

**A journey through learner language:
tracking development using POS tag sequences
in large-scale learner data**

Geraldine Maria Mark

Submitted to the University of Limerick for the degree of Doctor of Philosophy

Supervisors: Dr Anne O’Keeffe and Prof Pascual Pérez-Paredes

Submitted to the University of Limerick, April 2022

Abstract

This PhD study comes at a cross-roads of SLA studies and corpus linguistics methodology, using a bottom-up data-first approach to throw light on second language development. Taking POS tag n-gram sequences as a starting point, searching the data from the outermost syntactic layer available in corpus tools, it is an investigation of grammatical development in learner language across the six proficiency levels in the 52-million-word CEFR-benchmarked quasi-longitudinal Cambridge Learner Corpus. It takes a mixed methods approach, first examining the frequency and distribution of POS tag sequences by level, identifying convergence and divergence, and secondly looking qualitatively at form-meaning mappings of sequences at differing levels. It seeks to observe if there are sequences which characterise levels and which might index the transition between levels. It investigates sequence use at a lexical and functional level and explores whether this can contribute to our understanding of how a generic repertoire of learner language develops. It aims to contribute to the theoretical debate by looking critically at how current theories of language development and description might account for learner language development. It responds to the call to look at largescale learner data, and benefits from privileged access to such longitudinal data, acknowledging the limitations of any corpus data and the need to triangulate across different datasets. It seeks to illustrate how L2 language use converges and diverges across proficiency levels and to investigate convergence and divergence between L1 and L2 usage.

Acknowledgements and thanks

I owe a huge debt of thanks to my supervisors, to Anne O’Keeffe, who I have had the privilege of working with for almost 20 years, and to Pascual Pérez-Paredes, not just for their insight, generosity of time, wisdom, guidance and encouragement, but also for the laughs and friendship.

I gratefully acknowledge the funding received from Mary Immaculate College through the Mary Immaculate College Doctoral Award. Many thanks too for the nurturing and encouraging community at MIC through support from members of staff and fellow students.

My thanks to Claire Dembry, Ben Knight and Mark Brenchley of Cambridge University Press and Assessment, for granting access to the Cambridge Learner Corpus for this research.

Thanks also to friends, colleagues and family: to Jeanne McCarten, Mike McCarthy and Ron Carter (sadly missed) for insight, generosity, encouragement and friendship; to Niall Curry for the discussions, enthusiasm when reading drafts, wisdom and more laughs, and to Odette Vassallo for patience, encouragement and friendship. To my family and friends for all their support along the way. Lastly, thanks as ever to Bernard, Rory and Niamh, for the endless stream of encouragement, patience, tolerance, cake, gin, snacks and belief.

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

Signed:

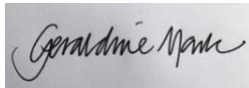
A rectangular box containing a handwritten signature in black ink. The signature is written in a cursive style and reads "Geraldine Nank".

Table of Contents

Abstract.....	2
Acknowledgements and thanks.....	3
Declaration.....	4
Table of Contents.....	5
List of Tables.....	8
List of Figures.....	12
Chapter 1 Introduction: Defining the landscape.....	15
1.0 Combining words to make meanings.....	15
1.1 Rationale for the study.....	16
1.2 LCR as description.....	20
1.3 Language acquisition or language development?.....	21
1.4 Using large-scale longitudinal data.....	22
1.5 Originality and relevance of the project.....	22
1.6 Research questions and summary.....	23
Chapter 2 Learner language development and learner corpora: the story so far.....	26
2.0 Introduction.....	26
2.1 Tracing theoretical underpinnings using learner corpora.....	28
2.2 Learner corpora and L2 development.....	30
2.3 L2 English developmental studies: complexity and accuracy of grammatical features.....	32
2.4 L2 developmental studies: n-grams, p-frames and multi-word sequences.....	36
2.5 L2 developmental studies: constructions.....	40
2.6 L2 developmental studies: using POS tags.....	42
2.7 Summarising: identifying the gaps.....	44
Chapter 3 Foundations and concepts.....	47
3.1 Language, frequency, structure and regularity.....	48
3.2 What is development?.....	55
3.3 Units of analysis.....	59
3.4 Summary.....	62
Chapter 4 From the bottom up: Data, tools and methods.....	64
4.0 Introduction: basic requirements.....	64
4.1 Largescale, longitudinal, levelled, homogeneous, tagged learner corpora.....	65
4.2 The CLC and the CLC sub-corpus.....	68

4.3 Mining the data: The tools and approach	74
4.4 Methodology discussed	81
4.5 Towards a methodology for bottom-up lexical and functional analysis	86
4.6 Summary	88
Chapter 5 Scanning the landscape: looking forward and looking back.....	89
5.0 Introduction	89
5.1 Frequency ranking and distribution: overall view	90
5.2 Overall view: a picture of convergence and divergence	92
5.3 Sequence types: A1 and C2.....	100
5.4 Overall sequence types: qualitative analysis of phrasal categorisation	107
5.5 Individual sequences: case study analysis of A1 and C2 #1 sequences.....	110
5.6 Scanning the landscape: general tendencies in POS tag sequence use	122
Chapter 6 Setting out	124
6.1 Focusing in: overall distribution A1, A2 and B1	126
6.2 A2 sequences: looking back and looking forward	128
6.3 A2 sequences: looking ahead to B1	131
6.4 A2 sequences: looking back to A1	135
6.5 Case study 1: Determiner + adjective + noun + preposition (DT JJ NN IN).....	139
6.6 Case study 2: PP MD RB VV pronoun modal adverb verb-base	150
6.7 A1 to B1: Setting out.....	158
Chapter 7 On the road, gathering pace from B1 to B2	159
7.1 Focusing in: overall distribution B1 and B2	160
7.2 B1 sequences	163
7.3 B2 sequences	170
7.4 Case study 1: pronoun + past-simple verb + to-inf + verb-base (PP VVD TO VV) ..	178
7.5 Case study 2: pronoun + past-simple verb + to-inf + verb-base (PP VVP TO VV)...	188
7.6 Beyond the 4-gram sequence: collocational patterning in case studies 1 and 2	196
7.7 Insights from comparing case studies 1 and 2: tense, context, register and theoretical alignment.....	201
7.8 B1 to B2: on the road	203
Chapter 8 Cruising: from C1 to C2.....	204
8.1 Focusing in: overall distribution C1 and C2	205
8.2 C1 sequences	208
8.3 C2 sequences	215

8.4 Case study 1: prep + noun + prep + det (IN NN IN DT)	223
8.5 Case study 2: -ed-form + prep + det + noun (VVN IN DT NN).....	230
8.6 Case study 3: det + noun + to-inf + verb-base (DT NN TO VV)	233
8.7 C1 to C2: Summary	240
Chapter 9 Discussion and conclusions: Mapping the routes	241
9.1 Recapping: aims of the study	241
9.2 RQ1 Is development in L2 writing observable through the frequency and distribution of POS sequences across proficiency levels?.....	242
9.3 RQ2 How does POS sequence usage develop across proficiency levels?	249
9.4 RQ3 Can existing frameworks for classification of language patterning account for a description of development in L2 writing?	253
9.5 Methodological and theoretical considerations.....	255
9.6 Current limitations and future avenues for investigation	258
9.7 Concluding remarks	260
References.....	261
Appendices.....	278
Appendix 1 Tasks at each exam level of the Cambridge mainsuite exams	278
Appendix 2 English Penn TreeBank tagset.....	279
Appendix 3 Sample of the master cohort of the top 1000 sequences at all levels and their rankings at other levels.....	282
Appendix 4 Lexical bundle classification (Biber <i>et al.</i> 2004)	283

List of Tables

Table 1.1 Development of adverb + adjective sequence across proficiency levels	17
Table 3.1 Top 10 ranked base verb form types and their raw token frequencies in the BNC written corpus.....	49
Table 4.1. Distribution of tokens across performance levels achieved and exams taken in the CLC main suite exam sub-corpus*	71
Table 4.2 Number of L1 backgrounds by level	72
Table 4.3 Top 5 POS tag sequences at A2 with frequency rankings at all other levels	79
Table 4.4 Top 10 4-gram POS sequences for each proficiency level in the CLC sub-corpus, by raw and per million word (PMW) frequencies	83
Table 4.5 relative distribution of Top 100 4-grams as types and tokens	84
Table 4.6 Top 20 4-gram POS tag sequences and lexical 4-grams in a sample of A2 data in the CLC sub-corpus	86
Table 5.1 Types and occurrences of POS 4-gram sequences per level	91
Table 5.2 Top 100 4-gram POS tag sequences as percentage of all 4-gram POS sequences ..	92
Table 5.3 Example of Top 10 sequences at A1 and their rank difference across all levels...102	
Table 5.4 Example of Top 10 sequences at C2 and their rank difference across all levels...103	
Table 5.5 Examples of A1 POS tag sequences not occurring at C2	105
Table 5.6 C2 POS tag sequences not occurring at A1, with lexical examples	107
Table 5.7 Structural classification of the top 100 4-gram POS sequences across levels: normalised occurrences (PMW)	109
Table 5.8 Breakdown of occurrences by level of .+pronoun+modal+verb	111
Table 5.9 Top 30 lexical exponents of the .+pronoun+modal+verb from A1 and C2.....113	
Table 5.10 Breakdown of occurrences by level of noun+preposition+determiner+noun.116	
Table 5.11 Top 20 most frequent lexical realisations of noun+preposition+determiner+noun at A1 and C2, categorised using Pattern grammar taxonomy (Hunston and Francis 2000)..118	
Table 5.12 Noun of noun pattern grammar meaning groups and examples from A1 and C2.	121
Table 5.13 (Semi)-fixed phraseological examples from the top 100 lexical exponents at C2	121
Table 6.1 Global scale descriptors for A1 A2 and B1 as defined by the Council of Europe	125
Table 6.2 Occurrences of POS 4-gram sequences across levels A1, A2 and B1	126
Table 6.3 Distribution of top 50 types across levels.....	126

