and the semantics of the noun in CPs 89

6. Investigation of evidence of use of CPs in corpora 91

6.1 Summary of corpora used 91

6.2 Approach to using corpora 92

7. Patterns of loss of the definite article 94

7.1 (the) place of 95

7.2 by (the) force of 95

7.3 on (the) ground(s) of 96

7.4 in (the) light of 97

7.5 at (the) risk of 98

8. Complementation patterns 99

8.1 collocates for at (the) risk of 100

8.2 in (the) process of 101

9. CPs that invariably include the definite article 103

10. Conclusion 104

Section 4A: Complex prepositions (CPs) and grammaticalization 107

1. Issues of variation, grammaticalisation and lexicalisation 107

2. Grammars 107

3. Dictionaries 108

4. Summary 108

Newly emerging subordinators in spoken/written English 110

1. Introduction 110

2. Questions of categorization 112

2.1 Gradience between grammatical categories 112

2.2 Complex subordinators 112

2.3 From adverbial adjunct noun phrase to subordinator. 114

2.4 Temporal adverbials as subordinators 115

3. Contexts of category change 117

3.1 Preposition + *that* 118

3.2 Zero Preposition + *that* 119

3.3 Preposition + Zero *that* 119

3.4 Zero Preposition + Zero *that* 120

4. Subordination in spoken and written genres 121

5. Investigation of data 122

5.1 Selection of corpora 123

5.2 ICE-AUS/ICE-GB/ICE-USA data 124

5.2.1 Instant/minute/moment 125

5.2.2 Day/time 127

5.2.3 Summary of ICE data 128

5.3 COCA/BNC data 128

5.3.1 Evidence for fully elliptical forms 128

5.3.2 Genre differences 130
5.3.3 Regional differences 131
6. Conclusions 132
Appendix 5.1 – ICE-GB Abbreviations for Word Classes, Functions and Categories 136
Chapter 6: Conclusion 137
1. Evaluating the criteria for assessing the lexicogrammatical status of multiword units .. 137
1.1 Fixity versus Variability 137
1.2 Semantic bleaching 140
1.3 Decategorialization 142
1.4 Syntactic reanalysis 144
2. Synthesizing the criteria 145
2.1 Established vs. emergent forms 145
2.2 Comparing fixity with other criteria: evaluating the set 147
3. Classification of multiword items in grammars and dictionaries 151
3.1 Grammatical categorization 152
3.1.1 Student and Learner grammars 153
3.2 Dictionary representation of multiword items 155
3.2.1 Grammatical labelling and subcategorization 156
3.2.2 Location of multiword items in dictionaries for learners and native-speakers 157
3.2.3 Dictionaries and grammatical gradience 159
4. Summary answers to research questions 160
5. Limitations of current study 161
6. Future research 162
Consolidated References 163

List of tables

Table 1.1: Lehmann's parameters/processes of grammaticalization 9

Table 1.2: Heine and Diewald's models compared 12

Table 1.3: Corpora used in the study 22

Table 2.1: The main attributes of TLVs according to Kearns (2002) 30

Table 2.2a: Coverage of light verbs in American, British, Australian and New Zealand L1 dictionaries 33

Table 2.2b: Coverage of light verbs in American, British, Australian and New Zealand L2 dictionaries 34

Table 2.3: The light verb construction in ICE-GB, ICE-AUS and ICE-NZ 36

Table 2.4: Frequency of have a look in ICE-GB, ICE-AUS and ICE-NZ 37

Table 2.5: Occurrences of *have a look* across ICE spoken categories 37

Table 2.6: The occurrence of pre-modified against unmodified light verbs in ICE-GB, ICE-AUS and ICE-NZ 39

Table 2.7a: Light verbs in the LOB and Brown corpora (adapted from Algeo 1995) 42

Table 2.7b: Light verbs in ICE-GB, ICE-AUS and ICE-NZ 42

Table 2.8: Choice of *have/take* in the ICE corpora 43

Table 3.1: Frequency of *a lot/lots of* in ICE-AUS, ICE-NZ, ICE-GB 60

Table 3.2: Noun complementation and verbal agreement for *a lot/lots of* in ICE-AUS, ICE-NZ, ICE-GB 61

Table 3.3: Verbal agreement with noun complement of *a lot of/lots of*: *existential there* construction preceding NNQ compared with verb following NNQ 62

Table 3.4: Verbal agreement with number of noun complement and/or NNQ for *a lot/lots of* 63

Table 3.5: Comparison of *a lot/lot of* with alternative singular/plural ONNQs 65

Table 3.6: Frequency of alternative singular/plural ONNQs in ICE-AUS, ICE-NZ, ICE-GB 66

Table 3.7: Collocations with *heap(s)/load(s)/bunch(es)* 68

Table 3.8: Noun complementation and verbal agreement for *a lot of/lots of* compared with alternative singular/plural ONNQs 70
Table 3.9: Verbal agreement with noun complement of alternative singular/plural ONNQs: existential *there's* preceding ONNQ compared with verb following ONNQ 71
Table 3.10: Verbal agreement with noun complement or ONNQ for singular/plural ONNQs 71
Table 3.11: NNQs common and specific too different regions 72
Table 4.1: Set of CPs listed in grammars, with or without definite article 89
Table 4.2: List of corpora used in this study 92
Table 4.3: Frequencies of CPs from grammars that can take *the* 93
Table 4.4: Top 20 most frequent R1 collocates for at (the) risk of in COCA 100
Table 4.5: Top 20 most frequent R1 collocates for in (the) process of in COHA/COCA 103
Table 5.1: Examples of complex subordinators 113
Table 5.2: Variant forms of possible emergent complex subordinators within ICE-AUS/ICE-GB/ICE-USA 125
Table 5.3: Variant forms of possible emergent complex subordinators within COCA/BNC (normalized per 1m words) 129
Table 5.4: zero prep. + moment/instant/minute + zero relativizer in spoken and written genres 130
Table 6.1a: Established and emergent forms by frequency in COCA and BNC 146
Table 6.1b: Scale of fixity for established and emergent forms in COCA and BNC 147
Table 6.2a: Criteria of fixity applied to established and emergent forms 147
Table 6.2b: Non-quantifiable criteria applied to established and emergent forms 148
Table 6.2c: Combined grammaticalization criteria applied to established and emergent forms 149
Table 6.3: Treatment of complex subordination in a native-speaker and learner dictionary 158

List of figures

Figure 1.1: Conception of language as separate elements 2

Figure 4.1: Normalised frequencies for *on (the) ground(s) of* by decade, in COHA 97

Figure 4.2: Normalised frequencies for *in (the) light of* by decade, in COHA 97

Figure 4.3: Normalised frequencies for *at (the) risk of* by decade, in COHA 98

Figure 4.4: Normalised frequencies for in (the) process of by decade, in COHA 102

Figure 5.1: Grammatical analysis from ICE-GB of an example of preposition + *that* 118

Figure 5.2: Grammatical analysis from ICE-GB of an example of zero preposition + *that* 119

Figure 5.3: Grammatical analysis from ICE-GB of an example of preposition + zero *that* 119

Figure 5.4: Grammatical analysis from ICE-GB of an example of zero preposition + zero *that* 120

List of appendices

Appendix 3.1 77

Appendix 5.1 136

List of abbreviations used

AmE	American English
ART	Australian Radio Talkback corpus
AusE	Australian English
BNC	British National Corpus
BrE	British English
COCA	Corpus of Contemporary American English
COHA	Corpus of Historical American English
CP	Complex Preposition
CS	Complex Subordinator
GloWbE	Global Web-based English corpus
ICE	International Corpus of English
LOB	Lancaster-Oslo/Bergen corpus
LongAm	*Longman Dictionary of American English*
LongBr	*Longman Dictionary of Contemporary English*
LSWE	Longman Corpus of Spoken and Written English
LV	Light Verb
MALD	*Macquarie Australian Learners Dictionary*
MD	*Macquarie Dictionary*
NNQ	Non-numerical Quantifier
NODE	*New Oxford Dictionary of English*
NZE	New Zealand English
NZOD	*New Zealand Oxford Dictionary*
OED	*Oxford English Dictionary*
ONNQ	Other Non-numerical Quantifier
RHD	*Random House Dictionary*
SBC	Santa Barbara Corpus
TLV	True Light Verb

Chapter 1: Preliminary discussion

1. Introduction

The research presented in this thesis is comprised of a set of research studies on language innovation and change, concentrating on synchronic variation which, as Aitchison argues (1991: 33ff) is a prerequisite for diachronic change. The focus is on multiword units that provide an environment where there is a potential disjunct between the grammatical role of the unit as a whole and the elements within it. The four types of multiword unit covered – light verbs, non-numerical quantifiers, complex prepositions and complex subordinators – have been chosen to represent two open grammatical classes (verb phrase and noun phrase), and two closed classes (preposition and subordinator).

This introductory chapter will start by summarizing diverse theoretical models: those where the grammatical system is seen as being relatively closed and resistant to change, as opposed to the open, fluctuating nature of lexis; and others where the grammar and lexis of language are seen as being interconnected and susceptible to change. The second section looks more closely at the processes involved in change – lexicalization and grammaticalization. In the third section, challenges to the description and analysis of multiword units in grammars and dictionaries are discussed. The fourth section presents the multiword units to be covered in the research, with a brief overview of previous literature and current research on each. The fifth section presents the choice of corpora and methodology used, and the sixth section summarizes the research questions.

The four peer-reviewed, published papers devoted to each of the multiword units under discussion follow the introduction as Chapters 2 to 5, with brief linking passages between them to highlight the common themes and present any relevant research that has become available since the papers' publication. The concluding chapter brings together the different aspects of the research, and addresses the initial research questions raised.

1.1 Lexis and grammar as separate entities

Traditional descriptions of language make a clear distinction between lexis and grammar. They view the lexicon as a fluid, open system, where words can shift referential and connotative meaning, or can be freshly minted, and can constantly recombine to form new

lexical units in the form of compounds or phrases. The grammar of a language interacts with these lexical choices to develop propositional meaning. Through morphology, grammatical meanings such as number and tense are added to the lexical content, and syntactic rules provide the tools to combine the lexical units meaningfully. There is a limited set of grammatical words which belong to relatively closed classes (preposition, determiner, pronoun, conjunction) in contrast to the open classes of lexical words (nouns, verbs, adjectives, adverbs).

This view is also the basis of transformational grammar: words are acted upon by rules, in Pinker's (1999) terms. Our innate grammatical rules act upon our mental lexicon to generate meaningful utterances. This conceptualisation of language is represented through the more complex model in Figure 1.1, where the further components of language are identified in separate boxes.

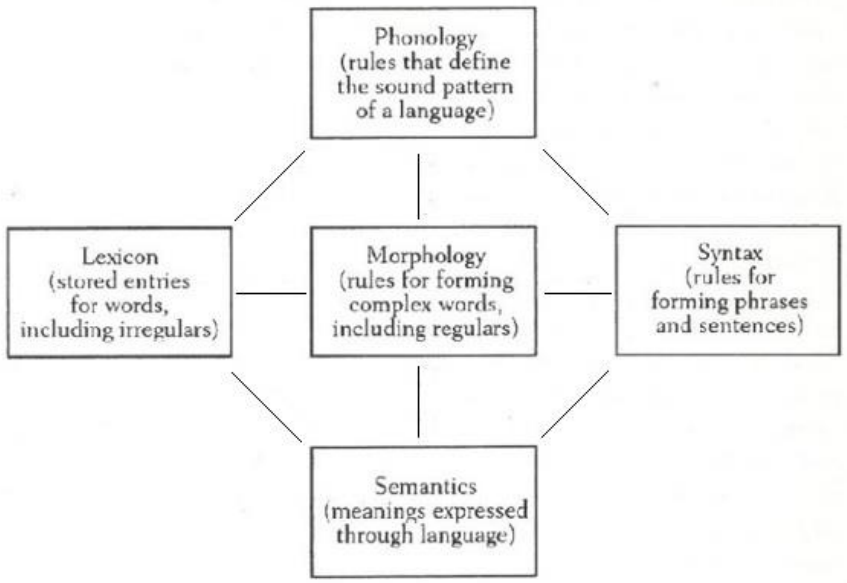

Figure 1.1: Conception of language as separate elements (adapted from Pinker, *Words and Rules*, 1999: 23)

Pinker's five part model includes elements other than lexis and grammar, with phonology and semantics added to show the complementary paths, on the vertical axis, for the composition of language through sounds and the expression of meaning. On the horizontal axis, the elements progress from the individual, stored items in the lexicon, to the morphological rules that are available to create complex words, followed by the rules

of syntax for forming larger constructions such as phrases and sentences. It should be noted also that this is a representation only of spoken language – orthography would also need to be added to represent written language.

The lexicon and syntax are placed at opposite ends of the horizontal scale, with the members of the lexicon including all items that are considered to be stored as individual units within the brain (including irregular, and therefore non-rule-based formations of words). Complex words are formed by morphological rules, and complex units of language, such as phrases or sentences, formed by syntax. This approach coincides with the traditional separate classification of language elements by lexicographers. Dictionaries focus on the lexicon, with grammatical class subsidiary to referential meaning, while grammars codify the syntactic functions of words and units, without engaging with their semantics.

1.2 The integration of lexis and grammar

Linguists' reactions to formal grammars tend to break down this compartmentalization of lexis and grammar, allowing for a much closer relationship between semantics and grammar. Halliday, in proposing the notion of "lexicogrammar", suggests that "grammar and vocabulary are merely different ends of the same continuum" (1994: 15). This is central to his conceptualization of a functional grammar in that he is "foregrounding its [grammar's] role as a resource for construing meaning" (ibid). This focus on meaning allows for the concept of "grammatical metaphor", where one grammatical class can be substituted for another, as in the process of nominalization, whereby a verbal process is replaced with a noun: *His eating of the cake* instead of *He ate the cake*.

A link between this more lexis-centred approach and transformational-generative grammar is made by "lexical functional grammar" which presents language as being made up of various structures (for instance "semantic structure", "morphological structure", "argument structure") which are "mutually constraining" (Kaplan & Bresnan, 1982). This interdependency replaces the need for transformations to provide a connection between deep and surface structure in Chomskyan theory.

Another kind of grammatical theory that combines grammar and lexis is cognitive grammar, which emphasizes the meaningfulness of grammar, rather than its acting as an abstract set of rules:

> Portraying grammar as a purely formal system is not just wrong but wrong-headed. I will argue, instead, that grammar is meaningful. This is so in two respects. For one thing, the elements of grammar – like vocabulary items – have meanings in their own right. Additionally, grammar allows us to construct and symbolize the more elaborate meanings of complex expressions (like phrases, clauses, and sentences). (Langacker, 2008: 3-4)

Cognitive grammar derives its conception of how language is structured from psychological approaches to how the brain processes language. Langacker speaks of his conception of grammar as emerging "organically from a comprehensive and unified view of linguistic organization characterized in terms of cognitive processing" (1987: 1). A similar conclusion on the close relation between grammar and lexis is reached by theorists of construction grammar, such as Fillmore et al. (1988) and others since, who highlight the fusion of grammatical and lexical meanings in the myriad constructions sanctioned by usage.

The advent of corpus linguistics has led some, such as Sinclair, to argue that the distinction between lexis and grammar is a false one, describing how the compilers of the *Cobuild Dictionary* found "There was in practice no clear distinction between grammar and lexis" (1987: 110). The use of corpora focuses on the surface of language, rather than postulating underlying rules. Arising out of this corpus-driven approach is "pattern grammar", which explains all language production in terms of words and their patterns of usage. In this view it is the lexicon and lexical meaning that governs the formation of all utterances, as explained by Hunston & Francis), "certain patterns 'select' words of particular meaning" (2000: 29). This is an extension of Firth's concept (1957) of colligation, which describes the tendency of certain grammatical categories to co-occur. In pattern grammar it is the semantic field that determines the grammatical structure of the language.

This theoretical model of a close integration between lexis and grammar opens to question the traditional concept of lexis as containing elements that are constantly evolving, in contrast to the relatively fixed set of grammatical or function words.

1.3 Emergent grammar

If we accept this closer relationship between meaning and structure, then the conventional dichotomy between closed-class function words (such as auxiliaries, conjunctions and prepositions) and open-class lexical words (such as nouns, verbs and adjectives) becomes harder to justify. Even closed classes admit new members over time, for example with the class of prepositions allowing additions such as participles like *concerning*, *including*.

This crossing of boundaries indicates that there are no clear divisions between classes, but rather a cline, or gradience. Aarts (2007) describes this interaction between different grammatical classes as "intersective gradience". He claims that there is also a scale within classes – where certain items are relatively core, or non-core (and therefore have greater potential for category change) – which he labels "subsective gradience". An example of these is the class of modals, where some are less central than others. For instance *ought* is considered a marginal modal compared to a prototypical model like *should* (see Denison, 2001: 124), because of its tendency to pattern as a main verb – as in the non-standard construction *didn't ought* as opposed to *shouldn't* (Quirk et al., 1985: 140).

Grammatical forms cannot be said to be fixed members of closed sets, but rather are constantly emerging. Hopper explains this concept of "emergent grammar" as arising out of discourse, and that it "is shaped by discourse as much as it shapes discourse in an on-going process" (1987: 142). It is out of discourse that multiword units arise that have the potential to cede the lexical properties of their individual elements to an integrated grammatical role.

The process whereby lexical items attain a grammatical function, and multiword units are invested with a single grammatical function, is known as grammaticalization. It is a phenomenon more accepted by grammarians with a descriptive approach to language structure (e.g. Quirk et al., 1985; Biber et al., 1999) than those whose more formal, structural approach is disinclined to allow for the creation of new grammatical elements (e.g. Huddleston & Pullum, 2002). As Hopper (1998: 148) remarks: "A wider view of grammaticalization demands a modification of our perspective on grammar, one which sees structure in language as intrinsically unfixed and unstable, in other words as emergent".

The emergence of new grammatical units is made possible by the indeterminacy in the distinction between lexical and grammatical identity, and by the potential for words to function in different grammatical roles, change category, or become part of a unit with a different grammatical function. For example: *since* can perform the function of a preposition (*since the meeting*), or a subordinator (*since we met*), words which start their life as nouns (e.g. *rubbish*) can expand their role to adjectives (*you're rubbish at football*) and verbs (*she rubbished his ability*); and lexical items can take on grammatical roles (e.g. *do* as a verb meaning "provide", as in *do the catering*, compared to its use as an auxiliary in forming negative and interrogative clauses).

How these changes occur is a question that merits further investigation. According to Auer and Pfänder "Emergent grammar […] is more interested in the vague boundaries of grammatical categories and units than in their prototypical centre" (2011: 7). In other words, they suggest that there are core units of lexicogrammar which are relatively fixed as to form and function. However beyond these central items there are variable boundaries where new forms and functions can emerge. Historical linguists have shown how indeterminate contexts contribute to the formation of multiword units and their acquisition of new grammatical functions (see further Section 2.3). This demonstrates the need to accept a degree of interplay between lexis and grammar, and to recognize the related processes of lexicalization and grammaticalization.

2. Lexicalization and Grammaticalization

2.1 Classification and criteria: Lexicalization

Just as the boundary between lexis and grammar is not necessarily clear-cut, the distinction between the processes of lexicalization and grammaticalization is a topic of debate. At the most basic level, they can be distinguished as the processes whereby a new lexical or grammatical unit (single or multiword) is formed from the juxtaposition of discrete lexemes or morphemes, rather than being derived from them. So, *saucepan* has been lexicalized as a single unit from *sauce* + *pan*, and multiword units like *telephone box* have their own lexical identity and meaning which are more than the sum of their parts. These are to be distinguished from units that have gained new grammatical functions, such as *because* (conjunction), *however* (adverb) or *in back of* (preposition), which have all been

composed from discrete elements. But as already indicated, this straightforward dichotomy between being lexicalization and grammaticalization may not apply to every example.

Lexicalizations can range from non-semantically integrated compounds such as *dining room*, to fully integrated compounds like *blackmail, mincemeat*. They are semantically opaque, like idioms such as *smell a rat*, but "opacity is not a necessary pre-requisite for lexicalization" (Bauer 1983: 49), although it is generally considered to be for idiomaticity.

The process of lexicalization is described by Bauer (1983) as developing from nonce formation to institutionalization to the final stage of acceptance as a fully lexicalized form. Within this process, he describes four different types of lexicalization: phonological, morphological, semantic and syntactic (although many examples may embody more than one type). The different types are summarized below (1983: 50-61:

- Phonological, e.g. *husband*, where the bound *hus* element has been isolated from the vowel shift to *house* in the free morpheme
- Morphological, e.g. nominalizations of adjectives ending in *-th*, like *warmth, length*, which are lexicalized because the *–th* suffix is no longer productive in English
- Semantic, where semantic information is either lost, e.g. *understand*, which retains none of the semantic content of its elements *under* and *stand*, or gained, as in *wheel chair/push chair*, which have the semantic markers of "for invalids" and "for children" respectively added to their constituent lexical elements.
- Syntactic, e.g. *out of the way*, which has the form of a prepositional phrase, but can function as an attributive adjective, as in *an out of the way place*

This last type looks more like producing examples of grammatical units, and here we run into a definitional problem that allows the proposition that "lexicalization is sometimes considered as the final stage of grammaticalization" (Ramat, 1998: 120).

This statement stems from a broad conception of lexicalization that regards any fusion of lexical elements as an example of lexicalization, regardless of the function of the unit formed. So, *because* is a lexicalization of the elements *by + (the) cause*. Brinton and Traugott (2005) argue that fusion ("the loss of original morphological boundaries") and coalescence (phonological change such as assimilation, reduction etc.) are processes where the borderline between lexicalization and grammaticalization are often disputed. They give examples such as composite predicates like *lose sight of*, which are fused as a

lexical unit, but can also be seen as grammatically productive (2005: 66-67). Examples of coalescence include phrasal discourse markers like *innit* (from *is it not*), which some scholars analyse as lexicalized because they involve the form becoming invariant, inseparable and morphologically opaque. They are more commonly described as resulting from grammaticalization as they "typically involve decategorialization and reanalysis as functional elements within the domain of discourse" (2005: 67).

With the processes of lexicalization and grammaticalization being so similar, some scholars distinguish them on the basis of the resulting collocation or word. Wischer, for example, argues that for lexicalisation "a specific semantic component is added", whereas for grammaticalization "specific semantic components get lost and an implied categorial or operational meaning is foregrounded" (2000: 364-5). This can be contrasted with Hopper's argument that "If grammar is not a discrete, modular set of relationships, it would seem to follow that no set of changes can be identified which distinctively characterize grammaticization as opposed to, say, lexical change or phonological change in general" (1991: 19). It therefore becomes almost impossible to construct a set of principles that can distinguish between grammaticalization and lexicalization.

2.2 Classification and criteria: Grammaticalization

What previous authorities have done is explore the features and processes that typically occur within grammaticalization, which are used as indicators of grammaticality, but are not exclusive to it. Lehmann (1985: 309) presents a set of six grammaticalization parameters and their associated processes, see Table 1.1.

Table 1.1: Lehmann's parameters/processes of grammaticalization

parameter	weak grammaticalization	— process →	strong grammaticalization
integrity	bundle of semantic features; possibly polysyllabic	— *attrition* →	few semantic features; oligo- or monosegmental
paradigmaticity	item participates loosely in semantic field	— *paradigmaticization* →	small, tightly integrated paradigm
paradigmatic variability	free choice of items according to communicative intentions	— *obligatorification* →	choice systematically constrained, use largely obligatory
scope	item relates to constituent of arbitrary complexity	— *condensation* →	item modifies word or stem
bondedness	item is independently juxtaposed	— *coalescence* →	item is affix or even phonological feature of carrier
syntagmatic variability	item can be shifted around freely	— fixation →	item occupies fixed slot

Not all grammaticalized items will display all of these characteristics, but Lehmann posits this as a complete set that different items possess to varying degrees. Other researchers present slightly different sets of features, with different focuses. So Boye and Harder (2012: 28-31) respond to Lehmann's list with the following features: ***boundness, phonological reduction, semantic reduction, closed-class membership*** and ***obligatoriness***. They overlap somewhat for example their "boundness" is basically equivalent to Lehmann's "bondedness", and their "obligatoriness" involves constraint in the choice of items (=obligatorification), while putting greater emphasis on some elements, such as the semantic reduction or bleaching of lexical elements. They comment that each of these are symptomatic of grammaticalization but can also be observed in lexicalized expressions.

An alternative set of "principles" was proposed by Hopper (1991: 21-2) who argues that Lehmann's parameters apply to an advanced rather than early stage of grammaticalization. For the diagnosis of potentially emergent grammatical forms and constructions Hopper proposes the following:

- Layering – where new "layers" can exist within a functional domain, such as tense/aspect/modality, and can coexist and interact with existing layers (e.g. future tense, *be about to* with *will, be going to be + –ing, be + to*)
- Divergence – where a lexical form grammaticalizes, say to an auxiliary, clitic or affix, while the original form remains as a lexical element (for example the French

word *pas* as a negative particle, from the noun meaning "step"). This is described as a special case of layering.

- Specialization – where within a functional domain, a set of words with slightly different semantic nuances may be possible, but this variety of choice narrows to produce a smaller set with more general grammatical meanings. Again *pas* is an example, with words like *mie* ("crumb") and *gote* ("drop") originally being used to reinforce *ne* as a negative, eventually reducing to the choice of *pas* as a general negative.
- Persistence – where traces of the original semantic meaning cling to grammaticalized forms. The future *will*, it is argued, has developed from the volitional intention or promise of a future event (*I will put them in the post*) to the non-semantically-related predictive in *I think most students will pass*. This is the other side of the coin to Boye and Harder's (2012) semantic reduction, where the focus is the loss of semantic weight through grammaticalization.
- Decategorialization – where forms lose the morphological markers and syntactic characteristics of full categories Noun and Verb to assume attributes characteristic of secondary categories such as Adjective, Preposition, etc. For example *considering* as a preposition (*Considering the bad weather, we had a good holiday*) can only take the present participle form, and cannot be qualified by an adverb.

It is clear from these divergent approaches and terminologies that there is no easily applied formula to distinguish grammaticalization from lexicalization, and that the criteria can only be used as indicators of possible grammaticality rather than definitive diagnostic tools. Hopper's principles are better designed for the kind of emerging multiword forms that this research will focus on, although some of the processes presented by Lehmann, such as fixation, will also bear discussion.

It should be noted that fixity is a factor for both lexicalization and grammaticalization, which isn't mentioned specifically in the criteria above as it pertains particularly to the formation of phraseological unification. Hudson employs the term "structuration" to distinguish between the processes of phrasal fixation and affixation (1998: 161). Fixity is not an easy thing to measure, and it could be argued that the potential for variability in a phrase disallows its categorization as a unit. However, frequency can be used to measure the degree of fixity of an item. As Boye and Harder

write "in order for an expression to come to be conventionalized as grammatical, it must be used with a certain frequency" (2012: 27) – though what that level of frequency is, remains to be seen.

While the principles above serve as useful indicators for the grammaticalization of individual items, Hopper puts forward another important consideration for judging the grammaticality of certain forms: Traugott's notion of "loss of discourse autonomy", especially in the case of decategorialization. A central feature of grammaticalization is that forms change their meanings from "propositional" to "textual" – that is, "from having a meaning independent of the text they change to having a meaning or function that is relative to the text or to some local construction" (1991: 31). In order to observe a change in function, we must look beyond independent units to the contexts in which they operate.

2.3 Linguistic and discourse contexts for category change

When considering context we need to look both at the linguistic and the discourse context where grammatical changes may occur.

2.3.1 Linguistic contexts

Several authors have discussed the linguistic contexts where a word or phrase can change grammatical category, amongst them Diewald (2002), Heine (2002), Denison (2010) and Traugott (2012). Denison looks at synchronic contexts where the distinction between adjectival and noun use of a word like *rubbish* (see Section 1.3) can be "neutralized" (2010: 110). Diewald and Heine take a diachronic approach, identifying types of context which relate to different stages of grammaticalization. Their models, as demonstrated below in Table 1.2, are slightly different in that Heine focuses on semantic changes, while Diewald's has more emphasis on morphological and structural aspects (Diewald, 2002: 117).

Table 1.2: Heine and Diewald's models compared (from Traugott, 2012: 230)

Heine	Diewald
	Stage 0: "normal" use
Stage I: "normal" use	**Stage I**: "untypical" context
Stage II: "bridging" context (pragmatic, semantic)	**Stage II**: "critical" context (multiple opacity: pragmatic, semantic, structural)
Stage III: "switch" context (Grammaticalization)	**Stage III**: "isolating" context (Grammaticalization: reorganization and differentiation)
Stage IV: conventionalization	

The common factor between Heine's "bridging" and Diewald's "critical" context is an element of ambiguity – in the classical sense of offering alternative interpretations, rather than the more general use of the word to indicate vagueness. Traugott illustrates the ambiguity inherent in the bridging/critical context in the development of *be going to* from verb of motion to future auxiliary in this example from 1692 "Mr Ennis…**is** this weeke **going to** try whither he cannot more quietly live among ye heathens in America" (2012: 236), where the presence of temporal expression makes the future tense interpretation a possibility. This is contrasted with the switch/isolating context of this 1699 example: "There is one Mr Colson I am shure my Lady has seen at diner with my Unckle **is going to** be married" where any concept of motion is "virtually ruled out" (ibid.).

2.3.2 Discourse contexts

The communicative context can be as important as the linguistic one in supporting grammatical change. It is widely accepted that innovation occurs more frequently in spoken interactions than in writing (Andersen, 2006; Traugott, 2008), and therefore linguistic evidence for emerging grammatical units and functions is likely to be discovered first in spoken data, before they are codified in the written medium. Specifically, Traugott looks at this in relation to WH- and ALL- pseudo-cleft constructions, such as *What/All I wanted was…*, finding that both had their origins in dialogic contexts. The spoken medium has also supported the development of semi-modals such as *be going to*, *have got to* and *had better*, according to Biber et al. (1999: 1051), with speech further contributing to their grammaticalization in the fusion of their elements by reduced pronunciation – orthographically represented by the spellings *gonna*, *gotta* and *better*.

In Traugott's terms, spoken discourse gives a context that promotes grammaticalization because such change is naturally the outcome of "speaker-hearer negotiation of meaning" (2012: 231). This view contrasts with the concept of change being primarily internal to language, which is suggested by those who support analogy as being the main mechanism of linguistic innovation, and we shall look at this process in the next section.

2.4 Analogy

The process of analogy is sometimes associated primarily with phonemic change within morphological paradigms. It can be a means either of narrowing of the range of variation (sometimes called levelling) – as in the regularization of *leapt* to *leaped* – or of extension, as exemplified by the variant past forms available for the verb *sing* – *sang* and *sung* (Bybee 2010: 66-8). The question of whether such morphological variability is rule-governed or by analogy is discussed by Bauer who suggests that rule-governed morphology may be associated with productivity, and analogy with creativity (2001: 92). Going beyond purely morphological concerns, a closer relationship between analogy and the processes of grammaticalization is suggested by Hopper and Traugott (2003), in a direct comparison between the effects of reanalysis and analogy:

> Reanalysis essentially involves linear, syntagmatic, often local, reorganization and rule change. It is not directly observable. On the other hand, analogy essentially involves paradigmatic organization, changes in surface collocations, and in patterns of use. Analogy makes the unobservable changes of reanalysis observable. (2003: 68)

Again, the development of the auxiliary *be going to* is given as an example of this relationship:

Stage I		be	going	[to visit Bill]
	PROG		Vdir	[Purp. Clause]
Stage II		[be going to]	visit Bill	
	TNS		Vact	
Stage III	[be going to]	like Bill		
	TNS	V		

The shift from progressive + directional verb to future auxiliary with a verb of activity is reanalysis, and the extension "of the directional class of verbs to all verbs, including stative verbs" (ibid.) is achieved by a process of analogy.

Hopper and Traugott's model is directly disputed by Fischer (2008), who argues that analogy has a much more central role. Adopting a cognitive approach, she claims that frequency of use automatizes both formulaic phrases like fixed collocations and idioms, and more abstract morphological and syntactic sequences such as the [NPS VP NPO] of English declarative sentences. Analogy should therefore "be seen as both a mechanism and a cause […] By means of analogy we may change structures and the contents of paradigmatic sets, but it is also analogy that causes the learner to build up more abstract types of schemas. In other words, in this learning model analogy is the primary force (and not reanalysis as argued in the grammaticalization model of Hopper and Traugott […]" (2008: 350). The relationship between the processes of grammaticalization and analogy in effecting language change is not clearcut, and it will be one of the purposes of this research to examine how they are manifested in the multiword units to be investigated.

3. Challenges for language description and analysis

The potential for the evolution of new grammatical units, or for grammatical entities to change categories offers potential difficulty of classification in dictionaries and grammars. We can see the problems caused for both lexicographers and grammarians particularly if we look at multiword units. These do not fit easily into the atomistic schema of dictionaries, which are built around the identities of individual words, or into the neat distinction between content and function words that clearly defines a set of words to be treated as lexical or grammatical. Many multiword units contain a combination of content and function words, as in multi-word verbs like *face up to*.

3.1 Forms of multiword units

For lexicographers, the first question that has to be asked is at what point does a particular collocation of words become a unit that needs defining in its own right, or recorded as a conventional phrase. Fixity is one criterion that can be used, and the frequency of such lexical units can be measured by using corpora. However, fixity does

not necessarily equate with familiarity – indeed the familiarity of certain idioms can lead to the kind of creative variation that adapts the notion of putting something "on the back burner" to moving something "to the front burner" as its importance increases.

Fixity is also a consideration for grammarians in deciding whether a multiword unit has achieved its own status as a grammatical entity. For example, within the set of phrases formed as P + NP + P, commonly called complex prepositions, there may be evidence of variation, for example *with/in regard(s) to,* or *in (the) face of*. For some grammarians this variation acts as proof that these units cannot be labelled as a word class with a unifying identity, while for others the variation simply indicates there may be more and less core members of a class. This latter stance is exemplified by Aarts' (2007) concept of "intersective gradience". The relative fixity of such items can be ascertained by using corpus evidence, both synchronic and diachronic, to discover whether there are concurrent major and minor forms, and whether there has been a historical shift towards a particular fixed form.

3.2 Lexical meaning/grammatical structure

It is not straightforward to decide what headword a multiword unit should be listed under in a dictionary, especially in the case of idioms (as opposed to collocations) which take on a semantic greater than the sum of their parts. Landau poses the problem as follows:

> "Should the idiom be placed under the first word, or the most important word? Sometimes the first word is variable, as in *shed* or *throw light on*. Sometimes it is not easy to say which word is more important, as in *hang fire*. Most dictionaries prefer to list idioms under the first word, though exceptions are common." (2001: 107).

Here, the variation in the choice of the verb *shed* or *throw* arises out of their synonymy, and therefore they add little semantic weight to the phrase, whereas the second example demonstrates the opacity of the individual elements of the unit, and therefore the difficulty in assigning them an individual headword.

The grammatical identity of multiword units may be dictated by the class of word that is acting as the head of the phrase, so the complex preposition *in face of* is headed by the preposition *in*. However some multiword units acquire grammatical functions that do not align with the head of the phrase. The identification of the head is not always straightforward, as is demonstrated by examples of ambiguous number agreement, as in "a

total of fifteen people was/were drowned" where the singular verb agreeing with *a total of* marks it as the head, whereas the plural verb marks the plural noun *people* as the head.

4. Examples of multiword units to be investigated

In order to study the variation within multiword units, and the contexts within which they may become lexicalized and/or grammaticalized, a set of units from four different word classes (open and closed), will be investigated. These all have disputed status as to their lexical/grammatical identity, as indicated in the literature, to be introduced for each of them.

4.1 Verb phrase – Light Verbs

Light verb phrases such *as have a look, take a walk* pose an interesting question of classification for linguists. Their status, "hovering…between grammar and lexis" (Algeo, 2006: 269) offers a useful example of the equivocal position of this multiword unit that makes it relevant to this research.

4.1.1 Previous literature

There is a lack of agreement as to what constitutes a light verb construction, which is reflected in the variety of terms to describe them, for example "light or insignificant verbs" (Jespersen, 1931), "composite predicates" (Cattell, 1984) and "expanded predicates" (Algeo, 1995). They are also questionable as to their lexical or grammatical status: "Traugott, for example, concludes that historically the development of composite predicates is 'best construed as lexicalization' (1999: 259). In contrast Huddleston and Pullum treat these forms as grammatically productive, including them in their chapter on clausal complements (2002: 290-6)" (Brinton & Traugott, 2005: 66).

Brinton and Traugott (2005: 130-2) argue that there are in fact two types of "composite predicate", which respectively exemplify processes of lexicalization and grammaticalization. The type represented by *lose sight of* is non-productive, very fixed, semantically idiomatized and non-compositional, and is therefore an example of lexicalization. On the other hand, composite predicates like *give an answer* or *take a look*

are productive in the sense that their template gives rise to many lexical types (e.g. *give a reply*, *take a walk*), and therefore constitute examples of grammaticalization.

As noted, by Halliday (1994: 147) among others, the prevalence of this construction can be ascribed to the potential flexibility given by the scope to modify the noun – *take another quick look*. This very flexibility detracts from the light verb's fixity as a unit, and prevents the possibility of one of the processes associated with grammaticalization – decategorialization of element(s) within the unit. The "nouniness" of the nominal element is emphasized by its potential to be premodified. In addition, it is less likely to undergo the process of semantic bleaching, as it carries the lexical weight of the unit, with the verb being semantically "light".

4.1.2 Current research

The light verb construction is well established both in English and in other languages[1], and therefore it is hard to justify its inclusion as an emergent form – although Leech et al. (2009) do find in their corpus study that it appears to be more established in British than American English. Because the question of the productivity of the *have a look* type relates to its identification as an instance of grammaticalization, this study will look at dictionary and corpus data to investigate instances of regional or colloquial productivity. While light verbs present an interesting case of the blurred line between lexicalization and grammaticalization, and equivocation between nominalization and verbalization, they perhaps do not have the potential for category change.

4.2 Noun phrase – Non-numerical quantifiers (NNQs)

Quantifiers have already been mentioned as a class where category change can occur (see the example of *certain* noted above, Section 2.3). Multiword, non-numerical quantifiers such as *a lot of* have been shown to have shifted status from partitives to quantifiers (see Traugott, 2008), and demonstrate several features the kind of grammaticalization processes we are concerned with here.

---

1 For example French (Abeillé, Godard & Sag, 1998), Hindi (Mohanan, 1994), Japanese (Grimshaw & Mester, 1988), Urdu (Butt, 1995)

4.2.1 Previous literature

As with light verbs, there are challenges of class membership and nomenclature for these non-numerical quantifiers. In Biber et al. (1999) they are separated into the categories *quantifying collectives, unit nouns* and *quantifying nouns*. Quirk et al. (1985) have *closed v. open class quantifiers* for a similar range of items.

For Channell (1994) and Drave (2002) their primary interest in them is as a type of vague language – their indeterminate quantity allowing for communicative strategies such as hedging. Channell does comment on the number ambivalence of *a lot of,* collocating with countables and uncountables. This opens up the question of delexicalization of the nominal element, from its primary lexical role as a descriptor, to functioning as a kind of quantifying collective.

In Traugott (2008), the limited development of *a deal of* as a quantifier is contrasted with *a lot (of)* and *a bit (of)*, which are well established, and have even progressed to an adverbial status where they can be "used as free adjuncts in response to questions, as can more prototypical degree modifiers like *quite* and *indeed*" (p.151).

4.2.2 Current research

The variable singular/plural forms off NNQs such as *a lot of/lots of* give scope for analysis of data on agreement to assess the headedness of this and other quantifiers. Combined with the potential bleaching of the lexical content of a noun like *heap* in the quantifying *a heap of/heaps of* there is the possibility of a decategorialization of the nominal element, which in turn could lead to a structural reanalysis from NP + postmodifier to complex determiner + noun phrase.

The productivity of the template *a* + N + *of*/plural N + *of* will be tested in corpus data to provide evidence for the grammaticalized status of the construction, as for light verbs. There is the possibility for evidence of grammaticalization through instances of specialization of quantifying nouns, along similar lines to Hopper's (1991) example of *pas* as a negative.