Table 6.4 Top 50 4-gram POS sequences at A2, and their rank differences at A1 and B1 ..	131
Table 6.5 Core sequences: A2 sequences which are closely ranked at both A2 and B1.	132
Table 6.6 Emerging sequences: A2 sequences which are higher ranked at B1 than A2 (with rank difference).....	133
Table 6.7 A2 sequences decreasing in ranking at B1 (with rank difference).	134
Table 6.8 Core sequences: A2 sequences which are closely ranked at both A2 and A1.	135
Table 6.9 Emerging sequences used more at A2 than A1	137
Table 6.10 Decreasing sequences used less at A2 than A1	138
Table 6.11 Breakdown of occurrences by level of determiner+adjective+noun+preposition	139
Table 6.12 Top 20 most frequent lexical realisations of determiner + adjective + noun + preposition at A1, A2 and B1	141
Table 6.13 Lexical breakdown of DT JJ NN IN across A1, A2 and B1 categorised according to a lexical bundle framework (Biber <i>et al.</i> 2004).....	148
Table 6.14 lexical breakdown of DT JJ NN IN across the top 10 lexical instances at all levels	150
Table 6.15 Breakdown of occurrences by level of pronoun+modal+adverb+verb-base.....	151
Table 6.16 lexical breakdown of PP MD RB VV across the top 20 lexical instances at A1, A2, B1	152
Table 6.17 Lexical breakdown of PP MD RB VV across the top 11 lexical instances at A2	156
Table 6.18 lexical breakdown of PP MD RB VV across the top 20 and top 40-60 lexical instances at B1	157
Table 7.1 Global scale descriptors for B1 and B2 as defined by the Council of Europe	159
Table 7.2 Occurrences of 4-gram POS tag sequences across levels B1 and B2.....	160
Table 7.3 Occurrences and percentage distribution of the top 50 types across B1 and B2 ..	161
Table 7.4 Top 50 4-gram POS sequences at B1, and their rank differences at A2 and B2 ..	166
Table 7.5 Core sequences: B1 sequences which are highly convergent in ranking at both B1 and B2.	167
Table 7.6 Emerging sequences: B1 sequences which are higher ranked at B2 than B1 (with rank difference).....	168
Table 7.7 B1 sequences decreasing in ranking at B2 (with rank difference).	170
Table 7.8 Top 50 4-gram POS sequences at B2, and their rank differences at B1 and C1 ..	173
Table 7.9 Core sequences: B2 sequences which are highly convergent in ranking at both B2 and C1.	174

Table 7.10 Emerging sequences: B2 sequences which are higher ranked at C1 than B2 (with rank difference).....	175
Table 7.11 B2 sequences decreasing in ranking at C1 (with rank difference).	176
Table 7.12 Ranking of PP VVP TO VV and PP VVD TO VV across all levels.....	178
Table 7.13 Breakdown of occurrences by level of pronoun+past-simple+to-inf+verb-base	179
Table 7.14 Top 20 most frequent lexical realisations of PP VVD TO VV at all levels.	182
Table 7.15 Pattern grammar classification of verb + to-inf.....	184
Table 7.16 Breakdown of the past simple verb form by level and grammar pattern.....	185
Table 7.17 Top 20 most frequent lexical realisations of PP VVD TO VV at all levels.	190
Table 7.18 Breakdown of the present simple verb form by level and grammar pattern.....	193
Table 7.19 Top 20 collocations N-1 preceding PP VVP TO VV	197
Table 7.20 Top 20 collocations N-1 preceding PP VVD TO VV	199
Table 7.21 Breakdown of occurrences by level of pronoun+past-simple+to-inf+verb-base	202
Table 8.1 Global scale descriptors for C1 and C2 as defined by the Council of Europe	204
Table 8.2 Occurrences of POS 4-gram sequences across levels C1 and C2.....	205
Table 8.3 Distribution of top 50 types across levels	205
Table 8.4 Top 50 4-gram POS tag sequences at C1, and their rank differences at B2 and C2	210
Table 8.5 Core sequences: C1 sequences which are highly convergent in ranking at both C1 and C2.	212
Table 8.6 Emerging sequences: C1 sequences which are higher ranked at C2 than C1 (with rank difference).....	213
Table 8.7 C1 sequences decreasing in ranking at C2 (with rank difference).	214
Table 8.8 Top 50 4-gram POS sequences at C2, and their rank differences at C1	217
Table 8.9 Core sequences: C2 sequences which are highly convergent in ranking at C1.....	219
Table 8.10 Emerging sequences: C2 sequences which are higher ranked at C2 than C1 (with rank difference).....	220
Table 8.11 C2 sequences decreasing in ranking at C2 in comparison with C1 (with rank difference).....	221
Table 8.12 Breakdown of occurrences by level of prep + noun + prep + det.....	223
Table 8.13 Top 20 lexical exponents for IN NN IN DT for all six levels	226
Table 8.14 Functional taxonomy for IN NN IN DT sequences	227
Table 8.15 Breakdown of functions for the top 25 B2, C1, C2 IN NN IN DT sequences	228
Table 8.16 Breakdown of occurrences by level of -ed form + prep + det + noun.....	231

Table 8.17 Top 20 VVN IN DT NN sequences at C1 and C2.....	233
Table 8.18 Breakdown of occurrences by level of det + noun + to-inf + verb base.....	234
Table 8.19 Top 20 C1 and C2 lexical exponents of det + noun + to + verb	235
Table 8.20 Functional categorisation of Top 20 DT NN TO sequences	238

List of Figures

Figure 3.1 Frequency distribution of the top 50 base verb forms in the BNC written corpus.	49
Figure 3.2 Example of writing from a 5-year-old L1 English user	56
Figure 3.3 Example of writing from an A2 L2 English learner	56
Figure 4.1 Original ‘hyper-text’ branching framework of the CEFR (Council of Europe 2001, p.2)	67
Figure 4.2 CEFR reference levels (Council of Europe 2018, p.34).....	67
Figure 4.3 Range of Cambridge English qualifications benchmarked to the CEFR https://www.cambridgeenglish.org/Images/22695-principles-of-good-practice.pdf	70
Figure 4.4 Breakdown of languages represented in the CLC mainsuite data across all levels	72
Figure 4.5 Distribution by L1 background across levels	73
Figure 4.6 Applying a generic bottom-up iterative approach for retrieving and analysing POS tag sequences	76
Figure 4.7 Front view from A1 to C2 and rear view from C2 to A1	78
Figure 4.8. Representation of front and rear view comparison on individual subcorpora.....	79
Figure 4.9 Retrieval and filtering process for investigation of POS tag sequences.....	86
Figure 4.10 Description of bottom-up POS tag sequence approach	87
Figure 4.11 Illustration of the common ground between this study, traditional SLA and LCR	88
Figure 5.1 Distribution of the top 100 sequences in normalised frequencies across all proficiency levels	90
Figure 5.2 Front view from A1 to C2 and rear view from C2 to A1	93
Figure 5.3 Top 100 A1 and C2 sequences: convergence and divergence across all levels (by percentage).....	94
Figure 5.4 Top 100 C2 sequences: convergence and divergence across all levels	95
Figure 5.5. Representation of front and rear view comparison on individual subcorpora.....	96
Figure 5.6 Top 100 A2 sequences: convergence and divergence across all levels.....	97
Figure 5.7 Top 100 B1 sequences: convergence and divergence across all levels	98
Figure 5.8 Top 100 B2 sequences: convergence and divergence across all levels	99
Figure 5.9 Top 100 C1 sequences: convergence and divergence across all levels	100
Figure 5.10 Distribution of sequence types: structural categorisation.....	108
Figure 5.11 Overall occurrences (PMW) of noun-based and verb-based sequences in the top 100 sequences at all levels	109

Figure 5.12 Percentage distribution of modal verbs in top 100 lexical exponents of .+pronoun+modal+verb all levels	112
Figure 5.13 Strength of collocation of <i>I must</i> and following verb in C2 data.	114
Figure 5.14 Distribution of noun form groupings of noun+preposition+determiner+noun sequence across the top 100 lexical realisations	119
Figure 6.1 Percentage convergence of the top 50 sequences at A2 with their rankings at all other levels	127
Figure 6.2 Extract from the ‘diagram’ group from Pattern Grammar https://grammar.collinsdictionary.com/grammar-pattern/n-of-n_5	143
Figure 6.3 Functional categorisation of lexical sequences of determiner + adjective + noun + preposition DT JJ NN IN	144
Figure 6.4 Distribution of forms for determiner and preposition positions in DT JJ NN IN across all levels	146
Figure 6.5 Distribution of hits of <i>you couldn't come</i> in the A1 data.	153
Figure 6.6 Sample of the concordance lines with <i>you couldn't come</i> in the A1 data.	154
Figure 6.7 Distribution of hits of all PP MD RB VV occurrences in the A1 data.	155
Figure 6.8 Distribution of hits of all PP MD RB VV occurrences in the A2 data.	155
Figure 6.9 Distribution of hits of all PP MD RB VV occurrences in the B1 data.	156
Figure 6.10 Concordance lines of B1 occurrences of PP MD RB VV from PET 2008	157
Figure 7.1 Percentage convergence/divergence of the top 50 sequences at B1 with their rankings at all other levels	162
Figure 7.2 Percentage convergence of the top 50 sequences at B2 with their rankings at all other levels	163
Figure 7.3 PMW frequency of all occurrences of PP VVD TO VV by level	179
Figure 7.4 PMW frequency of the first 1000 and 100 types of PP VVD TO VV by level ...	180
Figure 7.5 Percentage of all types of the first 1000 and 100 types of PP VVD TO VV by level	180
Figure 7.6 Overall breakdown of PP VVD TO VV in the top 100 types, across levels, by past simple verb form	182
Figure 7.7 Extract from the V to-ing grammar pattern	183
Figure 7.8 A context-function-form overview of the PP VVD TO VV sequence	187
Figure 7.9 PMW frequency of the first 1000 and 100 types of PP VVP TO VV by level	188
Figure 7.10 Percentage of all types of the first 1000 and 100 types of PP VVP TO VV by level	189

Figure 7.11 Overall breakdown of PP VVP TO VV in the top 100 types, across levels, by present simple verb forms	191
Figure 7.12 pronoun + present verb distribution of most frequently occurring verbs in top 100 types	194
Figure 7.13 Context-Function-Form-Cotext description	202
Figure 8.1 Percentage convergence of the top 50 sequences at C1 with their rankings at all other levels	206
Figure 8.2 Percentage convergence of the top 50 sequences at C2 with their rankings at all other levels	207
Figure 8.3 Concordance lines of the <i>in order for the</i> sequence	230
Figure 8.4 Top 10 C2 DT NN TO sequences distributed across A2-C2 levels	236
Figure 8.5 Distribution of top 10 C2 DT NN TO sequences across all levels.....	237
Figure 8.7 N to-inf pattern and ability meaning group	240
Figure 9.1 Core sequences across all levels	243
Figure 9.2 Emerging sequences across all levels	245
Figure 9.3 Decreasing sequences across all levels.....	247

Chapter 1 Introduction: Defining the landscape

1.0 Combining words to make meanings

Put simply, we use language to send and receive messages to and from each other, about people, things and places. Language is a shared social construct. We collectively cooperate in interpreting *intention* in language use, in understanding the relationship of the *forms* of language – whole words, parts of words, phrases – with their attributed *meanings*. This co-operation is characterised by both variation *and* commonality. Our language stores are both unique and common. They are the dynamic sum of our own individual linguistic experiences, while at the same time sharing a core, common understanding of conventions, of how to put the building blocks of language together to make meanings that are universally understood. Proficient users of a language seem to effortlessly produce and receive streams of words while successfully mapping these to a vast range of meanings in different contexts. This apparent simplicity trivialises the complex knowledge, experience and understanding of an intricate and dynamic web of words and functions, in which words, grammar and meanings are bound together. For second or multiple language (L2) users, the enormity of mastering this is not to be underestimated. As Tyler and Ortega put it:

“There is little question that learning a language is one of the most complex accomplishments humans achieve. This is true for the first language learner and perhaps even more so for the second language learner.” (2018, p. 3).

So how is it that second language learners come to know these shared conventions? A usage-based (UB) theory of language learning would argue that this happens in the same way as any other learning, through frequency and relevance of experience. We come to know linguistic conventions as we encounter and use them, and structural conventions emerge from this usage. We subconsciously tune into and count form-meaning regularities in the input, calculating and recalculating their frequency and distribution as we meet more and more of them. In UB terms these form-meaning mapped conventions are called ‘constructions’ (Wulff and Ellis 2018). Constructions exist at all levels of complexity and abstraction from morphemes (e.g. -s to make nouns plural), to words (e.g. dogs, houses), to phrases and idioms (gone to the dogs) and syntactic frames (Verb *to* Noun, e.g. go/return/commute/journey/escape to London / the sea / your favourite place). The process of abstracting structural conventions from linguistic input, and the building of a repertoire of

sequences and their related meanings, has been shown to be central to first language acquisition. However, an exploration of how such sequences develop in L2 users remains relatively uncharted.

When we learn a language we are ever-gathering a repertoire of choices, a polysemy of suitable ways to express a multitude of interactions and intentions between people, things and places. From a UB perspective, as noted, frequency is a key element in this learning process. It is predicted that language learners subconsciously acquire first the language that they come across most frequently in the input that they are exposed to. At the heart of this study is an attempt to explore a methodology for capturing whether such a developing repertoire is observable globally in L2 English. It will do this by looking at the frequency and distribution of part-of-speech (POS) tag sequences, in a large-scale corpus of L2 writing and then by examining the lexical and functional usage of these sequences. It seeks to investigate whether there are sequences that are common to learners at particular proficiency levels and how sequence usage develops through different stages of proficiency. In examining L2 language as ‘product’ in this way, it hopes to contribute to our understanding of the process of language learning.

This introductory chapter describes the background to this study and defines its scope.

1.1 Rationale for the study

The seeds for this study came from previous research with Dr Anne O’Keeffe, profiling grammatical development across proficiency levels in L2 English (O’Keeffe and Mark 2017). One output from this research was the creation of the English Grammar Profile (CUP), an online interactive resource describing usage of over 1200 grammatical features across six proficiency levels. Using the 55-million-word Cambridge Learner Corpus (CLC), we observed three-fold developmental growth at lexical, grammatical and functional levels. As learners moved through proficiency levels, they used more words and put more structures together in such a way that phrasal and clausal complexity was seen to grow alongside an increase in lexical repertoire. This growth, particularly at an advanced level, became less about using more and more syntactically complex features but more about putting known structural combinations to new uses and meanings, and syntactic contexts. Table 1.1 illustrates this, using the combination of adverb + adjective as an example.

We first saw growth at a lexical level. Learners at the A1 beginner level could typically produce lexical sequences such as *very + nice/good*, e.g. *My home is very nice, My teacher is*

very good, where *very* is used as an intensifier of *nice* or *good*. At the next ‘step up’ in proficiency, (A2) learners increased their repertoire of candidates both for the adverb slot and the adjective slot, producing lexical sequences such as *very/really/so + important/happy/expensive*. Moving through higher proficiency bands, at B2 and C1 levels, the lexical repertoire increased, and included use of fixed or semi-fixed, co-selected combinations, e.g. *painfully obvious*, *highly unlikely* with specialised pragmatic functions.

	Adverb + adjective combinations	Examples	Syntactic sequences
A1	very + nice/good	<i>My home is very nice.</i> <i>My teacher is very good</i>	my + noun + is + very + nice/good
A2	very / really / so + important / happy / expensive	<i>I'm so happy to see you.</i> <i>It is really important for me</i>	pronoun + be + adverb + adjective + to inf pronoun + be + adverb + adjective + <i>for</i> + noun
B2 C1	very / really / so / quite / almost / extremely + adjective fixed sequences painfully obvious highly unlikely	<i>It is painfully obvious that the internet is crucial nowadays</i> <i>It is highly unlikely that the goods can vanish from your warehouse without being noticed.</i> <i>I'm absolutely certain that some solutions can be found</i>	<i>It is</i> + adverb + adjective + that-clause pronoun + be + adverb + adjective + that-clause

Table 1.1 Development of adverb + adjective sequence across proficiency levels

Overall, development appeared on a cline: from a limited use of one or two options in each slot, to an increased use of multiple candidates for each slot, alongside use of formulaic sequences with a greater degree of fixedness between the lexical items.

Alongside this, systematic growth in the co-text surrounding these two slots was seen. As well as an increase in the lexical items, there was development in the phrasal and clausal patterning, as seen in the examples in the fourth column in Table 1.1. For example, at A1 the adjective clause consists of a premodifying adverb + adjective combination, used attributively, as a complement of *is*: *my + noun + is + very + nice*. At A2 the adjective clause was both pre- and post-modified: adverb + adjective + phrase/clause. Beyond the A levels, other forms of post-modification emerged, e.g. *be + adverb + adjective + that-clause*. On a functional level, learners used the same basic form to express more and more meanings. At A1, the structure was used to express a simple descriptive function. By A2 it was used to express intensifying emotions and opinions, and by B2/C1 the form was employed to express modal meanings. At these higher levels we also saw evidence of semi-fixed structures to express pragmatically specialised meanings, such as stance, e.g. *It's highly unlikely that ...*, *I'm absolutely certain that ...*

In summary, as learners became more proficient, they were able to put the same syntactic pattern to multiple uses, while becoming aware of the collocational and colligational limitations of the patterns. Forms were first used at lower levels with a limited range of lexis and functions. Then followed a period of stabilisation where a form reached its syntactic 'developmental endpoint' (after Thewissen 2013), where stabilised forms were put to use with a greater range of meanings. Overall we found resonance both with the work of Thewissen (2013) investigating development, and with a usage-based notion of the development of a syntactic slot and frame system, to a fully abstracted system of 'constructions' (Ellis *et al.* 2016) (See Chapter 3 for a fuller account of constructions).