4.3 Prepositional phrase – Complex preposition

The grammatical class of prepositions, as has already been posited (Section 1.3), has a stronger propensity than some other "closed" classes to admit new members. It is therefore a strong candidate for evidence of grammaticalization, and complex prepositions such as *in terms of*, *in front of* divide opinion as to their grammatical status.

4.3.1 Previous literature

The variability of the P + NP + P construction gives rise to different approaches to classification. Quirk et al. (1985) describe a set of 'indicators of separateness' that distinguish freely formed expressions that contain the PNP elements from fixed – and therefore, in their terms, grammaticalized – prepositional units such as *in spite of*.

Conversely, Huddleston sees these constructions only as lexicalized idioms at their most fixed, and states that there is a mismatch between their lexical cohesion and their grammatical function: "In *for the sake of the premier*, for example, *for the sake of* belongs together lexically, but grammatically the immediate constituents are not *for the sake of* + *the premier* but *for* + *the sake of the premier*" (1988: 126-7).

Corpus evidence of the fixedness of these units, and the decategorialization of the nominal element of certain complex prepositions, allows Hoffmann (2005) to argue against Huddleston's analysis, and affirm their status as grammaticalized units.

4.3.2 Current research

Hoffmann's study focussed primarily on high frequency, and therefore established and nonvariable items, such as *in terms of* and *in relation to*. This research will look at complex prepositions where there is some degree of variability, in the choice of preposition, as in *with/in regard to*, or the inclusion of the definite article, for example *in (the) light of*. This will allow for a comparison of more and less fixed forms in synchronic and diachronic corpora, to look for the possible recent emergence of new complex prepositions.

The prepositional identity of these units is strongly marked, so the likelihood of category change is small, but the representation of a range of core and less core members of the class in Quirk et al. (1985) indicates the possibility for subsective gradience within the category.

4.4 Complex subordinators

The morphological complexity of established subordinators such as *although* and *whenever* is an indication of the tendency of multiword units to unify with a grammatical function. Complex subordinators such as *in order that* and *as soon as* are commonly acknowledged, but there are more marginal members of the class to consider as candidates for inclusion.

4.4.1 Previous literature

The class of complex subordinator is not acknowledged by all grammars. Quirk et al. (1985) and Biber et al. (1999) show common nomenclature but only partial agreement on membership of a class. Huddleston et al. (2002) do not mention it at all.

There are historical examples of the grammaticalization of simple subordinators from complex units. For example, *while* originated in an adverbial phrase consisting of the noun – *hwile*, meaning "time" followed by the invariant subordinator *þe* (= "that"), and translatable as "at the time that". Historical studies by Edgren (1971) and Häcker (1999), have suggested that similar multiword units such as *the day (that), the minute (that), the moment (that)* also qualify as complex temporal subordinators, although their status is marginal. More recently, Brems and Davidse (2010) and Nykiel (2014) have looked at purposive conjunctions like *in hope(s) that* as examples of emerging complex subordinators.

4.4.2 Current research

This study will focus on the time adverbials such as *the day/moment/time (that)*, which have been identified as at least marginal examples of complex subordinators. Their potential shift in function from temporal adverbials to complex subordinators provides an ideal opportunity for the investigation of possible category change.

5. Corpus methodology

5.1 Using corpus data

As noted in the summaries above, corpus data has been used to investigate and categorize each of the multiword units under investigation in previous studies to some extent, but

the diversity of types of evidence makes it hard to compare the results directly. These studies will attempt to provide a standardized approach, based on usage and using balanced data that will allow us to draw conclusions about the grammatical status of these units. In the process we will look at the similarity or differences in the grammaticalization processes involved for each, in order to seek insights into how we can assess their emergence as lexical or grammatical units.

The different nature of the multiword units covered, both in terms of their ease of identification and their relative frequency, means that different kinds of corpora have to be used for the most efficient retrieval of relevant data. In addition, the choice of corpora used in previous studies to an extent dictates the choice of corpora for comparative purposes. The limitations of corpus methodology must also be acknowledged: that corpora can only provide a snapshot into the overall picture of so varied and rapidly changing an entity as the English language, and any conclusions have to be tempered by reference to the relative scope and structure of the corpora used.

Yet corpus data does provide us with substantial bodies of evidence on how language is actually used, on which fuller understanding of language change can be based – instead of relying on artificially generated examples.

5.2 Choice of corpora

The corpora will be selected on the basis of offering comparable data over different regions, genres and timeframes, to assess the level of structural variability of the items both synchronically and diachronically. As well as structural variability, the data will provide evidence of contexts of use within which category change is most likely to occur. The corpora to be used are summarised in Table 1.3 below.

Table 1.3: Corpora used in the study

Corpus	No. of words	Region	Period	Spoken	Written
International Corpus of English (ICE)					
ICE-AUS	1m	Australia	1990s	Y	Y
ICE-GB	1m	Britain	1990s	Y	Y
ICE-NZ	1m	New Zealand	1990s	Y	Y
ICE-USA	400,000	United States	1990s	N	Y
Santa Barbara	250,000		1988-96	Y	N
Other					
British National Corpus	100m	Britain	1980s-1993	Y	Y
COCA	425m	United States	1990-present	Y	Y
COHA	400m	United States	1810-2010	N	Y

This table includes a wide variety of types of corpora, and a diverse coverage of regions, period and genre. The ICE corpora have the advantage of being directly comparable, as they have been built to the same specifications (except for ICE-USA). The other, larger corpora offer more scope to investigate lower frequency items, and in the case of COHA the possibility of discovering emerging patterns of use over time, but are not structurally compatible, and therefore require some provisos when comparing the data they provide.

5.3 Corpus data

We will now look at the possibilities offered by this range of data across a variety of research criteria.

5.3.1 Relative frequency

The items under discussion have diverse frequency profiles, and will therefore require different approaches towards collecting relevant data:

- ***Light verb constructions*** are typically based on high frequency, general verbs such as *have, take, make, do* (Algeo, 1995), and therefore smaller corpora, such as the ICE corpora listed in Table 1.3, are needed to provide a manageable amount of data from which light verb uses can be extracted.
- The nouns around which ***non-numerical quantifiers*** are built can range from highly frequent, as *in a lot of*, *a number of*, to low frequency of highly specialized

uses, as in *a plethora of*, *a raft of*. For this study, the smaller ICE corpora will be interrogated, to allow analysis of the context of as wide a range of potentially quantifying nouns as possible, to distinguish their use as quantifiers from more semantically-specific functions.

- ***Complex prepositions*** do not tend to be high frequency items and so Hoffmann (2005) used the BNC to extract data on a broad range of current examples. This study will complement that data by looking at some of the lower frequency/more variable items that were not examined by Hoffmann, both in the BNC and the larger American corpus, COCA.
- The range of possible ***complex subordinators*** is much smaller than any of the other categories, but the temporal adverbial expressions that are the focus of this study contain high frequency nouns such as *time, day, minute*. Because the identification of the subordinative role of these expressions is highly dependent on context, the smaller ICE corpora will be used to identify the contexts where their subordinative functions may emerge, and then data from the larger corpora can then test these potential patterns of use.

5.3.2 Regional variation

Emergent forms are likely to arise at different rates within different varieties of English.

- ***Light verbs*** have received little attention in corpus studies, but Algeo (1995) looked at comparative frequencies in the British and American written corpora LOB and Brown. He found some regional differences, particular in the comparative use of the verbs *have*/*take*, which will be tested against the more recent data in a wider range of regions in the ICE corpora.
- ***Non-numerical quantifiers*** have not been formally studied using corpus methodology, but the corpus-based *Longman Dictionary of Contemporary English* labels, for example, *a bunch of* as American English, and *loads of* as British English, suggesting regional comparisons of the type and range of NNQs could prove productive.
- Hoffman's (2005) extensive study of ***complex prepositions*** looks only at British English, so American data will be interrogated to look for patterns in choice of particular CPs across the regions, and in variability of the P + NP + P unit to see if there are different degrees of grammaticalization in evidence.

- The only previous corpus study of any of the ***complex subordinators*** under discussion here was by Häcker (1999), who found that *the minute (that)* was used as a subordinator in Scots English, but there has been no systematic study of the temporal subordinators as a set. As these are the least attested of the multiword units in this study, the various corpora will be used to find evidence of any varying degrees of emergence across a broad spectrum of regional varieties.

5.3.3 Variation across different genres

Within the spoken/written heading in Table 1.3 are included a range of genres from newspapers to academic writing, scripted broadcast to spontaneous conversation, that offer a more nuanced discussion of genre than the basic spoken/written dichotomy.

- ***Light verbs*** are considered "highly colloquial" by Wierzbicka (1982), while Stein (1991) suggests they are used as a form of interpersonal courtesy rather than simply signalling informality. The range of genres provided in the chosen corpora will allow these competing claims to be tested across spoken genres, with more and less formal interactions, and against written genres where the need for interpersonal strategies would be expected to be less evident.
- ***Non-numerical quantifiers*** are also considered chiefly informal, by Quirk et al. (1985), but Channell (1994) finds specific uses for these vague quantifiers in academic writing. Neither of these viewpoints is supported by the use of frequency data to compare the different genres. This research will make use of corpus data to quantify their relative frequency of usage in spoken and written texts, and also help to establish the range and regional spread of phrases considered to be at the more colloquial end of the spectrum.
- ***Complex prepositions*** are generally considered to belong to the more formal registers of written style, although Hoffmann's (2005) characterization of *in terms of* as a discourse marker may point to distinctive, and perhaps more highly grammaticalized functions of other complex prepositions.
- Quirk et al. (1985) label some of the items they identify as ***complex subordinators*** as being generically limited, for example *forasmuch as* is considered formal. In general, we might expect subordinators to feature more strongly in written text than spoken as subordination is seen to be a feature of written text (O'Donnell, 1974). However this assumption is challenged by Beaman (1984) and

Halliday (1994) who challenge both the comparative syntactic complexity of written of spoken language, and the status of subordination as an indicator of complexity.

5.3.4 Diachronic changes

Most of the corpora used offer contemporary material, the only truly diachronic corpus being used is COHA, with a timespan of 200 years. However there is some scope for diachronic study of the development of these multiword units by comparison with previous studies, and potentially within the still expanding COCA corpus, that already covers a span of 20 years. The ICE data from the 1990s, on light verbs, can be compared with Algeo's data from corpora of the 1960s. Longer term developments are harder to track, because of the lack of sizable, structured corpora from before the 20th century. Hoffmann (2005) used a variety of corpus and non-corpus material to trace the grammaticalization of certain complex prepositions from Early Modern English to the present day. COHA will be used as a means of investigating similar developments, to see if newly grammaticalized complex prepositions are being formed.

The COHA corpus consists entirely of written texts, which makes it an appropriate data source for investigating complex prepositions (see comments above, about their formal nature), but less so for non-numerical quantifiers, which require spoken/written comparisons to test assertions about their comparative informal/informal nature (Quirk et al., 1985; Channell, 1994). Complex subordinators may well be a feature of written texts (O'Donnell, 1974), but the lack of previous corpus data on them means that cross-genre data is needed to test this theory. For both of these constructions, therefore, data from synchronic spoken/written corpora will be interrogated, with evidence for emergence as grammatical units sought in their comparative fixity of form, and in the potential for category shift where they appear in ambiguous grammatical contexts.

5.3.5 Evidence of indeterminate contexts from corpus data

Beyond the basic frequency comparison across regions/genres/time of frequency data for these multiword units, and the variations on them, is the consideration of the grammatical context within which they appear. Corpus data can provide examples of real usage where

the ambiguity demonstrated in, for example, different patterns of concord for a particular unit, or different roles implied by positions within the sentence, can be explored. This allows us to go beyond the use of examples (invented or from usage data) to exemplify standard usage, to test the boundaries of lexical units and their grammatical functions.

6. Summary – Research questions

There have been numerous studies of the nature of lexicalization and grammaticalization – some purely theoretical in nature (e.g. Hopper and Traugott, 2003; Brinton and Traugott, 2005), others making use of corpora, or a range of historical data, to assess and categorize the processes involved (e.g. Aarts, 2007; Partington, 1998). The literature also contains a large number of monographs and papers on particular words and multiword units that demonstrate facets of lexicalization and grammaticalization, and argue for the erosion of the notional boundary between lexis and grammar (e.g. Hoffmann, 2005; Brems, 2012).

However there is not, to my knowledge, a synthesized attempt to draw together a set of multiword units across different word classes, and analyse them using a consistent set of corpora in order to investigate the similarities and differences of their paths towards emerging as grammatical units. This research will aim to fill that gap in addressing the following questions:

i. How can synchronic corpus data be used to draw conclusions about the processes of grammaticalization for multiword units?
ii. Are the standard criteria for assessing grammaticalization more effective for some multiword units than others?
iii. How should marginal lexical and grammatical items be presented in descriptive grammars and dictionaries?

Chapter 2

Light verbs in Australian, New Zealand and British English (Paper 1)

Smith, A. 2009. Light verbs in Australian, New Zealand and British English. In P. Peters, P. Collins and A. Smith (eds.) *Comparative studies in Australian and New Zealand English: Grammar and beyond,* 139-56. Amsterdam/Philadelphia: John Benjamins.
https://doi.org/10.1075/veaw.g39.09smi

"He took a long drink on the bottle and a long eat on the bread."

Spike Milligan (*Puckoon*, 1963)

Light verbs in Australian, New Zealand and British English[1]

Adam Smith
Macquarie University

This paper examines regional and register differences in the use of the light verbs *give*, *have*, *make* and *take* across British, Australian and New Zealand English, to see whether statements in the literature such as the US preference for *take* can be supported. Primary and secondary materials were investigated, in the form of L1 and L2 dictionaries across the regions, and data from the ICE corpora for Britain, Australia and NZ. The dictionary data only partially confirmed regional differences between *take* and *have*, while the corpora showed a growing use of the light verb *have*, with Australian and New Zealand English leading the way. The corpora also demonstrated more frequent and more productive use of the construction in spoken than in written data, which allowed conclusions to be drawn about the interpersonal functions of light verbs.

1. Definition of "light verb"

Poutsma (1926) and Jespersen (1931) first identified the tendency of modern English to form verbal expressions with a noun complement, where the semantic content resides almost entirely in the noun: *have a look*, *take a rest*, *do a dance* etc. The construction "hovering between grammar and lexis" as Algeo (2006: 269) remarks, has provided problems of classification for grammarians and linguists ever since. The sheer range of terms that have been applied to the phenomenon gives an indication of its ambiguous status. For the verb there is "light or insignificant verb" (Jespersen 1931), "copula" (Curme 1935), "function verb" (Nickel 1968), "empty" or "stretched" verb (Allerton 2002); and for the construction "verbo-nominal phrase" (Rensky 1964), "complex verbal structure" (Nickel 1978), "composite predicate" (Cattell 1984) and "expanded predicate"

Notes

[1] With grateful acknowledgement to Pam Peters for her assistance and encouragement in the development of this paper, and to Peter Collins for his helpful comments.

(Algeo 1995). All of these encompass a wide range of constructions, the boundaries of which are not always clearly defined (although Algeo does distinguish between "core expanded predicates" and others (1995: 207)). In order to collect a compatible set of data for regional and register comparison, it is necessary first to establish some clear guidelines as to what is to be classified as a light verb.

The widest definition of light verbs is that they are "semantically 'light' in the sense that their contribution to the meaning of the predication is relatively small in comparison with that of their complements" (Huddleston & Pullum 2002: 290). This allows for constructions where the choice of verb might complement the noun without modifying the meaning (*ask a question*), or where a common verb is added to a noun where there is no equivalent single-word verb (*have mercy*). This definition also makes no stipulation about the form of the noun, so that *make a calculation*, *do the ironing* can be classified as light verb constructions.

At the other end of the scale, it is possible to classify the apparent noun complement as verb by limiting examples to those where the noun is exactly equivalent to the infinitive form of the verb (*have a* ***swim*** = *to* ***swim*** etc.). Wierzbicka (1982) took this approach in her study on the *have a V* frame. While this allows her to make some interesting observations on the semantic rules that govern the choice of verb in the *have* construction, the criteria for distinguishing between the identical forms as a verb or a noun are not always clearcut.

A more useful approach for the present study is provided by Kearns (2002) who divided the traditional class of light verbs into two categories, based on grammatical principles. Her distinctions between "true light verb"(TLV) and other types of light verbs which she termed "vague action verb" excluded nominalized forms of the verb as well as other variations where the noun complement is highlighted in some way.

Table 2.1: The main attributes of TLVs according to Kearns (2002)

Description	True Light Verb
1. The complement is headed by a N which is a stem form identical to a verb	Have a look = look Take a walk = walk
2. The complement NP must be indefinite	He gave a groan not Who gave the groan?
3. The complement NP cannot be subject of passive	* A groan was given by John
4. The complement NP cannot be the focus of a WH-question or modified by a relative clause	*Which groan did he give? *The groan (which) he gave startled me.
5. The complement cannot be pronominalized	*The deceased gave a groan at midnight and another one later.

In addition, she notes the particular identification between the stem noun and its corresponding verb in certain light HAVE constructions, where the noun use appears to be coined explicitly for the TLV construction in examples such as: *Can I have a /ju:z (*/ju:s/) of your pen* or *Can I have a lend/borrow of your pen?*

The light verb construction's tendency towards pre-modification[2] shifts the balance away from this verb as complement interpretation. Kearns also makes the point that modification often appears in TLV phrases to reinforce the "common use of these constructions as hedging strategies, e.g. *I just gave it a little poke and it exploded.*" (ibid). Evidence of pre-modified forms within the corpus evidence will be examined in the current study, but dealt with separately from the unmodified data.

Because this ambivalence between verbalization and nominalization appears to be at the heart of the construction, I shall follow Kearns's classification of the TLV (light verb = TLV in the rest of the paper) in identifying examples and corpus data. The light verbs covered are *give*, *have*, *make* and *take*, which are defined as the verb of the core expanded predicate by Algeo (1995: 208), and also, being the commonest, they provide the most corpus data for regional comparison.

2. Evidence for regional divergence

2.1 Research studies

The literature on light verb constructions is full of bold statements about its regional variability. "Examples of constructions with the gerund, which occur with higher frequency

[2] Observed by Jespersen "Such constructions also offer an easy means of adding some descriptive trait in the form of an adjunct: we had *a delightful bathe*, *a quiet smoke*, etc."

in American English, are ... I gave Gulliver's Travels a re-reading" (Nickel, 1968: 6). Another is "In these synonymous sets the selection of the light verb is often a matter of dialect – social or regional" (Live 1973: 33). An extended example is "In particular, the frequency of use, and hence the importance, of this construction in British or Australian English is far greater than in American English, which makes much greater use of the related *take a V* construction. In Australian English, in particular, the *have a V* construction constitutes a fundamental part of everyday talk." (Wierzbicka 1982: 756). Finally: "It appears that interdialectal variation is greater for the light HAVE construction, which is most robust in Australian English, common in New Zealand and British English, but limited in American English" (Kearns 2002). There have been few corpus studies to confirm or deny these claims.

The first research based on corpus material was published by Stein and Quirk (1991), using a corpus of British novels from the 1980s It therefore concentrated on the different semantic areas covered by three different light verbs, and was not able to make regional comparisons. In a separate paper, Stein (1991) had questioned Wierzbicka's contention (see above) that the *have + V* construction was a particularly informal one. Again, their corpus was purely composed of written material, so not suited to make comparisons of register – although one wonders what proportion of their light verb examples came from literary representation of speech.

A more extensive study, comparing British and American data in the Brown and LOB corpora, was carried out by Algeo (1995, 2006). He looked at the evidence for regional differences in the uses of the five light verbs *do*, *give*, *have*, *make*, *take*. While the examples he used were not all TLVs, he found no significant differences in frequency of use of four of them. However, "British [English] uses *have* as the verb of an expanded predicate nearly twice as often as American does and in about 1.75 times as many different constructions" (2006: 270). This modifies the accepted belief that AmE favours *take*, BrE *have* – Algeo quotes Quirk et al. saying that "when the eventive object collates with both *have* and *take*, *have* is typically British option, *take* the American" (1995: 211). Algeo's findings show that "The difference is not that American favours *take* but that British favours *have*" (213).

Both of these previous corpus studies concentrate on British and/or American written material. There have been no corpus studies of the use of light verbs in AusE and NZE, or of spoken material. This study will therefore focus on the spoken components of the Australian, New Zealand and British ICE corpora (ICE-AUS, ICE-NZ, ICE-GB) to examine whether there are regional differences between the varieties and test Wierzbicka's claim that the light verb construction is particularly informal.

First, however, to give a wider view of regional variation, let us look at the treatment of a set of light verbs using secondary evidence. Their coverage and labeling in a range of dictionaries can provide an indication of variation between regions and registers.

2.2 Dictionary evidence

Light verb constructions present a problem for lexicographers. As Algeo writes "Because it [the expanded predicate] is not exclusively either grammatical or lexical, it is likely to be treated inadequately in both grammars and dictionaries" (1995: 204). A selection of American, British, Australian and New Zealand dictionaries was therefore surveyed to see how far their treatment of light verbs was systematized, and if regional and register variations were covered by inclusion/exclusion and labeling. Dictionaries both for native speakers (L1) and second-language learners (L2) were included in the survey, as L2 dictionaries of comparable size are more likely to cover spoken idiom.

The following tables show a comparison of the dictionaries' treatment of the common light verb constructions *take/have a bath, break, holiday/vacation, look, shower, walk*.

Table 2.2a: Coverage of light verbs in American, British, Australian and New Zealand L1 dictionaries

	American (RHD)	**British (NODE)**	**Australian (MD)**	**NZ (NZOD)**
In definition				*bath* (take)
		holiday (take) – definition of *vacation* (chiefly US)	*holiday* (take)	*holiday* (take) – definition of *vacation* as verb
	shower (take)		*shower* (take) *vacation/holiday* (take/have) – definition of *vacation* as verb	
In example sentence	*bath* (take)	*bath* (take)		*bath* (take) – in sense of "suffer defeat" (colloq.)
		look (take)	*look* (have)	
As sub-headword		*bath* (take) – in sense of "suffer defeat, loss" (informal) *holiday* (take)	*bath* (take) – in sense of "suffer defeat" (colloq.)	

RHD = *Random House Dictionary,* 2nd Edition (1987)
NODE = *New Oxford Dictionary of English* (1998)
MD = *Macquarie Dictionary,* 4th Edition (2005)
NZOD = *New Zealand Oxford Dictionary* (2005)

There is quite uneven coverage of the 7 sample light verbs across the four L1 dictionaries. The Australian *MD* has the widest coverage, covering 5 of them either as a definition, in an example sentence or as a sub-headword. The American *RHD* has the smallest coverage with only 2. There is also a remarkable preference for *take* as the light verb across all the regions, with only the Australian dictionary offering *have* as an equal alternative with *take* in *take/have a holiday/vacation*, and as the chosen example for *have a look*. *Take a bath*, covered in *NODE*, *MD* and *NZOD*, is a particularly interesting example as it shows the light verb construction being used to introduce a new sense to the simple verb, and it is the only one marked as colloquial/informal.[3]

In Table 2.2b, 6 out of 7 of the light verb constructions chosen were found in at least one of the L2 dictionaries, with only *have/take a vacation* not covered. *LongBr* has the completest coverage with 6 of the 7, and v*acation* would be excluded as a US variant, in

[3] See Section 3.1.1 for treatment of the innovative colloquial use of light verb constructions

line with *NODE*'s labeling of it as "chiefly US". But across the regions there is greater coverage of the TLVs than in the L1 dictionaries. There is also a sign of regional preferences shown in *LongAm*'s consistent choice of *take* as the light verb, and *LongBr*/*MALD*'s presentation of *have*, at least as an option. *LongBr* even labels *have* as "Br" and *take* as "Am" in the case of *take*/*have a bath*/*shower*.

Table 2.2b: Coverage of light verbs in American, British, Australian and New Zealand L2 dictionaries

	American (LongAm)	**British (LongBr)**	**AUS (MALD)**
As definition			*bath* (have/take)
In example sentence	*bath* (take)	*bath* (take)	*bath* (have)
	break (take)	*break* (take)	*break* (have)
		holiday (have)	*holiday* (take)
	look (take)	*look* (have)	*look* (have)
	shower (take)	*shower* (take)	
	walk (take)	*walk* (take)	
As sub-headword		*bath* (have (Br)/take (Am))	
		break (have/take)	
		holiday (have)	
		look (have/take)	*look* (have) – in sense "see and pay attention to"
		shower (have (Br)/take (Am))	
		walk (have/take)	

LongAm = *Longman Dictionary of American English*, 2nd Edition (1997)
LongBr = *Longman Dictionary of Contemporary English*, 3rd Edition (1995)
MALD = *Macquarie Australian Learners Dictionary* (1997)

Across the L1 and L2 dictionaries there are varying ways of treating the light verb construction within the entry. Using it to define the simple verb demonstrates that it is regarded as exactly synonymous. Demonstration of its use through an example sentence gives no indication as to whether the light verb has been considered in choosing the example – although choice through frequency would have been an option for the Longman dictionaries where corpora were used to provide examples. Treatment of the construction as a distinct sub-headword occurs only in the British and Australian dictionaries, and only consistently in *LongBr*. This is actually an unconventional lexicographical approach, as the *LongBr* sub-headwords are exemplifying usage rather than drawing attention to a change in sense. The Australian dictionaries, on the other hand, do distinguish between the senses of both *have a look*, *take a bath* and the simple verb senses.

Overall, both sets of dictionaries confirm an American preference for *take* in common light verb constructions, while the British and Australian dictionaries at least present *have* as an option. Backing up Algeo's instinct, the uneven coverage of the construction across the dictionaries tends to imply that they have not been treated consistently (except in *LongBr*), although the greater coverage in L2 dictionaries acknowledges the prevalence of the construction in modern English, and the need to support L2 users in constructing sentences. While the Longman dictionary quotations would have been based on examples from corpora, it is not possible to say whether they were chosen to represent proportional tendencies of light verb choice.

3. Frequency of common light verbs in the ICE corpora

Many verbs go through the process of delexicalization, from those with quite a wide range of meaning such as *put*, *get*, *set*, to verbs like *shed*, *cast* and *throw* that are more semantically constrained (see Verde 2003). This study will focus on the main light verbs, or "core expanded predicates", as Algeo (1995) categorizes them: *give*, *make*, *have*, *take*.[4] These, being the most semantically general, have been found to be the most commonly used in light verb constructions, and are therefore likely to produce the most corpus data for regional and register comparisons.

As explained above, the Kearns classification of the TLV was used to collect a consistent set of data. Searches were made on the present and past forms of the four light verbs selected, with premodified constructions ("make a great impact"; "have a bit of a look") and monotransitive vs. ditransitive uses ("give a call" vs. "give her a call") noted.

The following results are taken from ICE-GB, ICE-AUS and ICE-NZ, with reference also made to ART, the corpus of Australian radio talkback. The ICE corpora were chosen because they are compatible in size and construction, and allowed for comparison between spoken and written usage.

[4] Algeo (1995) also looked at *do*, but found only 4 tokens in Brown, and none in LOB. Searches in the ICE corpora discovered a similarly limited range.

3.1 Spoken vs written

As noted above, Wierzbicka's instinct for *have a V* is that "the construction is highly colloquial" (1982: 757). Table 2.3 below gives an overview of the returns for the spoken and written components of the ICE corpora:

Table 2.3: The light verb construction in ICE-GB, ICE-AUS and ICE-NZ*

	ICE-GB	%	ICE-AUS	%	ICE-NZ	%
give (spoken)	2.3 (14)	40%	6.5 (39)	74%	6.0 (36)	53%
give (written)	3.5 (14)	60%	2.3 (9)	26%	5.3 (21)	47%
total	(28)		(48)		(57)	
have (spoken)	23 (138)	70%	44.0 (264)	80%	26.5 (159)	66%
have (written)	9.8 (39)	30%	11.3 (45)	20%	13.5 (54)	34%
total	(177)		(309)		(213)	
make (spoken)	9.5 (57)	59%	6.3 (38)	51%	6.3 (38)	68%
make (written)	6.5 (26)	41%	6.0 (24)	49%	3.0 (12)	32%
total	(83)		(62)		(50)	
take (spoken)	2.0 (12)	44%	3.2 (19)	51%	2.0 (12)	40%
take (written)	2.5 (10)	56%	3.0 (12)	49%	3.0 (12)	60%
total	(22)		(31)		(24)	

*Frequency/100 000 words is given to offset the different size of spoken and written components of the corpora, with raw figures in brackets. The percentages are based on the normalized frequencies.

This shows almost uniformly a higher incidence of the most common light verb constructions in spoken than in written English. The exceptions, where the frequencies are very close are equal, are in the comparatively lower frequency items for *give* and *take* in ICE-GB, and for *take* in ICE-NZ. Given that there are more spoken texts than written in the ICE corpora (300 spoken, 200 written), the raw figures have been normalized to show frequency per 100 000 words. This adjustment gives a higher proportional frequency in written texts for the examples above, and makes the difference minimal between spoken and written for *give* in ICE-NZ and *make*/*take* in ICE-AUS. It is therefore only *have* constructions that appear more often in speech across all three regions, with *give* more common in spoken AusE, and *make* more common in spoken BrE and NZE.

3.1.1 Informality vs. interpersonal courtesy

Even where there is higher frequency of light verbs in the spoken categories of the corpora, this does not necessarily indicate a greater degree of informality for the construction over the use of the simple verb. Stein (1991: 26) challenges Wierzbicka's characterization of the "informality" of *have* constructions, suggesting that the

construction is better described as a form of "interpersonal courtesy" rather than "reduced formality". We can test this assertion on the phrase *have a look*, which is by far the most common light verb construction across the corpora examined, and almost exclusively a spoken usage:

Table 2.4: Frequency of *have a look* in ICE-GB, ICE-AUS and ICE-NZ

	ICE-GB		**ICE-AUS**		**ICE-NZ**	
	Occurrences	**Freq./ 100 000**	**Occurrences**	**Freq./ 100 000**	**Occurrences**	**Freq./ 100 000**
have a look	43		116		79	
Spoken	42	7	113	18.8	76	12.7
Written	1	0.3	3	0.8	3	0.8

When we look at individual instances of the phrase, well over half of them are used as invitations to join the speaker in looking at something, as in:

(1) Let's just **have a look** at some of the headlines [ICE-NZ S1B-054:39]

(2) Thank you If Your Worship **has a look** at the evidence and Your Worship ah I submit has an easy job in this particular matter [ICE-AUS S2A-066:3]

(3) Let's **have a look** at that date [ICE-GB S1A-077:146]

It should be noted that each of these examples comes from a different category of the corpus (S1A - private dialogue; S1B – public dialogue; S2A – unscripted monologue), so that each represents a different communicative situation and different degrees of formality. The speaker is in a different power relationship with their listener: in (1) it is a teacher instructing students, in (2) a solicitor in court trying to gain the cooperation of the judge, and in (3) it is a personal conversation where there is no difference in status between the speakers. This suggests that the light verb construction, while typically prevalent in speech, covers a range of registers and interpersonal functions within the medium. A breakdown of the spoken categories for *have a look* confirms this spread:

Table 2.5: Occurrences of *have a look* across ICE spoken categories

	ICE-GB	**ICE-AUS**	**ICE-NZ**
S1A (Dialogue, private)	21	31	30
S1B (Dialogue, public)	14	18	21
S2A (Monologue, unscripted)	6	55	19
S2B (Monologue, scripted)	1	6	2

Only S2B is under-represented here. It is the scripted monologue category which might therefore be expected to bear more relation to written language, and make less use of the interpersonal resources of the light verb construction.

3.1.2 Hedging and other uses of pre-modification

The *have a look* examples not only provide interest by way of their distribution across the spoken categories. Another noticeable feature they demonstrate is the frequent use of hedging , a characteristic described by Kearns "the use of one of these TLV constructions trivializes or minimizes the denoted action or event, which is consistent with the common use of these constructions as hedging strategies" (p.5). The most common instance of hedging in the corpora is through the adverbial *just* as in (1) or the following example which is additionally hedged by *perhaps* before and *briefly* after the light verb construction:

(4) what i was going to do is just perhaps **have a look** briefly at
[ICE-NZ S2A-044:37]

Another means of hedging is by pre-modification of the noun complement, as in:

(5) We've just **had a really quick look** probably not very good uhm <,> at these single cell organisms that live in natural ponds
[ICE-GB S2A-051:110]

In this example the speaker is giving a tutorial, and is therefore using hedging as a means of mitigating the unequal teacher-student discoursal relationship. Another example from the same category in ICE-AUS shows the teacher using pre-modification to the opposite effect, to stress a point and give instruction as to what the students should focus on:

(6) There's a little more about Lizzie in this book and there's a a lot more about other people and one of the the things that you'll find the about this book that you it's the opportunity to **take a good long hard look** at what a stallholder actually uh is and looks like
[ICE-AUS S2A-048:115]

The purpose of pre-modification is therefore not clearcut in interpersonal terms. It also has the effect of emphasizing the nominal character of the complement and distancing it from the verb, dislocating the light verb construction as a discrete grammatical entity (in contrast to example (5) where the adverbial modification enhances the verbal status of the

have a look construction). In this light it would interesting to discover whether there are differences between spoken and written English in the amount of pre-modification used.

An overview of all the pre-modified constructions in the ICE corpora shows a subtle pattern in the difference in use for spoken and written English:

Table 2.6: The occurrence of pre-modified against unmodified light verbs in ICE-GB, ICE-AUS and ICE-NZ*

	ICE-GB		**ICE-AUS**		**ICE-NZ**	
	Pre-mod.	*Unmod.*	*Pre-mod.*	*Unmod.*	*Pre-mod.*	*Unmod.*
give	1.2 (7)	1.2 (7)	1.0 (6)	5.5 (33)	1.0 (6)	5.0 (30)
(spoken)	50%	50%	15%	85%	17%	83%
give	2.0 (8)	1.5 (6)	1.75 (7)	2 (0.5)	1.8 (7)	3.5 (14)
(written)	57%	43%	78%	22%	33%	67%
have	7.0 (42)	16.0 (96)	11.5 (69)	32.5 (195)	5.8 (35)	20.7 (124)
(spoken)	30%	70%	26%	74%	22%	78%
have	5.8 (23)	4.0 (16)	6.0 (24)	5.3 (21)	8.3 (33)	5.3 (21)
(written)	59%	41%	53%	47%	61%	39%
make	4.0 (24)	5.5 (33)	3.0 (18)	3.3 (20)	2.8 (17)	3.5 (21)
(spoken)	42%	58%	47%	53%	45%	55%
make	2.8 (11)	3.75 (15)	3.3 (13)	2.8 (11)	1.5 (6)	1.5 (6)
(written)	42%	58%	54%	46%	50%	50%
take	0.7 (4)	1.3 (8)	0.8 (5)	2.3 (14)	0.8 (5)	1.2 (7)
(spoken)	33%	67%	26%	74%	42%	58%
take	1.5 (6)	1.0 (4)	1.8 (7)	1.3 (5)	0.8 (3)	2.3 (9)
(written)	60%	40%	58%	42%	25%	75%

*Frequency/100 000 words is given to offset the different size of spoken and written components of the corpora, with raw figures in brackets. The percentages are based on the normalized frequencies.

These figures show that in general there are proportionally more pre-modified light verb constructions than unmodified ones in written texts, and less of them in spoken. Even when pre-modified forms are not the majority in written, as in *give* for ICE-NZ, the proportion still goes up from 17% to 33%. The only examples to contradict this trend are *take* in ICE-NZ, where pre-modification is lower in written texts (although there are very few occurrences), and *make* in ICE-GB, where there are equal proportions of pre-modified and unmodified in spoken and written.

The suggestion that pre-modification might be more prevalent in written texts is only partially supported by Stein and Quirk (1991) where their corpus of British fiction from the 1980s shows 3 times more modified than unmodified examples with *give*, but only 1:2 for *take*/*have*. They suggest that this skewed distribution is because "*Give* tends to be used with realized experience which is thus more prone to invite detailed description

and evaluation." (201), and give a clearly novelistic example from John Fowles: "She…gave him a still faintly doubting smile back" (ibid.). Of the ICE corpora, ICE-AUS shows a similar preference for pre-modified *give* in written texts, but the creative writing category (W2F) doesn't significantly swell the number of examples.

3.1.3 Colloquial productivity

The number of light verb constructions in spoken data, whether they are indicators of register or serve an interpersonal function, is just one sign that the construction is a feature of spoken language. Another is its capacity to produce creative extensions of conventional phrases or words. There is evidence in the ICE corpora of this happening in two ways:

a. Introduction of noun complements in a specialized sense within the light verb construction

The clearest example of this in the ICE data is the proliferation of alternative nouns based on the conventional collocation *have a/give (it) a try*. In ICE-AUS, the synonymous variations are *have a bash, a crack, a shot,* and in ART (radio talkback) there is also *have a lash* and *give (it) a go*. ICE-NZ has *have a crack, lash, give(it) a go* and *take a stab*, while ICE-GB has *make a stab*. All of these are marked as "colloquial" in the *Macquarie Dictionary,* and the *Oxford English Dictionary* (*OED*), and have the same sense "to try" or "make an attempt". Along the same lines are *take a nosey* (ICE-GB), *have a shoofty* (ART) – although the former could be said to be adding shades of meaning to the basic *have/take a look*. The phrase *take/have a shufti* is recorded in the *OED* as a now rare piece of military slang and is labeled as British slang in the *Macquarie Dictionary* (4th edition) with the spelling *shoofty*. *Take a nosey*, however, appears to be a nonce usage, highlighting the potential of the light verb construction to creatively expand usage. This use of the light verb construction to specify a variant sense for a familiar word shows what productive power is inherent within it in the spoken idiom at least.

b. Use of verb forms as noun complements

There are several examples of nouns within light verb constructions that hardly appear elsewhere as nouns, and tend to be categorized as highly informal or non-standard. Huddleston and Pullum claim that they are especially found in AusE, and cite some

examples: *Can I have a borrow/lend of your pen/a carry of the baby?* (2002: 296).[5] The *OED* at least supports the regional labeling of *lend* in this sense, describing it as "*Sc.* and *north. dial.*" as well as "*Austral.* and *N.Z. colloq.*", while the *Macquarie Dictionary* labels it "non-standard". These particular examples do not appear in the corpora, but there are four instances of *have a lend* in the sense of "to tease someone" in ART, and both ICE-AUS and ICE-NZ provide examples of *have a feed*, where the verb form *feed* is used to mean "meal".

All of the examples above, apart from *take a nosey*, appear in the spoken sections of the corpora, and *take a nosey* appears in a sample of reported speech from a novel, so it is at least imitating a spoken register. Of the others, virtually all appear in category S1A – private discussion, where we would expect the most informal speech. ICE-NZ appears to vary from the other ICE corpora in that it gives us *have a lash* in S1B (public dialogue), and *have a stab*, *have a crack* and *give it a go* in S2A (unscripted monologue). These appear less anomalous when we see that the S1B example is from an interview with a man talking about his criminal past, and all the S2A examples are from sports commentary where we might expect a more informal register than in other broadcast speech – especially as they often involve dialogic exchanges despite their classification as "monologue".

The comparative lack of BrE examples suggest that the light verb construction might be more productive in both AusE and NZE than it is for BrE, both for extending the sense of a noun in spoken idiom, and further breaking down the grammatical integrity of its element by employing a verb form as the noun complement.[6] From these particular instances, we shall move on to look at more general evidence of regional variation.