In this process of abstraction, it is asserted that a high percentage of language is stored as memorized wholes or formulae, clauses and clause structures that humans can retrieve as "automatic chains from the long-term memory" (Pawley and Syder 1983, p. 192). Studies of the sequential probabilities within the language system have illustrated how mastery of the language system involves not only knowing 'constructions' but also about knowing the strength of association within and around these sequences (Bybee 1998; Elman 2009; Ellis *et al.* 2015; Arnon and Christiansen 2017). Studies of formulae in L1 English have shown that up to 50% of language produced is formulaic (De Cock *et al.* 1998; Erman and Warren 2000). UB models assert that these formulaic sequences are learnt because they are both frequent and prototypical. This assertion is supported in Corpus Linguistic (CL) studies which, since the 1990s, through analysis of large bodies of text, have been providing

evidence of the existence of recurrent patterns of words and construction within language research. In pioneering CL studies, Sinclair described language usage in terms of the ‘idiom principle’, asserting that users of language have at their disposal “a large number of semi-preconstructed phrases that constitute single choices, even though they may be analyzable into segments.” (1991, p.110). This was in contrast to the ‘open choice’ in which units of language (words, phrases, clauses) are characterised as a series of slots and fillers, and “at each slot, virtually any word can occur” (1991, p. 109), with the only constraints on these slot choices being their grammatical category. (For further discussion of this see Chapter 3).

There is an abundance of research in the field of learner corpus research (LCR) on phraseology and formulaic language (Paquot and Granger 2012; Bestgen and Granger 2014). Research on the use of syntactic sequences in L2 writing has shown the existence of ‘constructions’ in the conventionalised form-meaning mapping sense, though studies are relatively scarce (Gries and Wulff 2005; Gilquin and De Knop 2016). A growing body of highly insightful, usage-based studies on the existence and use of verb argument constructions (VACs) in learner language (Ellis *et al.* 2016) is emerging (See Chapter 2). Such research has typically centered on single features or items, such as individual VACs or particular aspects of formulaic language (Ellis *et al.* 2016; Gilquin 2019; Römer and Berger 2019), items which have already been classified as ‘constructions’. As such this work has taken a top-down approach zoning in on pre-selected items, identified from previous corpus studies using L1 language data as a starting point for investigation. This brings me to a relevant point in this study.

‘Constructions’ and their usage in learner data can only be analysed if they are there. Within a usage-based framework, Ellis asserts that structures and their meanings emerge from usage (1998; 2011). If L2 users do make these structural generalisations, as our observations from the English Grammar Profile project suggest, is it possible to track the journey of emergence from words to sequences, from sequences to meanings, to pattern identification and abstraction? Using established known form-meaning mappings taken from L1 data, directs our gaze to what is under the spotlight. Can we find a way to look at the structuring and restructuring of language usage in L2 data as it happens without recourse to pre-selected known mappings? Developmental studies, taking a bottom-up approach, looking globally at structural sequences in large bodies of language *as they emerge* and develop in learner data are rare.

Access to the Cambridge Learner Corpus (CLC) affords this current research the privilege of a large body of proficiency-levelled written learner data. One of the challenges is to devise a methodology to best capture this journey of pattern-finding and abstraction of patterns, identifying ‘structural regularities’ (Ellis 2012) from a global perspective in large-scale data.

In this research I propose an approach that captures all sequences of POS tags in the CLC data. Following early corpus-based POS studies on small scale L2 corpora (Aarts and Granger 1998; Granger and Rayson 1998), I propose using POS tags to identify repeated structural sequences across proficiency levels, as a way to investigate abstraction of ‘structural regularities’. To do this I first explore the frequency and distribution of POS tag sequences, taking a corpus-driven, bottom-up approach (Tognini-Bonelli 2001), with no *a priori* list or preconceptions. For this reason, I use the term ‘sequences’ to refer to the combination of words or structural elements, not making the assumption that POS tag sequences are ‘constructions’. As chapters 2, 3 and 4 explain, this allows me to explore repeated POS tag sequences as regularities. I then draw on previous conceptual frameworks (e.g. lexical bundles (Biber *et al.* 1999), p-frames (Römer 2010), VACs (Ellis *et al.* 2016), grammar patterns (Hunston and Francis 2000) for mapping sequences to form patterns and meaning groups present in the data and to explore whether existing frameworks account for a description of any development identified.

In broad terms, this research is inspired by UB models of language acquisition, and aims to track the frequency and distribution of learner language at the part of speech level, and then to examine the lexical and functional manifestations of these sequences. The aim is to identify whether there are POS tag sequences that learners consistently rely on at differing levels of proficiency and the meanings they make from them, to investigate how POS tag sequence use differs and develops across proficiency levels, and to consider how a UB theory might account for any development.

1.2 LCR as description

The term ‘learner language’ refers quite generally to language produced by people learning a language other than their first language. Since the 1990s a growing body of studies on learner language has emerged through the use of learner corpora and learner corpus research (LCR). Over this time, the structural, morpho-grammatical studies which were typical of experimental second language acquisition (SLA) took a back seat in LCR against a foregrounding of frequency lists of keywords and lexico-grammatical patterning. As a result,

some thirty years on, we have at our disposal an abundance of descriptions of learner language, characterised by observations about the learner lexicon and phraseology, and in particular from a contrastive interlanguage analysis (CIA) perspective. Corpus studies on learner use of grammatical structures are comparatively scarce and those that exist focus on single linguistic items or closed grammatical classes (Gilquin and Granger 2015).

One of the key drivers in LCR claims to be to provide a greater understanding of SLA (Granger *et al.* 2015). LCR to date has provided us with insights on learner language as product, but has told us relatively little about the global acquisitional or developmental aspects of the process of this language learning (McEnery *et al.* 2019), from beginner to proficient users. In short we have an array of examples of the words and phrases that L2 learners use, and comparisons of these across cross-sectional datasets, but not much insight into what global development on a structural level looks like. This study aims to look closer at what the structural patterning in learner language can tell us about global language development.

1.3 Language acquisition or language development?

In her 2015 paper ‘Saying what we mean: Making a case for language acquisition to become language development’, Larsen-Freeman states that “There is no common end point at which all learners arrive” (2015, p.491). When we speak of ‘acquiring something’ there is an implication that at some point the acquisition is done or completed, that there is some kind of transfer. Relate this to language acquisition and the implication is that language is, in some way, a finite commodity, to be obtained. This assumption prevails in the classroom where there is often talk of having ‘done’ the past simple, or whether or not, at a given level of proficiency, learners ‘have’ a given structure. What does it mean to ‘have’ a structure? At what point in a language learner journey can we say that the language is ‘acquired’? It is not just learners of a language who do not reach a common end point but all users of a language, since language development is not static. Our novel daily encounters with language are new sources of evidence constantly contributing to the structuring and ‘restructuring’ (Ellis 2013) of our developing individual language systems, whether L1 or L2. We are in constant observation of the statistical occurrences of patterns in the language input, their collocational and colligational behaviours, their fixedness of usage and specificity of meaning.

This research seeks to explore if an investigation of POS tag sequences in large-scale learner language can contribute to our understanding of how this restructuring develops? What can it tell us about the language learning process? What implications does it have for teaching?

1.4 Using large-scale longitudinal data

More recently the analysis of learner corpora within SLA studies has begun to facilitate a growing body of usage-based research on the developmental nature of L2 acquisition (Ellis N. C. and Ferreira-Junior 2009a, 2009b; Ellis 2014; Tyler and Ortega 2016; Ellis *et al.* 2016). However, unlike in L1 acquisition studies, there is a dearth of longitudinal research designs in L2 studies. This is largely due to the lack of large-scale longitudinal L2 data, though its scarcity has led to the use of quasi-longitudinal data where variables such as year of study or proficiency level are used as a proxy for change over time. The reliability of this research design depends on the measurement of learners' proficiency, and increasingly, learner corpora which are linked to the Common European Framework of Reference (CEFR) are emerging as robust tools for research into L2 development. By using the CEFR levels (from beginner to advanced), data from different levels are comparable across variables such as proficiency, L1, age, etc. This study benefits from privileged access to the Cambridge Learner Corpus, a large-scale quasi-longitudinal learner corpus, of learner writing, from six levels of proficiency benchmarked to the CEFR.

1.5 Originality and relevance of the project

There exists a wealth of research into the cognitive and instructional process of L1 and L2 language acquisition. It is only recently, thanks to this scholarship (Ellis *et al.* 2016; Tyler and Ortega 2018, among others) that language corpora have begun to contribute to this.

Previous studies on L2 data using usage-based approaches focus predominantly on selection of established verb argument construction (VACs). In this study, I take advantage of the CLC dataset size as an opportunity to take a more open approach by looking at structural sequence development in general, from the bottom up. The aim is to see if there are overall structural patterns that might shed light into the developmental pathway(s) of learners from lower proficiency levels to higher proficiency levels, neither restricting the analysis to verb-centred categories, nor to specific levels or L1 backgrounds. The uniqueness, scale and quality of the data and the privileged access to it contribute additional originality to this project. Added to this, it is hoped that in response to a call for further analysis of learner data to be carried out this study will be timely in contributing to further understanding of this field.

This study attempts to trace a global development of form to meaning mappings as learners move from low to high levels of proficiency. It aims to examine how learners become proficient in employing a repertoire of exponents across a single sequence and a series of repertoires across a range of sequences. It seeks to investigate convergence and divergence between L2 usage at different proficiency levels in these respects and expects to throw light on the development process of language learning. It expects to point to exposure to frequencies in naturally-occurring language within and beyond the traditional classroom context as a key factor in the language development process.