3.2 Regional and temporal differences

3.2.2 Corpus comparisons

Light verb constructions are often considered to vary regionally (see Section 2 above), with the American preference for *take* over the British for *have* as the light verb the most

[5] Add to that Kearns' example cited above: *Can I have a /ju:z/ (*/ju:s/) of your pen?*

[6] Note that *have a lend* is also labeled as a Scottish and northern dialect usage.

widely cited variation of this evidence. Algeo's survey of the American Brown written corpus, and its British equivalent, LOB, showed the spread of light verb choice as follows:

Table 2.7a: Light verbs in the LOB and Brown corpora (adapted from Algeo 1995)

Summary of tokens/type		
	LOB	**Brown**
give	40/29	40/30
have	100/61	55/35
make	67/39	59/44
take	38/20	41/20

These figures, Algeo argues, do show regional variation, but it is not so much of US English preferring *take*, as British English favouring *have*. A direct comparison cannot be made with the ICE corpus findings for several reasons: LOB and Brown consist of exclusively written material; the texts are 30 years older than ICE, dating from the early 1960s, and Algeo used a much broader classification of what constitutes a light verb. However, an overview of the comparative frequency of the same light verbs in the ICE corpora offers an interesting contrast:

Table 2.7b: Light verbs in ICE-GB, ICE-AUS and ICE-NZ

Summary of tokens/type			
	ICE-GB	**ICE-AUS**	**ICE-NZ**
give	28/23	48/29	57/26
have	177/48	309/64	213/46
make	83/41	62/37	50/32
take	22/13	31/17	24/12

The most striking difference between these figures taken from 1990s data, as opposed to the earlier LOB/Brown corpora, is the higher ratio of tokens to types for *have* in the ICE corpora – reaching ratios of nearly 5:1 in ICE-AUS and ICE-NZ. This figure can be partially explained by the exceptionally high frequency of *have a look* (see above, Table 2.4). If we remove all these incidences, the adjusted number of tokens are: ICE-GB, 134; ICE-AUS, 193; ICE-NZ, 134. This now gives a consistent type/token ratio of about 3:1 across the regions, which is still higher than for any other light verb in ICE, or in LOB and Brown. There are several noun complements for *have* other than *look* that have multiple instances in the ICE corpora, such as *chat*, *effect*, *holiday*, *impact,* which account for this high ratio. It appears that *have*, as well as being the most frequent light verb across the ICE

regions also generates more multiple examples of constructions using the same noun complement – thus reinforcing the choice of *have* as the light verb.

The other light verbs do not have a significantly higher incidence in the ICE corpora than in LOB and Brown, and in fact *take* produces lower figures across the regions.

Algeo's broader categorization of the construction would tend to inflate the number of tokens/types, but this still supports his finding that *have* is the most common light verb outside the US. ICE-GB shows the same order of preference as LOB, as does ICE-AUS (*have, make, give, take*), while ICE-NZ has *give* higher than *make*).

The data suggests that there has been an increase in the use of *have* in light verb constructions in BrE over the 30 years between the compilation of the LOB and ICE-GB corpora, while the other light verbs show little sign of change, except in the case of *give* which appears to be less frequent. The even greater frequency of *have* in New Zealand, and particularly Australia, show the southern hemisphere varieties to be ahead in this trend.

3.2.3 Regional choices of have or take

It should follow, therefore, that in instances where there is a choice between the use of *have* or *take* with a particular noun complement, *have* will be the usual choice. The following table shows the occurrences in ICE of the light verbs checked for regional labeling in dictionaries in Tables 2.2a/2.2b above.

Table 2.8: Choice of *have*/*take* in the ICE corpora

	ICE-GB	ICE-AUS	ICE-NZ	Total
(have) bath	0	2	1	3
(take) bath	0	1	0	1
(have) break	5	4	3	12
(take) break	2	4	5	11
(have) holiday	10	5	5	20
(take) holiday	1	1	0	2
(have) look	43	119	79	241
(take) look	6	3	2	11
(have) shower	1	4	3	8
(take) shower	0	1	0	1
(have) walk	3	1	1	5
(take) walk	0	1	0	1

Again the preference for *have* in all regions is quite consistent, although the frequency is low in some cases. The only exception is with *break*, where *have* is preferred in ICE-GB, but is equal with *take* in ICE-AUS and lower in ICE-NZ. ART takes this trend even further by giving 15 examples of *take a break* but none with *have*. This might be an instance of the choice of verb changing the sense of the phrase – all the ART instances are to announce a pause for advertising, rather than being used in the more general sense of resting. It's also possible that this is a conventional formula borrowed from American talkshows, where *take* might be the more standard choice of verb. As Tables 2a/b above showed, this was one where no clear regional pattern was evident in dictionaries, and the corpus data backs up this lack of clear distinction.

4. Conclusions

The corpus evidence confirms the preference for *have* as a light verb in BrE, AusE and NZE that is indicated by its coverage in L2 dictionaries. In fact it gives reason to think that *have* may be even more prevalent than is suggested by the L2 dictionaries when they offer *take* as an equal alternative. The corpora show the L1 dictionaries' tendency to illustrate *take* as the light verb in all varieties to be somewhat misrepresentative. On the other hand, the L1 dictionaries do give an indication of the colloquial productivity of light verb constructions by recording variants on the simple verb use, as in the case of *take a bath*. It is possible that a more exhaustive study of the coverage of light verbs across these dictionaries would give an alternative picture of the use of light verbs in the different varieties, although the uneven coverage of the set of constructions chosen suggests that this is unlikely

Much more frequent use of light verbs in spoken than in written English was confirmed by the ICE corpora for Australia, New Zealand and Britain. However, this was not necessarily found to be an indicator of informality, but more of its usefulness for a range of interpersonal strategies. This appeared to be a particularly rich field of investigation, which warrants further research. Evidence for colloquial usage was found in the use of the light verb that extended the sense of the base verb, as in *have a bash/lend/feed, make a stab*, with AusE and NZE appearing to be more innovative than BrE

in this respect. Results were not definitive as to whether pre-modified light verb constructions are more typical to written than spoken texts.

Comparison of ICE-GB with Algeo's analysis of LOB data indicated that the use of *have* in light verb constructions had increased over the 30-year period between the corpora, while the use of other light verbs showed little or no movement. AusE and NZE again appear to be leading the way in the expanding use of light *have* constructions.

References

Algeo, John. 1995. "Having a look at the expanded predicate". In Bas Aarts and Charles F. Meyer, eds. *The Verb in Contemporary English: Theory and Description*. Cambridge: Cambridge University Press, 203-17.

Algeo, John. 2006. *British or American English? A Handbook of Word and Grammar Patterns*. Cambridge: Cambridge University Press.

Allerton, David. 2002. *Stretched Verb Constructions in English*. London: Routledge.

Cattell, Ray. 1984. *Composite Predicates in English*. (Syntax and Semantics 17). New York: Academic Press.

Curme, George. 1935. *A Grammar of the English Language*. Boston: D.C. Heath & Co.

Huddleston, Rodney and Geoffrey Pullum. 2002. *The Cambridge Grammar of English Usage*. Cambridge: Cambridge University Press.

Jespersen, Otto. 1931 *A Modern English Grammar*, vol. 6. Heidelberg: Carl Winters.

Kearns, Kate. 2002. "Light verbs in English". Unpublished Generals Paper, MIT.

Live, Anna H. 1973. "The *take-have* phrasal in English". *Linguistics* 95: 31-50.

Longman Dictionary of American English, 2nd ed. 1997. New York: Longman

Longman Dictionary of Contemporary English, 3rd ed. 1995. London: Longman.

Macquarie Australian Learners Dictionary. 1997. Sydney: Macquarie Library.

Macquarie Dictionary, 4th ed. 2005. Sydney: Macquarie Library.

New Oxford Dictionary of English. 1998. Oxford: Oxford University Press

New Zealand Oxford Dictionary. 2005. South Melbourne, Vic.; Auckland: Oxford University Press.

Nickel, Gerhard. 1968. "Complex verbal structures in English". *International Review of Applied Linguistics* 6: 1-21.

Oxford English Dictionary, 2nd ed. 1989. 20 vols. Oxford: Clarendon Press.

Poutsma, Hendrik. 1926. *A Grammar of Late Modern English*. Groningen: P. Noordhoft.

Random House Dictionary of the English Language, 2nd ed. 1987. New York: Random House.

Rensky, Miroslav. 1964. "Nominal tendencies in English". *Philologica Pragensia* 7: 135-50.

Stein, Gabriele. 1991. "The phrasal verb type 'to have a look' in Modern English". *International Review of Applied Linguistics* 29: 1-29.

Stein, Gabriele and Randolph Quirk. 1991. "On having a look in a corpus". In Karin Aijmer and Bendt Altenberg, eds. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London, New York: Longman, 197-203.

Verde, Maria. 2003. "Shedding light on SHED, CAST and THROW as nodes of extended lexical units". In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery, eds. *Proceedings of the Corpus Linguistics 2003 conference*. Page numbers? <http://ucrel.lancs.ac.uk/publications/CL2003/papers/verde.pdf> (5 May 2008).

Wierzbicka, Anna. 1982. "Why can you *have a drink* when you can't **have an eat*?". *Language* 58: 753-99.

Section 2A: Light verbs (LVs) and grammaticalization

The commonly occurring LV constructions found in the corpus data, such as *have a look* or *have a try*, or analogical formations based on them like *have a shufti* or *have a go*, have an indeterminate status as fixed lexical units and as grammatical phenomena. Therefore we must look more closely at how their status can be determined – both by applying some of the lexicalization and grammaticalization criteria presented in the introduction, and looking at how they are classified in grammars and dictionaries.

1. Issues of lexicalization and grammaticalization

The true light verb (TLV) phrases identified in Paper 1 (2009) are equivalent to the type of composite predicates identified by Brinton and Traugott (2005) that have light verbs like *make, take, give, have* and *do*, which they consider to show signs of grammaticalization (p.131). They contrast these with the type *lose sight of*, *pay attention to*, which they say are lexicalized as non-variable idioms. The major distinction they make between the two types is that those formed with light verbs are highly productive, while the lexicalized type are fossilized and non-productive. This productivity suggests the possibility of grammaticalization by analogy, but other criteria are not so clearly evident.

Some of Lehmann's (1985) parameters of strong grammaticalization such as 'integrity' and lack of 'syntagmatic variability' are not particularly evident in LV constructions, as they consist of 'bundles of semantic features' and, while they are structurally consistent (V + indef article + N), there are many potential variations within that structure. The most fixed forms are the lexicalized idioms, and *take a bath* in its idiomatic usage 'suffer a defeat/loss' is a good example of this type of usage from Paper 1. LV constructions whose meaning is equivalent to the simple verb – for example *have a look*/*look* – can be elaborated for descriptive purposes (e.g. *have a long hard look*). The prevalence of such instances where the noun element is pre-modified (especially in the spoken corpus data, shows their inherent flexibility rather than fixity of form.

There is a degree of fixity demonstrated in the choice of LV across different phrases, with *have* appearing to be extending its role in comparison to other LVs, particularly in the corpus data on AusE and NZE. This trend was not reflected in the dictionaries looked at, with *take* often supplied as the preferred LV, or offered as an

alternative. The limited diachronic range (30 years) given by the corpus data, indicated something of a narrowing of scope of choice of verb, but a much wider data set would be needed to demonstrate whether this is a historical trend towards greater fixity.

We might expect Boye and Harder's (2012) criterion of semantic reduction, where elements of a phrase lose their core semantic weight in the process of grammaticalization, to apply in the case of verbs that are described as 'semantically light'. It is true that the core sense of possession has been lost in the use of *have* as a LV, but the verb has already become so bleached by general use that it would be hard to argue that the LV construction had an influence on this process. Certainly as an LV, it hasn't reached the semantic reduction of its purely grammatical use in the present perfect.

Another criterion that Brinton and Traugott (2005) apply to support their classification of LV composite predicates as grammaticalized is decategorialization of the noun, and grammaticalization of the verb within the phrase. It is true that certain features of the noun category are lost within the LV construction, such as the ability to inflect as a plural, but the retention of the indefinite article 'a(n)' in all instances, and the tendency for adjectival premodification of the N element (*take a good long look*) serve to emphasize its status as a noun. Brinton and Traugott's description of the verb as grammaticalized is dependent on the interpretation of the choice of verb as having an aspectual role – as in *have* expressing a stative meaning, and *give*/*take* a dynamic meaning. This aspectual function was not evident in the LV constructions in this research: they tended to be interchangeable based on regional preference. In one type of LV construction discussed in Paper 1 (where verbs have been co-opted into the role of the noun, e.g. *have a lend, a borrow, a use* (/ju:z/)), they become an environment where words can switch (not lose) their categorial status. These are very particular instances, and should not be used to generalize about the LV construction as a whole.

2. Grammars

As discussed in Paper 1, issues of classification and nomenclature affect the consistency of treatment of LVs in grammars. It has already been noted that Huddleston and Pullum (2002) have a particularly broad interpretation of LV constructions, and include any instances where the verb might complement the noun without modifying its meaning, from

ask a question to *have mercy*, and *do the ironing*. Quirk et al. look at the category under the heading of "Eventive object" (1985: 750-2) which they describe as "A frequent type of object [which] generally takes the form of a deverbal noun preceded by a common verb of general meaning, such as *do, give, have, make, take*. This eventive object is semantically an extension of the verb and bears the major part of the meaning." (p.750). While this interpretation is very noun-centred, as the bearer of the phrase's meaning, their focus on the "common verb of general meaning" is much closer to Kearns's (2002) category of TLVs which I have focused on in Paper 1. Biber et al. (1999) also present this type as a distinct set of V + NP combinations (p.428). Within this set they distinguish between less and more idiomatic examples, for instance *make a bet* which just paraphrases the simple verb *bet*, as opposed to the informal idiom *take a leak* which changes the sense of the simple verb *leak*. Biber et al. use corpus data to look at the most frequently occurring idiomatic expressions with *have, make, take* + noun phrase, and find that both *have/take a look and* the variant *take a good look* feature strongly, emphasizing the tendency for the LV phrase to be modified (1999: 1028). So, despite differences in classifying them, LVs are quite well covered in the major grammars looked at, but as lexical idioms rather than grammatical units

3. Dictionaries

The coverage of the sample set of LV constructions in the L1 dictionaries examined was quite limited and several of the common examples were not attested at all. This is not particularly surprising as the lack of additional semantic content they supply, compared with the simple verb, makes them hard to define as independent lexical items. Where there was consistent coverage (except in the older *Random House Dictionary*, where the usage has not yet been registered) it was on the idiomatic use of *take a bath*, which clearly introduced a new meaning to the headword, and in each case it was listed under the noun, *bath*, rather than the verb. Conversely, this usage was not covered at all in the L2 dictionaries examined, presumably because the usage is figurative, and might be a distraction rather than useful to learners of English. These L2 dictionaries did, however, give a much better coverage of the set of LVs overall, and acknowledged the possible variation (sometimes regionally-based) in the choice of the verb. The fact that the verb

could be variable, as well as having little or no impact on the meaning of the phrase, meant that the preferred headword was the noun in every case for L1 and L2 dictionaries. This downplays each unit's grammatical function as a verb phrase, and serves to isolate them as individual cases, rather than showing how they congregate around these "common verbs of general meaning".

4. Summary

Having found little evidence of grammaticalization for this type of verb phrase, and only limited codification of them as a set of constructions with common features, the next published paper looks at the premodifying noun phrase, as exemplified by quantifiers such as *a lot of, loads of* (classified here as non-numerical quantifiers (NNQs)). I will examine their classification in a set of major English grammars, and use data from the ICE corpora again to investigate signs of variation across register and region. This type of noun phrase also offers potential evidence of grammaticalization, depending on whether the associated verb agrees with the singular/plural form of the quantifying noun, or the NNQ as a unit.

Chapter 3

Non-numerical quantifiers (Paper 2)

Smith, A. 2009. Non-numerical quantifiers. In P. Peters, P. Collins and A. Smith (eds.) *Comparative studies in Australian and New Zealand English: Grammar and beyond,* 157-180. Amsterdam/Philadelphia: John Benjamins.
https://doi.org/10.1075/veaw.g39.10smi

"What a Heap of Names does the poor Fellow call himself."

Samuel Richardson (*Pamela*, 1740)

Non-numerical quantifiers[1]

Adam Smith
Macquarie University

This paper looks at non-numerical quantifiers (NNQs), such as *a lot of, loads of*. The set of quantifiers to be discussed is first identified in relation to their description in major English grammars. Issues of variable noun complementation and verb agreement with the NNQ are identified as being of interest, along with the choice of quantifier and its collocations in different regions (Australian, New Zealand and British English) and registers. Corpus findings for *a lot/lots of* are compared with other NNQs (ONNQs), where the quantifying noun can be singular or plural indicating a level of delexicalization/grammaticization. Some regional variation was found, with the ONNQ *loads of* much more frequent in British English, and *heaps of* more frequent in Australian and New Zealand English. An additional set of NNQs was investigated for regional variance.

1. Introduction

The NNQ is a class of noun phrase that is rarely discussed – either for its grammatical properties or as an indicator of variation in English. This is surprising as phrases such as *a lot/lots of*, *a heap/heaps of* provide points of relevant interest on both fronts. Grammatically they raise problems of agreement with their associated verb ("heaps of food **was/were** on the table") and issues in relation to their noun complement (compare "a lot of **money/mistakes**"). The choice of singular or plural agreement provides a basis for regional comparison for collective nouns (see e.g. Hundt (2009); Levin 1998), and might well do so for NNQs also.

Another influence on number agreement in NNQs is what Reid calls "semantic weight" (1991: 269). The most straightforward use of numerical quantifiers will give regular agreement with the verb and the noun complement ("six dogs were barking";

Notes
[1] With thanks to Yasmin Funk (Macquarie University), for her work on helping to identify non-numerical quantifiers through corpus searches, and to Pam Peters for all her invaluable insights and assistance in the development of this paper.

"half a day was wasted"), whereas NNQs often contain extra semantic content that brings into question whether they are acting simply as a quantifier or have a descriptive purpose (see *heaps of* example above). To assess the degree to which particular NNQs are functioning as quantifiers or descriptors, it will be instructive to look at their agreement and collocational patterns across a range of corpora. In the process, it is hoped that regional differences in the choice of quantifier can be linked to the extent of their delexicalization.

Dictionary labeling can provide an indication of regional divergence in NNQs. *A bunch of*, for example, is sometimes marked as AmE, while *loads of* as more typically BrE. Dictionaries and grammars also note that many NNQs are more common in spoken than in written English, suggesting that a study of the different registers in which particular NNQs are used would be interesting.

Corpus research allows us to quantify the grammatical regularity of NNQs, and assess their regional, register and collocational divergence. This paper will use data from ICE-AUS, ICE-NZ and ICE-GB to look for patterns in the use of NNQs.

2. Classification of NNQs

The standard grammars do not provide a standard approach to classifying NNQs. Their divergent labeling and grouping of categories within the grammatical class problematizes the selection of a coherent group, which is necessary for a systematic corpus evaluation. I shall therefore look at the treatment of these quantifiers in three recent grammars, to give a definition of the term *non-numerical quantifier* for use in this study, and provide criteria for the selection of corpus searches.

The reference grammars selected are Quirk et al. (1985), Biber et al. (1999) and Huddleston and Pullum (2002). Both Quirk et al., and Huddleston and Pullum treat the kind of NNQs we have been looking at as a set, describing the quantity noun + *of* construction as "open-class quantifiers" and "non-count quantificational nouns" respectively.[2] Some of these take noncount nouns, some take plural count nouns, and

[2] The labeling of such nouns as "non-count" appears problematic when there are singular and plural alternatives in NNQs such as *a lot of/lots of; a heap of/heaps of.* The label is presumably being used to distinguish the quantitative use from the descriptive use where the sense of individual units is emphasised: *Two lots of paté* (ICE-GB), *Several heaps of leaves*.

some can take either. Biber et al. separate quantifying nouns into different categories according to whether they take plural count nouns (quantifying collectives), or noncount nouns (unit nouns). A summary of the grammars' classifications is provided below:

- *A Comprehensive Grammar of the English Language* (Quirk et al. 1985)
  - **closed-class quantifiers** (single-word postdeterminers, sometimes preceded by central determiners e.g. "a", "these", "that"); *much/a little* – only with noncount nouns; *many/(a) few/several* – only with plural count nouns
  - **open-class quantifiers** (phrasal quantifiers consisting of quantity noun + *of*, often preceded by indefinite article) *deal/amount of* – only with noncount nouns; *number of* – only with plural count nouns; *plenty/a lot/lots of* – with noncount or plural count nouns

- *Longman Grammar of Spoken and Written English*. (Biber et al., 1999)
  - **quantifying collectives** (e.g. *bunch of/ group of/set of*) – with countables
  - **unit nouns** (e.g. *bit of/piece of/slice of*) – with uncountables
  - **quantifying nouns** (e.g. *barrel of/heap(s) of /pint of/dozens of/load(s) of/armful of/pair of*) – with countable and uncountable

- *The Cambridge Grammar of the English Language*. (Huddleston & Pullum 2002)

Non-count quantificational nouns treated as a set with a noncount noun as head with *of* PP as **complement**. These have different patterns of complementation:

- Number of noun complement ("**oblique**") controls number of whole NP e.g. " a lot of work was done"/ " a lot of errors were made" (**number transparent**)
- Singular quantifying noun with singular oblique ("a great deal of work was done")
- Plural quantifying noun with plural oblique ("dozens of errors were made")

Despite the difference in terminology, it is clear that each of the grammars classifies NNQs according to the same principle – whether they regularly take a singular or plural noun, or are variable. In this paper I will be focusing on the variable NNQs, using Huddleston and Pullum's phrase "number transparent" to denote those that can take either a singular or plural noun complement (an "oblique" in their terms) – although the question as to whether it is the number of the noun complement or other factors that control the number of the associated verb will be questioned.

In order to ascertain a set of variable NNQs, corpus searches were carried out on a broad range of quantity noun + *of* constructions (see Appendix 3.1 for listing), excluding only items such as numerals that consistently take plural complements, or unit nouns (or partitives), which consistently take noncount complements. The partitive set is not a neatly circumscribed one, since quantifying nouns that are not unit nouns can be used partitively (*a lot of the cake was eaten*), and some unit nouns can function more widely than simply to "split up an undifferentiated mass" (Biber et al. 1999: 250). *A bit of*, for example, can sometimes express a large quantity of something, especially when qualified as in *quite a bit of time/money*, and can even be used as an intensifier with an adjective *she's a bit of alright*. Channell (1994: 99) labels her vague quantifiers "pseudopartitives", in order to compare their behaviour with standard partitive nouns. This study will exclude standard partitive nouns and uses, to concentrate on expressions of quantity that are ambivalent as to number.

3. Issues that apply to NNQs

3.1 Grammar: Verbal agreement and noun complementation

Number ambivalence is not restricted to NNQs. As Huddleston and Pullum write "Two of the most common overrides of the simple agreement rule are found with singular collective nouns and with the number-transparent quantification noun construction" (2002: 501). Reid (1991: 261) gives examples such as "Seventy years of Marxist doctrine now seems headed for the dustheap" to show how numerical quantification can present the same kind of mismatches as occur with collective nouns between the number of the noun and its associated verb. In an international survey on agreement, the sentence: *Six days of rain was/were not what we expected in the sunshine state* provided the "least decisive result": 52% of the respondents opting for singular (notional) agreement, and 48% for plural (formal) agreement. (Peters 1999: 6). The question hinges on whether the quantitative element is seen as a single span of time or as several successive days. A similar distinction could be made in the sentence *Heaps of food was/were on the table*. Are we to imagine simply a lot of food, or several distinguishable piles of food – at which point *heaps* is no longer a general quantifier, but a specific descriptor? If we take *heaps*, with all its lexical baggage to be the head of the noun phrase, then plural verbal agreement is natural,

whereas if the sequence *heaps of* is acting as a complex determiner, the singularity of *food* dictates the verbal agreement.

Heap can also be used as a quantifier in the singular form (like *lot*/*lots*, *load*/*loads* etc.), as in: *A heap of watches was*/*were stolen*. Here the descriptive possibilities of *heap* are less likely to come into play – delicate items like watches are not normally put in heaps. But a choice still has to be made over whether the verb agrees with the singular quantifier *heap*, or is affected by the proximity of the plural *watches*.

The different interpretations available in the use of *heap* suggest that there is a process of delexicalization for some NNQs. From their primary lexical role as a descriptor, they can become a kind of quantifying collective, thence to number transparency where the descriptive and quantificational functions assume varying significance according to context, to the final stage as a complex determiner such as *a lot of*, where the lexical content of the quantificational noun has no bearing on the number of the noun complement. Corpus data will help us trace patterns for NNQs to see if this process can be observed.

3.2 Collocation

The number of the noun complement was one of the main factors used by Biber et al. in their classification of "quantifying nouns" at large. These were classed either as "quantifying collectives" (see summary of classifications in Section 2 above), the most common of which were *bunch of* and *group of*, which were consistently used with plural nouns; and as unit nouns (e.g. *bit of, piece of*), which were consistently used with singular nouns. Using data from the Longman Corpus of Spoken and Written English, they found many of the quantifiers they looked at to be associated with particular entities. Typical collocations included *bunch of flowers, group of friends, bit of fun, piece of cake.* Yet some of these would behave less predictably as to the number of their noun complement when used as more general quantifiers (e.g. "a bunch of blood" (ICE-NZ), "a pack of nonsense"). Other categories of quantifying nouns, such as those denoting shape, were more flexible both in the number of their complement (*a heap of leaves*/*rubble; a pile of bricks*/*wood*) and their range of collocation. Such NNQs will repay investigation to see whether productivity

of collocation correlates with delexicalization where the emphasis is on quantification rather than description.

3.3 Semantic weight

The question of the semantics of individual quantifiers affecting their number agreement is discussed by Reid (1991). In dealing with a category that includes some of our NNQs, where a singular quantifier + *of* takes a plural complement (*a row of horns, an increasing number of men*), he suggests a scale of semantic categories with varying semantic weight that dictates whether the verbal agreement goes with the singular entity (in our terms the *quantifier*), or the plural complement. At the "light" end of the scale are "decimals and fractions with one as numerator" (e.g. *one fifth of Australian men are*…), while at the heavy end are "semantically specific words that do not imply a referential plurality (e.g. *presence, process, retrieval, weight*…)" (1991: 270).[3] In the middle of this scale, and therefore most unpredictable as to agreement, come the categories of "imprecise aggregate: *group, handful, host, spate*" and "precise aggregate: *team, band*…*number, series, sequence*" – many of which fall under the current classification of NNQ.

3.4 Variation

3.4.1 Regional divergence

While we might expect to find regional difference in terms of verbal agreement, following the example of collective nouns, there is also some dictionary evidence of regional divergence in the choice of NNQ. The *Longman Dictionary of Contemporary English*, based on data from the Longman Corpus, labels *a bunch of* and a *raft of* as "esp. AmE", while *loads of* and a *stack/stacks of* are labeled "esp. BrE". The fact that these phrases are also labeled as either colloquial or informal suggests that it is at the less standard end of the spectrum that these regional differences occur. By using spoken corpora we will be able to investigate the range of these idioms across regions, and whether AusE or NZE show higher frequencies of forms that are regionally marked as British or American.

---

[3] For example "Sheer weight of numbers **gives** us an advantage".

3.4.2 Register

The question of register also features heavily in the discussion of NNQs. Quirk et al. describe what they call "open-class quantifiers" as being "chiefly used informally" (1985: 264), and Channell writes that the "starting point" for her study of this class of "vague language" was Crystal and Davy's observation of the frequency of such phrases in conversation data (1994: 95). It might be expected that as a symptom of vagueness, non-numerical quantification belongs more naturally in speech than in writing. But Channell herself showed in an earlier study (1990) that some academics have quite specific uses in their writing for apparently vague expressions such as *a number of*. Corpus data can quantify their relative frequency of usage in spoken and written texts, and also help to establish the range and regional spread of phrases considered to be at the more colloquial end of the spectrum, such as *oodles of, heaps of, loads of*.

4 Previous corpus studies

There is very little by way of previous corpus studies of NNQs. Kennedy (1987) looked at the variety of types of quantification in English from a language learner's perspective, using a small corpus of journalistic and academic written English to identify subcategories of quantification, and then compared the relative frequencies of these subcategories in the scholarly texts of the Brown and LOB corpora. These categories were titled "specific", "non-specific" and "relative" quantities/degrees. The NNQ construction, as classified here, came in under several of the "non-specific" subcategories, such as "small quantities/degrees"/"large quantities/degrees"/"non-specific parts of a whole", so it is not possible to make direct comparisons with any of the figures given by Kennedy for the overall categories, and individual tokens were only recorded for the "approximation" category. Although American (Brown) and British (LOB) corpora were used, no regional inferences were drawn.

A study more focused on NNQs was conducted by Channell (1994), as part of her investigation into different types of vague language. Spoken and written corpora, including the Oxford Corpus of the English Language and the Birmingham collection of English Texts, were used to assess the frequency of certain NNQs such as *bags of, a load/loads of, a lot/lots of, oodles of, a bit of* in spoken and written English, and to analyze whether they

collocated more frequently with countable or uncountable nouns, where there was possible variation. All the corpora used were of BrE, and therefore no regional distinctions could be made. While the quantitative data presented by Channell is limited, her study gives a useful starting point to the discussion of NNQs.

Continuing with the theme of vague language, Drave (2002) used a corpus of conversation between native speakers of English and Cantonese to investigate the range and function of vague language in intercultural conversations. The only NNQ he studied was *a lot of*, the most frequent in the corpus. Drave focused on the collocational patterning around a set of examples of vague language, such as *about*, *stuff*, *thing*, to determine if there were any functional differences between the use of vague language by this set of native and nonnative English speakers.

In this study we use data from the spoken and written components of ICE-AUS, ICE-NZ and ICE-GB to investigate regional and register variation in the class of non-partitive quantifying constructions consisting of a quantifying noun followed by *of* (as defined above, Section 3). The areas considered will be:

- verbal agreement and noun complementation in the cases of common NNQs that can be either singular or plural in form (such as *heap(s), load(s)*, *lot(s)*)
- fixed and variable collocation with a set of quantifying nouns (*bunch*, *heap*, *load*)
- the regional and register distribution of a selection of lower frequency NNQs

5 Corpus findings

5.1 A lot/lots of

Both Channell and Drave state that *a lot of/lots of* is particularly frequent in their corpus material, and *a lot of* was found to be the most frequent NNQ in the ICE corpora; it is one of the quantifying nouns which may be singular or plural, and it can take either a singular or plural noun complement (i.e. is number transparent). Quirk et al. (1985: 262) note that it is chiefly used informally, and is therefore of interest with regard to register. This indication of plentiful data and variability make it a useful point to start from in a comparative corpus analysis.

5.1.1 Regional and register features

Channell found that there was a distinct register difference in the use of *lots of*, (confirming Quirk et al.'s observation about its informality). It was much more frequent in the Cobuild spoken corpus than in the written corpus, where its use was restricted to direct/reported speech or personal narrative (1994: 102). The overall frequency of *a lot/lots of* in each of the ICE corpora, along with the comparative figures between the spoken and written components, are shown in Table 3.1:

Table 3.1: Frequency of *a lot/lots of* in ICE-AUS, ICE-NZ, ICE-GB

	ICE-AUS			ICE-NZ			ICE-GB		
	spoken	**written**	**total**	**spoken**	**written**	**total**	**spoken**	**written**	**total**
(a) lot of	355	45	400	353	71	424	241	22	263
lots of	59	17	76	67	31	98	78	26	104

The difference in frequencies of both *lot of* and *lots of* between the spoken and written data is marked, with this NNQ notably more common in spoken than in written English across the regions. Both ICE-AUS and ICE-NZ show particularly high frequencies for *a lot of*, with ICE-GB showing a lower frequency, but higher for *lots of* than either ICE-AUS or ICE-NZ.

5.1.2 Grammatical features

Channell noted that *a lot of* collocated with countables and uncountables, but did not remark on any difference in frequency. There was a strong preference for collocating of *lots of* with countable nouns (*lots of /things/children),* while noncount complements were equally possible (*lots of money*) though less represented in her corpora. She made no comment on verbal agreement for either quantifier.

In this study, complements have been classified either as singular or plural, to concentrate on the grammatical form rather than semantic distinctions that may arise from the choice of count or noncount uses of the noun, as in "lots of time"/"lots of times". Table 3.2 shows how *a lot of* and *lots of* collocated with singular and plural complements, as well as showing the verbal agreement with the NNQ phrase, where it was the subject of the verb and the number of the verb was evident. The complement

people is treated separately as it is singular in form, although usually plural in meaning (it is treated as a plural complement by Channell).

Table 3.2: Noun complementation and verbal agreement for *a lot/lots of* in ICE-AUS, ICE-NZ, ICE-GB

	ICE-AUS				ICE-NZ				ICE-GB			
	CP	**CS**	**VP**	**VS**	**CP**	**CS**	**VP**	**VS**	**CP**	**CS**	**VP**	**VS**
(a) lot of	152	180	38	38	141	221	25	38	77	149	22	47
(people)	68	-	31	9	62	-	14	3	37	-	19	4
total	220	180	69	47	203	221	39	41	114	149	41	51
lots of	40	27	10	3	54	28	5	7	48	44	17	10
(people)	9	-	3	0	16	-	2	1	12	-	4	2
total	49	27	13	3	70	28	7	8	60	44	21	12

(CP= plural complement; CS= singular complement; VP= plural verb; VS= singular verb)

In the ICE corpora, *a lot of* took a singular complement more frequently than a plural one in all regions (the difference was most marked for ICE-GB (=1:2), and then, for ICE-NZ (=2:3)). However, the numbers of singular and plural complements become more even, as in Channell's findings, when *people* is added to the plural complements. *Lots of* took a plural complement more consistently, with *a lot of* being about half as common in ICE-AUS and ICE-NZ, but much more common in ICE-GB. These results show a slight tendency for the singular form of the quantifier to take a singular complement (most marked in ICE-GB), and for the plural form of the quantifier to take a plural complement (least marked in ICE-GB). So there is a tendency towards consistency, but not a clear pattern.

Different trends were detectable in terms of verbal agreement across the regions. ICE-AUS displayed a consistent preference for plural verbs with *a lot of and lots of* (especially if the figures for *people* are included), whereas ICE-NZ and ICE-GB both went the other way for *a lot of*, with a slight preference for singular agreement.

When we look more closely at examples from the corpora, we discover that a high proportion of the singular verbs used in conjunction with NNQs are instances of existential *there*. As Biber et al. note: "The subject status of existential *there* is also indicated by the strong tendency in conversation to use a singular verb regardless of the number of the notional subject" (1999: 994). We cannot therefore attribute the mismatch between the number of the verb and that of the complement to the singular *lot* in Table 3.3 below (1a/b).

Table 3.3: Verbal agreement with noun complement of *a lot of/lots of*: *existential there* construction preceding NNQ compared with verb following NNQ

		ICE-AUS	ICE-NZ	ICE-GB
lot				
1a	there's a lot of CP	8	11	5
1b	a lot of CP is	0	0	0
2a	there's a lot of CS	20 (74%)	11 (48%)	30 (81%)
2b	a lot of CS is	7 (26%)	12 (52%)	7 (19%)
		35	**34**	**42**
lots				
3a	there's lots of CP	2	6	4
3b	lots of CP is	0	0	0
4a	there's lots of CS	1	0	3
4b	lots of CS is	0	0	1
		10	**13**	**21**

All of the instances from 1a are examples with existential *there*, and coming predominantly from the spoken corpora are proof of Biber et al.'s assertion that "In fact, such examples [*there's* + plural noun] are somewhat more common in conversation than the standard constructions with plural verb plus plural noun phrase" (1999: 186).

The data for *lots of* is insufficient to show a clear preference, but the marked tendency for a singular verb to be attached to *lots of* + plural complement (especially in ICE-NZ, see 3a) again shows the influence of existential *there*.

Examples of mismatches between the number of the complement when the verb follows are not common, although there are some to be found in ICE-GB. The only example with *a lot of*, is a rather questionable one where the plurality of the verb is influenced by coordination of the singular NNQ phrase with a plural subject:

(1) Poorly shaded lights or **a lot of** movement **are** also undesirable.
[ICE-GB W2B-033:095]

The speech data provides good evidence for speakers of English being comfortable with *there's* + plural complement in quantitative statements (100% of the instances of *a lot/lots of* with a singular verb and plural complement in Table 3.3 have existential *there,* and they are all from the spoken components of the corpora). Yet the following example enacts the apparent uncertainty for one speaker over verbal agreement with *lots of*, with a self-correction within the same utterance:

(2) But there there there **there's lots of** deers and lots of rabbits
[ICE-GB S1A-006:260]

The stuttering repetition of *there* suggests hesitation by the speaker over what follows, and the normally zero plural *deer* is mistakenly regularized to *deers* to match *rabbits*. A few moments later the speaker has resolved the problem with this grammatically regularized statement:

(3) There **are lots of** deer and lots of rabbits
[ICE-GB S1A-006:264]

Even given the influence of existential *there* on the number of the verb, it is worth looking at the overall pattern of agreement between the NNQ, its complement and associated verb, to look for signs of variability. Table 3.4 presents the percentages and numbers for *a lot of* and *lots of* showing where the associated verb agrees with the number of the complement, with the NNQ, where there is no distinction (i.e. both the NNQ and the complement are either singular or plural), or there is no agreement with either (e.g. *there's lots of ideas*).

Table 3.4: Verbal agreement with number of noun complement and/or NNQ for *a lot/lots of**

	ICE-AUS	ICE-NZ	ICE-GB
a lot of			
plural verb + plural complement	38 (50%)	23 (38%)	21 (31)
singular verb + sNNQ	8 (11%)	9 (15%)	5 (7%)
singular verb + singular complement + sNNQ	30 (39%)	28 (47%)	40 (60%)
plural verb + singular complement + sNNQ	0	0	1 (2%)
lots of			
singular verb + singular complement	1 (10%)	0	4 (19%)
plural verb + pNNQ	0	0	1 (5%)
plural verb + plural complement + pNNQ	7 (70%)	7 (50%)	12 (57%)
Singular verb + plural complement + pNNQ	2 (20%)	7 (50%)	4 (19%)

sNNQ = singular NNQ; pNNQ = plural NNQ
*includes existential constructions

Here we can see for *a lot of* that, where there is a mismatch between the number of the NNQ and the complement, the number of the complement is most likely to dictate the number of the following verb in ICE-AUS, and it is reasonably common in ICE-NZ (38%) and ICE-GB (31%). For these last two, there is a high proportion of NNQ phrases where the number of the complement and the NNQ match, and it is therefore not

possible to distinguish which element dictates agreement. However, if we add the examples for *a lot of* + *people*, we find a very consistent use of the plural verb across the regions (ICE-AUS, 31/40 instances, ICE-NZ, 14/17 and ICE-GB 19/23) which makes the complement the main determiner of verbal agreement across the regions, i.e. giving semantic (notional and proximity) rather than formal grammatical agreement.

We will now look at the set of NNQs found in the corpora that can vary between singular and plural forms, to see if they show any similar tendencies to the most frequent NNQ *a lot/lots of*.

5.2 ONNQs that have a singular or plural quantifying noun

The only vague quantifiers that Channell identified as being able to take a singular or plural form are *a load/loads of* and *a mass/masses of* (1994: 101-2). Neither of these were particularly frequent, so for the purpose of this study we will look at a larger set of NNQs with the singular/plural alternative, to see how they compare with *a lot/lots of*. As noted earlier, NNQs are an open grammatical set. We therefore applied certain criteria to select a coherent set of other NNQs (ONNQs) for comparison:

- No partitives were included, which also excluded measure nouns such as *cup/cups*.
- Numerical uses were included where the number did not refer to a precise amount, as in *a couple of/dozens of*.
- Only those were included that could take either a singular or plural noun complement (i.e. are number transparent), thus excluding some relatively frequent NNQs such as *group*, *set*, which consistently take plural complements.
- Only examples that had a quantitative function were included. Uses such as *clouds of gas*, which could be interpreted either as descriptive or quantitative were excluded, as were others such as *two piles of leaves*, where the number obviously takes over any quantifying function that would exist in examples such as *piles of homework*, and those with an adjectival qualifier that puts focus on the descriptive function of the word – *a small pile of leaves*.
- For the purposes of identifying a set of NNQs that could take either a singular or plural noun, only those that took a singular and a plural form in at least one of the ICE regions were included.

See Table 3.6 for a full list of the other quantifiers (ONNQs) inspected.

5.2.1 Regional and register features

Regional and register frequencies across the corpora for the plural and singular forms of these NNQs are shown in Table 3.5, compared with data from Table 3.1 for *a lot/lots of*.