1.6 Research questions and summary

This PhD study is situated at a cross-roads where elements of second language acquisition studies intersect with a corpus linguistics methodology which uses a bottom-up data-first approach to shed light on second language development. Using POS tag sequences as a starting point, searching the data from the outermost syntactic layer available in corpus tools, it is an investigation of grammatical development in learner language across the proficiency levels in the 52-million-word CEFR-benchmarked Cambridge Learner Corpus. It takes a mixed methods approach, first examining the frequency and distribution of POS tag sequences by level, identifying convergence and divergence, and secondly looking qualitatively at form-meaning mappings of these sequences at differing levels. It seeks to observe if there are sequences which characterise competence levels within the CEFR and the transition between levels and explores whether an analysis of the accumulation of their use at a lexical and functional level can contribute to our understanding of how a generic repertoire of learner language develops. It aims to contribute to the theoretical debate by looking critically at how current theories of language development and description might account for learner language development. It responds to the call to look at large-scale learner data, and benefits from privileged access to such longitudinal data, acknowledging the limitations of any corpus data and the need to triangulate across different datasets. It seeks to illustrate how L2 language use converges and diverges across proficiency levels. This is explored using the following research questions:

RQ1 Is development in L2 writing observable through the frequency and distribution of POS sequences across proficiency levels?

RQ2 How does POS sequence usage develop across proficiency levels?

RQ3 Can existing frameworks for classification of language patterning account for a description of development in L2 writing?

1.6.1 Overview of study

- This study is an investigation of development in learner language, taking a descriptive, observational approach to learner data.
- It aims to bring together elements of second language acquisition studies and corpus linguistics methodology, using a bottom-up data-driven methodology.
- It responds to the call to look at large-scale learner data, and benefits from privileged access to such longitudinal data, acknowledging the limitations of any corpus data and the need to triangulate across different datasets.
- It offers a methodology for approaching large-scale proficiency levelled learner data.
- It aims to examine how learners become proficient in employing a core repertoire of exponents across a single grammatical pattern and a growth in this repertoire across a range of grammatical patterns and across different datasets.
- It seeks to investigate convergence and divergence of usage between levels with respect to form and usage.
- It seeks to contribute to the theoretical debate by looking critically at how the evidence from the learner data might align with current theories of language development.
- It expects to point to exposure of frequency of language use in naturally-occurring language within and beyond the traditional classroom context as a key factor in the language development process.

1.6.2 Summary of chapters

In this first introductory chapter, *Defining the landscape*, I have set out the motivation and context for this research and briefly touched on some of the previous research in which it is situated. As the main title of the thesis implies (a journey through learner language) the study attempts to describe language development as a dynamic process along a pathway. It acknowledges that the departure points of language learning are as many and varied as there are language users but that there is global convergence along this pathway that can be observed and tracked. Chapter 2 (*Learner language development and learner corpora: the story so far*) describes previous relevant research within learner corpus research in relation to studies of development, and reveals the prevalence of phraseological studies, of a contrastive

descriptive nature, with a focus on learner output, illustrating a need for engagement between corpus linguistics and second language development to contribute to our knowledge and understanding of the learning process. Chapter 3 (*Foundations and concepts*) outlines considerations when approaching the data, through methodological and theoretical concepts. It deals with the diverse terminology found in previous research, and considers the range of units of analysis through which to view the data. It gives accounts of usage-based theory, pattern grammar, emergent grammar, idiom principle, constructions, continuous and discontinuous sequences of language, of grammar patterns, lexical bundles, p-frames and n-grams and their relevance to language learning.

Chapter 4 (*From the bottom up*) gives a detailed account of the data and methods, offering a novel bottom-up, data-driven approach for measuring frequency and development in large-scale learner data. Chapter 5 (*Scanning the landscape*) showcases the methodology in practice, illustrating an overall global perspective on development from the lowest proficiency level up - a front view - and from the highest proficiency level back - a rear view. Picking up on the journey metaphor the front view looks at what is up ahead in the road, from the perspective of a lower level learner, and the rear view considers what is left behind along the road as proficiency increases. Continuing with the journey metaphor, chapter 6 (*Setting out*) offers a description of the low level learner perspective, the starting point for language learning, and characterises the development from A level to B level. Chapter 7 (*On the road*) describes development from B level to C level, in which there is evidence of an established pathway where a great deal of linguistic territory is negotiated. In the final results chapter 8 (*Cruising*), I take a detailed look at development within the C level and beyond where the learner has built a wide and versatile linguistic repertoire and negotiates the landscape with dexterity and skill.

Chapter 9 brings the study to a close discussing the findings, insights, limitations and their implications.

Chapter 2 Learner language development and learner corpora: the story so far

2.0 Introduction

In its relatively short history, since the 1960s, the field of Second Language Acquisition (SLA) has explored and theorised, proven and counter-proven how and to what extent learners use second languages (R. Ellis 2021). Over time, and across opposing views, focus has been given to whether answers can ever be fully found to how languages are learnt or acquired, consciously or subconsciously. Given the cognitive nature of the process and the many other complexities involved, such as individual differences (Robinson 2002), issues of L1 transfer (Odlin 1989; MacWhinney 1992; Kellerman 1995), the role of L2 instruction (Long and Robinson 1998; Norris and Ortega 2000), task effects (DeKeyser 2001), it is hardly surprising that debates continue to rage. The elusive status of knowledge about how individuals actually learn or acquire languages, and to what degree, is reflected in studies that are sometimes conflicting or inconclusive. For example, research into the psychological processes of grammar learning has, on the one hand, generated a model for automaticity in the learning process (Logan 1990) meanwhile other studies partially support this but also show variation from the model (Palmeri 1997; Robinson 1997). Others look at the role of task in the process of learning and application of grammar rules and find this also to be a variable in attainment (DeKeyser 2001). From a usage-based perspective Ellis and Ferreira-Junior (2009a, p.188) refer to a myriad of factors raised in the literature in relation to the associative learning of morpho-syntactic constructions, such as form, frequency and salience; factors relating to prototypicality, generality, redundancy and surprise value; factors relating to the contingency of form and function; and factors relating to learner attention, such as automaticity and transfer.

There is, without doubt, a wealth of research into the cognitive and instructional process of L1 and L2 language acquisition but as yet language corpora have played a relatively small part (Ellis 2019b; McEnery *et al.* 2019). The use of learner corpora to address SLA research questions is long overdue (as noted by Gilquin and Granger 2015, and Myles 2015, among others) but it is an area which has recently been gaining momentum, bringing together the largely product-oriented focus of learner corpus research (LCR) to shed light on language learning processes (Granger 2021). Both SLA and LCR have learner language as their focus of study and yet their methods of analysis and objectives have rarely converged. This may be

largely due, as Gilquin and Granger (2015) note, to the fact that learner corpora are built for the purpose of addressing many research questions which have often not yet been conceived at the time of the data gathering, whereas, in contrast, SLA is hypothesis-driven and therefore takes a top-down approach, collecting data or designing experiments to test hypotheses. Driven principally by the work of Nick Ellis and associates (Ellis 2002; Ellis and Simpson-Vlach 2009; Ellis and Ferreira-Junior 2009a, 2009b; Ellis *et al.* 2015), and a cognitive turn in SLA, there has been an awakening to the usefulness of corpora in emerging usage-based (UB) theories in Second Language Acquisition (Ellis *et al.* 2015; Ellis 2019a; Pérez-Paredes *et al.* 2020) (See Chapter 3 for further discussion of UB theory). UB theorists, for whom frequency and distribution of language use are key, began to see the value of interrogating large bodies of data to investigate theories of language acquisition, moving from the small-scale hypothesis building of SLA and testing to exploring hypotheses and generalising from them on a larger empirical base (Myles 2015). Alongside this there has been a call for greater attention to SLA research questions in LCR (Hasko and Meunier 2013; McEnery *et al.* 2019). Learner corpora present a natural starting point for exploring frequency and distribution. Ultimately, in this call, there is a plea to engage beyond the contrastive paradigm, and a focus on error analysis, which characterises much of learner corpus scholarship to date, and to explore learner language usage in its own right. Römer and Garner (2022) argue that LCR can offer insights into some key areas of SLA research focus. They identify six core areas of SLA research focus, two of which concern an understanding the process of SLA and its development. It is this coming together of analysis of L2 usage through large-scale learner data and the relevance of findings for theories of language development that is at the heart of this study. It offers an approach for building a developmental picture by studying change in the written products themselves, across proficiency levels, using the powerful tools afforded by corpus linguistics (Durrant *et al.* 2021).

With this in mind, in this chapter I briefly chart the historical route taken through the early years of LCR in analysing learner language, while also looking at the types of data that have been used in learner corpus studies. I describe relevant studies that engage specifically with learner language development, touching on descriptions and definitions of development. I set out the pathways taken so far in the field and identify an additional pathway, a gap in the scholarship which this study explores.

2.1 Tracing theoretical underpinnings using learner corpora

2.1.1 *Learner errors and interlanguage*

Beginning in the 1960s, the *empirical* study of learner language came in and out of focus, across related frameworks of contrastive analysis, error analysis and interlanguage. In the 1990s, Corpus linguistics (CL), began to play an important role in keeping learner language in focus, especially through the work of Granger, and in advancing theoretical frameworks for its analysis (Granger 1996; Jarvis 2000; Gilquin 2008). Before this, early work on learner language is associated with structuralist scholars, especially Pit Corder and his 1967 work ‘The significance of learner’s errors’ which asserted the need to focus on learner errors not as ‘bad habits’ to be eradicated, but as a window into the learning process. Errors were seen as evidence of the learner’s strategies and processes. Learner language was described in various terms as ‘an approximative system’ (Nemser 1971) and ‘transitional competence’ (Pit Corder 1967, 1981). The emergence of the notion of ‘interlanguage’ (Selinker 1972) was the term that was eventually adopted and this marked a crucial conceptual milestone whereby learner language was given a name and independent status. Core to a broad definition of interlanguage is that it views learner language as an autonomous system, involving the building of a mental system of rules by the learner, and that what results is both different from the learner’s L1 and the target language system (Tarone 2018). And yet noticeably, despite its autonomous status, all three of these early terms offered in the early days hint at something unfinished, approximate, transitional.