Table 3.5: Comparison of *a lot/lot of* with alternative singular/plural ONNQs

	ICE-AUS			ICE-NZ			ICE-GB		
	spoken	**written**	**total**	**spoken**	**written**	**total**	**spoken**	**written**	**total**
(a) lot of	355	45	400	353	71	424	241	22	263
lots of	59	17	76	67	31	98	78	26	104
Total	414	62	476	420	102	522	319	48	367
Singular ONNQ	48	23	71	21	36	57	31	20	51
Plural ONNQ	38	13	51	50	32	82	36	25	61
Total	86	36	122	71	68	139	67	45	112

The most striking contrasts with the *a lot/lots of* data is that there is a much more even spread of ONNQs between the regions. ICE-NZ is just ahead, followed by ICE-AUS and then ICE-GB. There is not the same weighting of frequencies towards the spoken component of the corpora. While there are still higher frequencies of ONNQs in spoken data across the regions, the ratio between spoken and written is much closer, particularly in ICE-NZ.

With regard to the ratio of singular and plural forms, again the effect is one of a more even distribution. Whereas the singular *a lot of* is the dominant form in all regions, ONNQs appear to offer more flexibility as to the choice of singular or plural form, with the plural forms actually the more frequent option in ICE-NZ and ICE-GB. Table 3.6 gives a breakdown of the ONNQs selected following the criteria in Section 5.2:

Table 3.6: Frequency of alternative singular/plural ONNQs in ICE-AUS, ICE-NZ, ICE-GB

	ICE-AUS			ICE-NZ			ICE-GB		
	spoken	**written**	**total**	**spoken**	**written**	**total**	**spoken**	**written**	**total**
bag	2	0	2	0	1	1	2	0	2
bags	1	1	2	0	0	0	0	0	0
band	0	3	3	1	2	3	2	2	4
bands	0	0	0	0	1	1	2	4	6
block	7	2	9	5	3	8	1	2	3
blocks	1	4	5	3	4	7	0	1	1
body	3	4	7	3	7	10	3	3	6
bodies	1	0	1	1	0	1	0	1	1
bunch	9	1	10	5	5	10	3	1	4
bunches	1	0	1	0	2	2	0	0	0
crowd	3	2	5	0	0	0	0	1	1
crowds	0	0	0	1	0	1	0	4	4
dose	4	3	7	0	2	2	2	0	2
doses	0	1	1	0	1	1	0	0	0
dozen	2	0	2	0	2	2	0	0	0
dozens	3	2	5	3	4	7	4	3	7
heap	0	0	0	1	2	3	0	0	0
heaps	21	0	21	38	4	42	2	1	3
host	0	3	3	2	1	3	1	2	3
hosts	0	0	0	0	0	0	0	1	1
load	7	1	8	1	1	2	10	1	11
loads	1	1	2	0	2	2	16	4	20
mass	4	0	4	0	4	4	5	8	13
masses	3	1	4	0	0	0	5	1	6
pile	2	4	6	2	2	4	1	0	1
piles	1	0	1	1	3	4	3	0	3
quantity	3	0	3	0	3	3	1	0	1
quantities	3	3	6	0	7	7	3	5	8
stack	2	0	2	1	1	2	0	0	0
stacks	2	0	2	2	1	3	2	0	2
	86	**36**	**122**	**71**	**68**	**139**	**67**	**45**	**112**

This list shows the variety of ONNQs available. While *a lot/lots of* gives no more specific sense than "a large amount of" (see Channell 1994: 101, 106), we have several different categories here:

- Apparently numerical (*dozen*) or precise (*dose*) which are used to refer to a large or unspecified number, e.g. *dozens of celebrities*, *a dose of discipline*
- Non-specific quantifiers that often have to be qualified to give a sense of the amount referred to: *a small/large body of work*
- Nouns that are used generally to denote a large amount (*crowd, host, mass, quantity*)
- Nouns that are used generally to denote a collection of objects of unspecified size, but are used as quantifiers always to express a large amount (*bag, heap, load, stack*)

– The fact that they can be used either singularly or plurally shows that they can all be conceptualised either as an undifferentiated mass or a collection of discrete items.

The low frequencies of individual ONNQs in Table 3.6, in comparison to the figures for *a lot/lots of*, indicate that their primary functions are not as quantifiers. They can however be number transparent, and are therefore interesting to compare with *a lot/lots of*. Some of the more frequent ONNQs such as *set of*, *group of*, which are excluded from Table 3.6 because they showed no signs of number transparency, only taking plural complements, are listed in Appendix 3.1.[4]

In contrast to *a lot/lots of* it is clear that these NNQs are not restricted in the same way to the spoken genre. Some such as *band*, *blocks*, *body* and *host* in particular show a higher frequency in written material. The following examples show them being used in both bureaucratic (4) and technical (5) written contexts, with (6) appearing in the business correspondence category:

(4) Directly related to these problems of responsiveness and sensitivity were a **host of** 'technical' problems…
[ICE-AUS W2A-012:13]

(5) The alternatives are, therefore, discussed, where possible, in terms of **blocks of** energy corresponding to baseload generation at 250 MW.
[ICE-NZ W2A-038:10]

(6) As discussed I am introducing a new **band of** charges, which will be applicable for performance times, as against rehearsal times.
[ICE-GB W1B-021:091]

Table 3.6 also draws attention to some regional difference in the use of ONNQs. *Heaps of* appears to be a distinctively antipodean quantifier, with 21 and 42 occurrences in ICE-AUS and ICE-NZ respectively, with only 3 in ICE-GB. Conversely, *loads of* appears almost exclusively in the British corpus, with 20 in ICE-GB as against 2 each in ICE-AUS and ICE-NZ (although ICE-AUS has a similar frequency for the singular form). The other ONNQ showing a distinctive regional grouping is *bunch of*, often identified as typically American

[4] Note that these are both classified by Biber et al. as "quantifying collectives", which only take countable nouns. Likewise *bunch* which, conversely, was found to be number transparent in the corpus data.

(see Section 1), which is more than twice as frequent in ICE-AUS and ICE-NZ as it is in ICE-GB.[5] In order to look more closely at these regional divergences, we will assess the collocations of *bunch*, *heap* and *load*, to see if there are differences in usage that reflect the differences in frequency.

5.2.2 Collocation with heap(s)/load(s)/bunch(es)

These ONNQs form a useful set for investigation not only because of their regional differences, but also because they each possess, in their primary senses, an element of physical description that more abstract quantifiers such as *lot/lots* do not. Table 3.7 gives an inventory of collocations for each of the ONNQs for the regions under discussion:

Table 3.7: Collocations with *heap(s)/load(s)/bunch(es)*

	ICE-AUS	ICE-NZ	ICE-GB
heap	*metal*	*clothing, cushions, papers, stuff*	
heaps	*cash, character, friends, HECS, junk, kids, nationalities, money, people* (2), *places, plastic surgery, stuff, students, times* (2), *work* (2)	*action, apostrophes, energy, files, halyard, hassles, leaves, lights, mending, money* (2), *movies, music, noise, overtime*(2), *people* (7), *places, profits, pubs, presents, rouge, rubbish, seats, spray & rain, shit, stars, support, tapes* (2), *things* (2), *time, wine, work, zeros*	*bodies, hours, relations*
load	*crap* (3), *filters, money, people, rubbish* (2)	*casks, nonsense*	*advantages, bread, engineers, fun, nonsense, numbers, rubbish* (2), *sensitivity, uniforms*
loads	*calcium, cards & letters*	*ammonia & phosphate, food*	*blanks, books* (2), *films, jobs, money* (2), *people, photos, rabbits & guinea pigs, rails, sentences, space, stuff* (2), *texts, things* (2)
bunch	*Aborigines, clods, cows, kids, lads, men & women, people* (2), *recruits, words*	*blood, guys, lefties, live wires, mountains, no-hopers, people* (3), *warriors*	*dorks & bubbleheads, flowers, people, roses*
bunches	*flowers*	*flowers, greenery*	

It is interesting to note, first of all, the difference between the uses of *heap* and *heaps*. While the collocations for the singular form are all objects that can form a physical heap, *clothing*, *cushions*, *metal*, the collocations for ICE-AUS and ICE-NZ give a much wider range, from the purely abstract *time*, to objects that have an abstract and a concrete sense,

[5] There are also 7 instances of *heaps*, and 12 of *bunch* in ART, with no examples of *loads* as a quantifier, further confirming the regional distinctions.

such as *money*, and to animate entities such as *kids*, *people*. The ICE-GB collocations show some of that range, but the reference to *bodies*, as in example (7) below, is referring to piles of human corpses, and therefore foregrounding the descriptive rather than quantitative sense of *heaps*:

(7) I was walking by **piles of heaps of** bodies that had been torched cos there was these black charred embers of grotesque <,> figures you know
[ICE-GB S2A-050:152]

Load(s) doesn't show the same distinction between the singular and plural forms in the ICE-GB data, with *load* and *loads* both giving examples of animate, inanimate and abstract objects. ICE-AUS and ICE-NZ have a much more limited range, with inanimate objects (*casks*, *filters*) again emphasizing the descriptive (as do forms not included here such as *bus loads*, *car loads*, *truck loads*, which appear in both corpora). Both ICE-AUS and ICE-NZ do, however give a dismissive connotation to *loads* with the collocations *crap*, *nonsense*, *rubbish*.

(8) They reckon this thing's cost them about six thousand dollars and I reckon that's **a load of** crap
[ICE-AUS S1A-030:51]

Interestingly ICE-GB shows a similar variation happening with *bunch*. Alongside the very literal collocations of flowers, roses, is the evocative *dorks and bubbleheads*, mirroring the sardonic tone of *clods*, *lefties*, *no-hopers*. Perhaps this is a transition that many NNQs go through, from being purely descriptive to a colloquial means of bunching things/people together in an offhand, dismissive way, to achieving a neutral sense of quantity.

5.2.3 Grammatical features

To draw comparisons between *a lot/lots of*, and the number-flexible ONNQs, we will treat the latter as a set. Table 3.8 compares the figures from Table 3.2 for noun complementation and verbal agreement (with any indeterminate plurals such as *people*, *data* removed).

Table 3.8: Noun complementation and verbal agreement for *a lot of*/*lots of* compared with alternative singular/plural ONNQs

	ICE-AUS				ICE-NZ				ICE-GB			
	CP	**CS**	**VP**	**VS**	**CP**	**CS**	**VP**	**VS**	**CP**	**CS**	**VP**	**VS**
(*a) lot of*	152	180	38	38	141	221	25	38	77	149	22	47
lots of	40	27	10	3	54	28	5	7	48	44	17	10
ONNQ (singular)	37	32	5	17	32	22	5	4	20	24	4	10
ONNQ (plural)	28	22	6	3	35	34	12	6	30	29	8	2

(CP= plural complement; CS= singular complement; VP= plural verb; VS= singular verb)

There is a remarkable consistency between the AusE and NZE preference for a plural complement with a singular ONNQ, and the data for ICE-GB shows a much higher proportion of CPs than for *a lot of*. This suggests an even higher degree of grammaticization for the ONNQs than for a lot of, where the CS is preferred across the regions.

Again the plural forms of the ONNQs are again less frequent (as for *lots of*) but to a lesser degree. They show a greater tendency than *lots of* to take a singular complement, reinforcing the case for grammaticization of ONNQs.

With verbal agreement, the numbers for singular and plural verbs with a singular ONNQ are quite even across the region, in contrast to the preference for a singular verb in ICE-NZ/ICE-GB for *a lot of*. Plural ONNQs take a plural verb more regularly than does *lots of*, so we have contrasting trends whereby the singular ONNQ appears more likely to take a plural complement, but the verb agreement is unpredictable, whereas plural ONNQs take plural complements and plural verbs quite consistently. As with a *lot*/*lots of*, it will be instructive to look more closely at the impact that the existential *there* construction has.

Table 3.9: Verbal agreement with noun complement of alternative singular)/plural ONNQs: existential *there's* preceding ONNQ compared with verb following ONNQ

		ICE-AUS	ICE-NZ	ICE-GB
Singular				
1a	there's sONNQ of CP	1 (33%)	1 (50%)	1 (33%)
1b	sONNQ of CP is	2 (67%)	1 (50%)	2 (67%)
2a	there's sONNQ of CS	6 (86%)	2 (67%)	3 (60%)
2b	sONNQ of CS is	1 (14%)	1 (33%)	2 (40%)
		10	5	8
Plural				
3a	there's pONNQ of CP	7 (88%)	4 (100%)	1 (100%)
3b	pONNQ of CP is	1 (13%)	0	0
4a	there's pONNQ of CS	0	1 (50%)	0
4b	pONNQ of CS is	0	1 (50%)	1 (100%)
		8	6	2

sONNQ = singular ONNQ; pONNQ = plural ONNQ
CP = plural complement; CS = singular complement

If we compare the results in Table 3.9 to those of Table 3.3, there is not quite the same degree of exclusivity for *'s* as the singular verb with a plural complement. Whereas Table 3.3 had no returns for 1b in any of the regions, the singular verb following the singular ONNQ + plural noun complement is actually more common than *there's* introducing it in both ICE-AUS and ICE-GB. There is also one example of *there's* with the plural ONNQ. Again we could attribute this to the fact that these ONNQs are less delexicalized/degrammaticized than *a lot/lots of*.

Table 3.10 gives the overall patterns for agreement between other singular/plural ONNQs and the noun complement, including the existential *there's* construction.

Table 3.10: Verbal agreement with noun complement or ONNQ for singular/plural ONNQs

	ICE-AUS	ICE-NZ	ICE-GB
singular verb + sONNQ	4 (29%)	4 (44%)	4 (33%)
singular verb + singular complement + sONNQ	3 (21%)	2 (22%)	3 (25%)
plural verb + singular complement + sONNQ	7 (50%)	3 (33%)	5 (42%)
plural verb + plural complement	0	0	0
singular verb + singular complement	0	2 (13%)	1 (11%)
plural verb + pONNQ	0	3 (20%)	1 (11%)
plural verb + plural complement + pONNQ	2 (20%)	6 (40%)	6 (67%)
singular verb + plural complement + pONNQ	8 (80%)	4 (27%)	1 (11%)

sONNQ = singular ONNQ; pONNQ = plural ONNQ

The clear contrast between these figures and those for *a lot* in Table 3.4 is that there is a greater tendency for the number of other singular ONNQs to dictate verbal

agreement, although the complement still has a higher percentage. This suggests that these ONNQs have more lexical weight – *a load* or *a heap* retaining a physical quality that *a lot* doesn't. This lexical weight is lost in the plural form, where the proportion of verbal agreement with the ONNQ is much lower. The figures for plural ONNQ are inconclusive as to agreement with complement or ONNQ, but they do show an even more marked tendency for inconsistency than that already seen for *lots of* – although again this is heavily influenced by existential *there*, as in:

(9) And there was heaps **heaps of** kids like you go to Carnarvon Gorge [ICE-AUS S1A-067:308]

Although the numbers that demonstrate verbal agreement with these plural ONNQs are quite small, there is evidence here of the tension between the descriptive and the quantitative roles of these less frequent ONNQs.

6 NNQs with singular or plural forms only

This paper has concentrated on a subset of NNQs – those that can be either singular or plural in form – in order to focus on questions of agreement. The full inventory of NNQs discovered through corpus searches (see Appendix 3.1) spans a much wider range. The NNQs listed are all non-partitive, and were selected on the basis that they had quantitative uses that went beyond their conventional descriptive use, e.g. *a parcel of shares* (ICE-AUS S2A-031:125), *a spot of coffee and porridge* (ICE-NZ W1B-003:423) . The lists may not be exhaustive, but they are representative of the range of NNQs found. There was a remarkable consistency between the regions of number and range of quantifiers (ICE-NZ had the most NNQs, with 70, then ICE-AUS with 67, ICE-GB, 58). Table 3.11 below shows the only points of regional divergence.

Table 3.11: NNQs common and specific to different regions

ICE-AUS only	ICE-NZ only	ICE-GB only	ICE-AUS,NZ	ICE-NZ,GB
dollops (1)	*dash* (1)	*wadges* (1)	*clump*(s) (2,3)	*smattering* (1,3)
	droves (1)		*mob* (1,1)	*spot* (2,1)
	gaggle (1)		*oodles* (1,1)	
	swag (2)		*raft* (2,3)	

Clearly the numbers are not large enough to make regional comparisons. There are however particularly AusE and NZE words, such as *droves*, *mob* and *swag*, and it is

therefore not surprising that they don't appear in ICE-GB. Conversely, *oodles* is one that Channell highlighted in her BrE data (1994: 103), so its appearances only in ICE-AUS and ICE-NZ English here are not representative.

While the overall figures in Appendix 3.1 do point to a greater use of NNQs in spoken than written English, they are by no means exclusive to informal communication. Consider the following examples for *raft of*:

(10) Male unionists were intent on the exclusion of possible competition; they managed to achieve a **raft of** restrictions to women's employment opportunities at the same time as they successfully excluded immigrant competition by means of the White Australia Policy.
[ICE-AUS W2A-017:73]

(11) There was a a **raft of** of basically policy givens that we were working within
[ICE-NZ:S2A-047:20]

(12) The use of in situ concrete for floors in Australia generates a whole **raft of** falsework design associated with table forms.
[ICE-NZ:W2A-040:92]

Two of these examples are from the "learned information" category of ICE, and are being used in a quite formal, or technical context (in the case of (12)). While (11) comes from the spoken medium, it is clearly being used as a piece of political jargon, and therefore within a specialised discourse. *Raft of* has no more specific sense than *lot of*, and yet its imprecision does not disqualify it from a formal setting. Channell (1990) looked at some examples of academics in the field of economics – where we would expect numerical precision to be particularly important – who indicated "that they recognize the inherent vagueness [of non-numerical quantifiers] and know how to exploit it for particular communicative purposes" (1990: 103). For example, the choice of the NNQ *a number of* was explained by the academic who had used it to indicate "this is an area where a considerable amount of work has been done and there is no monopoly of interpretation…There's at least two, 'cause I think I'd have put two if there's only two, and I think the word *a number* also indicates there's no front-runner. I think if there was about 25 of them I'd have started introducing classifying things in there – 'a great number',

'a vast number'…". Note also the prevalence of existential *there's* with a plural in the expert's reported utterance.

7 Conclusions

This study has shown several areas of interest in the study of NNQs that would repay study beyond the corpora used here:

Process of grammaticization

The figures for *a lot/lots* of showed a high degree of variability in the choice of a singular or plural complement, while verbal agreement was quite regular – when the influence of existential *there* was allowed for – demonstrating this NNQ's role as a complex determiner. The singular/plural ONNQs looked at developed rather lower degrees of number transparency, but more variability as to verbal agreement, suggesting a lesser degree of delexicalization/ degrammaticization.

Classifications of NNQs modified by corpus evidence

The grammars looked at provided a template which allowed this study to target the most likely types of NNQ to give evidence of variation as to the number of the noun complementation and verbal agreement. Therefore partitives were excluded, and Biber's class of "quantifying collectives" not looked at in detail. Corpus evidence did, however, provide motivation to reassess his labeling of some NNQs, with *bunch of* in particular showing signs of number transparency that suggest it does not always function simply as a collective.

Regional differences

Corpus evidence confirmed *load(s) of* as a typically British NNQ, and found that the AmE-marked bunch of is more popular in AusE and NZE than in BrE. *Heap(s)* was also found to be a particularly antipodean NNQ, with few examples of its usage in ICE-GB. NZE showed itself to be particularly productive both in the overall range of NNQs used, and in the range of collocations found with the subset *bunch/heap/load*.

Register differences

While the figures for *a lot of/lots of* confirmed Quirk et al.'s contention that NNQs are more frequent in spoken than written English, some NNQs revealed more subtle relationships to register. Examples such as *band/block/host/number* of showed evidence of use in written texts of carrying more specialised meanings than simply that of a vaguely large amount. Even where vague quantity appeared to be the only sense, as in *raft of*, the formal context in which this NNQ is used suggests a more purposeful approach to numerical imprecision than mere informal vagueness or overstatement.

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Channell, Joanna. 1990. "Precise and vague quantities in academic writing". In Walter Nash, ed. *The writing scholar: Studies in the Language and Conventions of Academic Discourse*. Beverly Hills: Sage, 95-117.

Channell, Joanna. 1994. *Vague Language*. Oxford: Oxford University Press.

Drave, Neil. 2002. "Vaguely speaking: a corpus approach to vague language in intercultural conversations". In Pam Peters, Peter Collins and Adam Smith, eds. *New Frontiers of Corpus Linguistics. Papers from the 21st International Conference on English Language Research on Computerized Corpora.* Amsterdam, New York: Rodopi, 25-40.

Huddleston, Rodney and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hundt, Marianne. 2009. "Concord with collective nouns in Australian and New Zealand English". In P. Peters, P. Collins and A. Smith (eds.) *Comparative Studies in Australian and New Zealand English: Grammar and Beyond,* 207-224. Amsterdam/Philadelphia: John Benjamins.

Kennedy, Graeme. 1987. "Quantification and the use of English: A case study of one aspect of the learner's task". *Applied Linguistics* 8(3): 264-86.

Levin, Magnus. 1998. Concord with collective nouns in British and American English. In Hans Lindquist, Staffan Klintborg, Magnus Levin and Maria Estling, eds. *The Major Varieties of English.* Växjö University, 193-204.

Longman Dictionary of Contemporary English, 4th ed. 2007. Harlow: Pearson Education.

Peters, Pam. 1999. "Differing on agreement: A report on Langscape 3". *English Today* 15(2): 6-9.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Reid, Wallis 1991. *Verb and Noun Number in English: A Functional Explanation*. London, New York: Longman.

Appendix 3.1

	ICE-AUS	Spoken	Written	ICE-NZ	Spoken	Written	ICE-GB	Spoken	Written
1	bag	2	0	bag	0	1	bag	2	0
2	bags	1	1	bags	1	2	bands	2	4
3	band	0	3	band	0	1	batch	1	4
4	batch	2	0	batch	1	0	block	1	2
5	block	7	2	block	5	3	blocks	0	1
6	blocks	1	4	blocks	3	4	bodies	0	1
7	bodies	1	0	bodies	1	0	body	3	3
10	body	3	4	body	3	7	bunch	3	1
11	bunch	9	1	bunch	5	5	chunks	1	2
12	bunches	1	0	bunches	0	2	cluster	1	0
13	chunks	0	1	chunk	4	1	clusters	0	1
14	clump	0	1	cluster	0	5	clutch	0	1
15	clumps	1	0	clusters	0	1	couple	97	29
16	cluster	2	4	clutch	1	1	crowd	0	1
17	clusters	1	2	couple	148	51	crowds	0	4
18	clutch	2	0	crowds	1	0	cup	10	3
19	clutches	0	2	cup	13	5	cups	1	0
20	couple	187	36	cups	1	2	deal	26	12
21	crowd	3	2	dash	1	0	dose	2	0
22	cup	20	7	deal	13	26	dozens	4	3
23	cups	4	1	dose	0	2	drop	1	0
24	deal	21	9	doses	0	1	flock	0	1
25	dollops	0	1	dozen	1	2	flood	1	1
26	dose	4	3	dozens	3	4	group	38	32
27	doses	0	1	drop	1	1	groups	10	21
28	dozen	2	0	droves	0	1	heaps	2	1
29	dozens	3	2	flocks	0	2	hordes	2	0
30	drop	0	1	flood	0	4	host	1	2
31	flocks	2	0	gaggle	0	1	hosts	0	1
32	flood	1	3	gang	1	1	load	10	1
33	gang	1	0	gangs	0	1	loads	16	4
34	gangs	0	1	group	33	51	lot	241	22
35	group	47	17	groups	7	16	lots	78	26
36	groups	8	5	heap	1	2	lumps	1	0
37	heaps	21	0	heaps	38	4	mass	5	8
38	hordes	0	1	horde	2	0	masses	5	1
39	host	0	3	hordes	0	1	myriad	0	2
40	load	7	1	host	2	1	number	123	115
41	loads	1	1	load	1	1	numbers	19	10
42	lot	355	45	loads	0	2	parcels	0	1
43	lots	59	17	lot	353	71	pile	3	0
44	lump	0	1	lots	67	31	piles	3	0
45	mass	4	0	mass	0	4	pinch	1	1
46	masses	3	1	mob	0	1	quantities	3	7
47	mob	1	0	myriad	0	2	quantity	2	2
48	myriad	1	2	number	113	200	scrap	0	2
49	number	87	69	numbers	9	32	set	20	34
50	numbers	8	15	oodles	0	1	sets	6	7
51	oodles	0	1	pack	4	0	smattering	1	0
52	parcel	1	0	parcel	0	2	spate	0	2

	ICE-AUS	Spoken	Written	ICE-NZ	Spoken	Written	ICE-GB	Spoken	Written
53	pile	3	5	pile	1	2	spot	2	0
54	piles	1	0	piles	1	4	stacks	2	0
55	pinch	0	1	pinches	1	0	swathe	0	0
56	quantities	3	3	quantities	1	13	swathes	0	1
57	quantity	1	0	quantity	0	7	touch	2	3
58	raft	1	1	raft	1	2	wadges	0	1
59	scrap	0	1	scrap	1	0			
60	scraps	0	1	scraps	0	1			
61	set	31	32	set	26	43			
62	sets	1	6	sets	6	7			
63	spate	1	3	smattering	0	3			
64	stack	2	0	spate	1	0			
65	stacks	2	0	spot	1	0			
66	swathe	0	2	stack	1	1			
67	touch	2	3	stacks	2	1			
68				swag	0	2			
69				swathe	1	0			
70				touch	2	6			
	Total	**932**	**329**		**883**	**651**		**752**	**381**

Section 3A: Non-numerical quantifiers (NNQs) and grammaticalization

Relatively little has been written on NNQs to suggest that they are multiword units – hence the need for a term to cover them, provided in Paper 2 (2009). But more recent work by Langacker (2010) and Brems (2012) has argued for the grammatical status of some of the more common instances of them. Brems presents the coalescence of words in the forms *buncha*, *loadsa* and *heapsa* as evidence of movement towards a more advanced stage of grammaticalization (2012: 94), although she admits that these forms are only found in informal, internet data. No evidence of their use was found in the corpora used in this study.

1. Issues of lexicalization and grammaticalization

The productivity of the quantifying N + *of* template – demonstrated in Paper 2 – suggests that many of the NNQs are created by a process of analogy, but sometimes the line between quantifying and descriptive noun is hard to draw. One clear means of distinguishing was the use of premodifying adjectives as in *a small pile of leaves* where the nature of the quantification has to be supplied in a way that more canonical quantifiers like *many* or *a lot of* don't require. Some of the NNQs investigated, like *a heap/heaps of* and *a load/loads of* appeared to be in an intermediate stage between functioning as descriptors (*a truck-load of coal* vs. *loads of money*).This invokes Hopper's (1991) concept of 'layering', where the more grammaticalized usage can coexist with the lexical one.

The presence of premodification also draws attention to both the question of semantic weight of the noun, and its categorial status. Use of adjectives within potential NNQ constructions both modifies the semantic content of the noun, and emphasizes its status as a noun. If we consider also that in a number of cases the noun can be either singular (with preceding indefinite article), or plural (with appropriate inflection) – in both cases the category is clearly marked. There is therefore little evidence of decategorialization within this multiword unit.

Having said this, when indicators of category such as the plural do not have an effect on the corresponding complement and subsequent verb agreement, then this can be an indicator of grammaticalization. The NNQs looked at in Paper 2 tended to contain nouns or a type that, according to Reid (1991) is most unpredictable as to agreement (see

Paper 2, 3.3, p.51) so don't dictate whether verbal agreement goes with the quantifier or the complement. Paper 2 found several examples where the number of the complement, rather than the number of the quantifying noun, dictated the number of the verb, thus suggesting that the NNQ is functioning as a complex determiner. In these instances, the whole unit undergoes syntactic reanalysis, rather than individual elements within it losing or changing grammatical status (compare the unmarked verbs becoming nouns as in LV constructions). This is a very clear indicator of grammaticalization in some established NNQs like *a lot* or *lots of*, and there are signs of it for others such as *heaps of* and *loads of*.

2. Grammars

The grammars looked at in Paper 2 (Quirk et al., 1985; Biber et al., 1999; Huddleston & Pullum , 2002) presented similar problems of non-convergent nomenclature as were found for LVs, with each reference having a different term for the essentially similar items they described. What was common to each grammar, though, was that the role of each quantifier was described in terms of complementation patterns, whether it went with count or non-count nouns. In this respect, unlike the LVs which were listed as variously idiomatic units, NNQs are described in relation to their grammatical relationships, with Quirk et al. in particular arguing that the expressions function as complex determiners (1985: 264), as suggested here.

3. Dictionaries

Dictionaries were not systematically surveyed in Paper 2, but the *Longman Dictionary of Contemporary English* provided some attestation of regional preferences for some of the more colloquial NNQs, such as a *bunch of* (AmE) and *loads of* (BrE), and the corpus data pointed to *heaps of* as an antipodean quantifier. This regional labelling was not found in the L1 dictionaries examined, with the exception of *bunch of*, labelled as "chiefly N.Amer" by *Oxford Dictionary of English* (3rd ed., 2010). Evidence for the grammaticalized status of the most common NNQ, *a lot of* or *lots of*, is to be found in both L1 and L2 dictionaries. The *Oxford Dictionary of English*, somewhat unfortunately, classifies it as a pronoun, but one of the examples given – "there are a lot of actors in the cast" – clearly demonstrates verbal agreement with the complement, and therefore supports its analysis as a determiner. The

Collins Cobuild Dictionary (3rd ed. 2001), distinguishes the use of *a lot of* and *lots of* as quantifiers from *lot* as a pronoun ('I learned *a lot*') and as an adverb ('I like you *a lot*').

4. Summary

The first two published papers investigated examples of open class multiword units – the verb phrase in Paper 1 and the noun phrase premodifier in Paper 2 – and found quite different outcomes in terms of grammaticalization. While the LV construction was relatively unfixed, and indeed appeared to encourage variation, particularly in the premodification of the N element, the NNQs looked at were able to be distinguished as to their descriptive or quantifying function on the basis of whether they allowed premodification or not. Moreover, there was evidence for reanalysis of the noun phrase as a complex determiner, depending on whether the number of the quantifying noun had an effect on the agreement of the verb – and thus grammaticalization of an open class unit into a closed class.

The next two published papers will apply the same corpus methodology to two closed class multiword units: complex prepositions like *in spite of*, and a potential class of complex subordinators of the type *the moment, the instant*. Each of these units contains a N element, and therefore has the potential to exhibit similar characteristics, that will assist in identifying whether there are common features of grammaticalization exhibited across different grammatical classes.

Chapter 4

Complex prepositions and variation within the PNP construction (Paper 3)

Smith, A. 2013. Emergent complex prepositions: variability within the PNP construction. In H. Hasselgård, J. Ebeling, and S.O. Ebeling (eds.) *Corpus perspectives on patterns of lexis,* 153-174. Amsterdam/Philadelphia: John Benjamins. https://doi.org/10.1075/scl.57.12smi

"This granulo-pulpy matter was in process of being converted into ova."

Charles Darwin (*Journal,* 1845)

Complex prepositions and variation within the PNP construction

Adam Smith

This study looks at the variations within preposition + noun + preposition (PNP) sequences such as *at (the) risk of,* commonly classified as complex prepositions (CPs). The current literature suggests that the more indivisible the structure, the more grammaticalised the unit. Representations of complex prepositions within contemporary grammars indicate that the most common intruder within the fixed PNP sequence is the definite article. Synchronic and diachronic corpus studies were carried out to assess how fixed the form with the definite article is, and whether any CPs have shown a recent tendency to lose it. Decategorialisation was found to be only a minor factor for the CPs investigated, with a combination of semantic and grammatical factors featuring in the grammaticalisation process.

1. The status of the CP as a grammaticalised unit

There is disagreement in the grammatical literature as to whether certain preposition + noun + preposition (PNP) sequences, such as *in front of, on behalf of, in relation to* constitute grammaticalised units or are simply examples of idiomatic expressions. The majority opinion accepts their grammatical status, from Sweet onwards, who identifies a word sequence such as *by means of* as a sub-category of prepositions, terming it a "group-preposition" (1891: 134-5). Contemporary grammars such as Quirk et al. (1985) and Biber et al. (1999) include them within the category of complex prepositions (CPs). A dissenting view is expressed by Huddleston and Pullum (2002), who regard possible variations such as "*on behalf of* my son" and "*on my son's behalf*" as proof that these units are divisible, and therefore cannot be said to be functioning as grammaticalised units.

The potential for variation within the unit is acknowledged both by those who view the construction as a grammaticalised unit, and those who see it simply as a more or less free idiomatic expression. Huddleston and Pullum show variations as proof of non-grammaticality: what others categorise as CPs differ from free expressions only in that "they are in varying degrees idiomatic" and "they do not permit the full range of syntactic

manipulation that applies with free expressions" **(**2002: 619**)**. Those that accept CPs as a class see variations as points on the continuum between free expression and grammaticalised unit:

> In the strictest definition a complex preposition is a sequence that is indivisible both in terms of syntax and in terms of meaning. However, there is no absolute distinction between complex prepositions and constructions which can be varied, abbreviated, and extended according to the normal rules of syntax. (Quirk et al. 1985: 671)

Biber et al. also stress this sense of gradience: "Some multi-word sequences function semantically and syntactically as single prepositions…The distinction between complex prepositions and free combinations is a matter of degree." (1999: 75). We shall look first at the kinds of variation that are considered possible for the PNP sequence.

2. Types of variation

Quirk et al. present a range of possible variations, with *in spite of* used as an example of the most invariable type of PNP sequence, and *on the shelf by* as a free combination at the other end of the scale:

a. Prep2 can be varied
   on the shelf at (the door) [but not: **in spite for*]
b. noun can be varied as between singular and plural
   on the shelves by (the door) [but not: *in spites of*]
c. noun can be varied in respect of determiners
   on a/the shelf by; *on shelves by (the door)* [but not: **in a/the spite of*]
d. Prep1 can be varied
   under the shelf by (the door) [but not: **for spite of*]
e. Prep + complement can be replaced by a possessive pronoun
   on the surface of the table ~ on its surface
   [but: *in spite of the result ~ *in its spite*]
f. Prep2 + complement can be omitted
   on the shelf [but not: **in spite*]
g. Prep2 + complement can be replaced by a demonstrative
   on that shelf [but not **in that spite*]
h. The noun can be replaced by nouns of related meaning
   on the ledge by (the door) [but not: **in malice of*]
i. The noun can be freely modified by adjectives
   on the low shelf by (the door) [but not: **in evident spite of*]

(Quirk et al 1985, 671-2)

This list of 'indicators of separateness' is used by Quirk et al. to make the case for CPs as grammaticalised units, by arguing that examples such as *in search of* with two possible variations (f and h) is considered more grammaticalised than *in defence of* with four variations (c, e, h, i) are still closer on the scale to *in spite of* than *on the shelf by*. Huddleston and Pullum use similar examples of variation within the lexical bundle to disprove the class. Neither of these authorities treats the types of variation differently – in Quirk et al. it is the number of potential variations that a PNP construction allows, rather than which ones, that determine how cohesive the unit is.

This approach does not seem entirely consistent. Can a variation like *in/on behalf of*, where the choice of initial preposition appears to be no more than a regional difference between British and American English, be given the same status as a) *in case of/in the case of*, where the definite article changes the **meaning** of the unit, or b) *in lieu of/in lieu*, where the loss of the second preposition changes the **grammatical function** of the lexical bundle from prepositional to adverbial? Denison argues that grammarians are making difficulties for themselves by treating these variations as equivalents: "ICE-GB2, one of the most carefully and consistently tagged of corpora, takes *behalf* as part of a complex preposition when followed by *of*, but as N when preceded by a possessive" (2010: 120).

3. Variation and grammaticalisation

The correlation between fixity of form and grammaticalised status, as opposed to the variability inherent to free expressions is a key element of grammaticalisation theory. This is expressed by Traugott and Heine as follows:

> The study of grammaticalization therefore highlights the tension between relatively unconstrained lexical expression and more constrained morphosyntactic coding, and points to relative indeterminacy in language and to the basic non-discreteness of categories.
>
> (1991: 1)

Traugott and Heine are here referring to an indeterminacy between lexical and grammatical status, and the potential for words to move between grammatical classes. We have already seen this indeterminacy within the PNP sequence, with the potential for

internal variance providing an indicator as to whether a grammaticalised unit has been formed.

In the case of complex prepositions, the clearest examples of grammaticalisation are ones where the noun element is fused with a grammatical one, as in the formation of the preposition *despite*, or *instead of*. This process is an example of decategorialisation, as defined by Hopper as one of his "five principles of grammaticization" (1991: 22), whereby the traditional category Noun can lose its grammatical status both by loss of grammatical markers such as articles, and by assuming a secondary function, for example as a preposition – he gives the example of *in (the) face of* – rather than as a participant in a discourse (1991: 30). Hopper and Traugott present the reanalysis of the PNP construction into a CP as a paradigmatic example of grammaticalisation through decategorialisation (2003: 106-111), and previous studies on the grammaticalisation of CPs have also focused on this feature.

3.1 Previous studies on the grammaticalisation of PNP sequences

This theme of grammaticalisation through decategorialisation is a feature of studies on complex prepositions such as Bordet and Jamet (2010), and the small set of studies that demonstrate the process using corpus data. A common approach has been to look at how CPs have grammaticalised over time by the loss of internal variation and lexical content.

A group of CPs, with equivalent meanings – *instead of, in place of, in lieu of* – were investigated by Schwenter and Traugott (1995) using the Helsinki Corpus of the English Language, the Toronto Corpus of Old English and the online *Oxford English Dictionary* (OED). They found that each of these CPs had undergone a movement from the purely locative use indicated by the noun to express the concept of replacement, and "in the process, the semantics of location was weakened or 'bleached'" (1995: 260). *Instead of* is more fully grammaticalised than either of the other two, the Old English *stede*, "place" (as in *homestead*) having lost its nominal identity by becoming fused with the preceding preposition, and being able to be used in a wider variety of environments, for example "with a gerund as in *Sam watched TV instead/*place/*lieu of studying*" (ibid.: 246). *In lieu of*, on the other hand, appears to have developed a specialised sense of compensation, and is therefore often found in financial or legal contexts, "e.g. *hold a person in lieu of*

taxes/*bond*/*bail*" (ibid.: 258). All three, according to Schwenter and Traugott, show signs of grammaticalisation through the loss of the definite article, and movement towards abstraction of meaning: "they have become fixed phrases in which the article is no longer available in the locative NP", and "[they] have undergone semantic shifts from more concrete to more abstract meaning" (ibid.: 246).

The inclusion of the definite article in P+ NP+ P constructions as a sign of an earlier concrete use of the noun, which becomes more abstract over time, is also discussed by Akimoto: "in P + NP + P phrases, the definite article *the* was often attached to the nouns in early English and 18th century English. This seems to indicate that the phrases were used more literally than they were used now." (1999: 227)

Akimoto's corpus data is limited to literary material from the 16th to the 19th centuries. A more systematic use of corpora is demonstrated by Hoffmann (2005) who focuses mainly on a set of the 30 most frequent (and therefore almost entirely non-variable) CPs in the British National Corpus (BNC). He gives the example of *in view of*, where, in parallel to the move from a literal to a more abstract meaning of *view,*

> the nominal element of the construction over time loses the features that define its categorical status as a noun. For example, in the complex prepositional use of *in view of*, *view* cannot occur in the plural or with a determiner, nor can it be premodified by an adjective. The noun *view* has thus undergone the process of decategorialization. (2005: 56)

An alternative view on the inclusion of the definite article is presented by Petré, Davidse and Van Rompaey (2012), who use corpus data from the 56 million word Collins Wordbanks Online English corpus to argue that, in the case of the "way-noun" complex prepositions such as *on the road to, on the way to* that they are discussing, the presence of a determiner is not necessarily an indicator of a less grammaticalised state. They argue that these three constructions challenge the view of decategorialisation as merely a loss of nominal features such as determiners, as the way-noun is preceded by a determiner in the majority of cases. This calls into question the implication both from Quirk et al.'s indicators of separateness, and the diachronic studies of the movement of CPs towards fixity summarised above, that the presence of the determiner within the PNP sequence is necessarily an indicator of lack of cohesion.

In order to investigate variability within the PNP unit, particularly in relation to the determiner, this study will first look at the canonical representation of CPs within

contemporary grammars, then at corpus data to see how far this representation is reflective of usage.

4. Representation of CPs in contemporary grammars

The grammars used for this study are those already mentioned: Quirk et al. (1985), Biber et al. (1999) and Huddleston and Pullum (2002). They are chosen as being major contemporary grammars with divergent positions on the status of complex prepositions, and therefore potentially offering a range of types and forms of CPs.

Overall there are 75 different CPs listed, with agreement on quite a high number of them (42 of them being common to all three of the grammars, and, 12 common to two of them, with 21 unique items. Of these unique items, Huddleston and Pullum list the highest proportion (13), which is not surprising given their position that this is an arbitrary group of idiomatic phrases rather than an identifiable set belonging to a grammatical class. It is also noteworthy that several of their unique items stand out as examples of fossilisation, where the noun is unlikely to appear outside of the particular PNP sequence it is presented in here (e.g. *at the behest of, under the aegis of, under the auspices of*).

The types of variation offered by these lists include variation of P1 (e.g. *in/on behalf of*), P2 (e.g. *in respect of/to*), the inclusion of the indefinite article (only two instances – *with a view to* and *as a result of*), and the possible inclusion of the definite article. While it is interesting that several types of variation are included in these lists of examples, this study will focus on the status of the definite article, so the expressions that include the indefinite article are excluded from Table 4.1.

Quirk et al. and Huddleston and Pullum indicate variable inclusion of the definite article with brackets, while Biber et al. list both forms where they suggest inclusion is variable. Table 4.1 follows the practice of using brackets to show possible inclusion of the definite article. The grammars agree on the CPs that have this potential variation, where they list them, except in the case of f*or the sake of* – which is not considered variable by Biber et al., but is by the other two grammars.

Table 4.1: Set of CPs listed in grammars, with or without definite article

Invariably with def. article (12)	Possibly with def. article (7)	Invariably without def. article (54)
at the behest of *at the expense of* *at the hands of* *in the case of* *in the event of* *in the name of* *on the matter of* *on the part of* *on the strength of* *under the aegis of* *under the auspices of* *with the exception of*	*at (the) risk of* *by (the) force of* *for (the) sake of* *in (the) face of* *in (the) light of* *in (the) process of* *on (the) ground(s) of*	*at loggerheads with, at odds with, at variance with, by dint of, by means of, by reference to, by virtue of, by way of, for lack of, for want of, from want of, in accordance with, in addition to, in aid of, in back of, in behalf of, in case of, in charge of, in common with, in comparison with, in compliance with, in conformity with, in consequence of, in contact with, in contrast to, in exchange for, in favour of, in front of, in league with, in lieu of, in line with, in need of, in place of, in quest of, in reference to, in regard to, in relation to, in respect of, in respect to, in return for, in search of, in spite of, in step with, in terms of, in touch with, in view of, on account of, on behalf of, on pain of, on top of, with effect from, with reference to, with regard to, with respect to*

* *in case of* and *in the case of* are listed as separate CPs

The way the CPs are represented in the grammars suggests that the majority of canonical examples occur invariably without the definite article, but there are some that appear invariably with the definite article (although there is not complete agreement on which they are – see example of *for (the) sake of*), and some that can occur with or without it. This broad agreement on the list of CPs that are invariable without the definite article is notable given the different methodological approaches of the grammars – Biber et al. being explicitly corpus-based, whereas Huddleston and Pullum make no claim to representativeness in their choice of examples. This impression will be tested when we look at corpus material, but for the moment we will focus on examples where there appears to be variability in the inclusion of the definite article.

If we see the intrusion of the definite article as an indicator of a less grammaticalised status, then the CPs in the middle column (where the definite article is optional) appear the most likely to be in a process of change, whereas the apparent fixity of the left-hand column suggests either that the grammaticalisation process has stalled for these cases, or there is some other reason for the retention of the definite article.

5. The definite article and the semantics of the noun in CPs

One possible reason for the retention of the definite article is suggested by the one example that is listed both with and without it: *in case of, in the case of*. If we consider

examples (a) and (b) below, we can see that the inclusion or non-inclusion of the definite article can have an effect on the meaning of the CP:

(1) *in (the) case of:*

a. "*in the case of* Japan and the other East Asian economies" (= "in the example of")

b. With doctors standing by *in case of* a severe reaction (= "in the event of something happening")

This appears to be a clear example of the definite article indicating a concrete use of the noun *case* (=example), whereas the omission of the definite article precedes a less specific possibility. The distinction between the concrete and the more abstract seems to go the other way if we look at another CP where the removal of the definite article would alter the meaning, *on the part of*.

(2) *on (the) part of*

a. "Is that wrong *on the part of* pro-union supporters?"

b. "We join our hero *on part of* his journey."

Here the concrete sense goes with (b) = "section", rather than the agentive usage in (a). So in these cases, we cannot invoke the argument made in previous studies that the loss of the article necessarily accompanies an abstraction of the noun element of the PNP – and is therefore a form of grammaticalisation.

This change of meaning is also available with the introduction of the definite article to some of the CPs where it is not listed as a possible inclusion, e.g. *in front of, in back of, in charge of, on top of.*

For example:

(3) a. in back of the building (= "behind the building")

b. in the back of the building (= "in the back part of the building")

This is equivalent to the different meanings for *in front of/in the front of; on top of/on the top of*, which seem to indicate a specific use for the definite article with these topographical nouns. Huddleston and Pullum point out this change of meaning for *in front of* to

distinguish structural manipulations of a PNP sequence from changes that affect the semantics of the noun element (2002: 620).

We therefore have a range of possible reasons for the inclusion/exclusion of the definite article:

1. the CP is undergoing a process of decategorialisation towards a fixed form without the definite article
2. the use of the definite article adjusts the meaning of the phrase
3. the CP is fully grammaticalised with the definite article, and shows no signs of moving to the archetypal CP with no determiner

In order to assess which of these explanatory categories the CPs that can take the definite article in Table 4.1 belong to, it is necessary to look at corpus data to quantify current usage, and chart diachronic change towards a more or less fixed form.

6. Investigation of evidence of use of CPs in corpora

6.1 Summary of corpora used

Large corpora are necessary to capture a range of CPs, which do not tend to be high frequency items. Hoffmann made use of the BNC for his synchronic data and a combination of texts from the Gutenberg Project and the OED quotation database to provide his diachronic data. For this study, the BNC data is supplemented by the larger (and still growing) Corpus of Contemporary American English (COCA)[1], which was not available to Hoffmann, and gives the opportunity to observe regional differences as well as giving a fuller picture of contemporary usage. The Corpus of Historical American English (COHA) is another recently constructed corpus that can give us a picture of any diachronic development in individual CPs over the past 200 years. While this is not as large a time frame as that covered by Hoffmann's study, COHA does offer the advantage of being a structured corpus[2] that covers a variety of written genres, and can therefore give a more complete picture than the snapshots available in a quotation database, or the

[1] COCA (http://corpus.byu.edu/coca/) and COHA (http://corpus.byu.edu/coha/) were both compiled by Mark Davies at Brigham Young University, and the version of the BNC consulted was the BYU-BNC (http://corpus.byu.edu/bnc/)

[2] Note, the comparative coverage of different decades in terms of number of texts and range of genres is not consistent throughout COHA, presenting the possibility of skewed frequencies for the first half of the 19th century but from the 1870s onwards coverage is reasonably consistent.

purely literary texts of the Gutenberg Project. Table 4.2 lists the corpora used in this study, along with their size and time period.

Table 4.2: List of corpora used in this study

COHA	COCA	BNC
c. 400 million words	c. 425 million words	c. 100 million words
1810 - 2010	1990 - present	1980s – 1993

Although the timespan offered by COHA is inadequate to show some of the longer term grammaticalisation processes demonstrated by Hoffmann, it may be able to show more recent changes than were available in Hoffmann's data, especially as the loss of the determiner can happen very quickly (see for example *in front of*, (2005: 90-92)).

6.2 Approach to using corpora

All of the CPs listed in the grammars were checked for frequency in the BNC and COCA, both with and without the inclusion of the definite article, to establish current usage in the major varieties of English. A selection from these was then made from these where the inclusion of the definite article was of comparatively high frequency, and did not affect the meaning of the CP in the ways demonstrated above (Section 5). These selected CPs were then investigated in COHA, to look for evidence of diachronic change.

Of the 54 listed as occurring without the definite article (see Table 4.1), there were some that showed evidence of the inclusion of the definite article, but most at a very low frequency. The difference between the form with and without the definite article was statistically significant for all but *for lack of, in league with, in place of*. Examples show that *in the league with* is not functioning as a complex preposition (e.g. "They are the only team *in the league with* a worse record at home than on the road"), and that P1 in *for the lack of* is usually dictated by the preceding verb or noun (e.g. "to compensate *for the lack of* money"/"one reason *for the lack of* interest"). Therefore only *in (the) place of* will be added to those under consideration where the definite article is a possibility.

We will look first at the overall frequencies for these two CPs, as well as those for which the grammars considered the definite article likely or possible. It appears that current usage largely reflects the picture presented in the grammars, in respect to which

of them is likely to include the definite article. Table 4.3 shows the normalised frequencies for these in BNC and COCA.

The CPs are ordered by frequency of their appearance in the BNC with the definite article, and the figure for the dominant form is shown in bold for both corpora. *In the case of* and *on the part* of figure prominently and, as we have already seen, fulfil a role that differs from their equivalent CP without the article. All of those that are represented in the grammars as invariably containing the definite article clearly show that this is the dominant form in the corpus data. It is noticeable that the strongly fossilised forms (e.g. *under the auspices of, at the behest of*), while being among the lowest frequency items are also among the least variable, with almost no examples without the definite article – and therefore show no indication that they have progressed towards the more highly grammaticalised form.

Table 4.3: Frequencies* of CPs from grammars that can take *the*

	BNC (/1m words)		COCA** (/1m words)	
	with the	without the	with the	without the
in the case of	**49.03**	3.40	**20.62**	2.76
in (the) light of	**18.57**	1.30	3.27	**10.24**
on the part of	**15.95**	0.48	**13.36**	0.26
in (the) face of	**15.18**	0.64	**14.37**	0.16
at the expense of	**11.34**	0.01	**6.71**	0.01
in the event of	**10.97**	0.01	**3.14**	0.03
in (the) process of	**9.34**	0.26	**9.03**	0.08
for (the) sake of	**9.12**	0.03	**7.79**	0.05
with the exception of	**8.01**	0.02	**6.26**	0.05
in the name of	**7.60**	0	**9.54**	0.78
on (the) ground(s) of	**5.02**	3.21	**1.48**	0.48
at the hands of	**3.97**	0.01	**3.52**	0.01
under the auspices of	**2.78**	0	**1.45**	0.02
on the strength of	**2.31**	0.01	**1.20**	0.01
under the aegis of	**1.11**	0	**0.53**	0.01
at (the) risk of	0.99	**2.38**	1.08	**2.86**
at the behest of	**0.75**	0	**0.85**	0
in (the) place of	0.70	**8.04**	0.51	**3.85**
on the matter of	**0.58**	0	**0.37**	0
by (the) force of	0.31	**0.47**	**0.38**	0.37

* Frequencies normalised to account for different sizes of corpora

** Figures for COCA all from searches carried out 28/5/12

The corpus data adds to the list of CPs that are most likely to include the definite article with three of those that were represented in the grammars as less fixed (see Table 4.1) – *for the sake of*, *in the face of* and *in the process of*. The four highlighted CPs in Table 4.3 –*on (the) ground(s) of, at (the) risk of*, *in (the) place of* and *by (the) force of* – are the ones where the difference in frequency between the PNPs with and without the article was not found to be statistically significant at a level of <0.0001 in all cases, except for *at risk of*, *by force of* and *on grounds of*. The lack of a significant distinction between the form in these instances demonstrating that the unit is not yet fixed. The preference for inclusion/omission of the article is not always consistent across regional varieties. There is:

- a preference for the form without the article *at risk of* (COCA 2.86 to 1.08, BNC 2.38 to 0.99)
- a slight preference for *by force of* in the BNC (0.47 to 0.31), with a slight preference for *by the force of* in COCA (0.38 to 0.37) – although the frequencies are particularly low for this CP
- only a slight preference for *on the ground(s) of* over *on ground(s) of* in the BNC (5.02 to 3.21), although the ratio is higher in COCA (1.48 to 0.48)
- a strong preference for the form without the definite article for *in (the) place of* in the BNC (8.04 to 0.7) with the same preference slightly less marked in COCA (3.85 to 0.51)

In addition, the clear regional disparity shown in a strong preference for *in the light of* in the BNC (18.57 to 1.30), with an almost as strong preference for *in light of* in COCA (10.24 to 3.27), makes this PNP a candidate for further investigation.

7. Patterns of loss of the definite article

In order to investigate these indications of change in the forms of these CPs, we will take a similar approach to Hoffmann's – looking at current usage to see if there is an emphasis on a concrete or abstract sense of the noun within the CP, and using data from COHA to plot any diachronic tendency for decategorialisation of the noun element[3]. We can then assess whether these semantic and grammatical changes naturally accompany each other, as has been proposed both by Hoffmann (2005: 56), and Schwenter and Traugott – for one of the CPs under consideration here, *in place of* – "the absence of the definite article

[3] As we do not have an appropriate diachronic corpus for British English, the BNC data will not be used in this part of the study.

is paralleled by an extension of the construction to a different set of discourse contexts, dealing with more abstract conceptions of place" (1995: 255).

7.1 (the) place of

The definite article is still present in well over a quarter of the instances in COCA (61 out of 224) where it has an abstract ("instead of") rather than a locative sense. It is therefore not a definitive marker of the concrete sense of the noun (e.g. "sat in the place of honour"), but the form without the definite article is almost always used in the substitutive sense (e.g. "meagre excuses in place of a real response"), thus confirming Schwenter and Traugott's (1995) description of it as an example of decategorialisation going alongside abstraction of the noun element.

7.2 by (the) force of

The frequencies for *by (the) force of* in COHA are too low to show definitive patterns of decategorialisation, but there are signs of the definite article being used to distinguish concrete from abstract senses of the noun in COCA. About one third of the uses of *by the force of* in COCA (59 out of 167) refer to literal forces such as gravity, natural elements (wind, earthquakes, waves) or physical impacts or explosions:

(4) Four generations used to live under the same roof, a family tree shattered and splintered *by the force of* the sea.
(COCA 2011 SPOK PBS_NewsHour)

These types of forces do not collocate at all with the *by force of* form, although a range of collocates from the more physical *arms* to the more abstract *will* are available for forms of the CP both with and without the definite article, for example:

(5) The EU takes pride in being a civilian power that expands *by force of* example, rather than *by force of* arms. (COCA 2009 ACAD ForeignAffairs)

(6) Italy's European Affairs Minister Rocco Buttiglione said " the war may have been a mistake, " adding: "Terrorism can not be defeated only *by the force of arms*."
(COCA 2004 NEWS CSMonitor).

It may be that the definite article is losing ground overall, but maintains a function in distinguishing the physical senses of *force* from others.

7.3 on (the) ground(s) of

On (the) grounds of offers a starker contrast between the concrete and the abstract, with an example like:

(7) I have a clinic *on the grounds of* our long-term care institution.
(COCA 2011 SPOK NPR_TalkNation)

as opposed to

(8) "But to exclude a pupil *on grounds of* race or ethnicity would be illegal, and that is what the court decided had occurred." (COCA 2010 ACAD Commentary).

Example (7) refers to a location, while (8) has lost all sense of ground as a physical space, and is equivalent to the (much more frequent) CP *on the basis of*. Over half of the instances of *on the grounds of* in COCA (296 of the 582) have the locative sense, indicating that the abstraction of the noun is not particularly advanced in American English.[4]

If we consider both criteria for decategorialisation possible in this CP – loss of the definite article and the option to be inflected – the evidence in COHA paints a contradictory picture. See Figure 4.1.

On the ground of appears as the dominant form at the start of the 19th century, but has gradually been replaced, with *on grounds of* and *on the grounds of* jockeying for position as the favoured form in the middle of the 20th century. The fact that the plural form of the noun has now almost completely replaced the singular suggests a form of decategorialisation in that choice over inflection of the noun has been lost, although it is unusual that the plural form has become the fixed one in this case. However, the rise of the definite article (not only in the locative use of the CP) reinforces the status of the noun, posing a question as to what process is taking place. In fact the current preference is for the most categorically marked form possible – the inflected form of the noun together

[4] In the BNC the concrete sense is almost non-existent, with only 4 out of 388 instances that are locative

with the definite article makes. It therefore does not conform to the standard descriptions of the decategorialisation process.

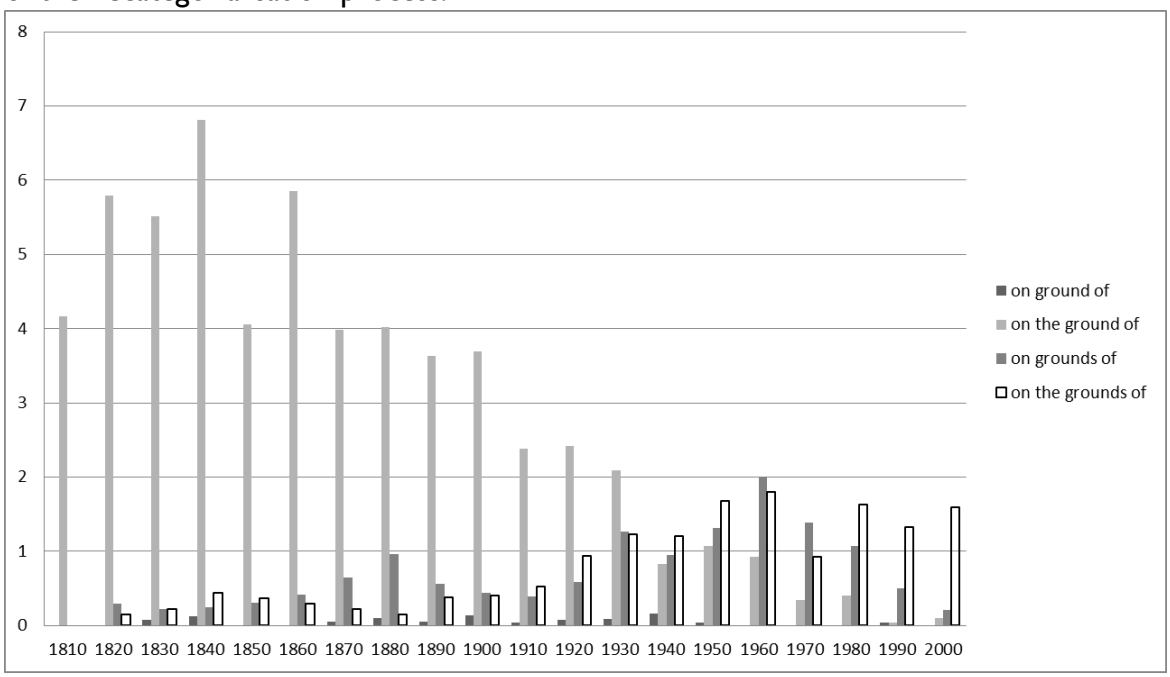

Figure 4.1: Normalised frequencies for *on (the) ground(s) of* by decade, in COHA

7.4 in (the) light of

COHA confirms the loss of the definite article for *in (the) light of* shown in the frequency data from COCA, see Figure 4.2 below:

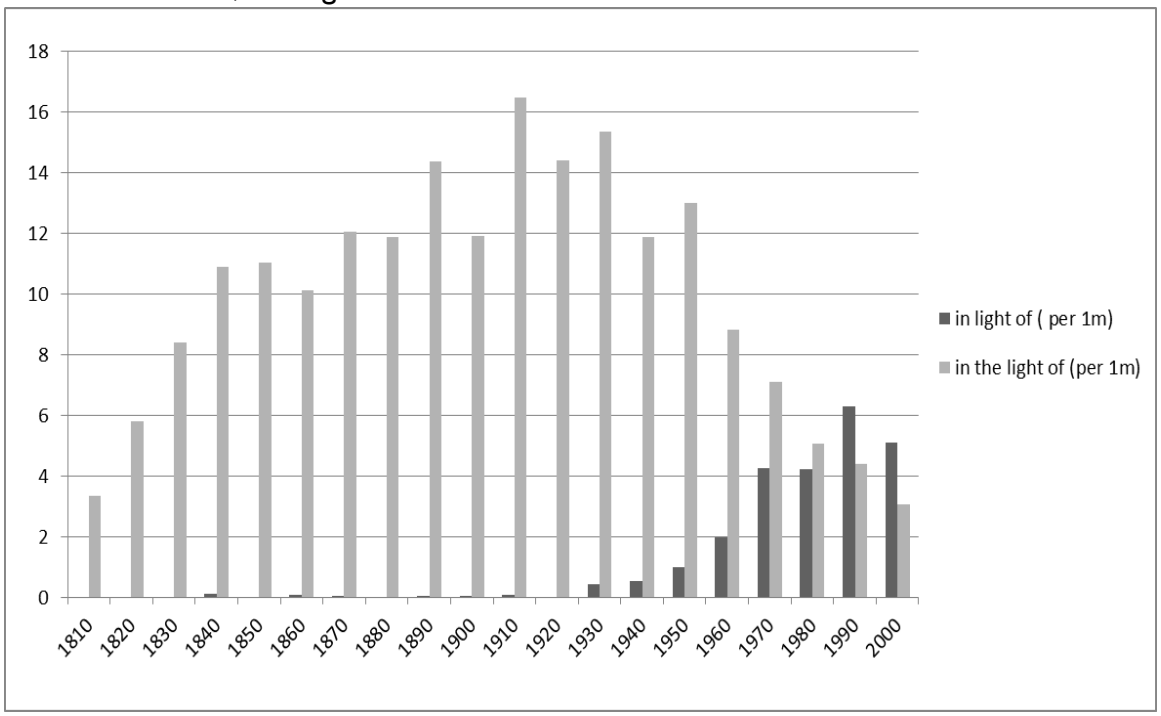

Figure 4.2: Normalised frequencies for in (the) light of by decade, in COHA

This does appear to be an example of the kind of decategorialisation process observed by Hoffmann in examples such as *in view of*, but a more recent one, for which his study did not have the data. The concurrent movement from concrete to abstract sense of the noun that accompanies the loss of the definite article can be seen with examples without the definite article. For example, the most common collocate for *in the light of* is *day*, with 75 instances (5% of overall occurrences of *in the light of*) as opposed to 1 for *in light of day* (0.02%) in COCA. It therefore appears that decategorialisation and abstraction of the noun are strongly related in the American use of this CP.

7.5 at (the) risk of

A similar pattern of loss of the definite article is shown in the COHA data for *at risk of*, see Figure 4.3:

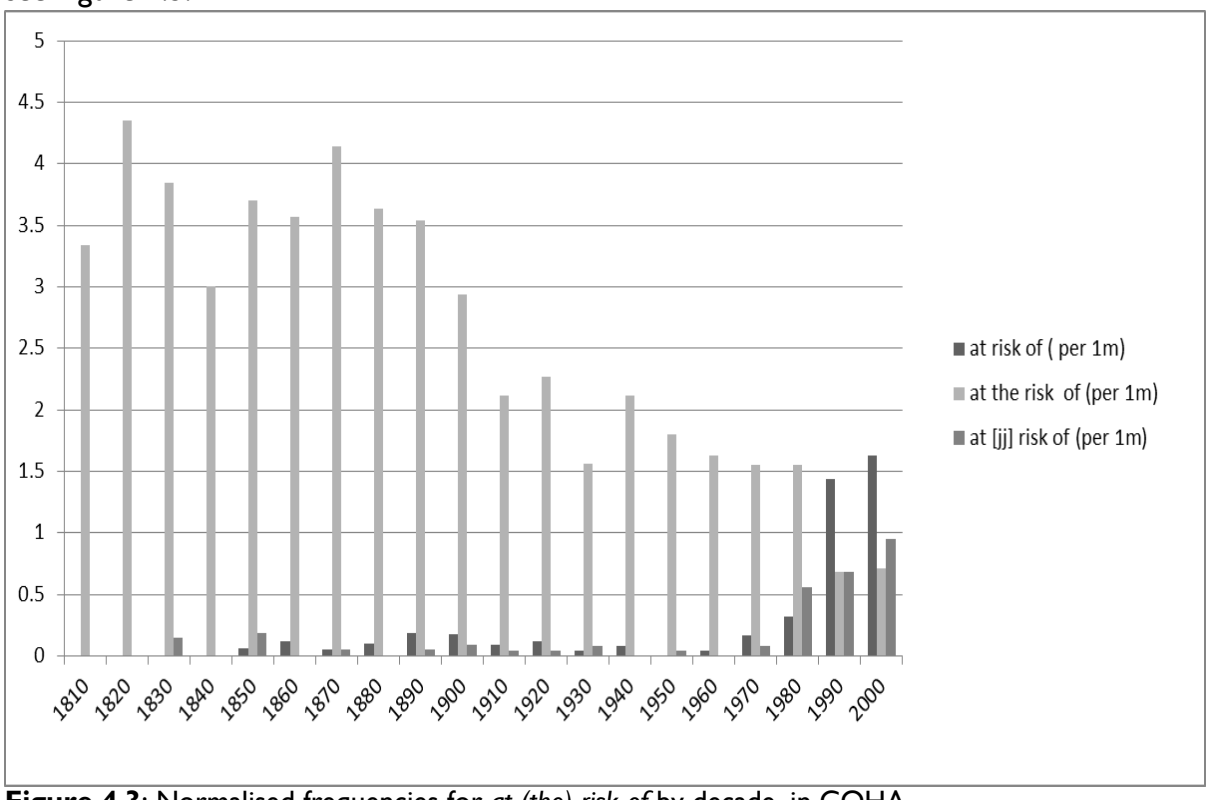

Figure 4.3: Normalised frequencies for *at (the) risk of* by decade, in COHA

While there is some evidence for the form *at risk of* as early as the 1850s, its rise and dominance over *at the risk* of appears to have only happened in the last two or three decades, as with *in light of*. This picture of possible decategorialisation is however complicated by the rise in adjectival premodification of *risk* within the CP over the same period. This is counterintuitive because, as noted above, adjectival premodification is one

of the features that defines the status of the nominal element of a CP within the category of a noun. The strong tendency for *at risk of* to contain adjectival premodification such *at increased risk of*, *at high risk of,* is confirmed in current usage with the frequency of adjectival premodification being higher than the presence of the definite article in COCA (1.24 to 1.08), as is the case in COHA.

There are also no apparent signs of the noun becoming more abstract, or of the definite article highlighting a more physical sense of *risk*, as was observed with *by the force of*. On this evidence, a*t risk of* does not appear to be a clear example of decategorialisation in the terms set out in Section 3 above.

According to the corpus data investigated, only two of the five CPs for which there is no statistically significant difference in the frequencies with and without the definite article demonstrate unambiguously a process of decategorialisation (*in (the) place of* and *in (the) light of*). We therefore need to look into the possibility of other factors being involved in the grammaticalisation process.

8. Complementation patterns

The retention of the determiner in "way-noun" CPs such as *on the road to, on the way to* is noted by Petré, Davidse and van Rompaey (2012), and they argue that this set of CPs have developed specialised aspectual uses that are more indicative of their grammatical status than categorial markers such as the definite article or possessive determiner. If we look at the complements of the 19 CPs where the definite article is a possible/likely inclusion (listed in Table 4.3), we find that the majority have noun phrase complements, but in two cases – *at (the) risk of* and *in the process of* – the most frequent type of complement is the gerund, as in:

(9) Or, *at the risk of* stating the obvious, try just eating a small snack before you run and saving lunch for when you get back. (COCA 1993 MAG MensHealth)

and

(10) Two other families from Iowa are *in the process of* adopting from the same orphanage. (COCA 2010 SPOK ABC_GMA)

This tendency towards a more verbal type of complementation could indicate a specialised grammatical function for these particular CPs.

8.1 collocates for at (the) risk of

The data for *at (the) risk of* in COCA shows that the gerundial complementation is particularly strong for the definite article – 72% of R1 collocates being gerunds, but with the figure still being quite high (46%) for *at risk of*. Table 4.4 shows the most frequent right collocates for each form, with the prevalence of the gerund clearly indicated for both, but with some interesting points of divergence.

Table 4.4: Top 20 most frequent* R1 collocates for at (the) risk of in COCA

at risk of		***at the risk of***	
BEING	130	SOUNDING	67
LOSING	64	BEING	40
BECOMING	51	LOSING	19
DEVELOPING	48	THEIR	14
STARVATION	30	OVERSIMPLIFICATION	10
HAVING	24	REPEATING	9
NOT	23	APPEARING	9
CONTRACTING	22	HIS	9
EXTINCTION	22	MAKING	8
EARLY	20	ALIENATING	7
DROPPING	19	GOING	7
FAILING	16	OVERSIMPLIFYING	6
DYING	16	OFFENDING	6
HEART	14	STATING	6
EXPOSURE	12	GETTING	6
INFECTION	11	HAVING	5
HIV	10	INCURRING	4
ARREST	9	RAISING	4
FALLING	9	YOUR	4
GETTING	9	INDULGING	3

* Raw frequencies are shown. All frequencies for COCA from searches on 28.5.12

At the risk of demonstrates a more limited range of complements than *at risk of*, with the only high frequency non-gerundial complements for *at the risk of* being the nonverbal noun oversimplification – which is also here found as the verbal noun *oversimplifying* – and the possessive determiners *their*, *his* and *your*. These complements are almost invariably

followed by *life* (*at the risk of his/her/their live/lives*) so this appears to be an example of a relatively fixed idiom rather than proof of the grammatical flexibility of the construction.

At risk of has a greater variety of complements, with a range of deverbal nouns (*starvation, extinction, exposure, infection*) as well as nouns premodified by adjectives (*early retirement, heart disease*) that are further removed from the verbal. Schwenter and Traugott argue in the case of *instead of* that it is more fully grammaticalised than its semantic equivalents *in place of, in lieu of* because it "occurs in a wider number of environments" (1995: 246), such as with the gerund, or as an adverb. On the evidence in COCA, *at the risk of* appears to be a specialised use of the CP, primarily with a gerundial complement, while *at risk of* is a more flexible and therefore more fully grammaticalised form.

8.2 in (the) process of

In the process of appears to belong uncontroversially to the set of CPs that is fixed with the definite article. We can see from Table 4.3 that the contemporary corpora favour very strongly the form with the definite article (BNC 9.34 to 0.26 and COCA 9.03 to 0.08). However COHA shows signs of development, with *in process of* being the slightly preferred form through to the start of the 20th century, while *in the process of* has clearly taken over since.

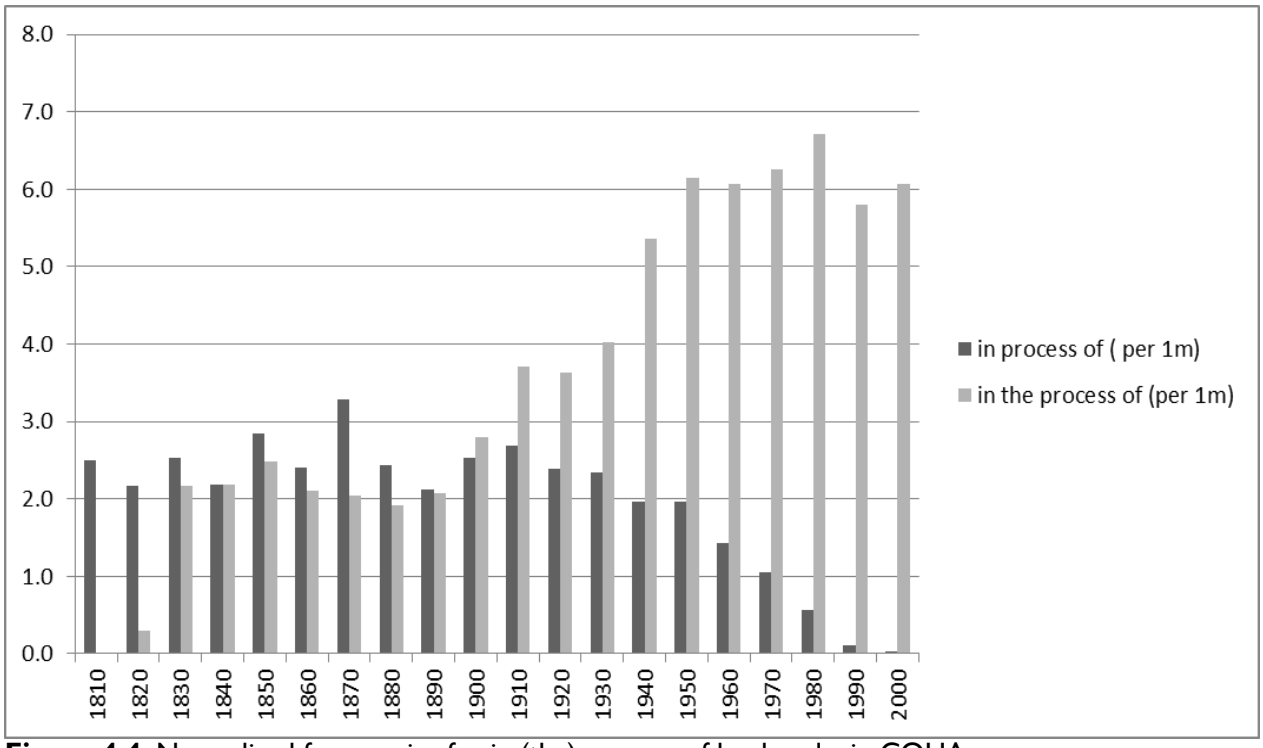

Figure 4.4: Normalised frequencies for in (the) process of by decade, in COHA

To help explain this adoption of the definite article, we can look at the range of complements for *in (the) process of* in COHA, and compare them with those showing current usage in COCA – see Table 4.5.

The frequency of *in process of* in COHA is very heavily influenced by the idiom *in process of time*, which accounts for 27% of instances in the corpus as a whole. This idiom is unattested in COCA (which only has 34 instances of *in process of*), and there has been a shift towards gerundial complementation with the CP that includes the definite article. There are only 3 gerunds in the top 20 R1 collocates for *in process of* in COHA, going up to 13/20 for *in the process of*, and then 20/20 for that form of the CP in COCA. It is therefore even more limited in its range of complements than *at the risk of*.

Table 4.5: Top 20 most frequent* R1 collocates for in (the) process of in COHA/COCA

in process of (COHA)	*Total 698*	*in the process of (COHA)*	*Total 1687*	*in the process of (COCA)*	*Total 3953*
TIME	189	BEING	121	BEING	252
CONSTRUCTION	63	BECOMING	45	DOING	110
BEING	48	MAKING	33	DEVELOPING	89
FORMATION	42	GETTING	25	TRYING	78
DEVELOPMENT	27	BUILDING	24	BECOMING	74
ERECTION	23	DEVELOPMENT	21	GETTING	72
BECOMING	11	EVOLUTION	20	MAKING	66
CHANGE	10	FORMATION	18	BUILDING	51
BUILDING	8	MANUFACTURE	17	CREATING	49
MAKING	7	CONSTRUCTION	15	CHANGING	48
NEGOTIATION	6	LEARNING	15	WRITING	46
COMPLETION	6	DOING	15	LEARNING	44
PUBLICATION	6	DEVELOPING	14	ACQUIRING	30
SOLUTION	6	PRODUCTION	14	ESTABLISHING	30
DEMOLITION	5	CHANGING	13	MOVING	30
EVOLUTION	5	CREATING	10	WORKING	30
EXECUTION	5	GROWING	10	BUYING	28
ORGANIZATION	5	BUYING	9	COMING	24
EXTINCTION	4	CONVERTING	8	ADOPTING	21
PREPARATION	4	DIGESTION	8	FORMING	21

* Raw frequencies are shown. All frequencies for COCA from searches on 28.5.12

This concentration on the gerund also emphasises the aspectual role of *in the process of*. Rather like Petré, Davidse and van Rompaey's "way-nouns", it focuses both on the progressive nature of the action, as well as its eventual goal – what they term the "telic-progressive" aspect – although the progress is not spatial but procedural in this case. It also has a strong parallel with the subordinating function of the simple preposition with the non-finite clause (e.g. "after trying") which further supports its role as a prepositional unit.

It is possible that the definite article-less form of this CP may develop with a broader range of grammatical contexts, as appears to be the case for *at risk of*, but the dominant form in current usage, *in the process of,* has fixed around a specific grammatical function.

9. CPs that invariably include the definite article

The corpus data does not show us signs of development in the other CPs where the definite article is the strongly preferred option, or present clear patterns to explain why

these CPs maintain the definite article. As mentioned above, there are the examples of *in the case of* and *on the part of* for which the loss of the article signals a change of meaning. There are several CPs in this group that form a congruent set in that the N element is a fossilised noun that rarely or never appears outside of the PNP construction: *at the behest of, under the aegis of, under the auspices of*. However, we can see from Table 4.1 that there is a similar set of CPs containing fossilised nouns that are invariable without the definite article: *at loggerheads with, by dint of, in lieu of*. If there has been any development in the form of these CPs, it is outside the date range covered by COHA, and Schwenter & Traugott (1995: 258) argue that, for *in lieu of* at least, there has been no development, having been calqued into English from the French phrase *en lieu de*, which also does not contain the article.

10. Conclusion

The implication of Quirk et al.'s (1985) scale of cohesion, and the demonstration of decategorialisation in the corpus studies of Schwenter and Traugott (1995), Akimoto (1999) and Hoffmann (2005) is that complex prepositions can exist in varied forms, but that there is a tendency towards a fixed form that excludes nominal markers such as adjectival premodification and determiners. Contemporary grammars stress this fixity in their examples of CPs, but allow for the possibility of variation, chiefly in the form of the inclusion of the definite article. The corpus evidence we have looked at confirms the fixity of some of these forms with definite articles, and that the definite article can have a specific semantic or grammatical function that differentiates its use from the form without the article.

Returning to the explanatory categories presented at the start of this study, only *in place of* and *in light of* demonstrated the conventional signs of decategorialisation, with parallel abstraction of the noun component in the American English data. In examples such as *in (the) case of, on (the) part of* and *in (the) front of*, the definite article had the role of changing the semantics of the CP. It also had a semantic role for *by (the) force of* and *on (the) grounds of*, by foregrounding more concrete senses of the noun element, although the presence of the definite article does not always correlate with concreteness over abstraction – as in the example of *at risk of*. In the majority of cases where a definite

article was included, the corpus data indicated that these were fixed forms with no sign of loss of the article.

The high incidence of gerundial complementation for *at (the) risk of* and *in the process of* was indicated a specialised grammatical role for CPs that are not conventionally classified as highly cohesive. This finding supports Petré, Davidse and van Rompaey's contention that there could be a role for certain complex prepositions as aspectual markers. With the development of new diachronic corpora such as COHA, the potential for investigating the process of grammaticalisation is expanded. We can start to look beyond the variations within structural composition, exemplified by Quirk et al.'s indicators of separateness, to the context and function of these developing units, and recognise that there can be multiple paths towards grammaticalisation.

Acknowledgements

With grateful thanks to the two reviewers of this paper, for their careful reading and constructive suggestions, and to Emeritus Professor Pam Peters for her continual inspiration and support.

Corpora

British National Corpus (BNC) <http://corpus.byu.edu/bnc/> (3 October 2012).
Corpus of Contemporary American English (COCA) <http://corpus.byu.edu/coca/> (3 October 2012).
Corpus of Historical American English (COHA) <http://corpus.byu.edu/coha/> (3 October 2012).

References

Akimoto, Minoje 1999. Collocations and idioms in Late Modern English. In *Collocational and idiomatic aspects of composite predicates in the history of English*. [Studies in Language Companion Series 47], Laurel Brinton & Minoji Akimoto (eds), 207-238. Amsterdam: John Benjamins.

Biber, Douglas, Johannson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education.

Bordet, Lucie & Jamet, Denis. 2010. Are English prepositions lexical or grammatical morphemes? *Cercles*, *Occasional Papers*: 1-26.

Denison, David. 2010. Category change in English with and without structural change. In *Gradience, Gradualness and Grammaticalization* [Typological Studies in Language 90], Elizabeth Closs Traugott & Graeme Trousdale (eds), 105-128. Amsterdam: John Benjamins.

Hoffmann, Sebastian. 2005. *Grammaticalization and English Complex Prepositions: A Corpus-Based Study*. *Approaches to Grammaticalization*, Vol I [Typological studies in language 19(1)]. London: Routledge.