2.1.2 *Descriptions of contrast*

Descriptions of interlanguage have an inherently contrastive focus, typically manifested either through L2 comparison with L1 ‘targets’ (L2:L1) or with L2 outputs from learners with a variety of L1 backgrounds (L2:L2). From this, the field of Contrastive Interlanguage Analysis (CIA) emerged (Granger 1996, 2015). The pioneering work of Sylviane Granger, and her associates, on the International Corpus of Learner English (ICLE) project (Granger 1994) from the early 1990s, brought a new intensity and rigour to the study of interlanguage because of the possibilities it opened up for the large-scale contrastive analysis of learner language. Within CL, the International Corpus of English (ICE) project offered a solid corpus framework for this contrastive work because ICLE was built within the design matrix evolved for the ICE project. It continues to play a central role in the field today. Initially the first iteration of ICLE began with 2.5 million words, from 11 L1 backgrounds growing in a

second version to around 3.7 million words from 16 L1 backgrounds; a more recent web-hosted version has grown to 5.5 million words from 25 L1s (Granger *et al.* 2020). Working within the design framework of ICLE, many new learner corpora were built with comparability in mind (Tono and Díez-Bedmar 2014). The international comparability of these data from so many different sources of learner language continues to have an immense impact on the field.

2.1.3 A dynamic system?

However, the view of interlanguage through such a comparative lens led to Bley Roman's notion of 'comparative fallacy' (1983) and the distorted notion of an idealised target L1, leading to descriptions of L2 usage in deficit terms of overuse and misuse. Controversially, one of Selinker's claims in this regard is that learner language fossilises, that it stops developing and never reaches a point where it aligns with L1 usage (1972). Larsen-Freeman counters this with a dynamic view of no one fixed homogenous target or end point (2005), viewing L2 language as distinct from the target language where "there will never be complete convergence between the two systems." (Larsen-Freeman 2006, p.592). Still widely used, the term 'interlanguage' has adopted a variety of meanings. The interpretation of the term used in this study aligns with a dynamic view of learner language as an ever-changing system, more in line with Swain's definition of 'linguaging' as "the process of making meaning and shaping knowledge and experience through language" (Swain 2006, p.98). Larsen-Freeman makes a case for dispensing with the term 'language acquisition' and replacing it with 'language development', reflecting the dynamic, changing nature of language (2015, p.491).

The sophistication of analytical approaches that came with LCR brought with it a broadening of the linguistic features under scrutiny along with analysis of a wider range of outcome variables affecting L2 learning and production (e.g. effects of task, L1 background, time spent in L2 context) (Granger *et al.* 2015). LCR became instrumental in moving the research gaze out from the morphological and syntactic focus at the heart of SLA studies, to include investigation and description of learner language at a phraseological and discursal level, mostly through Contrastive Interlanguage Analysis (CIA) and Contrastive Error Analysis (CEA) approaches. However as Granger herself describes (2015), when revisiting CIA some twenty years on, comparatively fewer CIA studies have centred on grammatical features, and those that have, have taken a lexically-based approach (Gilquin 2002; Aijmer 2002; Callies

2008). Grammatically-oriented research has tended towards single focus closed classes (Aijmer 2002; Diez-Bedmar and Papp 2008) and less frequently on POS-led studies (Aarts and Granger 1998; Granger and Rayson 1998; Tono 2000; Gilquin 2018).

The methodological landscape in LCR has been characterised by *corpus-based*, contrastive, cross-sectional, quantitative studies, applied to analysis of predominantly written advanced learner English (Callies 2015). This has resulted in a wealth of *description* of learner language dominated by *performance*, defined by overuse, underuse and misuse (in a CIA and CEA tradition) with respect to a perceived L1 norm, and has overshadowed longitudinal work in pursuit of descriptions of learner language *development* across proficiency levels (McEnery *et al.* 2019). This dominance comes partly from the CIA-driven focus in LCR, alongside issues of, on the one hand, the relative ease of certain types of written data collection and, on the other, the lack of dense, large-scale and truly longitudinal data that would afford a developmental view across proficiency levels. Another reason for the dominance comes from L2 corpus design, as the following section outlines.

2.2 Learner corpora and L2 development

2.2.1 Types of corpora

As noted above, learner corpora were not built with language acquisition hypotheses in mind; they were built to capture learner language as a variety, as part of the ICE project, and as a result we end up what Granger describes “all-purpose learner corpus” rather than “purpose-built” (Granger 2021, p.246). This approach has allowed the compilation of large corpora, but has not always meant that the data meets the needs of the enquiry. As McEnery *et al.* point out, learner corpora are numerous and growing but their variety is limited (2019). (See <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html> for an up-to-date list of available learner corpora). Early learner corpora have been criticised for inadequate background documentation (Myles 2015, 2021), though more recently through detailed corpus metadata, and the ability to filter corpora along a range of variables (e.g. L1 background, age, gender, task, proficiency) researchers have more control of these variables and can target specific SLA hypotheses (e.g. Römer 2019). For analysis of development, Myles outlines a need for “rich L2 corpus data which reflects the underpinning linguistic system of a specific learner or group of learners, at a specific point in time or at a range of different points in time.” (2015, p. 313). She also argues for spoken corpora as a preferred source of data for exploring development suggesting that written corpora are a better

representation of a learner's ability to memorise strings of language and explicit linguistic knowledge where revisiting and editing in the writing process allows for conscious reflection. Spoken corpora, by contrast, she maintains, "gives a better window into implicit knowledge" (Myles 2015, p. 314) because of the pressures of online production (See Chapter 3 for a brief discussion of implicit and explicit knowledge). Controlling for proficiency is key, and yet most learner corpora are cross-sectional, (typically data collected at one point in time from groups with multiple L1 backgrounds) often representing one proficiency level, particularly at upper-intermediate or advanced level (Myles 2021). Truly longitudinal data, charting change in language usage from the same individuals or group at multiple time points, capturing beginner to advanced levels of proficiency, while ideal, are scarce. As a result many studies have used cross-sectional corpora consisting of different learners or groups of learners over *different* proficiency levels (Meunier 2015a) described as pseudo-longitudinal (Johnson and Johnson 1999) or quasi-longitudinal (Granger 2002; Thewissen 2013). In these terms, in the absence of truly longitudinal data, change in proficiency level becomes a proxy for time (see also Chapter 3 for a discussion of development).

However, using such an approach to gather a longitudinal, developmental view is hindered by the slipperiness in defining proficiency. The criterion for calibration of level of proficiency of most learner corpora is the students' institutional status (i.e. the learner's position within an educational setting or institution, typically their year of study, Callies *et al.* 2014) and/or self-reported level. One of criticisms levelled at applying institutional criterion is that students at the same institutional level may not have a uniform proficiency level (Pendar and Chapelle 2008; Díez-Bedmar 2012; Tono and Díez-Bedmar 2014). More recently there has been an emergence of the use of the Common European Framework of Reference (CEFR) as a standardising measure for comparing learner corpus data development across a series of quality-driven attainment levels (see Hawkins and Buttery 2009, 2010; Díez-Bedmar 2012; Hawkins and Filipović 2012; Negishi *et al.* 2013; Thewissen 2013; Harrison and Barker 2015; Tono 2013; O'Keeffe and Mark 2017; Römer and Berger 2019; Gablasova *et al.* 2019a, 2019b) While not without its critics, the CEFR provides a more objective benchmark for proficiency than age or institutional level and has provided a means of tracking development through pseudo-/quasi-longitudinal corpora such as the EF Cambridge Open Language Database (EFCAMDAT), the Trinity Lancaster Corpus (TLC) and the Cambridge Learner Corpus (CLC), all of which are collections of data from different learners, at different times, across different levels of proficiency.

2.2.2 Focus of analysis

Having looked briefly at the suitability of learner corpora for studies of development, in the following sections I now consider where the focus of learner corpus studies has been. As already noted, early LCR has been characterised by detailed analysis of frequency lists of words, collocations, and lexico-grammatical patterning, compared either across different L1s or against the target L2. As a result, we have a wealth of description of learner language output as product characterised by observations about the learner lexicon and phraseology at static points in time. Corpus studies with a focus on structural or syntactic development have tended to look at single linguistic items or closed grammatical classes (Gilquin and Granger 2015). Those that explore global development are comparatively scarce. In the following sections, I review some of the research that has used learner corpora to contribute to an understanding of L2 English, as a window into the process of language learning. While using a diverse range of approaches and targets of focus all these studies have a unifying factor which is that they look at development of L2 language usage. The range of focus of previous relevant research can be broadly categorised by an eclectic mix of:

- Complexity and accuracy of grammatical features, e.g. Tense, aspect and modality, relative clauses
- Phraseology: n-grams, p-frames, lexical bundles, and multi-word sequences
- Constructions (Verb argument constructions)
- POS tags

What is immediately striking about these categories is the diversity of foci. Within this range of studies comes a raft of approaches and methods which may reflect what Durrant *et al.* posit to be a lack of a comprehensive and mutually agreed definition of the theoretical constructs in focus. This, they argue, results in studies of writing development “aiming at multiples fuzzy targets which are not only moving but being pushed in different directions by different people” (2021, p. 206). In sections 2.3 to 2.6 I take a descriptive look at some of these relevant studies and conclude by identifying the gaps in the research that this study is addressing.