Hopper, Paul J. 1991. On some principles of grammaticalization. In Traugott & Heine (eds), 17-35. Amsterdam: John Benjamins.

Hopper, Paul J. & Elizabeth C. Traugott. 2003. *Grammaticalization*, 2nd edn. Cambridge: CUP.

Huddleston, Rodney & Pullum, Geoffrey. 2002. *The Cambridge Grammar of English Usage*. Cambridge: CUP.

Petré, Peter, Davidse, Kristin & Van Rompaey, Tinne. 2012. On ways of being on the way: lexical, complex preposition and aspect marker uses. *International Journal of Corpus Linguistics* 17(2): 229-258.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik. Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Schwenter, Scott A. & Traugott, Elisabeth Closs. 1995. The semantic and pragmatic development of substitutive complex prepositions in English. In *Historical Pragmatics*, Andreas Jucker (ed.), 243-273. Amsterdam: John Benjamins.

Sweet, Henry. 1891. *A new English grammar – Logical and Historical*. Approaches to Grammaticalization, Vol I [Typological studies in language 19(1)], Oxford: OUP.

Traugott, Elizabeth Closs & Heine, Bernd. 1991. Introduction. In *Approaches to grammaticalization, Vol 1,* Elisabeth Closs Traugott & Bernd Heine (eds), 1-14. Amsterdam: John Benjamins.

Section 4A: Complex prepositions (CPs) and grammaticalization

CPs consisting of the form P + NP + P are uncontroversial as commonly occurring sequences. Their status as lexical or grammatical units is however open to question, and hence whether they are classed as a type of complex preposition, or simply a prepositional phrase to which an additional P (+NP) is attached.

1. Issues of variation, grammaticalisation and lexicalisation

Amid their individual patterns of variation, the chief factor confirming the grammatical status of CPs was found to be fixity, with low frequency items such *in lieu of* and high frequency items such as *in terms of* demonstrating equally low degrees of possible internal variation. Paper 3 (2013) gave evidence that even CPs that retain a determiner as a categorial marker of the N element of the CP (such as *in the process of*) have become almost entirely fixed in this form, and therefore decategorialization must be questioned as a crucial marker for grammaticalization in this instance. Bleaching of the core, concrete sense of the noun within CPs gives a more generalised or abstract sense within the construction (as for *in spite of*), suggesting the lexicalization of the unit.

However, there is no indication of abstraction in the set of very frequent locational CPs, such as *in front of*, where the core sense of the noun is central in indicating its prepositional function. Nor is there any suggestion that the elements of the CPs are undergoing reanalysis, with the N element serving to give a more specific function to the very general prepositions used (often *in* + *of*). Historically, there is some evidence of the formation of CPs by analogy (Hoffmann, 2005), but it does not appear to be a class that is generating emerging members. Only *in terms of* can be said to be a recent addition, so the class is not currently productive.

2. Grammars

CPs did not suffer from the same problems of identification through nomenclature as did the items covered in Papers 1 and 2. The grammars differ only on whether it was correct to identify them as grammatical units with a prepositional function (Quirk et al., 1985 and Biber et al., 1999), or as lexical idioms with a PNP construction (Huddleston & Pullum, 2002). There is broad agreement on identification of typical members of the set across the

grammars, and Quirk et al. attempted to provide a scale from free combinations of the PNP sequence to core members, based on their "indicators of separateness" (1985: 671-2). This scale unfortunately problematizes the categorisation of the class as a whole by including examples like *on the shelf by*, that are clearly not functioning as a prepositional unit.

3. Dictionaries

CPs are invariably treated as lexical phrases, as runons to the entry of the key noun and with no acknowledgement of their grammatical role within the L1 and L2 dictionaries examined. In some examples, the bleaching of the N element creates a problem of definition, as is the case with *in terms of*, which is often given a rather convoluted and literal definition, as in the *New Zealand Oxford Dictionary*'s (2005) "In the language peculiar to; using as a basis of expression or thought". More abstract uses are acknowledged in the L2 dictionaries, and their prepositional function is acknowledged by the use of other CPs to define them, for example *in relation to*, or even the simple preposition *as* in the *Macquarie Learners Dictionary* (1999). The same dictionary also calls into question the integrity of the unit with the sub-entry 'think (or talk) in terms of', where the CP is absorbed into the verb phrase.

4. Summary

As we have observed, the CP is a disputed class, mainly because of the internal variation it can allow. The final published paper in this thesis will investigate a set of multiword units that have not been previously identified as examples of grammaticalization – temporal phrases like *(at) the moment (that/when)*, that have the potential to function as complex subordinators. Paper 4 will look particularly at the context they are found in across the corpora interrogated, with their position within the clause and their loss of a preceding preposition of following relativizer creating a context in which their grammatical function may be reanalyzed.

Chapter 5

Newly emerging subordinators in spoken/written English (Paper 4)

Smith, A. 2014. Newly emerging subordinators in spoken/written English. *Australian Journal of Linguistics* 34(1), 118-138.

"Most things, which are necessary to acompany the doing of a good Action, lose their Advantage, and Grace, the moment that Action is over."

Pliny the Younger *(Epist. & Panegyrick ,* trans. Henley et al, 1724*)*

Newly emerging subordinators in spoken/written English

Adam Smith

Macquarie University

Contemporary grammars such as Biber et al. and Quirk et al. acknowledge the class of complex subordinators such as *as far as, in case, in order (that).* However, there is no consensus on membership of this class, and temporal adverbial expressions such as *(at) the moment (that)* can be seen as borderline cases. This paper will argue that the emergence of the fully elliptical form – *the moment* – with zero preposition and zero *that* in an ambiguous context allows the reanalysis of an adverbial adjunct introducing a relative clause as a subordinator. Corpus data from the Australian, British and US ICE corpora is used to demonstrate which of a set of these temporal adverbials are most likely to be emerging in this subordinator role. The frequency and range of different types of subordinator in the spoken corpora are compared with written genres such as fiction where narrative (and therefore temporal subordination) is a feature. These findings are tested against similar genres in the larger BNC (British English) and COCA (American English) corpora. Written English, and fiction in particular, was found to be more productive of these new subordinators than spoken English.

Keywords: *complex subordinator, corpus linguistics, intersective/subsective gradience, grammaticalization, spoken/written genres, variation*

1. Introduction

At first sight, the word class of subordinator, or subordinating conjunction, does not look like a good candidate for emerging forms. Function words such as auxiliaries, conjunctions and prepositions are conventionally described as belonging to closed, unproductive systems, unlike the open classes of lexical words such as nouns, verbs and adjectives, see for example Biber et al. (1999: 56). In addition, there is no consensus on how to define the class. Biber et al. describe subordinators as "words which introduce (mainly finite)

dependent clauses" (1999: 85). The membership of the class is more severely restricted to those that can only take finite complement clauses by Huddleston and Pullum – giving just *that*, *whether and* nonconditional *if* (2002: 1011-16). They classify most traditional simple subordinators (e.g. *like*, *while*) as prepositions, because they can be used with both nonfinite and finite clauses. This ambiguity of function and classification is acknowledged by Biber et al., among others, who give the examples of *since, before* and *after* functioning either as prepositions or subordinators (1999: 77). Examples 1a-d below demonstrate the varying uses of *since*:

1a. I haven't seen him since the Adelaide meeting	PREPOSITION
1b. I haven't seen him since the meeting in Adelaide	
1c. I haven't seen him since meeting in Adelaide	
1d. I haven't seen him since we met in Adelaide	SUBORDINATOR

Sentences 1a and 1b are both prepositional uses of *since*, introducing a noun phrase, while 1c and 1d show *since* introducing a subordinate clause with a nonfinite (*meeting*) or finite (*met*) verb. The grammatical type of the complement (nominal or verbal) determines the grammatical function of *since*, which we identify with the word class of preposition or subordinator respectively.

Yet the *-ing* form is itself ambiguous as a verbal noun (gerund) or present participle of the verb, depending on the context. In instances where it can be interpreted as either nominal or verbal, as in *She did her exercises before dancing* it affects the grammatical status of *before* (as preposition or subordinator). If we interpret *dancing* as a verbal noun (e.g. *before the dancing lesson*) *before* is a preposition, but seen as a present participle (e.g. *before performing a dance*) *before* is a subordinator. The ambivalent grammatical function of *before* in this context is a good example of the kind of the indeterminacy where authors such as Diewald (2002) have argued that new grammatical functions can emerge.

In addition to the indeterminate grammatical use (or 'underdetermined' (Denison 2010)) of simple function words like *before*, there are also clear examples where verb phrases have been grammaticalized to become used as complex subordinators, as with *providing (that)/provided (that)* moving from being nonfinite verbs to a conjunctive role. Some grammarians allow that noun phrases such as *the moment* can also effectively be used as complex subordinators in modern English. This opens up the possibility for other

new members of this supposedly closed class. This paper will investigate the process for the emergence of new subordinators, focussing on one particular type of complex subordinator.

2. Questions of categorization

2.1 Gradience between grammatical categories

As remarked above, grammarians are not in complete agreement over a canonical set of simple subordinators, partly because the words concerned can also take on the functions of other word classes. In addition to the subordinator/preposition ambiguity, there is also overlap between subordinators and adverbials (see Biber et al. 1999: 86-87[1]). The distinction between subordinator and coordinator is blurred in some cases. Quirk et al. describe a gradience from the central coordinators *and, or* to the subordinators *if, because*, with conjuncts like*, however* and *so* demonstrating the properties of both coordinators and subordinators to greater and lesser extents. (1985: 927-928).

Such a gradience between categories is not generally allowed for by Huddleston and Pullum (2002), who in borderline cases 'sever the categorical Gordian knot in an Aristotelian fashion by deciding that the elements in question must belong to one or another category' (Aarts 2007: 67). Aarts terms the kind of gradience he believes to occur between grammatical categories 'intersective gradience', while that which can occur within categories he calls 'subsective gradience' – some words are more central or core to a word class than others which can be peripheral.

In the case of complex subordinators, we can observe the effects of both of these types of gradience, with different categorization of the potential sets acknowledged.

2.2 Complex subordinators

Complex subordinators, as construed by Quirk et al. (1985) and Biber et al. (1999) are multiword units that function in the same way as simple subordinators, i.e.

1. to link a main and subordinate clause
2. to mark a clause as subordinate to an adjacent higher level clause.

[1] Where the adverbial use of *since* is also demonstrated: "She had not heard one word from him since".

Quirk et al. note that 'The distinction in form between the simple and complex subordinators is in part orthographic, since some of the simple subordinators are internally (that is morphologically) complex' (1985: 998). The existence of these morphologically complex subordinators such as *although, whereas, whenever* are an indication of the tendency towards grammatical unity of multiword units. The complex subordinators listed by Quirk et al. (1985) and Biber et al. (1999) are summarized in Table 5.1 below, with those that are common to both grammars listed in the centre, while those unique to each are listed on either side.

This list reveals a good deal of common ground between these two authorities (as shown in the centre column), which indicates that there is a set of complex subordinators that are core to the class. By contrast, most of those that are unique to Quirk et al. are given labels that show a limited context of use, e.g. *insofar that* 'formal, rare', *save that* 'literary', *forasmuch as* 'formal'), and are therefore peripheral to the class in terms of subsective gradience. Others that appear only in Quirk et al. are expanding a set that are common to both grammars, e.g. *assuming, considering, excepting (that)* and supplement the examples *granting, providing, seeing, supposing (that)*, which are also in Biber et al., and are therefore not unique as types. The remainder are marked in bold in Table 5.1, and several of these belong to a set that Quirk et al. describe as marginal, on the basis that there is not always a clear distinction between what is a grammatical unit, and what is a free expression: 'As with complex prepositions, it is difficult to distinguish categorically between complex subordinators and free syntactic constructions' (1985: 1001).

Table 5.1: Examples of complex subordinators

	Quirk et al. only	Common	Biber et al. only
Always formed with final *that*	*insofar that, in the event that, save that*	*but that, in that, in order that, such that*	
Formed with optional final *that*	*assuming, considering, excepting, except,* ***for all***	*given, granted/-ing, provided/-ing, seeing, supposing, now, so*	***on condition, the moment***[a]
Always formed with final *as*	*forasmuch as, inasmuch as, insofar as, insomuch as*	*according as, as far as, as long as, as soon as*	
Others		*as if, as though, in case*	***even if, even though, no matter (+ wh-* word*)***

[a] Not included in this list are the adverbs *directly* and *immediately* followed by an optional *that*, on the basis that there was no corpus evidence for them as complex subordinators in my data. While there was some evidence for the adverb on its own followed by a finite verb, meaning "as soon as", there were no examples of this subordinator use when followed by *that*.

These marginal complex subordinators are categorized by Quirk et al. (1985) in 4 types, including several that feature in Table 5.1:

- a subordinator with a preceding or following adverb like *if only, just as, even if*
- participle forms such as *supposing (that), provided (that)*
- prepositional phrases ending in *the fact that,* such as *due to the fact that*
- noun phrases that commonly function as temporal adverbials, such as *the moment (that)*

These examples do not constitute a clearly definable set, with diverse reasons given for their comparative marginality or centrality to the class of complex subordinators. For example, *even if* is considered a subordinator in its own right because this complex form modifies the meaning of the simple subordinator *if*, whereas *due to the fact that* is classed as more marginal because it is considered a 'stylistically clumsy' equivalent to simple subordinator (*because*). The temporal adverbial type is considered particularly marginal because of the degree of internal variation that it allows. The choice of temporal noun can be varied (*instant*, *minute* and *time* are presented as alternatives to *moment*), the noun element can be premodified by an adjective, and the noun phrase can be preceded by a preposition and/or followed by a relativizer. Below in its most elaborated form, is the structure of these temporal adverbials realised through prepositional phrases followed by a restrictive relative clause:

(At/from) the (precise) moment (that/when) I saw him…

Its syntactic analysis is much less certain when we have the least elaborated version indicated by Quirk et al.:

The moment I saw him…

Here the omission of the preceding preposition and the following relativizer creates an ambiguous context where the adverbial adjunct introducing a relative clause could be reclassified as a subordinator, and is therefore an example of category change or intersective gradience.

2.3 From adverbial adjunct noun phrase to subordinator.

Within the possibilities demonstrated by Quirk et al.'s example above, only the determiner *the* is a nonvariable element. The variation includes the possible use of a

preposition/zero preposition at the start of the phrase, the option of adjectival premodification of the nominal element of the phrase, which itself can vary between semantically similar nouns, and the use/nonuse of a relativizer at the end of the phrase all presented as optional variants. All this potential variation leads Quirk et al. to describe this type of marginal complex subordinators unequivocally as 'more like free syntactic constructions than like complex subordinators' (1985: 1002).

Similar criteria are invoked by Kortmann (1997) in his study of adverbial subordinators across European languages. He identifies a set of 'ideal' adverbial subordinators (71-75) from which the string P + NP + COMP/REL *in the event that* is excluded because of the inclusion of the definite article, possible modifiability of the head noun, inclusion of the complementizer; whereas *on condition (that)*, *in case* are included because the definite article has been lost, the head noun cannot be modified and the complementizer is optional.

The presence or lack of any of these elements is not necessarily essential in justifying the categorization of a string as a free expression or a fixed grammatical unit. For example, if we look at the class of complex prepositions, the historical loss of the definite article can demonstrate a process of decategorialization of the nominal element leading to a grammaticalized form (*in the spite of* becoming *in spite of*). However, as Smith (2013) has argued, there are also examples of grammaticalized forms that retain the definite article, for example *at the behest of*. In the case of complex subordinators, the point at which the distinction between complex subordinator and free expression can be made is not clear cut. While Biber et al. include *the moment (that)*, as one of their set of example complex subordinators, Quirk et al. classify it as only marginal to the class. This disparity merits further investigation. The layering of more and less grammaticalized examples is a recognised phenomenon in emergent grammar (Hopper 1991, Denison 2010).

2.4 Temporal adverbials as subordinators

The examples of complex subordinators given in Table 5.1 above, demonstrate the range of grammatical classes that can be sources for subordinators. This supports the variability that Hopper and Traugott acknowledge: 'Clause linkage markers have their sources in nouns, verbs, adverbs, pronouns, case morphemes (including prepositions and

postpositions), derivational prefixes, and in phrasal combinations of these' (2003: 184). They give the example of a noun phrase as an adverbial of time that has become grammaticalized to become a simple subordinator. W*hile* originated in an adverbial phrase translatable as 'at the time that' which consisted of the noun – *hwile*, meaning 'time' followed by the invariant subordinator *þe* (= 'that'):

2. & wicode Þær Þah wile Þe man Þa burg worhte & getimbrode.
   And lived there that time that one that fortress worked on and built
   'And camped there at the time that/while the fortress was worked on and built'
   (Chron A [Plummer] 913.3)

This phrasal expression with the loss of the preceding definite article and following complementizer became reduced to the simple conjunction *wile*. Over the course of time it has added to its temporal meaning a concessive sense, as in contemporary English 'while you like peaches, I like nectarines' (2003: 90-91). Biber et al. found that this concessive use of *while* was a feature of academic writing, and that it was much more commonly used as an expression of time in conversation and fiction, with the spread between the two semantic categories being more even in newspapers (1999: 849-850). The development of specialized uses of a subordinator within particular text genres is a feature that will be looked at later (in Section 4).

In the example of *while* there is no evidence of a preceding preposition, as in Quirk et al.'s presentation of variables, but the loss of *that*, reclassification of the nominal element as a subordinator (= decategorialization), and subsequent shift from a concrete reference (a particular time) to an abstract one in the form of concessive (= subjectification) show how an expression with many of the attributes of the marginal complex subordinators such as *the moment (that)* has become fully grammaticalized.

A precedent like this allows for the possibility of similar development with other subordinators from temporal noun phrases. Biber et al. are not the only authors to note the use of noun phrases such as *the moment (that)* as temporal conjunctions. Other researchers who acknowledge their use are Edgren, who lists *the day, the time, the minute*,

the moment both with and without a following relativizer and preceding preposition (1971: 28-29). Häcker finds *the minute (that)* as a temporal subordinator in Scots English[2], using corpus evidence (1999: 95), but marks *the minute (that)* and *the moment (that)* as borderline cases whose function as a subordinator is uncertain (1999: 211), in Present-day Standard English – presumably influenced by Quirk et al.'s analysis.

It is clear that none of the temporal noun phrases noted above have developed to the extent that *while* has, to become a simple subordinator, and it is unlikely that we will find any examples of them shifting semantic domain (time, condition, reason, place – see Quirk et al. 1985: 774-775). However, the degree to which they allow internal variation will shed some light on what extent they belong to the marginal subordinator type, described by Quirk et al., or more properly belong to a set of complex subordinators which are closer to the core of the word class, as indicated by Biber et al.

3. Contexts of category change

The potential variability within these phrases indicated by Quirk et al. is an argument against their fixity as grammatical units. A more grammatically based argument would be that the presence of a preceding preposition or following relativizer determines the grammatical function to a considerable extent, whereas the absence of these elements presents an ambiguous, or in Denison's terms 'neutralized' (2010: 110) context that presents the potential for grammatical reanalysis.

We shall be looking at various corpora to determine the possible emergence of a set of temporal adverbial phrases as subordinators. It involves intersective gradience of which there is little theoretical discussion, as Huddleston acknowledged in (1976: 101) 'the transformational literature contains comparatively little work on what are traditionally known as adverbial clauses', and the situation has not changed. Fortunately, one of the corpora we will use, ICE-GB 'one of the most carefully and consistently tagged of corpora' according to Denison (2010: 120), provides grammatical analyses of each of its text units.

---

[2] Kortmann notes an influence from Irish on Hiberno-English for "adverbial subordinators" such as *the time (that)* (1997: 65), and that a large number of phrasal subordinators came into Middle English as borrowings, particularly from Old French, e.g. *for as much as*<OFr *pourtantque* (1997: 299-300) and therefore did not undergo a process of grammaticalization in English.

We will make use of some examples of these to demonstrate the ambiguity that arises in real language data.

The following examples focus on the analysis of the strings that we are concerned with, such as *the time, the day, the instant*, also showing the full text unit that provides the linguistic context. The four examples are ordered (a) to (d) in terms of the relative indeterminacy they provide for the temporal noun phrases with which we are concerned. The first two examples are from spoken ICE-GB data, the third and fourth from written data. The generic context is noted in each case.

3.1 Preposition + *that*

(3a) Full text unit*: But that seems to have been the last moment at which it was common in which it was indeed allowed because* ***by the time that*** *we get into the mid seventh century* (S2A-060)

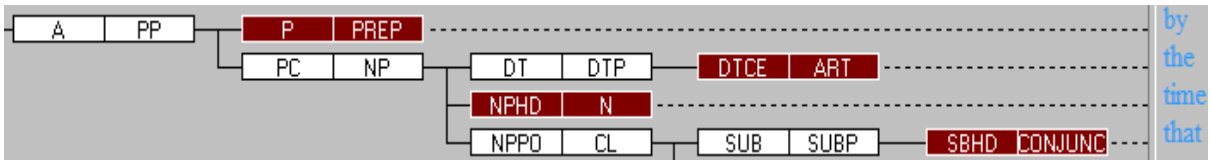

Figure 5.1: Grammatical analysis from ICE-GB of an example of preposition + *that*[3]

This example is from the spoken section of ICE-GB (in the "Unscripted Demonstrations" text category). Figure 5.1 shows the syntactic tree generated by the grammatical parsing of the relevant string. Characteristically of such commentary on the run, the matrix clause that we would expect to be introduced by the adverbial phrase is suspended (see the footnote[4] for full context). In example (3a) by *the time* is analysed as a prepositional phrase used as adverbial adjunct, prefaced by the subordinator *because*, and followed by another subordinating conjunction *that*. Its potential status as a complex subordinator is therefore undermined both by the preposition emphasizing its adverbial function in the suspended matrix clause, and the presence of *that* to supply the subordinating role.

[3] See Appendix 5.1 for a full list of the grammatical abbreviations used in ICE-GB

[4] Full context: "but that seems to have been the last moment at which it was common in which it was indeed allowed because by the time that we get into the mid seventh century and uhm we're looking here at a coin of Justinian the Second so from the six eighties uhm we find quite a major shift from symbolic representations to absolutely accurate physical representations"

3.2 Zero Preposition + *that*

(3b) Full text unit: *All* ***the time that*** *you're there, you're having to build up the muscles that you have got working for you because to push a wheelchair you need muscle power* (S1A-003)

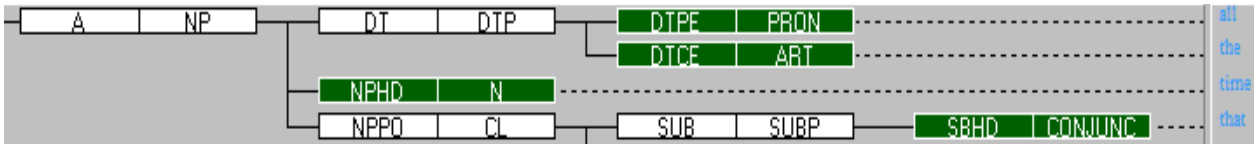

Figure 5.2: Grammatical analysis from ICE-GB of an example of zero preposition + *that*

This example comes from the "Private Conversations" text category of ICE-GB. Again the temporal string is part of an adverbial adjunct, this time at the start of a clause introduced by *all* as a predetermining pronoun (DTPE PRON) (see Figure 5.2). The adverbial adjunct noun phrase is clearly separated from the following subordinate clause by the presence of *that*, which is analysed as a conjunction at the head of a subordinate phrase (SBHD CONJUNC). There is no potential for *the time that* to be read as a unified conjunctive phrase.

Both of the first two examples include the relativizer/complementizer *that* functioning as a subordinating conjunction. The following examples have zero *that*, and therefore no clear marker of subordination.

3.3 Preposition + Zero *that*

(3c) Full text unit*: In the family photograph album there is a photo which Cathy's father took* ***on the day*** *she left home to go to university*. (W2F-019)

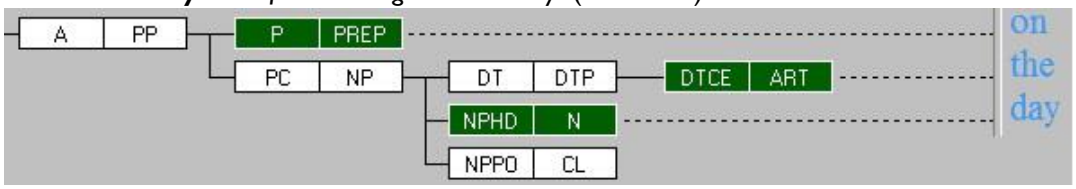

Figure 5.3: Grammatical analysis from ICE-GB of an example of preposition + zero *that*

In this example from fiction writing, the preposition *on* marks the start of an adverbial prepositional phrase (A PP) with *the day* as the prepositional complement (PC) (see Figure 5.3). Again the adverbial status of the phrase in connection with the matrix clause is marked by the presence of the preposition. But in the absence of *that* there is nothing to mark the status of the following subordinate clause.

3.4 Zero Preposition + Zero *that*

(3d) Full text unit: *When a suitable victim tries to answer the incoming call, the ringer will stop* ***the instant*** *the phone is picked up.* (W2B-032)

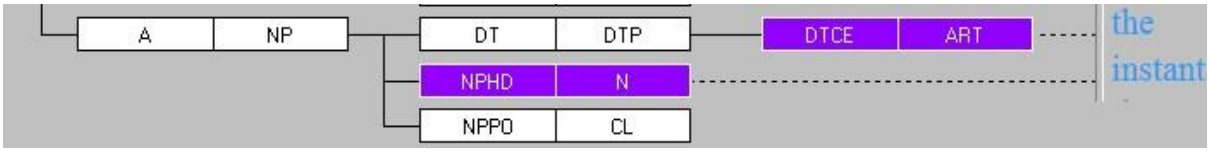

Figure 5.4: Grammatical analysis from ICE-GB of an example of zero preposition + zero *that*

In this fully elliptical example from the 'non-academic' written category, we note that the second adverbial adjunct is realized by a noun phrase (A > NP) (see Figure 5.4). Without the preceding preposition, its connection with the matrix clause is unmarked, and its semantic connection of temporality with the following clause is exactly like that of *when* which heads the subordinate clause at the start of the same sentence. Syntactically, *instant* serves as the head of the noun phrase (NPHD), introducing a postmodifying clause (NPPO CL). The fact that *the instant* is not clause-initial (as in some examples discussed below, Section 5.2.1) contributes to the ambiguity of its function, on the boundary between main clause and the embedded relative clause.

The zero preposition/zero *that* construction (example (3d)) therefore clearly offers the most potential for ambiguity of analysis, and therefore possible reclassification of the temporal noun phrase from adjunct to subordinator. In this context, its syntactic affiliation is at its most equivocal (Denison, 2010), allowing structural reanalysis in which the new complex subordinator phrase can emerge. In the other three contexts the presence of *that* after the noun phrase or of a preceding preposition (or both) prevent the analysis of the temporal phrase as a subordinator.

The subordinative role of *the instant* in example (3d) is made additionally ambiguous by its position within the sentence. As noted above (Section 2.2) the clause being initiated by a subordinating conjunction is one of the formal indicators of subordination. The example above has *the instant* as subordinate-clause-initial but embedded within the syntax. Example (3d) has *the instant* as subordinate-clause-initial but embedded within the syntax. There is no overt marker that it is initiating a clause, and therefore its role as a subordinator can only be said to be emergent. Peters argues that for the example "Allan's dad stopped to talk to the cop, the way he always did", separation of *the way* from the main clause, by means of punctuation, is a clear indicator of

its independent grammatical role as a subordinator (2012: 136). Even stronger evidence of subordinate-clause-initial use occurs when the adverbial phrase prefaces the second of two coordinated clauses joined by *and*/*but*/*or*, as in:

> (4) ICE-AUS, S1A-019(B): 76). And it was really good except one of her students one of my friends got into the Conservatorium and **the moment** that happened she took it all seriously'

Or most conclusively, as sentence initial: '**The moment** I heard your snooty twang on the phone I knew it was love.' (see example (11) below). In these cases, its role as a subordinator is unmistakable, and the indicators listed above will be used in querying the corpus data as indicators of fully emerged subordinator use of the temporal adverbial phrases under investigation.

4. Subordination in spoken and written genres

The syntactic complexity across these spoken and written examples, with their multiple levels of embedded subordination, opens up the question of genre having an influence on the use of subordination. We have already noted (Section 2.3) Biber et al.'s (1999) observation that different uses of the subordinator *while* are typical of different text genres. The temporal use found most commonly in conversation and fiction could be accounted for by narrative sequencing typical of these genres, whereas the strong evidence for concessive use of *while* in Academic writing is indicative of the genre's greater reliance on argumentation.

A feature of subordination (generally) is that it is not evenly distributed between spoken and written discourse, and it has therefore been used as an indicator of the comparative syntactical complexity of spoken and written language. Previous studies have produced contradictory evidence on this front, with some linguists finding that written texts are syntactically more complex than spoken in that they exhibit a greater number of subordinate structures (e.g. O'Donnell, 1974), while spoken texts are syntactically simpler as they rely more on coordinate structures (e.g. Kroll, 1977, Chafe, 1982). Conversely, Halliday argues that in spoken English 'the sentence structure is highly complex, reaching

degrees of complexity that are rarely attained in writing (1994: xxiv), and that subordination is associated with the real-time production constraints characteristic of speech, where there is less opportunity than in writing to elaborate through precise lexical choice (see Biber, 1988: 170). Beaman (1984) had earlier challenged the assumption that subordination implies complexity, and finds textual evidence to show that it does not necessarily correlate with written language, or coordination with spoken. She notes from her data that, for example, classifying *and* purely as a coordinator, as other studies have done, does not necessarily reflect its function in the text: 'A large percentage of the *and's* in the spoken narratives function analogously to subordinate time adverbials or time adjuncts' (1984: 76). Moreover, subordination cannot be treated as a functionally unified construct, with different types of subordination represented to varying degrees in different mediums: 'as opposed to the greater frequency of time adverbials in the written narratives, the spoken stories show a greater percentage of both condition and reason adverbial subordinate clauses' (1984: 77).

If Beaman's data based on spoken and written stories is representative, then the emerging temporal subordinators under investigation here may be more likely to be discovered in written than spoken sources. On the other hand the use of zero relativizers is strongly associated with both fiction and conversation (Biber et al. 1999). They comment further: '*That* as complementizer or relative pronoun is relatively infrequent in conversation because the zero alternative is so often selected for these functions in that register, but not so much in the written ones' (Biber et al. 1999: 351). So the elliptical forms of these subordinators (with zero preposition/zero relativizer), which we have argued to be the strongest indicators of full emergence as subordinators, may be more frequent in spoken data, as found by Häcker (1999) and Peters (2012). The generic dimension suggests a need to look at both written and spoken data in researching roles of temporal subordinators.

5. Investigation of data

This study will focus on the set of possible complex temporal subordinators identified by Quirk et al.'s list of variable marginal complex subordinators: *the instant, the minute, the moment, the time* (see Section 2.2 above), along with *the day* – identified by Edgren as

another possible member of this set (see Section 2.4). Various corpora will be used to identify both the amount of variability within each collocation – testing Quirk et al.'s argument that they are closer to free expressions than grammatical units – and assessing the extent to which the elliptical (zero preposition/zero *that*) form is dominant, suggesting its emergence as a fully realized subordinator. In addition, other indicators of their status as a subordinator will be taken into account, notably, their ability to appear at the start of the clause, which is one of the main formal indicators of subordination, according to Quirk et al. (1985: 997).

5.1 Selection of corpora

In order to assess the form and function of the complex subordinators under investigation, a range of corpora were selected. Firstly, the Australian, British and US components of the ICE corpora (ICE-AUS, ICE-GB, ICE-USA[5]) which provide recent texts (around 20 years old) over a range of spoken and written genres in three major varieties of English. The range of subcategories within each genre gives coverage of a range of styles from the informal (conversations, personal letters) to the much more formal (public speeches, court proceedings, academic prose). ICE-AUS and ICE-GB can be directly compared as they have been designed to cover the same range of texts, and have the same ratio of spoken to written material (0.6m : 0.4m). Any divergences may point to a development in one variety or the other, and the hypothesis that, for example, *the minute*, might belong to a regional variant of British English (see Häcker, Section 2.4), can be tested.

The other major variety of English, American English, is represented in the ICE-USA written corpus (c.400,000 words), supplemented by the Santa Barbara corpus (SBC) of conversational English (c. 249,000 words, 1988-96). While the written component of the corpus mirrors the other ICE corpora in size and range of texts, the SBC is smaller than the spoken components of ICE-AUS and ICE-GB, and does not cover the same range

[5] These corpora were accessed from a range of sources: ICE-AUS from the Australian National Corpus website, ICE-GB from the CD-ROM published by the Survey of English Usage, and the spoken and written sections of ICE-USA from their respective websites – see "Corpora Used" at the end of the paper.

of spoken genres, consisting only of conversations. For the sake of brevity, this constructed corpus will be referred to as ICE-USA.

The ICE corpora are comparatively small, which allows for textual analysis of the potentially more frequent items (e.g. *the day*, *the time*), as well as some preliminary data on the comparative frequency of items across genres. These results will then be compared with frequency analyses of the much larger British National Corpus (BNC – 100 million words, 1980s-93)), and Corpus of Contemporary American English (COCA – 450 million words, 1990-2012), to give an indication of the frequency of the fixed and variable forms of these temporal subordinators across a similar range of genres. Both of these larger corpora have a range of spoken and written material, and their written components (though not the spoken) contain some parallel text genres. Unlike the ICE corpora, however, they do not have parallel structures, date ranges or sizes, so some care has to be taken when comparing results.

5.2ICE-AUS/ICE-GB/ICE-USA data

Preliminary searches were carried out in ICE-AUS, ICE-GB and ICE-USA to ascertain the variability within the subordinator use of the temporal expressions. Given the relatively small size of the corpora (1m words, or less in the case of ICE-USA), it was possible to search on the nominal element of the phrase (*moment*, *time* etc.) and extract the instances where they appeared to function as part of a subordinate phrase. In this way, it was possible to exclude examples such as (5) – following a copula verb, (6) – where it is in apposition to a preceding temporal expression, and (7) – where it is the object of the preceding verb:

(5) This was **the day (that)** war began in the Gulf.
(6) It is pouring today, **the day (that)** I wanted to do my washing.
(7) I remember **the last time (that)** I broke up with him.

The results are summarized in Table 5.2 below, with items ordered from least variable (*the instant*) to most variable *(the time*).

Table 5.2: Variant forms of possible emergent complex subordinators within ICE-AUS/ICE-GB/ICE-USA

	Nominal element premodified	Zero prep, zero relativizer	Prep + zero relativizer	Zero prep + *that/when*	Prep + *that/when*
instant					
ICE-AUS	-	**1**	1	-	-
ICE-GB	-	**1**	-	-	-
ICE-USA	-	**1**	-	-	-
minute					
ICE-AUS	-	**1**	-	2	-
ICE-GB	-	**4**	-	-	-
ICE-USA	1	**2**	-	-	-
moment					
ICE-AUS	1	**4**	-	-	-
ICE-GB	2	**4**	-	-	-
ICE-USA	1	**-**	-	1	-
day					
ICE-AUS	-	**0**	4	1	1
ICE-GB	1	**2**	4	-	1
ICE-USA	1	**1**	3	-	-
time					
ICE-AUS	11	**-**	36	1	4
ICE-GB	19	**-**	25	2	5
ICE-USA	16	**1**	17	-	1

Although the figures are not large overall, we can see the tendency here for most variability of expression around the nouns with the most polysemy and least specific time reference (*day, time*), and the least variability around the more focussed time words *instant, minute, moment*. We will therefore look at the data from these sets separately.

Another factor that was taken into account was the number of subordinate-clause-initial occurrences of the elliptical form (see Section 3 above). If we compare this to the ambiguous context noted for *the instant* in example (3d) , we can argue that the more evidence there is of subordinate-clause-initial use, the more unambiguous is the role of these temporal phrases as subordinators.

5.2.1 Instant/minute/moment

The majority of the instances of these three temporal expressions in the three corpora are with the fully elliptical, zero preposition/zero *that* form, with some evidence of nominal premodification and the inclusion of a following relativizer. The least frequent of the three expressions is *the instant*, with one occurrence in each corpus for the fully elliptical form, as in:

(8) ICE-USA. W2F-009#37:1 Big fluffy snowflakes had started to fall, turning to water **the instant** they hit the windshield just in time to be swept away by the wipers.

None of the examples of *the instant* occur as subordinate-clause-initial, highlighting its more marginal status.

The minute is most strongly attested in ICE-GB, in its elliptical form and with no variation, suggesting that this is the most firmly established as a complex subordinator in this variety (although we should be wary of comparing frequencies, particularly with the smaller ICE-USA corpus). The AusE examples all come from the same text, but with variability that demonstrates the speaker's uncertainty over the correct usage, and indicates that it is not yet a fixed expression for them:

(9) ICE-AUS. S2A-044(A):77-87. Uhm and I remember being at art school that uhm being sort of familiar with all the the high theories of feminism was was very chic uh very you know was very encouraged, but **the minute that** you tried to translate those theories into a piece of work or **the minute** you tried to make them accessible, uhm you were there was this sort of embarrassed silence and awkward glances and things and and it was uh incredibly derided … Uh **the minute that** you are relevant and you are articulate and you are effective, than you're called reductive, you know.

It's notable that the clause-initial occurrences here both include the subordinator *that*, whereas the fully elliptical form only appears mid-clause – 'or the minute you tried'. The only example of *the minute* occurring at the start of the clause, in ICE-GB, is:

(10) ICE-GB. S1A-044: 383. But **the minute** he came on the stage

Allied to its higher frequency in the BrE data, this indicates that the use of *the minute* as a subordinator may be more advanced in this variety.

The moment appears equally strong in ICE-AUS and ICE-GB, but has very little representation in ICE-USA. There are four instances each of the fully elliptical form in the Australian and British data, the only other variable being the premodification of the noun. It is interesting to note that none of the examples where the noun is premodified is

subordinate-clause-initial, whereas two out of the four for ICE-GB and three out of the four for ICE-AUS for the fully elliptical form are subordinate-clause-initial, for example:

> (11) ICE-AUS. W2F-001:104 Thank God. Thank Jesus, Mary and Josephus, she reads! Rose, you're a lovely girl. **The moment** I heard your snooty twang on the phone I knew it was love.

Clause-initial occurrences like this confirm the role of the elliptical form as a subordinator.

With the small amount of data for this set, it is hard to make comparisons between different text genres. It is to be noted, however, that almost all the examples of *minute* come from spoken texts, the only exception being *at the last minute* in the Fiction category in ICE-USA. There is nothing inherent in the noun *minute* to suggest that it might be a more colloquial usage than *instant*/*moment*, but we will look further for generic differences in the larger corpora in the distribution of these constructions.

5.2.2 Day/time

In contrast to the set of temporal expressions discussed in the previous section, *day* and *time* provide only limited evidence of the elliptical form, and much more internal variability – justifying Quirk et al.'s description of them as being closer to free expressions than complex subordinators. Of these variations, the most common for both is PREP + NP + zero *that* (although *time* does also have high frequencies for the premodified noun form). Evidence for the elliptical form of *the day* was very limited. One of the three examples as a potential subordinator was:

> (12) ICE-GB. W2D-004 101. In this case the automatic period of interruption of employment will end **the day** you started work.

The other two instances come from Spoken (ICE-GB) and Fiction (ICE-US) categories, as is typical of the temporal subordinators discussed in 5.2.1. It's therefore interesting to note this example is from an employment manual, where we would expect the tenor of discourse to be more formal than the contexts where these expressions are typically found in the ICE data.

5.2.3 Summary of ICE data

On the basis of ICE evidence, it appears that *the moment, the minute* and *the instant* are the most likely candidates as emergent subordinators, both through their relative frequencies, and their appearance at the start of a clause (particularly in the case of *the moment*). The noun *day* presents more cases with a preposition *(*Table 5.2*)* than elliptical cases, which suggests it still functions as part of a prepositional phrase within the matrix clause, not as a subordinator. *Time*, the most general in meaning of all of these, hardly even appears without a preposition. The motivation for *the time* to develop as a complex subordinator is hardly there with the ready availability of simple temporal subordinators *when* and *while* (although the latter did develop from a similar construction in Old English, as we have already seen).