2.3 L2 English developmental studies: complexity and accuracy of grammatical features

In chapter 1, I describe how O’Keeffe and Mark (2017) used the 55-million-word Cambridge Learner Corpus (CLC) to profile learner use of multiple grammatical features, traditionally covered in English language teaching classroom contexts, across six proficiency levels. In

this pseudo-longitudinal study, we observed development as an expanding repertoire of lexis and functions and pragmatic competence. As proficiency increased, learners put syntactic patterns to multiple uses, using an increasing lexical range, alongside displaying a greater awareness of the collocational and colligational limitations of a given pattern, as well as an understanding of specialised pragmatic meanings. Using the same corpus, Hawkins and Filipović (2012), and Hawkins and Buttery (2010) identified a series of ‘criterial features’, properties that were seen to characterise and point to L2 proficiency, at each of the levels of the CEFR as evidenced in the CLC. Their aim was initially to discover these properties at the level of lexis and grammar in order to identify “a set of linguistic features which provide the necessary specificity to CEFR's functional descriptors for each of the proficiency level” (Hawkins and Buttery 2009, p.159). These features are framed in positive or negative terms compared to their exemplification in L1 usage; where a feature corresponds with L1 usage (in the BNC) it is said to be a positive linguistic property and where it does not it is said to be a negative linguistic property. Different distributions of positive and negative properties distinguish different levels of proficiency. Negative linguistic properties demonstrate error types at a given level and these were seen to decrease as proficiency increased, particularly from B2 to C1 levels, which indicated development between these levels.

Murakami and Alexopoulou (2016) also used the CLC to evaluate the long-held view that there is a universal order of acquisition for English morphemes (Brown 1973; Dulay and Burt 1973). Using a subcorpus of the CLC, from seven L1 groups across five proficiency levels, they explored the development of six most frequently studied morphemes, from morpheme studies, (articles, past tense *-ed*, plural *-s*, possessive *'s*, progressive *-ing*, and third-person *-s*). Their findings demonstrated the role that large-scale corpora in LCR can play in examining SLA hypotheses. They concluded that there was a strong L1 influence in the accuracy of the morphemes, which affected different morphemes in different ways, and refuted the universal order of acquisition theory.

Using the 33-million-word EFCAMDAT, another large pseudo-longitudinal corpus, Alexopoulou *et al.* (2015) took a Natural Language Processing (NLP) approach to following the development of relative clauses, as an exemplar to demonstrate how large datasets can be used to study developmental trajectories across proficiency levels, playing a key empirical role in SLA research. Their findings indicate L1 effects and show how different types of relative clauses increase with proficiency. At 55 million words and 33 million words respectively, the CLC and EFCAMDAT are considered relatively large in LCR. These

studies demonstrate Granger's assertion that learner corpora of naturally-occurring language can facilitate studies that "lay claim to greater representativeness" (2009, p. 16) than previous SLA studies that relied on an experimental approach involving tasks such as acceptability judgements and gap-fills, with a small number of participants.

Thewissen (2013) provides an important 'crossover' study in which she looked longitudinally and contrastively at sample lexical and grammatical items, moving away from year of study as her cross-sectional point (in favour of the CEFR) and tracking learner development across four proficiency levels (B1, B2, C1, C2) specifically in relation to accuracy. She tracked the developmental pathways of error types in an error-tagged sample of the ICLE (comprising 223 learner essays from three L1 backgrounds – French, Spanish and German, amounting to 150,000 tokens) and observed strong progress (in terms of error decrease) between B1 and B2 levels. Contrary to Hawkins and Filipović (2012), she observed a plateauing of progress in relation to errors between B2 and C2 levels which she posits may "hide qualitative development" (Thewissen 2013, p.87). This is highly relevant to this study and is in line with O'Keeffe and Mark (2017) who found development in lexical and functional repertoire across proficiency levels when examining grammatical structures qualitatively at different levels of proficiency.

In another study also tracking errors, Meunier and Littré (2013) adopted an experimental approach alongside analysis of the Longitudinal Database of Learner English (LONGDALE) to study tense and aspect development in 38 L1 French students, with written contributions of one argument essay per year across three years, noting a decrease in errors over the three years. Using another small-scale longitudinal dataset Vyatkina (2013) tracked the developmental complexity of syntactic structures (e.g. coordinate and complex nominal structures per clause) in the same two (German L1) beginner learners, over four semesters. Combining POS tagging with manual checking and annotation, and concordance software, she identified points where target structures emerged in the data. Important to note here is that the developmental profiles of the two learners observed were different, with one learner showing a greater development than the other. Replications of these studies using larger data sets covering a wider range of learners, from different L1 backgrounds and would be welcome to be able to generalise from these findings.

In two studies using the Spanish component of the International Corpus of Crosslinguistic Interlanguage (ICCI) (Tono and Díez-Bedmar 2014) comprising 17,034 tokens, Pérez-Paredes and Díez-Bedmar (2019) and Díez-Bedmar and Pérez-Paredes (2020) also use a

combination of methods to measure syntactic complexity, across a range of age groups (grades 8 to 12). In the 2019 study they combine POS keyword analysis with automatic statistical complexity analysis software (TAASSC) (Kyle 2016) to look at noun phrase complexity and syntactic sophistication through analysis of verb argument construction (VACs). The 2020 study concentrates on the noun phrase, combining manual analysis with a TAASSC approach (Kyle 2016). Both studies reveal the affordances offered by different research methods and point to the analysis of complexity of the noun phrase as being “of great interest ... in terms of identifying development milestones in language acquisition” (Pérez-Paredes and Díez-Bedmar 2019, p. 101). Their findings also include the importance of “countable nouns, prepositional phrases, verbs and general adverbs in defining the transition from lower to higher secondary school” learning (ibid.).

In line with this focus on phrasal complexity, Biber and Gray (2011, 2016) offer an innovative framework which highlights the phrase and “compressed phrasal structure” as an equally important indication of grammatical complexity and development as clausal structure and dependence (Biber *et al.* 2020a). Alongside the phrasal complexity they point to the role of register and register awareness in the developmental process. The compressed phrasal structure takes centre stage in development as learners become more aware of its importance in writing. Biber *et al.* (2011, 2020a) offer five hypothesised stages of development which indicate a general trend towards a decreased use in dependent clause complexity and an increased use of phrasal complexity (from finite complement clauses to pre and post modified noun phrases). They argue that studies of development should encompass analysis of register, a focus on individual grammatical complexity features, including compressed phrasal structures as well as clause dependency (2020a, 2020b). They call for descriptions of writing development that include frequently used devices that mark the phrasal compressions such as premodification of nouns with attributive adjectives, and prepositional phrases as post-modifiers (e.g. *increase in inflation rates*). Several studies have borne this out, among which Staples *et al.* (2016), using the BAWE corpus (Heuboeck *et al.* 2008) found that complement clauses, relative clauses and adverbial clauses decreased in university academic writing while noun phrase modification increased. They point out the lack of the types of verb-based embedded clauses which are traditionally associated with ‘advanced’ writing. Many of these studies, as much of the research into structural complexity, tend to be dominated by learners at the higher end of the proficiency and with academic language production as the focus.

2.4 L2 developmental studies: n-grams, p-frames and multi-word sequences

2.4.1 Continuous sequences

In recent years LCR studies have been largely focused on frequently recurrent word sequences in L2 data. These sequences are referred to variously as lexical bundles, formulaic sequences, clusters, or multi-word units or multi-word sequences, phrasicon, n-grams, p-frames, coming under the umbrella of phraseology, as detailed in the comprehensive overview of L2 studies offered by Paquot and Granger (2012). This area of research has been driven by evidence of the fixedness or semi-fixedness of multi-word sequences in language and of the importance of formulaicity in gaining proficiency (Ellis 1996; Ellis and Simpson-Vlach 2009; Forsberg-Lundell 2021; Pawley and Syder 1983; Wray 2002). Among this scholarship are notable studies which combine quantitative and qualitative approaches to look at lexical bundle frequency and functional distribution. Many of these studies adopt the structural and functional taxonomy in the *Longman Grammar of Spoken and Written English* (LSWE) (Biber *et al.* 1999, 2004), encompassing a taxonomy of three functions (referential, stance marking and discourse organising). Simpson-Vlach and Ellis (2010) combine a frequency approach with a mutual information (MI) approach to measure the stretch of collocation between three-, four- and five word sequences as a way to observe the psychological validity of the sequences. Their findings focus on three categories of lexical phrases, those that are typical of (1) academic speech (2) academic writing (3) speech and writing which are then categorised using the Biber *et al.* (2014) taxonomy. Chen and Baker (2010) compared the use of recurrent lexical bundles in academic writing in both L1 and L2 writing in terms of their structures and functions. They examined data from L1 and L2 student essays as well as L1 expert writers. Adopting a hybrid quantitative approach with a detailed analysis of expanded concordance lines, they found that the use of structural and functional features in the lexical bundles of L1 and L2 writing were similar in the student writing, with a tendency to use more verb-based bundles than L1 expert writers who demonstrated a wider range of noun-based structures. However, these insights reflected point in time data and as such were not indicative of development. In a subsequent study, Chen and Baker (2016), they took a developmental perspective, this time benchmarking L1 Chinese data from the Longman Learner Corpus (LLC) to CEFR proficiency levels, and examined four-word lexical bundles across three sets of data, totalling just over 200,000 words, representing B1, B2 and C1 levels. Overall, they found that it was lower level learners (B1) who demonstrated more use of verb-based bundles, closer to conversational bundles,

reflecting functions of personal interaction and quantity, whereas the higher levels learners used bundles that were more characteristic of academic prose, with a higher proportion of noun and preposition-based bundles, reflecting a more impersonal tone. They argued that, at B2 level, learners start to become sensitive to the bundles that index differences in formality (Chen and Baker 2016). However development beyond C1 level is not investigated.