Because of the limited data available through the ICE corpora, the tendencies that appeared need to be tested in the larger corpora of British and American English described in Section 5.1. *The minute* as a complex subordinator appears to be more common in British English than elsewhere, and *the instant* appears to be the least common, which allied to its non-appearance at the start of a clause indicates that it is less developed as a subordinator than *the minute* or *the moment*. With the greater frequencies available across genres in COCA and BNC, we will also be able to test some of the general statements about subordination outlined in Section 4 above.

5.3 COCA/BNC data

5.3.1 Evidence for fully elliptical forms

There are no sufficiently large, multi-genre corpora currently available for contemporary Australian English, so the corpora chosen for this part of the study compare British (BNC) and American (COCA) English. These corpora have not been built on the same design, like the ICE corpora, so they are different sizes, cover different periods of time[6], and have different genres within the general categories of spoken and written. The BNC has a much wider range of spoken sub-genres (including informal conversations, meetings, radio shows and phone-ins), while COCA's spoken material is composed purely of unscripted

[6] COCA is periodically updated, so is still expanding. Data in this paper is based on searches carried out on or before 15/3/2013.

conversation broadcast on TV and radio. The range of written material in the BNC is also broader.

If we look first at the range of variation allowed by each of the three temporal expressions shown by the ICE data as being likely emergent subordinators, we find that the fully elliptical form is the most common for all three in both BNC and COCA (see Table 5.3). Identifying these forms was more problematic for the larger corpora than for ICE. Because of the size of these corpora, more restrictive searches were used in BNC and COCA to extract a manageable amount of data. Seeing that the majority of subordinators introduce a finite clause, searches for the elliptical form were carried out with the nominal form preceded by zero preposition and followed by personal/impersonal pronoun, to exclude those followed by a relativizer. Examples such as 5-7 above, where the temporal expression is not acting as a subordinator, were excluded, and subordinate-clause-initial examples were found by searching on the phrase preceded by punctuation (commas, full-stops etc.)or by the coordinators *and*/*but*/*or,* as discussed in Section 3 above (under example (3d)).

Table 5.3: Variant forms of possible emergent complex subordinators within COCA/BNC (normalized per 1m words)

	Zero prep, zero relativizer	Prep + zero relativizer	Zero prep + *that/when*	Prep + *that/when*
moment				
BNC	2.16	1.53	0.4	0.29
COCA	4.03	1.95	0.56	0.12
minute				
BNC	1.35	0.19	0.01	0.02
COCA	2.63	0.36	0.14	0.02
instant				
BNC	0.25	0.05	0.01	0.04
COCA	0.79	0.13	0.05	0.04

For *the minute* and *the instant* the dominance of the fully elliptical form is clear. In the case of *the moment*, while still the majority form in both varieties, there are quite large numbers of the form preceded by the preposition – even when obviously standalone adverbial phrases such as *at the moment* and *for the moment* are removed.

5.3.2 Genre differences

In order to look at the possibility of differences across genre, subcategories that were not common to both corpora were removed from the data collected. These were the written categories Non-Academic and Miscellaneous in the BNC, reducing the overall size of the corpus from around 100m to around 70m words. The size of COCA was not affected.

The generic distribution of the fully elliptical forms of the three least variable temporal subordinators according to the ICE data – *the instant, the minute* and *the moment* – are shown in Table 5.4. The figures have been normalized to show the returns per 1m words, and the temporal expressions ranked from highest to lowest frequency.

Table 5.4: zero prep. + *moment/instant/minute* + zero relativizer in spoken and written genres

	Total	Spoken	Written (overall)	*Fiction*	*Magazine*	*Newspaper*	*Academic*
moment							
BNC	2.60	2.10	3.83	9.47	2.00	2.64	0.50
COCA	4.03	2.60	4.40	10.64	4.03	2.13	0.86
minute							
BNC	1.69	3.00	1.47	4.18	0.63	0.55	0.06
COCA	2.63	3.49	2.40	4.96	2.46	1.96	0.26
instant							
BNC	0.33	0.10	0.37	1.06	0.06	0.18	0.06
COCA	0.79	0.13	0.97	2.79	0.68	0.32	0.11

In accordance with the ICE findings, *the moment* is the most frequently occurring of the three across both corpora, followed by *the minute* and then *the instant*. In each case, both the overall totals and the totals for spoken and written texts show a higher frequency in the American data than the British, which is in contrast to the stronger representation of these forms in ICE-AUS and ICE-GB compared to ICE-USA. It is possible that this reflects a diachronic development rather than a regional difference in the use of these temporal subordinators, as the COCA data is more recent than the BNC – but contemporary British data would be needed to confirm this. The proportion of clause-initial (and therefore unambiguously subordinating) examples was found to be around 40% of the overall total for each of the temporal subordinators under investigation. However, there tended to be a much higher proportion in spoken than in written genres (particularly for *the minute*, at over 60% in each corpus). We will first look at the genre differences, and then discuss any regional differences that appear.

The highest frequencies for each phrase are in the written data, backing up Beaman's (1984) findings for temporal subordination in general, that it is more common in written than in spoken genres. For each of these complex subordinators, the frequency is highest in the Fiction category of both corpora – which is not surprising, given the reporting of past events through temporal sequences typical of the genre. It does appear to be a feature of narrative rather than dialogue, as an inspection of the search results didn't reveal a large proportion of them to be within quotation marks. So the use of direct speech in this category apparently does not particularly boost these frequencies (although quotation marks are not necessarily the only marker of direct speech in fiction). It is not only Fiction that supports these constructions for *the moment* and *the instant*, with the Magazine category also showing consistently higher frequencies than spoken categories, and Newspaper usually on a par with or higher than spoken. Many of the examples from the Magazine category are in personal, reflective articles, which therefore have a strong narrative element to them.

The frequencies for the Academic category are clearly lower than the other categories across the board. This reflects Biber et al.'s findings for *while* – that the Academic genre is less likely to display temporal subordination, and that there is no sign of a semantic shift (comparable to that of *while* to show concession) which might increase the incidence of these subordinators in a genre where the use of argumentation is more prevalent than narrative structures.

The minute shows a slightly different pattern to the other two complex subordinators, with its incidence in Spoken being higher than the overall figures for Written genres (although Fiction maintains the highest frequency). This could indicate that the fully elliptical form is still emerging in the informal context for *the minute*, whereas it is more entrenched for *the moment* and *the instant*, hence the higher number of occurrences in written genres.

5.3.3 Regional differences

As noted above, the COCA data gives higher frequencies in AmE for all the temporal subordinators under investigation, in contrast to the indication from the ICE corpora that

BrE and AusE might be ahead. This could be an effect of the AmE data from COCA being more recent, and perhaps the BNC data presents an earlier developmental stage.

We might expect to find more evidence of the emergent forms in the less formal genres, e.g. Speech and Fiction. The disparity between frequencies is less marked in the spoken and fictional data for each of the temporal subordinators (except *the instant* in Fiction), in comparison to the Magazine and Newspaper categories, while the low frequencies for both in the Academic category makes comparisons difficult. The one exception is *the moment* in the Newspaper category, which actually gives higher frequencies in BNC than COCA. This could be a result of the relative homogeneity of the Newspaper category in COCA (only 10 broadsheet newspapers represented), whereas the BNC has a greater range of sources and registers within the category, with tabloid newspapers included as well as broadsheets. On this evidence, these temporal subordinators are more widely used in British than American English.

6. Conclusions

The corpus evidence for the emergence of the temporal expressions *the instant*, *the minute* and *the moment* as complex subordinators is strong, both in ICE and the much larger BNC and COCA. The dominance of the zero preposition/zero relative form for each is an indication of their use in ambiguous contexts where category change is most likely, according to authorities such as Diewald (2002). They are to be distinguished from other potential complex temporal subordinators such as *the day, the time*, which show a stronger tendency to maintain their status as temporal adverbials –indicated by the fact that they are usually found in full prepositional phrases in the ICE data.

This set of phrases demonstrates an intersective gradient scale, with *the day* and *the time* less likely to shift category, and *the instant, the minute* and *the moment* much more likely. They also demonstrate subsective gradience. Quirk et al. (1985), unlike Biber et al.(1999), describe them as being marginal to the class of complex subordinators. But the corpus data shows clear evidence of their movement towards a subordinative role – with *the moment* the most advanced of the three, and *the instant* the least. The example of *while* gives evidence that such temporal expressions can develop from adverbial phrases to become core members of the subordinator class.

The strong tendency for specific temporal expressions to occur at the start of a subordinate clause following a syntactic boundary marked by punctuation or a coordinator makes their subordinator role unambiguous. Many others can be identified through the underlying syntax. These contexts are the matrix for intersective gradience, involving both category change and structural change (Denison 2010).

Further evidence for the emergence of these temporal expressions as complex subordinators was found in the range of genres where they occurred. While Quirk et al. (1985) described some marginal complex subordinators as generically limited: *save that* 'literary', *forasmuch as* 'formal', these expressions occurred in their fully elliptical form across most everyday written forms, and particularly in Fiction, though with little evidence of use in the more formal Academic genre.

Less evidence was found for their use in spoken genres, which is somewhat surprising given the preference for zero *that* in general in conversation attested by Biber et al. (1999). *The minute* was an exception to this finding, which is a possible indication of more informal, and therefore less established usage. This preference was backed up by a higher incidence of clause-initial occurrences in the spoken categories of both BNC and COCA, suggesting that *the minute* is more established as a subordinator in spoken than in written texts.

The regional preferences indicated by the ICE data, particularly for *the moment* in British and Australian English over US English, were not supported by the larger BNC and COCA. The US data consistently gave higher frequencies for each of the temporal expressions under investigation than occurred for the BrE data. This difference could be explained as a diachronic one, with the more recent COCA data indicating the growing emergence of these forms. More investigation is required, with temporally equivalent regional corpora. The recent publication of the 1.9 billion word GloWbE: Corpus of Global Web-Based English, covering 20 different English-speaking countries, will allow these types of comparisons.

Corpora used

British National Corpus (BNC). corpus.byu.edu/bnc

Corpus of Contemporary American English (COCA). corpus.byu.edu/coca

ICE-AUS www.ausnc.org.au

ICE-GB (Release 2). 2006. CD-ROM. Survey of English Usage, University College London.

ICE-USA downloaded from ice-corpora.net/ice

Santa Barbara corpus downloaded from www.linguistics.ucsb.edu/research/santa-barbara-corpus

References

Aarts, Bas. 2007. *Syntactic Gradience: the nature of grammatical indeterminacy* Oxford: Oxford University Press.

Beaman, Karen. 1984. 'Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse' in D Tannen (ed.) *Coherence in spoken and written discourse* Norwood, N.J.: Ablex. pp.45–80.

Biber, Douglas. 1988. *Variation across speech and writing* Cambridge: Cambridge University Press.

Biber, Douglas, Johannson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education.

Chafe, Wallace. 1982. 'Integration and involvement in speaking, writing, and oral literature' in D Tannen (ed.) *Spoken and Written Language: exploring orality and literacy* Norwood: Ablex. pp. 35–53.

Denison, David. 2010. 'Category change in English with and without structural change' in E Traugott and G Trousdale (eds) *Gradience, Gradualness and Grammaticalization* Amsterdam/Philadelphia: John Benjamins. pp. 105–128.

Diewald, Gabriele. 2002. 'A model for relevant types of contexts in grammaticalization' in I Wischer & G Diewald (eds) *New Reflections on Grammaticalization* Amsterdam: John Benjamins. pp. 103–120

Edgren, Eva. 1971. *Temporal Clauses in English* Acta Universitatis Upsaliensis (Studia Anglistica Upsaliensa 9) Uppsala: Uppsala University.

Häcker, Martina. 1999. *Adverbial Clauses in Scots: a semantic-syntactic study* Berlin, New York: Mouton de Gruyter.

Halliday, Michael. 1994. *An introduction to functional grammar* 2nd edition. London: Edward Arnold.

Hopper, Paul. 1991. 'On some principles of grammaticization' in E Traugott & B Heine (eds) *Approaches to Grammaticalization* Vol. I. Amsterdam: John Benjamins. pp. 17–36.

Hopper, Paul & Traugott, Elizabeth Closs. 2003. *Grammaticalization* 2nd edition. Cambridge: Cambridge University Press.

Huddleston, Rodney. 1976. *An Introduction to English Transformational Syntax* London: Longman.

Huddleston, Rodney & Pullum, Geoffrey. 2002. *The Cambridge Grammar of English usage* Cambridge: Cambridge University Press.

Kortmann, Bernd. 1997. *Adverbial Subordination: a typology and history of adverbial subordinators based on European languages Berlin* New York: Mouton deGruyter.

Kroll, Barbara. 1977. 'Combining ideas in written and spoken English: a look at subordination and coordination' *Discourse Across Time and Space* Southern California Occasional Papers in Linguistics No. 5. Los Angeles: U Southern California. pp. 69–108.

O'Donnell, Roy. 1974. 'Syntactic differences between speech and writing' *American Speech* 49: 102-110.

Peters, Pam. 2012. 'Emergent conjunctions' in S Chevalier & T Honegger (eds) *Words, Words, Words, Philology and Beyond : Festschrift for Andreas Fischer on the occasion of his 65th birthday* Tübingen: Francke. pp. 127–144.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik. Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Smith, Adam. 2013. 'Emergent complex prepositions: variability within the PNP construction' in H Hasselgård, J Ebeling & SO Ebeling (eds) *Corpus Perspectives on Patterns of Lexis* Amsterdam/Philadelphia: John Benjamins. pp. 153–174.

Appendix 5.1 – ICE-GB Abbreviations for Word Classes, Functions and Categories

Word Classes

Adjective (ADJ)
Adverb (ADV)
Article (ART)
Auxiliary verb (AUX)
Cleft it (CLEFTIT)
Conjunction (CONJUNC)
Connective (CONNEC)
Existential there (EXTHERE)
Formulaic expression (FRM)
Genitive marker (GENM)
Interjection (INTERJEC)
Noun (N)
Nominal Adjective (NADJ)
Numeral (NUM)
Preposition (PREP)
Proform (PROFM)
Pronoun (PRON)
Particle (PRTCL)
Reaction signal (REACT)
Verb (V)

Functions and Categories

Adverbial (A) [Function]
Adjective Phrase (AJP) [Category]
Adjective Phrase Head (AJHD) [Function]
Adjective Phrase Postmodifier (AJPO) [Function]
Adjective Phrase Premodifier (AJPR) [Function]
Adverb Phrase Head (AVHD) [Function]
Adverb Phrase (AVP) [Category]
Adverb Phrase Postmodifier (AVPO) [Function]
Adverb Phrase Premodifier (AVPR) [Function]
Auxiliary Verb (AVB) [Function]
Central Determiner (DTCE) [Function]
Clause (CL) [Category]
Cleft Operator (CLOP) [Function]
Conjoin (CJ) [Function]
Coordinator (COOR) [Function]
Detached Function (DEFUNC) [Function]
Determiner (DT) [Function]
Determiner Phrase (DTP) [Category]
Determiner Postmodifier (DTPO) [Function]
Determiner Premodifier (DTPR) [Function]
Direct Object (OD) [Function]
Discourse Marker (DISMK) [Function]
Disparate (DISP) [Category]
Element (ELE) [Function]
Empty (EMPTY) [Category]

Functions and Categories (cont.)

Existential Operator (EXOP) [Function]
Floating Noun Phrase Postmodifier (FNPPO) [Function]
Focus (FOC) [Function]
Focus Complement (CF) [Function]
Genitive function (GENF) [Function]
Imperative Operator (IMPOP) [Function]
Indeterminate (INDET) [Function]
Indirect Object (OI) [Function]
Interrogative Operator (INTOP) [Function]
Inverted Operator (INVOP) [Function]
Main Verb (MVB) [Function]
Nonclause (NONCL) [Category]
Notional Direct Object (NOOD) [Function]
Notional Subject (NOSU) [Function]
Noun Phrase (NP) [Category]
Noun Phrase Head (NPHD) [Function]
Noun Phrase Postmodifier (NPPO) [Function]
Noun Phrase Premodifier (NPPR) [Function]
Object Complement (CO) [Function]
Operator (OP) [Function]
Parataxis (PARA) [Function]
Parsing Unit (PU) [Function]
Postdeterminer (DTPS) [Function]
Predeterminer (DTPE) [Function]
Predicate Element (PREDEL) [Category]
Predicate Group (PREDGP) [Function]
Prepositional (P) [Function]
Prepositional Complement (PC) [Function]
Prepositional Modifier (PMOD) [Function]
Prepositional Phrase (PP) [Category]
Provisional Direct Object (PROD) [Function]
Provisional Subject (PRSU) [Function]
Stranded Preposition (PS) [Function]
Subject (SU) [Function]
Subject Complement (CS) [Function]
Subordinator Phrase Head (SBHD) [Function]
Subordinator Phrase Modifier (SBMO) [Function]
Subordinator (SUB) [Function]
Subordinator Phrase (SUBP) [Category]
Tag Question (TAGQ) [Function]
Particle To (TO) [Function]
Transitive Complement (CT) [Function]
Verbal (VB) [Function]
Verb Phrase (VP) [Category]

Chapter 6: Conclusion

1. Evaluating the criteria for assessing the lexicogrammatical status of multiword units

The lexicogrammatical items examined in this research all present problems of categorization and description for dictionaries and grammars as to what extent they function as units. And if they do, are they lexical or grammatical units? Various criteria for assessing the grammatical status of items were introduced in Chapter 1. In concluding this research, we revisit these criteria in relation to canonical examples of the multiword units discussed in Papers 1 to 4, so as to compare their relative value for each type of unit.

1.1 Fixity versus Variability

The concept of fixity is not a straightforward one, as noted in Chapter 1, where it was discussed as a relative, not an absolute criterion, and the means of differentiating between degrees of fixity were questioned. When considering the unity of an expression, both the integrity of its syntactic structure and the degree of collocational restriction within it have to be taken into account. The use of corpus data, as in the four preceding papers, allows the evidence of actual relative frequencies – not simply the theoretical possibility of variant forms – to influence our assessment of the status of the unit. The four constructions examined here display different degrees of fixity.

Light verb (LV) phrases are the least fixed of the constructions investigated. While there is an underlying syntactic structure of V + indefinite determiner + N, the corpus data shows that this structure is often interrupted by premodification of the N element. The collocational choices are restricted by the limited range of general verbs that Algeo (1995) lists as the core expanded predicates: *give, make, have, take*. Nor are these interchangeable, with certain nouns collocating exclusively with one (*make a record*), or allowing limited choice (*have*/*take*/*make** *a look*, while the use of *give* changes the structure of the phrase by requiring a direct object ('I'll give it a look') and can modulate the sense of the phrase, making it a non-'light' use of the verb ('she gave me a disconcerting look'). The corpus data used in the current research supported earlier research on regional preferences (Algeo, 1995), in that *have* was preferred over *take* in LV constructions in BrE,

and indeed there appears to have been an increase in the use of *have* if we compare the LOB data with the ICE-GB data – as discussed in Paper 1 (2009). This preference was even more pronounced for AusE and NZE, indicating a possible movement towards a narrowing of choice of the light verb, and therefore towards fixity (at least in the southern hemisphere).

However, if we look at *have/take a look*, by far the most frequent individual LV construction, its frequency didn't correlate with fixity of structure – with several examples of premodification found across the corpora – although there was a clear regional preference for the *have* form across the British, Australian and New Zealand data investigated. LV phrases therefore cannot be deemed to be showing signs of lexicalization/ grammaticalization through movement towards fixity.

For NNQs, the basic syntactic structure of indefinite determiner + quantifying noun + *of (a lot of),* or in some cases, plural quantifying noun + *of (lots of)* is found to be less likely to contain premodification of the nominal element (see Paper 2, 2009). But where it does, for example *there was a **big** heap of leaves* as opposed to *there was a heap of leaves*, the adjectival premodification highlights the descriptive rather than the quantifying function of the noun in that instance, and therefore the integrity of the unit as a quantifier is undermined. The variation in syntactic structure that allows for either singular or plural quantifying nouns is only available in a subset of NNQs, notably *a lot/lots, bunch/bunches, heap/heaps, load/loads of*.

As with the corpus data on light verbs (on *have/take a look*), one NNQ was notably more frequent than the others – *a lot/lots of*, with the frequency of the singular form being considerably higher than the plural form. This type of variation, of the inflectional ending of the noun element within that phrase, marks its grammatical category, and will be discussed further below under *decategorialization* (1.3). There was no collocational variation within the phrase, with no alternatives to the indefinite determiner *a* preceding, or *of* following the quantifying noun. The NNQs in this study therefore demonstrated a greater degree of fixity than all the LV constructions investigated.

Fixity is an important consideration for complex prepositions (CPs), with possible structural and collocational variations within the unit cited specifically by Quirk et al. (1985) as indicators of their status as a grammatical unit or a freely combined form.

Structural "indicators of separateness" in the P1 + N + P2 sequence include P + its complement being replaced by a possessive pronoun (*in stead of* [NP] / *in its stead*), P2 being omitted (*in lieu of [NP]* / *in lieu*), and P2 + complement being replaced by a (premodifying) determiner (*in case of [NP]* / *in that case*). While these possible variations are proof for grammarians such as Huddleston and Pullum (2002) that there are no syntactic grounds for recognizing the class of complex preposition, Denison (2010) argues that they should be analysed separately, with the status of one construction pre-empting the syntax of the other. In the case of *behalf*, it is part of a complex preposition when followed by *of* (*on behalf of*), but an independent noun when preceded by a possessive (*on her behalf*). Another structural variation, the potential for the inclusion of a determiner before N (*at the risk of, with a view to*) will be discussed further in Section 1.3, on decategorialization. Unlike NNQs, the variation between the singular and the plural form of the N element in CPs is extremely rare. The majority only take the singular form, with the few that take the plural tending to do so invariably (e.g. *in terms of*). One rare exception, *on (the) ground(s) of*, appears only to have gained the plural inflection over the last century, according to the corpus-obtained data from this study (see Paper 3, 2013)[1].

Collocational variations can be found in the choice of preposition before and after the noun (*with/in respect to/of*). However, very few CPs offer a choice, and in general the range of prepositions used is restricted (*in* + *of* being by far the most common). Overall there are more possibilities of variation in CPs than NNQs, but these are restricted to a small set. The majority of CPs show a similar level of fixity to NNQs, and much less variation than demonstrated by the LV construction.

Of the four types of construction looked at, the complex subordinators (CSs) showed the greatest degree of fixity as to their syntactic structure (see Paper 4, 2014). Those examined were the temporal expressions *(at) the minute/moment/instant (that)*, which showed the most evidence for grammaticalization as subordinators (when used without either a preceding preposition or following relative pronoun). That is to say they are more structurally and collocationally fixed than the prepositional phrases on which they are based. Others in the same temporal set, such as *(on) the day (that)* and *(at) the*

[1] And is also probably affected by the idiomaticity of the plural form generally. Compare *What are the grounds of his argument*? with *What is the ground of..?*

time (that) still show some propensity for premodification, even in their full elliptical form (*the very day, the last time*). Even in the form P + NP + relativizer, of temporal adverbials such as in *at the moment that*, the data show little variation in choice of preposition. Once reduced to their fully elliptical form, *the moment*, their syntactic structure is firmly fixed.

With these four case studies, we can see a scale of fixity across the different cases. We shall look at some specific examples from each in Section 2.1 below, to gauge the relative fixity of each type, and its relation to the other criteria looked at.

1.2 Semantic bleaching

A second criterion for phrasal unity raised in the Introduction is semantic bleaching (or "reduction" in Boye and Harder's terms (2012)), where the individual elements of a multiword unit can be seen to have lost their individual semantic identity, and meaning is attached to the phrase as a whole, not the elements within it. This may result in a unit where the overall sense is not deducible from its constituents (which are therefore semantically opaque), as is the case with pure idioms. In other cases, the semantics of the component elements may become generalized or abstracted within the context of a phrase.

LV constructions are relatively transparent in their semantics, with the main semantic weight carried by the noun, which specifies the action of the phrase. Possible slight modifications may be given by the choice of verb, but the highly polysemous nature of the verbs typical of the construction means that they tend to be semantically "light" and therefore do not influence the overall sense (*have/take a look*). In the data examined in Paper 1, semantic opacity was only found for a subset of LV phrases such as *take a bath* ("be defeated"), or *have a lend of* ("to tease"), *have a lash* ("to try"). These are distinct from the standard examples where the sense of the noun determines the meaning of the unit. They are lexicalized as idioms where the meaning of the phrase cannot be interpreted from its components. This productivity of the whole construction through idiomatization is to be distinguished from the process of bleaching, where individual elements gradually lose their semantic specificity.

All the NNQs looked at in Paper 2 contained a noun element that had some sense of quantity attached to it – usually large (although some words that don't carry a specific

sense of quantity, like *bag*, come to express a large amount as quantifiers). Typically, they show signs of moving from specific, descriptive qualities to more general, quantitative ones. This was particularly evident in those where a singular noun had a plural complement (*a lot of/load of/heap of things)* where the items are conceptualized as an undifferentiated mass rather than a one or more collections of discrete items. A large range of quantifying nouns was found in the corpus data, although only a small subset of these could be said to be functioning strictly as a quantifier, with many of them still retaining descriptive features. It follows that the degree of semantic opacity for these was low. Even those examples that functioned primarily as quantifiers still retained the semantic properties that distinguished a large/small amount (e.g. *lot, bunch, heap, stack* vs. *drop, pinch, scrap, touch*).

In the more abstract CPs such as *in terms of/in relation to,* the generalization of the nominal element certainly shows a higher degree of semantic opacity than for either LVs or NNQs. There is evidence for a more abstract use of concrete nouns such as *view* and *place* in their use within CPs *in view of, in place of*, in the data looked at in Paper 3. However, this does not apply to the locational CPs such as *in front of*, which are structurally highly fixed but semantically transparent. There is therefore no direct correlation here between semantic bleaching and formal unity.

The CSs examined in Paper 4 formed a uniform semantic set, all being temporal subordinators, and therefore would be expected to contain elements that preserve a general sense of time, just as the nominal elements of the NNQs preserve a sense of quantity. In fact they are less bleached than the NNQs, with the relative time period embedded in the choice of noun. Yet one could argue that the noun within subordinator phrase is serving a more generalized purpose in this syntactic role than the specific temporal reference of the nouns *day* or *moment*.

Semantic bleaching is not consistently evident across the four sets of items in this study, with the main evidence of semantic loss being in the form of idiomatization of the LV phrase (e.g. *take a bath*), and in NNQs and CPs where some abstraction of the core sense of the key noun has occurred. This is perhaps evidence that it is a later stage of the grammaticalization process – Brinton and Traugott link bleaching to the establishment

through frequency of the grammaticalized construction (2005: 30), and therefore not fully realized in these emergent forms.

1.3 Decategorialization

Decategorialization of the word's grammatical class is one of the standard indicators of grammaticalization cited, for example, by Hopper (1991) and Brinton and Traugott (2005). It involves the loss of morphological markers and syntactic characteristics of grammatical classes, such as plurals of nouns or verbal tense inflections. These lexically open classes are considered to be the most likely candidates for decategorialization, as part of a process of becoming closed categories such as conjunctions or pronouns (Hopper and Traugott, 2003: 106-7). As each of the constructions investigated in this study has a noun as one of its elements, we might expect to find decategorialization as a consistent indicator of grammaticalization across each one.

There is no evidence of decategorialization in the LV category, with the full repertoire of tense inflections available for the verb element. The noun marked by the indefinite article *a*, and often reinforced by adjectival premodification: *he was taking a long hard look at it*. The construction does create an environment for grammatical indeterminacy – the nominal elements tending to share the same form as their verbal equivalents (*look, walk*, etc.), and, conversely, conventionally verbal forms taking on nominal functions (*have a **use**, have a **lend***). This is more an indicator of (particularly colloquial) flexibility than a loss of categorial status.

NNQs also show little sign of decategorialization, with the consistent inclusion of the determiner *a* to introduce the noun, or plural inflections preserved in several cases *lots/heaps/loads of*. However, the meaningfulness of these plural inflections is undermined by the "number transparency" (Huddleston and Pullum, 2002) of these nouns, in that they are able to take either a plural or a singular complement. We can argue that category reinforcement by adjectival premodification, as in *a big pile of books*, emphasizes the descriptive quality of the noun over its quantifying role, so there is some limitation on the full expression of the category implied by NNQ usages. But to call this decategorialization would be stretching the definition.

Decategorialization appears to be a much clearer factor in the development of CPs, with the process clearly stated by Hoffmann:

> the nominal element of the construction over time loses the features that define its categorical status as a noun. For example, in the complex prepositional use of *in view of*, *view* cannot occur in the plural or with a determiner, nor can it be premodified by an adjective. The noun *view* has thus undergone the process of decategorialization. (2005: 56).

However, the data in Paper 3 showed that decategorialization is not a prerequisite for fixity as a unit for some CPs, with *at the risk of, in the process of* and *on the grounds of* all retaining categorial markers, and showing no signs of losing them.

For CSs, the example of *while* developing from an earlier form that included the definite article and conjunction (*þa hwile þe,* "the while that") indicates that decategorialisation might well play a part in the ongoing grammaticalization of the temporal adverbials discussed in this study, but in their current status, reduced to noun phrases, the noun category is still clearly marked by use of the determiner in *the day*, *the moment* etc. However, if we consider their role as noun phrases in syntax, we're looking at non-categorial functions of the noun phrase.

Decategorialization is the syntactic equivalent to semantic bleaching. While the erosion of meaning of a component within a phrase is an indication of the lexicalization of the phrase as a unit of meaning, the loss of individual grammatical identity of elements within a phrase points to grammaticalization of the unit as a whole. This clear correlation between the erosion of the function of an individual item within a phrase, and the grammaticalization of the phrase makes it a very useful indicator, as argued for examples like *in spite of*, where the N element has lost categorial markers such as plural inflection, a determiner, or any potential for adjectival premodification. However, this correlation does not work so well in all cases. In their study of the similar string *in (the) hope(s) of* as a CS, Brems and Davidse argue that "forms do not have to be fully decategorialized to function as a complex subordinator or have it as one of its analyses" (2010: 112). Certainly, amongst the potentially grammaticalized items investigated here, there is inconsistent evidence of it, with only some clearly emergent CPs showing signs of it. As a diagnostic of grammaticalization, decategorialization therefore has to be treated with some caution.

1.4 Syntactic reanalysis

Syntactic reanalysis is the ultimate proof of a unit's functional, as well as structural integrity. Although it is not equivalent to grammaticalization, as Hopper and Traugott (2003: 58) observe, since some reanalyses occur as lexicalizations (e.g. the preposition *up* becoming a verb or a noun), it is the main mechanism whereby grammaticalization occurs. There are many examples of phrasal items in English that have shifted category. For example *be going (to)* as a purposive construction becoming a semi-modal auxiliary, with this use becoming phonologically reduced in the form *gonna*. None of the phrases examined in this study have reached the stage of fusion to a single word in this way, but there was some evidence for syntactic reanalysis.

LV constructions of the type under consideration in Paper 1, for example *have a look, take a walk*, would appear on the surface to offer the potential for grammatical reanalysis because of the exact correspondence between the form of the noun and its verbal equivalent (*look*/*look*, *walk*/*walk*). In fact this is not the case, with the elements within the construction maintaining their categorial identity (inflection for the verbal element, determiners and premodification for the nominal one), and the construction as a whole functioning as a full verb phrase, despite the semantic influence of the noun. The syntactic class of CPs also shows no signs of reanalysis, except in the special case *in terms of,* which Hoffmann (2005) suggests shows tendencies towards development as a discourse marker.

There was evidence for reanalysis of the NNQs investigated in Paper 2 from quantifying noun phrase to complex determiner, especially for those where the noun shows number transparency. Verbal agreement with the number of the quantifying noun indicates that it is the head of the noun phrase (*heaps of soil were…*). Whereas if the verbal agreement is instead with the complementary noun (*a lot of people are…*) then the NNQ can be reanalysed as a complex determiner. Unsurprisingly, the most established NNQ, *a lot of,* is the one that demonstrates the highest degree of number transparency, both with the plural and singular form of the quantifying noun, but particularly when it is in the singular. Other candidate NNQs, such as *a heap/load of*, were less likely to take on a complex determiner role in the singular, but in their plural form they were more likely to be followed by a singular verb, and thus number transparent. They are therefore emerging candidates for grammaticalization through reanalysis.

The strongest case for syntactic reanalysis was observed in the temporal CSs *the instant/minute/moment*. The presence of these elliptical forms in the corpus data examined in Paper 4, without either a preceding preposition or following relativizer to determine their status as adverbial adjuncts, offers the conditions of ambiguity which allows their analysis as subordinators. Of the constructions examined, *the moment* demonstrated the strongest evidence for reanalysis in terms of the frequency of occurrence in its fully elliptical form, and at the start of a clause, where its subordinating role is unambiguous.

While it may only be occasionally observed in the emergent constructions that are the focus of this research, syntactic reanalysis provides powerful evidence for the tendency of a type of phrasal unit to move towards a more grammaticalized status.

2. Synthesizing the criteria

Each of the criteria that have proved useful in this study for assessing the level of grammaticalization of a construction depends on different kinds of linguistic evidence. Fixity, or syntactic integrity, can be assessed by comparing the frequency of fixed and variable forms. Semantic bleaching relies on qualitative analysis, where elements within a phrase can be seen to be losing, or having diluted, specific senses. Decategorialization and syntactic reanalysis are both objective indicators of categorial change, the former at the individual word level, the latter at a phrasal or clausal level. The question thus arises as to whether these different criteria can be synthesized as a means of demonstrating a scale of grammaticalization, from emergent to established, across different types of phrases.

2.1 Established vs. emergent forms

Overall frequency can be used to establish both the fixity of a phrase, relative to its possible variants, and its establishment relative to other similar phrases that may be established or emergent. In the corpus data looked at in Papers 1 and 2, two categories (LV and NNQ) had a particularly frequent example, i.e. *have a look* and *a lot of*. These can therefore be used as exemplars of established forms, as described by Bybee (2010), to be compared with less frequent phrases in the same categories that might be considered to be emerging. There were less obvious candidates for CPs in Paper 3 (where there were several high frequency examples), and CSs in Paper 4 (where there are no canonical

examples of the temporal type investigated). The exemplars of the established forms were taken from Quirk et al. (1985), which uses *in spite of* as the most grammaticalized CP, and from Mittins et al. (1970) and Peters (1993, 2012) *the way* as a CS. The latter consists of the same elements as the CS's of time investigated here. Table 6.1a compares frequencies of the established forms and their possible variations with the most grammaticalized examples of emergent forms found in the current study.

While COCA and BNC are not equivalent corpora, in size or genres covered, as noted in Papers 3 and 4, they are large enough to provide a general sense of the comparative frequency of these phrasal units in the major varieties of AmE and BrE. The figures given are the raw frequencies for each construction, with its fixed form set against a search allowing for any variant before the nominal element. The percentages give the relative frequencies of the fixed and variable forms.

Table 6.1a: Established* and emergent forms by frequency in COCA and BNC

		COCA	**%**	**BNC**	**%**
LV	***have a look***	1105	**65%**	1909	**91%**
	*have a * look*	598	**35%**	193	**9%**
	have a bath	74	**69%**	143	**87%**
	*have a * bath*	34	**31%**	21	**13%**
	take a look	8523	**82%**	395	**66%**
	*take a * look*	1874	**18%**	205	**34%**
	take a bath	268	**78%**	25	**83%**
	*take a * bath*	75	**22%**	5	**17%**
NNQ	***a lot of***	152681	**99%**	14440	**99%**
	*a * lot of*	2057	**1%**	187	**1%**
	a heap of	439	**80%**	126	**72%**
	*a * heap of*	113	**20%**	49	**28%**
CP	***in spite of***	7049	**100%**	2692	**100%**
	*in * spite of*	0	**0%**	0	**0%**
	in the process of	4097	**96%**	892	**93%**
	*in the * process of*	190	**4%**	66	**7%**
CS	***(-prep) the way + subj. pron.***	18185	**88%**	2227	**86%**
	*(-prep) the * way + subj. pron.*	2411	**12%**	361	**14%**
	(-prep) the moment + subj. pron.	1131	**89%**	261	**86%**
	*(-prep) the * moment + subj. pron.*	136	**11%**	42	**14%**

*Established forms in bold.

The status of the established forms is confirmed often by a higher frequency than for the emergent form, but a high proportion of the fixed form in relation to the variable

form is also evident for the emergent examples. This is clearly represented for the LVs regionally where the more frequent verb is used (*have* in AmE, *take* in BrE), but less so for the dispreferred verb in each region. An average scale of fixity across the corpora for each item can be shown as follows:

Table 6.1b: Scale of fixity for established and emergent forms in AmE and BrE (COCA/BNC)

%	100	95	87	83	76
	in spite of ***a lot of***	*in the process of (96/93)*	***the way*** *(88/86)* *the moment (89/86)* ***V + a look*** *(82/91)*	*V + a bath (78/87)*	*a heap of (80/72)*

These percentages for both the established and emergent multiword units place most of them in the top "completed" category of linguistic variants using synchronic (apparent-time) data. This category is derived from the diachronic corpus-based research of Nevalainen and Raumolin-Brunberg (2003), which gives a set of categories for the gradual diffusion of an incoming variant based on the percentage of use. The completed category is identified as over 85%, which most of the items in Table 6.1b meet, with the exception of *V + a bath* and *a heap of*, which would both be classified as "nearing completion" (66% to 85%) on their criteria[1] (2003: 55).

2.2 Comparing fixity with other criteria: evaluating the set

If we combine the frequency-based scale for fixity with the other relative features of grammaticalization discussed above, we can get an overall picture of how far along the grammaticalization scale each multiword unit has progressed, both in its established and emergent form. Based on the purely numerical data we get the following scale for fixity:

Table 6.2a: Criteria of fixity applied to established and emergent forms

	High degree of grammaticalization ←					Lower
Fixity (Syntactic integrity)	**in spite of** **a lot of**	in the process of	**the way** the moment	**V a look**	V a bath	a heap of

[1] Note also that *a heap of*/*heaps of* span either side of the central point in Brems' "scale of grammaticality" based on the relative frequency of quantifier and non-quantifier uses in the contemporary Collins WordbanksOnline corpus (2012: 211).

As a set, it is the CPs that demonstrate the highest degree of grammaticalization according to fixity, with both the established *in spite of*, and the emergent *in the process of* featuring at the more fixed end of the scale. For NNQs, we have *a lot of* and *a heap of* at opposite ends of the scale, with the dominance of the established form perhaps inhibiting the emergence of any particular alternative (although Paper 2 demonstrated that the class as a whole is productive). There was no distinction between the two CSs, in the middle of the scale, and the LV set was least fixed, both appearing towards the lower end of the scale.

In general, as might be expected, the emergent forms are less fixed than the established ones, although, as demonstrated in Paper 3, a lack of variation can be demonstrated in some of the less canonical forms for CPs such as *in the process of*. To give a broader picture of how the emergent forms feature in comparison to the established ones on a scale of grammaticality, the non-quantifiable, qualitative criteria from Papers 1-4 – semantic bleaching, decategorialization and syntactic reanalysis – also have to be included. Table 6.2b provides a less finely-graded (3-part rather than 6-part) scale based on the analyses in this study, where the highest degree of grammaticalization is based on strong evidence of each criterion, the middle grade on some evidence, and the lowest grade on little or no evidence.

Table 6.2b: Non-quantifiable criteria applied to established and emergent forms

	High degree of grammaticalization ← Lower		
Semantic bleaching	**in spite of** **a lot of** **the way**	in the process of a heap of the moment	**V a look** V a bath
Decategorialization (of N element)	**in spite of**		in the process of **a lot of** a heap of **the way** the moment **V a look** V a bath
Syntactic Reanalysis	**the way** the moment	**a lot of** a heap of	**V a look** V a bath **in spite of** in the process of

Again, the more invariable and established forms tend to congregate at the high end of the scale for each criterion, but in Table 6.2b we start to see the grouping of particular items at different stages along the scale, sometimes showing a very sharp divide in degree of grammaticalization, sometimes demonstrating graduated distinctions. In order to see how well each criterion functions as a discriminator of grammatical status, Table 6.2c, below, combines the quantitative and qualitative criteria, and show the different categories colour-coded to demonstrate where they cluster along the scale for the different criteria.

Table 6.2c: Combined grammaticalization criteria applied to established and emergent forms

	High degree of grammaticalization ← Lower					
Fixity (Syntactic integrity)	**in spite of** **a lot of**	*in the process of*	***the way*** *the moment*	***V a look***	*V a bath*	*a heap of*
Semantic bleaching	***in spite of*** ***a lot of*** ***the way***		*in the process of* *a heap of* *the moment*		***V a look*** *V a bath*	
Decategorialization (of N element)	***in spite of***				*in the process of* ***a lot of*** *a heap of* ***the way*** *the moment* ***V a look*** *V a bath*	
Syntactic Reanalysis	***the way*** *the moment*		***a lot of*** *a heap of*		***V a look*** *V a bath* ***in spite of*** *in the process of*	

Key: LV NNQ CP CS
(bold = established form, italic = emergent)

Table 6.2c demonstrates considerable variation, both in the degree that the different grammaticalization criteria are in evidence across types of construction, and between established and emergent forms of each. The most consistently grouped are the LVs, which are low on the scale for both *V a look* and *V a bath* (as would be expected from the lack of evidence for grammaticalization found in Paper 1). The only distinguishing feature is the slightly higher level of fixity for the established form *V a look*, as shown by the frequency data in Table 6.1a above.

Fixity is not, on its own, conclusive proof of grammaticalization, as it can equally be evidence for lexicalization (see Introduction, 2.1 p.13). It should be noted for LVs that while the N element does not show signs of decategorialization (with the determiner still

present, and the potential for premodification still available), Brinton and Traugott (2005) argue for extended aspectual function of the verb in this LV form of the composite predicate: "Composite predicates began to be associated with telic aspect, that is, the situation is perspectivized as an accomplishment or an achievement, as in *take a walk* vs. *walk"* (2005: 130). The verb has therefore grammaticalized and, they argue, has the potential to become a derivational prefix, similar to *be* in *bedevil*. There is no corpus evidence for such a development, which would require a decategorialization of the N element.

The most consistent evidence for grammaticalization is found for the established CP *in spite of*, with the emergent CP *in the process of* coming close to matching it only in terms of fixity. Even this example does not satisfy all the criteria, being at the low end of the scale for syntactic reanalysis, with CPs occurring only as prepositional phrases, with no ambiguity around their grammatical role. Syntactic reanalysis is most strongly evident for CSs, with both the adverbial phrases *the way* and *the moment* clearly taking on a subordinator role with the loss of a preceding preposition and following relativizer. Evidence of syntactic reanalysis is also available for NNQs, with the loss of number agreement between N element in NNQ and that of the noun complement allowing for the analysis of the unit as a complex determiner rather than a premodifying noun phrase. This development is less consistently observable than for the CSs, and is therefore placed lower on the scale. On this basis, syntactic reanalysis is a criterion that is clearly useful as a discriminator between degrees of grammaticalization, even though it doesn't apply to all examples.

The least effective criterion as a discriminator in Table 6.2c is decategorialization of the N element. Only *in spite of* has lost the categorial marker of its N element, with all the other examples retaining determiners. One could argue that within the remaining examples there are different degrees of decategorialization, with most forms not allowing plural inflection for example, whereas the NNQs do. However, standard definitions of decategorialization (e.g. Hopper and Traugott, 1993) do not distinguish between types of categorial marker as being more or less indicative, and so we cannot separate them on that basis. It could well be that decategorialization is more evident in more advanced stages of grammaticalization (contrary to Hopper's assertion, see Introduction, 2.1 p.12),

as when complex units become single words (e.g. *þa hwile þe* to *while; in the stede of* to *instead*). It therefore does not figure so prominently in this set of emergent, multiword units.

Semantic bleaching works better as a discriminating criterion because it is more evenly distributed across the examples investigated. Granted that the boundaries between meanings are not clearcut, and therefore do not lend themselves easily to representation in a scale. Yet distinctions can be made on the basis that in the most bleached forms (*in spite of, a lot of, the way*) the core sense of the N element is definitely eclipsed by the sense of unit. At the other end of the scale, the core sense of N is preserved in both LVs, but there is at least some generalization of the sense of N in the middle group (*process, heap, moment*).

What is clear from Table 6.2c, and from this study in general, is that a combination of several criteria is needed to demonstrate the emergence of multiword units as grammatical units, and that the force of these criteria may differ across different types of construction. For these multiword units, each of which features a noun at its core, decategorialization proves to be the least useful criterion as a means of distinguishing degrees of grammaticalization, while fixity, semantic bleaching and syntactic reanalysis all prove useful as discriminators, albeit in different combinations for different units.

3. Classification of multiword items in grammars and dictionaries

Each of the multiword constructions researched in this study presents challenges of categorization and description for grammars and dictionaries. They struggle for recognition because of their questionable status as discrete units with a defined lexical/grammatical class, and in some cases the fact that they are relatively new, emerging forms. They draw attention to wider issues of how far usage data can influence the description of the grammar of the English language and to what extent grammatical labelling in dictionaries can be used where lexical definitions are inadequate to explain their functions.

3.1 Grammatical categorization

In comprehensive grammars, the treatment of the emergent items investigated in this research is to present them as marginal in their relation to the core grammatical categories, and within these marginal groups to suggest some items are more central than others. However, this is not demonstrated systematically across the grammars looked at. In some cases, the construction is not recognized as a grammatical class at all, as with Huddleston and Pullum's (2002) rejection of CPs. The centrality or not of particular instances of complex units can generally be only referenced by whether they are included as examples in the grammars involved.

Only in one case – Quirk et al.'s (1985) treatment of CPs – is a scale presented where items shown to be more or less central to the class. This scale itself is problematic, as it only represents items in terms of their cohesion, or fixity, and includes questionable criteria that open to debate the validity of the class as a whole (see Paper 3, 2013). A more consistent application of a wider range of criteria, based on corpus data (as in Table 6.2c above), would allow a more accurate representation of these emergent categories. This would apply equally to multiword items that spanned several criteria, such as the example of syntactic reanalysis above, giving a more accurate picture of the potential for movement across classes than is currently demonstrated even in the corpus-driven grammar of Biber et al.

The use of statistics from the corpus designed specifically for Biber et al.'s grammar project – the Longman Corpus of Spoken and Written English (LSWE) – allows comparison of synchronic usage across different regions and categories of spoken and written English. For example, when discussing what are irregular or regular forms of the past tense, Biber et al. use percentage criteria similar to those cited from Nevalainen and Raumolin-Brunberg (2003) above to establish which forms can be considered more regular.

The LSWE corpus statistics are not used very often in Biber et al. to make such observations about individual usage trends. More commonly they present data on the comparative formality/informality of usage, or on collocational patterns, such as those observed for different types of NNQs (1999: 252-3). So, where there are potential emergent forms within categories, there are no data-driven criteria for demonstrating

which should be included as examples, and whether they are more central or marginal to the class.

The grammars looked at in this study adopt quite different approaches to the grammatical description of language. Huddleston and Pullum (2002) present a categorial view of language, with fixed, discrete grammatical categories – as exemplified by their denial of CPs as a grammatical class. Gradient phenomena are much more evident in both Quirk et al. and Biber et al., but they do not have a consistent approach as to how they should be classified – as central, emerging, or marginal members of categories. It would be helpful to the users of grammars both to have it overtly stated whether the grammatical reference they are using adopts a categorial or a descriptive approach, and what analytical tools and principles are being used to show membership of, and movement between classes.

One current English grammar that states specifically that it is descriptive is Aarts' *Oxford Modern English Grammar* (2010). In the Preface, the author states that "The account of grammar presented in this book is descriptive, not prescriptive" and the potential for emerging grammatical categories is also acknowledged – "the English language is not a static entity, but is continually subject to inevitable change, which is reflected in its lexis and grammar". (Preface, 2010). The Aarts grammar is designed as a general grammar for non-specialists, but also specifically for "undergraduate and postgraduate students of English language and linguistics". Let us therefore compare it with other student/learner grammars to see if their approach to language description differs from that presented in the comprehensive grammars discussed so far.

3.1.1 Student and Learner grammars

The three recent student grammars to be compared in this section are the *Oxford Modern English Grammar* (Aarts, 2010), *A Student's Introduction to English Grammar* (Huddleston & Pullum, 2005) and the *Cambridge Grammar of English* (Carter & McCarthy 2006). Of these, only Huddleston and Pullum (2005) state overtly in the title that it is designed for students. As already noted, Aarts (2010) states university-level students as an intended audience, and Carter and McCarthy include learner features such as the use of potential

error warnings to indicate areas where learners of English have been observed in their learner corpus to have difficulties (2006: 16).

The use of corpora is one aspect that the Aarts (2010) and Carter and McCarthy (2006) share, although they are employed quite differently – Aarts using ICE-GB simply as a means of providing authentic examples (2010: 6), whereas Carter and McCarthy's grammar is "informed by the corpus" (2006:6). This means that not only are illustrative examples taken from the corpus, but choices about the inclusion of items, including multiword units, described as word clusters (2006: 15), are based on corpus frequencies. Information is also included about the formality/informality of terms according to their use in spoken or written English. Huddleston and Pullum (2002) do not make overt use of corpora in their student or their general grammar, despite the acknowledgement of the Brown, Lob and ACE corpora in the preface.

The coverage of the multiword units discussed in this research varies across these grammars. Only CPs are covered by all of them, with Huddleston and Pullum including them to dispute their status (2005: 146-7) – as they do in their larger comprehensive grammar. Both Aarts (2010) and Carter and McCarthy (2006) supply lists of typical CPs, although their lists differ and there is no evidence of a systematic approach to the selection of examples – such as the use of frequencies. Aarts does extend the discussion of the grammatical function of CPs further by presenting a subset of them that "take clauses as Complements (predominantly *-ing* participle clauses" and calls them "conjunctive complex prepositions" (2010: 157).

The other three multiword units we are concerned with here, LVs, NNQs and CSs are not covered by Aarts or Huddleston and Pullum (2005) but are all in Carter and McCarthy (2006). It is possible that their inclusion is influenced by the use of corpus data for the selection of items, although this is not specifically stated. A list of CSs is supplied (2006: 559), although this doesn't include the temporal subordinators discussed in Paper 3, and is not chosen according to frequency of items. LVs are presented as "Delexical expressions, using verbs such as *do, get ,give, make, take*" (although not *have*), which "enable a verb-type meaning to be expressed in a following noun object" (2006: 784). NNQs are classed as partitives, and in fact Carter and McCarthy state that "a lot of/lots of are not determiners (they are best seen as partitive noun phrases in their own right"

(2006: 371). This is contrary to the argument set out in Paper 2, but Carter and McCarthy offer no further discussion of this classification, despite their use instead of the determiners *much* and *many*.

These last two items (LVs and NNQs) point to the problems that nomenclature can present. If we looked for "light verbs" and "non-numerical quantifiers" in the grammar's index, we wouldn't find them at all. A lexical index, which is included in Aarts (2010), and incorporated within the general index of Carter and McCarthy (2006) goes some way to overcoming the handicap to finding items that might be labelled differently according to different grammatical descriptions by allowing the user to search for typical, established examples such as *a lot of*. This option is not available in Huddleston and Pullum (2005), where the index is entirely based on categorial labels – although they do provide a Glossary to help the student with the grammatical labels used.

It is not surprising to find that there is limited coverage of the multiword items investigated in this research, both in the student and comprehensive dictionaries. They tend to be marginal in terms of membership of particular grammatical classes, and lower in frequency. The one example that occurs with notable frequency, LVs, demonstrates more characteristics of lexicalization than of grammaticalization, and therefore is not a core candidate for treatment in grammars. What this set of items does demonstrate, however, is a lack of systematic methods for dealing with emergent items that are of potential interest both to the learner of English, and the general reader seeking to understand the grammatical structure of the language. Grammars could give a more complete picture by:

- Using corpora to give sense of most commonly occurring examples of constructions
- Representing gradience based on multiple criteria for core and marginal members of classes
- Presenting of terminology/nomenclature so as to show how classes interrelate, and explaining alternative terms/classifications

3.2 Dictionary representation of multiword items

Dictionaries are more open to the possibility of language change than grammars, but only on the lexical level. The semantics of individual words govern the organization of dictionaries, so that multiword units and their evolving grammatical usages are

marginalized. In addition, the apparatus for describing the grammatical function of items that do not belong to clearly defined categories is limited in dictionaries. We will look first at how dictionaries present grammatical information, and then at how well served are some marginal examples.

3.2.1 Grammatical labelling and subcategorization

The traditional means of presenting grammatical information in a dictionary is through labelling. Yallop (1996) points out the limited range of grammatical description available in the category labels of most general native-speaker dictionaries, with extended grammatical information given only by some learners dictionaries in the form of subcategorical labels. He gives the example of the word *say* which in the native-speaker *Macquarie Dictionary* is given definitions according to the subcategory labels of *v.t.* (transitive verb), *v.i.* (intransitive verb), as well as *n.* (noun). In the advanced learners dictionary, the *Longman's Dictionary of Contemporary English),* extra grammatical information is given with the coding [S;U] for the noun, which tells us that *say* as a category of noun that is singular and uncountable. Some learners dictionaries go the other way, for example the *Collins Today's English Dictionary* (1995), abandoning all grammatical labelling, and providing instead the grammatical context of the word within its full-sentence definitions, as in the verbal definition "When you *say* something, you make a comment or statement, or ask a question" as opposed to the nominal one, "If you *have your say*, you give your opinion on something".

The question of subcategories is also raised by Partington who looks at how dictionaries deal with problematic terms such as the adverbs *rather* and *somewhat*, which he says are "notoriously difficult for dictionaries to deal with largely because they do not have a clear denotational meaning" (1998: 33) unlike, say, an adverb like *cautiously*. They belong to a subcategory of modifiers called "downtoners" by Quirk et al. (1985), but this distinction is not noted by dictionary labelling, where they are generally classified simply as adverbs. Because their function depends on their discoursal context, they can be hard to define consistently too. For example, the *Cobuild Dictionary* (1987) gives *somewhat* as a synonym of *rather* (but not vice versa), but defines *rather* as "to a certain limited, or slight extent" while *somewhat* has a different emphasis: "to a fairly large extent or degree".

Often multiword items with a grammatical function do not get labelled at all in dictionaries, particularly when their status has yet to be generally acknowledged. Of the items discussed in this study, only *a lot of/lots of* is labelled as a grammatical item, and there is inconsistency in its treatment. While the *Collins Cobuild Dictionary* (3rd ed., 2001) distinguishes the quantificational use from the pronoun and adverbial use, the *Oxford Dictionary of English* mislabels a determiner use as a pronoun – as noted in Section 3A, 3 (p.71).

3.2.2 Location of multiword items in dictionaries for learners and native-speakers

In addition to the problems of labelling, multiword units present particular challenges to the lexicographer as to where they should be placed in a dictionary. Even where they contain content words that have lost their semantic salience in the process of grammaticalization, as with CPs like *in spite of*, the content word still tends to have the function of acting as the headword for such items. Most headwords in dictionaries are single-word items, with compounds and phrases derived from them usually presented towards the end of the entry. Yet some multiword items can be given as headwords in their own right. Grammatical items commonly presented in such a way include *all round*, functioning as an adverb, adjective or preposition, or the conjunction/adverb *in case.*

The treatment of multiword items as phrasal instances of a particular headword – normally the content word within the phrase – presents particular problems for the set of items under discussion here. In the case of LVs, where the lexical content is carried by the noun, and the grammatical function by the verb – the question is raised as to what headword the LV belongs to. Learner dictionaries typically find more practical solutions to this particular problem than do general dictionaries. For example, the *Collins Cobuild English Dictionary for Advanced Learners* (3rd ed., 2001) provides a usage note to supplement the entry for common LVs like *have*:

> **Have** is used in combination with a wide range of nouns, where the meaning of the combination is mostly given by the noun.

The entry then gives an explanation of usage with an example, and follows it with a list of examples where the verb is used with variety of nouns in the same way:

> You can use *have* followed by a noun to talk about an action or an event, when it would be possible to use the same word as a verb. For example, you can say 'I had a look at the photos' instead of 'I looked at the photos.' Examples*: I went out and had a walk around…, I'll have a think about that…*etc.

This approach avoids the danger of isolating these multiword units within their separate entries, although it does not got to the lengths of providing a pull-out page for a grammatical feature, as, for example, the *Cambridge International Dictionary of English* does for grammatical features like auxiliary verbs (1995: 84-5).

The tyranny of the alphabet within the organization of dictionaries can have the effect of obscuring common grammatical features– even when they are tacitly acknowledged by similar treatment within the entry. See Table 6.3 below.

Table 6.3: Treatment of complex subordinators in a native-speaker and learner dictionary

***New Zealand Oxford Dictionary,* 2005 (NZOD)**	***Longman Dictionary of Contemporary English,* 5th ed. 2009 (LDCE)**
instant *n.***1**. a precise moment of time, esp. the present (*I told you the instant I heard*). **minute** *n*. **3c**. (foll. by *clause*) as soon as (*call me the minute you get back*) **moment** *n*. **3.** an exact or particular point of time (*I came the moment you called*)	**instant 2. the instant (that)** as soon as something happens: *The instant I saw him, I knew he was the man from the restaurant.* **minute 11. the minute (that) sb does sth** as soon as someone does something**:** *Tell him I need to see him the minute he arrives.* **moment 5. the moment (that) sb does sth** as soon as someone does something*: He said he'd phone you the moment he got home.*

In the native-speaker dictionary NZOD, the definitions do not point to the common subordinator function that is clear from the illustrative examples, or show that the headword is part of a multiword item. Only the entry for *minute* acknowledges a grammatical function, both by stating that it is followed by a clause, and, more implicitly, by using another CS, *as soon as,* to gloss the expression. There is a much more uniform treatment of the three items in the learner dictionary LDCE, with each being glossed by the same CS, and being treated as a phrase. There is, however, an inconsistency between the definitions that implies that *minute*/*moment* are likely to be followed by an agent performing an action ("someone does something"), whereas instant is simply followed by an even ("something happens"). This distinction is not borne out by the examples used.

It is true that such inconsistencies only become evident once the items have been identified as a set, which neither dictionary allows the user to do. However, I would argue that such identification would allow both the user to understand the grammatical function of these items better, and the lexicographer to develop a more systematic approach to describing them.

3.2.3 Dictionaries and grammatical gradience

Clearer categorization of all these phrases by their grammatical function would aid the lexicographer in treating the items consistently, and the dictionary user in being able to use them correctly. Yallop (1996) argues that grammatical labelling emphasizes the notion of discrete grammatical categories, which is at odds with the corpus-driven notion supported by linguists like Halliday and Sinclair, that grammar is probabilistic rather than categorial. The concept of gradience illustrated by this research also provides an argument against the application of a rigid classifying system.

As an alternative, Yallop proposes the use of illustrative (corpus) examples to present the range of typical uses. These are now extensively used in corpus-based learners dictionaries such as the Collins Cobuild and Longman series. It's possible that the future of dictionaries, for non-native speakers at least, may be based on a model like the *Louvain English for Academic Purposes Dictionary* where the dictionary is customizable according to the user's academic discipline and mother tongue background. The integration of a corpus into the search tools of the dictionary also allows users to identify a wider set of commonly occurring phrases and collocations than are identified in standard learner and general dictionaries. (Granger and Pacquot, 2010).

The examples of emergent CSs given above demonstrate that illustration alone is not necessarily sufficient to show where there are common functions across diverse items, particularly multiword ones. With this usage-based approach, learner dictionaries appear to present emergent grammatical uses more consistently than general dictionaries do, but the grammatical role tends to be implied by analogy with equivalent constructions rather than overtly stated. Also, this consistency of treatment isn't observable by the user because the items are not systematically linked.

Based on the patterns found in the data in this research, and the evidence provided by dictionaries, possible enhancements of their description of the grammatical features of language could include:

- A more sophisticated dictionary labelling system so as to mark grammatical subcategories, and multiword items. These labels might include ones like CP, CS, as used in this research, and a clear description of the labelling system would also be needed to avoid problems of nomenclature noted in 3.1.1 above. Such a system may well be unwieldy in a print, relying on the user to move between entry and label key, but it would be straightforward to implement in an online dictionary.
- Fuller grammatical explanations, independent of individual headwords, with examples that demonstrate similarity between category members. These could be presented as explanatory notes, in the preface or inserts, or additional links in an online environment, which would function as a means of bringing together items from different parts of the alphabet, and linking to each entry.
- Use of corpus data to determine membership of category based on frequency, invariability, equivalent structures [also for grammars]. Currently corpus data is used in dictionaries such as *Longman's Dictionary of Contemporary English* for such information as the difference in frequencies of items in spoken and written registers, which helps the user with appropriate context of usage, but doesn't assist in showing the systemic links between items.

4. Summary answers to research questions

Let us now review the three research questions posed in the introduction to this thesis.

i. How can synchronic corpus data be used to draw conclusions about the processes of grammaticalization for multiword units?

The corpus data provided evidence of variation across different instances of the multiword items investigated, in terms of their form and their context of use, which indicated different degrees of grammaticalization. Diachronic studies, even over quite a short timespan, are important to demonstrate structural changes such as the loss of the definite article for particular items such as *at risk of* (see Chapter 4), and over the longer term can be used to find historical shifts of context from ambiguous "critical" ones to "isolating"

contexts that show grammaticalization to have occurred (see Chapter 1, Section 2.3). But synchronic data can show structural variation across sets of items to point to potentially emergent grammaticalized forms, and can quantitatively demonstrate whether there is a preponderance of instances in which items occur in critical contexts where changes in grammatical function can potentially occur.

ii. Are the standard criteria for assessing grammaticalization more effective for some multiword units than others?

As shown in Section 2.2 of this conclusion, sets of criteria are needed to demonstrate grammaticalization effectively, but decategorialization proved to be the least prevalent in the multiword units that were found to have reached the highest degree of grammaticalization. Fixity and semantic bleaching were found in the widest range of items, although it should be noted that assessment of the degree of bleaching is rather subjective, and the CPs investigated showed quite variable signs of semantic bleaching, with some highly frequent and invariant CPs of location, such as *in front of*, retain the core sense of the noun. Syntactic reanalysis was a strong indicator of grammaticalization, and provided evidence for two of the four items investigated to be functioning as grammatical units.

iii. How should marginal lexical and grammatical items be presented in descriptive grammars and dictionaries?

The multiword items considered in this research can all be said to be marginal both to the lexis and grammar of English, and their treatment in grammars and dictionaries, as summarized in Section 3 of this conclusion, reflects this. Methods for improved coverage of such items have been suggested in Sections 3.1.1 and 3.2.3 above. These include the more systematic use of corpus data to identify emergent grammatical units, the expansion of grammatical terminology, and a more comprehensive labelling system to identify and cross-reference groups of items.

5. Limitations of current study

This research project has sought to develop a framework for assessing the status of grammaticalization of a set of noun-based multiword units. To do this it has looked at variation, between regions and registers, over a range of different corpora. Only one of these corpora, COHA, provided diachronic data, to show the movement towards fixity of

a number of CPs. Ideally, the historical development of each of these units could be demonstrated, across a range of registers and regions (not simply the written, American sources that COHA provides). Spoken data, in particularly, is particularly useful in showing the emergence of newly grammaticalizing units, and there is a lack of historical spoken corpora to provide this information. For lower frequency items, such as NNQs other than *a lot of*, and CSs, the ICE corpora provided only limited data, that could be supplemented by larger corpora – although those available did not have the parallel structure as the ICE corpora, so limited comparisons could be made across the larger corpora.

6. Future research

There is scope for future research into these and other N-based multiword units (for example the type nouns *sort of, kind of*, see Davidse and Brems, 2008) as well as multiword units with verbal functions (for example *the idea is* as an intention marker, see Krug and Schützler, 2013) to investigate the degree to which they have become grammaticalized, and the contexts in which grammaticalization occurs. Ideally, to meet the practical applications outlined in this conclusion, this research would be based on similarly structured corpora that allowed comparison over regions, register and time. This would then allow categorization and labelling of the items to be based on large volumes of complementary data so that the synchronic variability within multiword items can be properly calibrated within other notions of grammatical gradience and grammaticalisation. In the absence of suitable existing corpora, the recently-compiled GloWbE suite of corpora, providing large amounts of internet text across a range of English-speaking regions, provide a promising resource for the analysis of variation across regions and more formal and informal categories of online text, where new grammatical units might be expected to emerge.

Consolidated References

Aarts, Bas. 2007. *Syntactic Gradience: The Nature of Grammatical Indeterminacy*. Oxford: Oxford University Press.

Abeillé, Anne, Danièle Godard and Ivan A. Sag. 1998. Two kinds of composition in French complex predicates. In Erhard Hinrichs, Andreas Kathol and Tsuneko Nakazawa, eds. *Complex Predicates in Nonderivational Syntax*. New York: Academic Press, 1-41.

Aitchison, Jean. 1991. *Language Change: Progress or Decay*, 2nd edn. Cambridge: Cambridge University Press.

Akimoto, Minoji. 1999. Collocations and idioms in Late Modern English. In Laurel Brinton and Minoji Akimoto, eds. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. [Studies in Language Companion Series 47] Amsterdam: John Benjamins, 207-238.

Algeo, John. 2006. *British or American English? A Handbook of Word and Grammar Patterns*. Cambridge: Cambridge University Press.

_____1995. Having a look at the expanded predicate. In Bas Aarts and Charles F. Meyer, eds. *The Verb in Contemporary English: Theory and Description*. Cambridge: Cambridge University Press, 203-217.

Allerton, David. 2002. *Stretched Verb Constructions in English*. London: Routledge.

Andersen, Henning. 2006. Synchrony, diachrony and evolution. In Ole Nedergaard Thomsen, ed. *Competing Models of Linguistic Change*. Amsterdam/Philadelphia: Benjamins, 59-90.

Auer, Peter and Stefan Pfänder. 2011. Constructions: Emergent or Emerging? In Peter Auer and Stefan Pfänder, eds. *Constructions: Emerging and Emergent*. Berlin/Boston: Walter de Gruyter & Co., 1-22.

Bauer, Laurie. 1983. *English Word Formation*. Cambridge: Cambridge University Press.

_____ 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.

Beaman, Karen. 1984. Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse. In Deborah Tannen, ed. *Coherence in Spoken and Written Discourse*. Norwood: Ablex, 45–80.

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

Biber, Douglas, Stig Johannson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Bordet, Lucie and Denis Jamet. 2010. Are English prepositions lexical or grammatical morphemes? *Cercles*, *Occasional Papers*: 1-26.

Boye, Kasper and Peter Harder. 2012. A usage-based theory of grammatical status and grammaticalization. *Language* 88(1), 1-44.

Brems, Lieselotte. 2011. *Layering of Size and Type Noun Constructions in English*. Berlin, Boston: Mouton de Gruyter.

_____ 2012. The establishment of quantifier constructions for size nouns: a diachronic case study of *heap(s)* and *lot(s)*. *Journal of Historical Pragmatics* 13(2), 202-231.

Brems, Lieselotte and Kristin Davidse. 2010. Complex subordinators derived from noun complement clauses: grammaticalization and paradigmaticity. *Acta Linguistica Hafniensia* 42(1), 101-116.

Brinton, Laurel J. and Elizabeth Closs Traugott. 2005. *Lexicalization and Language Change*. Cambridge: Cambridge University Press.

Butt, Miriam. 1995. *The Structure of Complex Predicates in Urdu*. Stanford, California: CSLI Publications.

Bybee, Joan. 2010. *Language Usage and Cognition*. Cambridge: Cambridge University Press.

Cattell, Ray. 1984. *Composite Predicates in English*. (Syntax and Semantics 17). New York: Academic Press.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen, ed. *Spoken and Written Language: Exploring Orality and Literacy* Norwood: Ablex, 35–53.

Channell, Joanna. 1994. *Vague Language*. Oxford: Oxford University Press.

_____ 1990. Precise and vague quantities in academic writing. In Walter Nash, ed. *The Writing Scholar: Studies in the Language and Conventions of Academic Discourse*. Beverly Hills: Sage, 95-117.

Curme, George. 1935. *A Grammar of the English Language*. Boston: D.C. Heath & Co.

Davidse, Kristin, Lieselotte Brems and Liesbeth De Smedt. 2008. Type noun uses in the English NP. *International Journal of Corpus Linguistics* 13(2), 139-168.

Denison, David. 2001. Gradience and linguistic change. In Laurel J. Brinton, ed. *Historical Linguistics 1999: Selected papers from the 14th International Conference on Historical Linguistics, Vancouver, 9-13 August 1999* (Current Issues in Linguistic Theory 215) Amsterdam and Philadelphia PA: John Benjamins, 119-144.

_____ 2010. Category change in English with and without structural change. In Elizabeth Closs Traugott and Graeme Trousdale, eds. *Gradience, Gradualness and Grammaticalization* (Typological Studies in Language 90). Amsterdam/Philadelphia: John Benjamins, 105-128.

Diewald, Gabriele. 2002. A model for relevant types of contexts in grammaticalization. In Ilse Wischer and Gabriele Diewald, eds. *New Reflections on Grammaticalization*. Amsterdam: John Benjamins, 103–120.

Drave, Neil. 2002. Vaguely speaking: A corpus approach to vague language in intercultural conversations. In Pam Peters, Peter Collins and Adam Smith, eds. *New Frontiers of Corpus Research*. Amsterdam, New York: Rodopi, 25-40.

Edgren, Eva. 1971. *Temporal Clauses in English* Acta Universitatis Upsaliensis (Studia Anglistica Upsaliensa 9). Uppsala: Uppsala University.

Fillmore, Charles, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: the case of *let alone. Language* 64(3), 501-538.

Firth, John, R. 1957. *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Fischer, Olga. 2008. On analogy as the motivation for grammaticalization. *Studies in Language* 32(2), 336-82.

Giacalone-Ramat, Anna. 1998. Testing the boundaries of grammaticalization. In Anna Giacalone-Ramat and Paul J. Hopper, eds. *The Limits of Grammaticalization*,. Amsterdam/Philadelphia: John Benjamins, 107-128.

Granger, Sylviane and Magali Paquot. 2010. Customising a general EAP dictionary to meet learner needs. In Sylviane Granger and Magali Paquot, eds. *eLexicography in the 21st Century: New Challenges, New Applications*. Louvain-la-Neuve: Presses universitaires de Louvain, 87-96.

Grimshaw, Jane and Armin Mester. 1988. Light Verbs and θ-Marking. *Linguistic Inquiry* 19(2), 205-232.

Häcker, Martina. 1999. *Adverbial Clauses in Scots: A Semantic-Syntactic Study*. Berlin, New York: Mouton de Gruyter.

Halliday, Michael. 1994. *An Introduction to Functional Grammar*, 2nd edn. London: Edward Arnold.

Heine, Bernd. 2002. On the role of context in grammaticalization. In Ilse Wischer and Gabriele Diewald, eds. *New Reflections on Grammaticalization.* Amsterdam: John Benjamins, 83–101.

Hoffmann, Sebastian. 2005. *Grammaticalization and English Complex Prepositions: A Corpus-Based Study*. London & New York: Routledge.

Hopper, Paul J. 1987. Emergent grammar. *Berkeley Linguistics Society* 13, 139-157.

_____ 1991. On some principles of grammaticization. In Elizabeth Closs Traugott and Bernd Heine, eds. *Approaches to Grammaticalization*, Vol I. Amsterdam/Philadelphia: John Benjamins, 17-35.

_____1998. The paradigm at the end of the universe. In Anna Giacalone Ramat and Paul J. Hopper, eds. *The Limits of Grammaticalization*. Amsterdam/Philadelphia: John Benjamins, 147-158.

Hopper Paul J. and Elizabeth Closs Traugott. 2003. *Grammaticalization*, 2nd edn. Cambridge: Cambridge University Press.

Huddleston, Rodney. 1976. *An Introduction to English Transformational Syntax.* London: Longman.

Huddleston, Rodney. 1988. *English Grammar: An Outline*. Cambridge: Cambridge University Press.

Huddleston, Rodney and Geoffrey Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hudson, Jean. 1998. *Perspectives on Fixedness: Applied and Theoretical*. Lund: Lund University Press.

Hundt, Marianne. 2009. Concord with collective nouns in Australian and New Zealand English. In Pam Peters, Peter Collins and Adam Smith, eds. *Comparative Studies in Australian and New Zealand English: Grammar and Beyond,*. Amsterdam/Philadelphia: John Benjamins, 207-224.

Hunston, Susan and Gill Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins.

Jespersen, Otto. 1931. *A Modern English Grammar*, vol. 6. Heidelberg: Carl Winters.

Kaplan, Ronald and Joan Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In Joan Bresnan, ed. *The Mental Representation of Grammatical Relations*. Cambridge, Massachusetts: MIT Press, 173-281.

Kearns, Kate. 2002. Light verbs in English. Unpublished Generals Paper, MIT.

Kennedy, Graeme. 1987. Quantification and the use of English: A case study of one aspect of the learner's task. *Applied Linguistics* 8(3), 264-86.

Kortmann, Bernd. 1997. *Adverbial Subordination: A Typology and History of Adverbial Subordinators Based on European Languages.* Berlin, New York: Mouton deGruyter.

Kroll, Barbara. 1977. Combining ideas in written and spoken English: A look at subordination and coordination. *Discourse Across Time and Space,* Southern California Occasional Papers in Linguistics No. 5. Los Angeles: U Southern California, 69–108.

Krug, Manfred and Ole Schützler. 2013. Recent change and grammaticalization. In Bas Aarts, Joanne Close, Geoffrey Leech, Sean Wallis, eds. *The Verb Phrase in English*. Cambridge: Cambridge University Press.

Landau, Sidney I. 2001. *Dictionaries: The Art and Craft of Lexicography*, 2nd edn. Cambridge: Cambridge University Press.

Langacker, Ronald W. 1987. Foundations of Cognitive Grammar: Theoretical Prerequisites. Vol. 1. Stanford: Stanford University Press.

_____ 2008. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.

Leech, Geoffrey, Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Lehmann, Christian. 1985. Grammaticalization: Synchronic variation and diachronic change. *Lingua e Stile* 20(3): 303-18,

Levin, Magnus. 1998. Concord with collective nouns in British and American English. In Hans Lindquist, Staffan Klintborg, Magnus Levin and Maria Estling, eds. *The Major Varieties of English.* Växjö University, 193-204.

Live, Anna H. 1973. The *take-have* phrasal in English. *Linguistics* 95, 31-50.

Longman Dictionary of American English, 2nd edn. 1997. New York: Longman

Longman Dictionary of Contemporary English, 4th edn. 2007. *Harlow: Pearson Education.*

Longman Dictionary of Contemporary English, 3rd edn. 1995. London: Longman.

Macquarie Australian Learners Dictionary. 1997. Sydney: Macquarie Library.

Macquarie Dictionary, 4th edn. 2005. Sydney: Macquarie Library.

Mittins, William H., Mary Salu, Mary Edminson and Sheila Coyne. 1970. Attitudes to English Usage. London: Oxford University Press.

Mohanan, Tara. 1994. *Argument Structure in Hindi*. California: CSLI Publications.

Nevalainen, Terttu and Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. New York: Longman.

New Oxford Dictionary of English. 1998. Oxford: Oxford University Press

New Zealand Oxford Dictionary. 2005. South Melbourne, Vic.; Auckland: Oxford University Press.

Nickel, Gerhard. 1968. Complex verbal structures in English. *International Review of Applied Linguistics* 6, 1-21.

Nykiel, Jerzy. 2014. Grammaticalization reconciled: functionalist and minimalist insights into the development of purpose subordinators in English. *Language Sciences* 42, 1-14.

O'Donnell, Roy. 1974. Syntactic differences between speech and writing. *American Speech* 49, 102-110.

Oxford English Dictionary, 2nd edn. 1989. 20 vols. Oxford: Clarendon Press.

Pacquot, Magali. 2015. Lexicology and phraseology. In Douglas Biber and Randi Reppen, eds. *The Cambridge Handbook of English Corpus Linguistics*.. Cambridge: Cambridge University Press, 460-79.

Partington, Alan. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Peters, Pam. 1993. Corpus evidence on some points of usage. In Jan Aarts, Pieter de Haan and Nelleke Oostdijk, eds. *English Language Corpora: Design, Analysis and Exploitation* Amsterdam: Rodopi, 247-256.

Peters, Pam. 1999. Differing on agreement: A report on Langscape 3. *English Today* 15(2), 6-9.

Peters, Pam. 2012. Emergent conjunctions. In Sarah Chevalier and Thomas Honegger, eds. *Words, Words, Words, Philology and Beyond : Festschrift for Andreas Fischer on the occasion of his 65th birthday* Tübingen: Francke, 127–144.

Petré, Peter, Kristin Davidse and Tinne Van Rompaey. 2012. On ways of being on the way: lexical, complex preposition and aspect marker uses. *International Journal of Corpus Linguistics* 17(2), 229-258.

Pinker, Steven. 1999. *Words and Rules*. University of Michigan: Basic Books.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Poutsma, Hendrik. 1926. *A Grammar of Late Modern English*. Groningen: P. Noordhoft.

Random House Dictionary of the English Language, 2nd ed. 1987. New York: Random House.

Reid, Wallis. 1991. *Verb and Noun Number in English: A Functional Explanation*. London, New York: Longman.

Rensky, Miroslav. 1964. "Nominal tendencies in English". *Philologica Pragensia* 7: 135-50.

Schwenter, Scott A. and Elizabeth Closs Traugott. 1995. The semantic and pragmatic development of substitutive complex prepositions in English. In Andreas Jucker, ed. *Historical Pragmatics*. Amsterdam: John Benjamins, 243-273.

Sinclair, John. 1987. *Collins Cobuild English Language Dictionary*. Glasgow: HarperCollins.

_____1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press

Stein, Gabriele. 1991. The phrasal verb type "to have a look" in Modern English. *International Review of Applied Linguistics* 29, 1-29.

Stein, Gabriele and Randolph Quirk. 1991. On having a look in a corpus. In Karin Aijmer and Bendt Altenberg, eds. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London, New York: Longman, 197-203.

Sweet, Henry. 1891. *A New English grammar – Logical and Historical*. Approaches to Grammaticalization, Vol I [Typological studies in language 19(1)]. Oxford: OUP.

Traugott, Elizabeth Closs. 1999. A historical overview of complex predicate types. In Laurel Brinton and Minoji Akimoto, eds. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*,. Amsterdam/Philadelphia: John Benjamins, 239-260.

_____ 2008. "All that he endeavoured to prove was…": On the emergence of grammatical constructions in dialogual and dialogic contexts. In Cooper and Kempson, eds) *Language in Flux: Dialogue, Coordination, Language Variation, Change and Evolution*,. London: Kings College Publications, 143-177.

_____ 2012. The status of onset context in analysis of micro-changes. *Language and Computers* 76, 221-255.

Traugott, Elizabeth Closs & Heine, Bernd. 1991. Introduction. In Elizabeth Closs Traugott and Bernd Heine, eds. *Approaches to Grammaticalization, Vol 1*. Amsterdam: John Benjamins, 1-14.

Verde, Maria. 2003. Shedding light on SHED, CAST and THROW as nodes of extended lexical units. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery, eds. *Proceedings of the Corpus Linguistics 2003 conference*, 859-868. <http://ucrel.lancs.ac.uk/publications/CL2003/papers/verde.pdf> (5 May 2008).

Wierzbicka, Anna. 1982. "Why can you *have a drink* when you can't **have an eat*?". *Language* 58, 753-99.

Wischer, Ilse. 2000. Grammaticalization versus lexicalization: "Methinks" there is some confusion. In Olga Fischer, Anette Rosenbach and Dieter Stein, eds. *Pathways of change: Grammaticalization in English*. Amsterdam/Philadelphia: John Benjamins, 355-370.

Yallop, Colin. 1996. Grammatical information in dictionaries: how categorical should it be? In Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström and Catarina Röjder Pampehl, eds. *EURALEX '96 Proceedings: Papers Submitted to the 7th EURALEX International Conference on Lexicography in Göteborg, Sweden*, Vol. 2. Göteborg: Göteborg University, 507-511.