Vidakovic and Barker (2010) took both a quantitative and qualitative approach to examining four-word lexical bundles in 100 written texts from the Cambridge Skills for Life data (part of the Cambridge exam suite), across proficiency levels A1 to C1. Their results revealed that higher proficiency levels used a wider range of bundles and used them more frequently than at lower levels. Their functional analysis showed an increase in stance indicating and discourse organising use as proficiency increased. Staples *et al.* (2013) also used exam data Test of English as a Foreign Language (TOEFL iBT) to also look at lexical bundle frequency and usage across three proficiency levels (loosely described as low, medium and high). Across all levels they found stance-indicating bundles were most prevalent, and these tended to reflect the immediate context and topics of the exam prompts. In an additional level of analysis they looked specifically at variability of fixedness within bundle slots. Unlike Vidakovic and Barker (2010) their results showed a decrease in frequency of fixed bundles at higher levels which they propose was linked to a lower level reliance on bundles which came directly from the exam task prompt (Staples *et al.* 2013). This contributed evidence to support a developmental sequence in some aspects of formulaicity, as proposed by Ellis (2002), in which learners move from a heavy reliance on formulaic patterning at lower levels to ‘self-constructed’ sequences (Ellis 2002, p. 145) as proficiency increased (Staples *et al.* 2013). This suggests a move from formula to a slot and frame system (Ellis 1996, 2002). In this usage-based developmental model there is also a further step of ‘abstraction’, in which formulaicity plays a key role, increasing with proficiency (Ellis *et al.* 2016). This observation is also corroborated by Lenko-Szymánska (2014) who in a study using data from the ICCI six L1 backgrounds, spanning levels A1 to B2, exploring 3-gram lexical bundles found that formulaicity increases with proficiency. She found that bundles containing verb fragments were used at lower levels whereas bundles containing noun and prepositional phrases were seen at higher levels of proficiency. The issue as to whether these are stored as pre-fabricated units continues to be debated (Forsberg Lundell 2021).

Notwithstanding, studies on recurrent word sequences in L2 language have made a substantial contribution to our understanding of how learners put words together in sequences

and what these sequences are used for (Paquot and Granger 2012). With respect to the present study, two main issues emerge from this, the first relates to what constitutes a meaningful combination of words and the second to the fact that any structural generalisations made taking this approach (e.g. Chen and Baker 2010, 2016) are done first through a lexical lens, and run the risk of missing broader generalisations, the ‘structural regularities’ (Ellis 2013), discussed in chapter 1, that might emerge from a structure-first approach.

Following Durrant and Schmitt (2009), Granger and Bestgen (2014) combine strength of collocational patterning (based on association scores) with a POS tag approach to compare four pre-selected ‘syntactic bi-grams’ (Seretan *et al.* 2004), “directly adjacent words” (Granger and Bestgen 2014, p.3) across L1 data with L2 data from three ICLE L1 subcorpora at two proficiency levels (aligned broadly to CEFR B and C levels). Granger and Bestgen (2014) found that the lower level was characterised by high frequency collocations and that a distinguishing feature between the two levels was an increased use of adjective + noun combinations in the C level data. They acknowledge the limitations of proficiency level coverage and call for investigation across all levels of proficiency, including A levels but fall short of acknowledging that B and C levels as categories are broad and hide a complexity of difference and development within them (Hawkins and Buttery 2009, 2010; Díez-Bedmar 2012; Hawkins and Filipović 2012; O’Keeffe and Mark 2017). In a subsequent study, also following Durrant and Schmitt (2009) Bestgen and Granger adopted a CollGram technique, an automated process to analyse bigrams in the Michigan State University (MSU) Corpus of L2 writing taking both a longitudinal and pseudo-longitudinal approach to look at development of the two-word collocations over time. The corpus is made up of 171 essays from 57 participants. Time in this instance is one semester. Their findings showed a decrease in the number of high frequency bigrams in the longitudinal study, which they propose may be explained by UB theory and a progressive “deconstruction of multi-word units ... to more complex units like collocations and idioms” (Bestgen and Granger, 2014, p. 37). This study also underlines the need for larger corpora as well as highlighting the limitations that a focus on 2-word sequences reveal, aligning with Biber (2009a), and calling for further investigation of sequences longer than two elements.

2.4.2 Discontinuous sequences

The studies reviewed so far in this section have looked at continuous strings of words. Studies of discontinuous sequences in learner data are scarce. These sequences, recurrent strings in which not all words are fixed, are variously referred to as collocational frameworks, lexical frames phrase frames or p-frames. In a small corpus of L1 English writing, Renouf and Sinclair (1991) examined what they term “collocational frameworks” (e.g. *a + ? + of*) and identified common candidates to fill the variable (?) slot, from a selection of these frameworks. Studies, such as Renouf and Sinclair (1991), Biber (2009b) and Römer (2010), take what Biber subsequently terms a ‘hybrid’ approach, first extracting all common lexical bundles (in Biber 2009b’s case from the *Longman Spoken and Written Corpus*) and then investigating those four-word patterns for discontinuous sequences (e.g. *in the ? of*) which they call lexical frames. As Gray and Biber (2015) point out, the bundle is the starting point, and assumes that all discontinuous sequences or lexical frames will correspond to one frequently occurring continuous bundle. To investigate this further Gray and Biber (2013) adopt a more corpus-driven approach in which they identify all recurrent discontinuous bundles from scratch, bypassing the bundle extraction, and explore the extent to which each slot is variable. They then compare these to the discontinuous sequences in the hybrid study (Biber 2009b). They note that the existence of frequently occurring discontinuous sequences which are not connected to any particular bundle, thus demonstrating the need to examine discontinuous sequences in their own right as linguistic building blocks. Their findings reveal that the frames that appear most frequently in academic writing consist of function words (e.g. *in the ? of, the ? of the*) (Gray and Biber 2013).

In a quantitative and qualitative study of L2 writing, across five proficiency levels, Garner (2016) examines the German subsection of the EFCAMDAT for p-frames. Taking the hybrid approach, first extracting four-word n-grams or bundles, the bundles are filtered for sequences that are similar except for one word in the same slot; for example *as well as the, as far as the, as soon as the*, are combined into the 4-frame *as ? as the* (2016, p.38). After further filtering, which included the removal of any frames which did not constitute “meaningful units”, the p-frames are categorised into fixed, variable or highly variable types, and their structural (after Gray and Biber 2013) and functional (after Biber *et al.* 2004) characteristics classified. Garner’s findings showed that more proficient learners introduced more variability into their frame usage. A distinct increase was seen in the variability between B2 and C1 learners. Garner concludes that lower proficiency level learners rely

more on fixed type frames, whereas higher level learners employ a greater range of phraseological items. Taking a usage-based perspective, they account for these results by proposing that higher level learners would have had more exposure to English, in a wider variety of communicative contexts, and therefore would have encountered more p-frame exemplars with the effect of “entrenching p-frames in the learners’ linguistic inventories” (2016, p. 49).

This section has described studies that move along a lexical-structural continuum, moving in the direction of structure, in an attempt to get at structural regularities. As discussed, studies of learner language have been dominated by a phraseological approach, have tended to rely largely on small data sets, of higher level learners or with a restricted range of L1 backgrounds. Those that are moving towards structural generalisation are still driven first by lexis. Section 2.5 is concerned with developmental research which site the construction at the heart of the analysis.

2.5 L2 developmental studies: constructions

In chapter 1, I briefly described how, from a UB perspective, conventionalised pairings of forms and meanings (Langacker 1987), termed ‘constructions’, are central to both L1 and L2 development. Constructions exist at all levels of abstraction ranging from morphemes, e.g. affixes like *in-* in *incredible*, to words to phrases to syntactic frames, such as the ditransitive construction, give something to someone, carrying a meaning related to ‘transfer’. Over the past twenty years studies examining development in learner data through construction use have been emerging with increasing frequency. Early research centred around small-scale longitudinal data. Various notable studies investigating negation (Eskilden 2012; Eskilden and Cadierno 2007), constructions with *can* (Eskilden 2009), questions (Eskilden 2015) and motion constructions (Li *et al.* 2014) centered around the same two Spanish learners, over four years.

Subsequent studies examining verb argument constructions (VACs) became synonymous with the work of Nick Ellis and associates. VAC studies are centred around the verb and its complementation patterns. The verb is seen as the predominant predictor of sentence meaning over other word classes because it is central to basic human experiences (someone causing something, moving something somewhere, doing something, having something, affecting something, changing something) and carries a heavy meaning load. Perek notes that “it is precisely for this reason that more so than other content words, verbs are rarely uttered in

isolation but are usually accompanied by certain other words” (e.g. the direct or indirect object) (2015, p. 1). In two of these early VACs studies, Ellis and Ferreira-Junior (2009a, 2009b), investigated three VAC types VL (verb locative, e.g. *go somewhere*) VOL (verb object locative, e.g. *put something somewhere*) and VOO (verb object or ditransitive, e.g. *give someone something*) in seven learners with L1 Italian and Punjabi. Their findings illustrated how acquisition of the VACs was affected by their frequency and their prototypicality. (Following a Zipfian-like distribution, for each VAC there was one ‘prototypical’ verb which had the lion’s share of the occurrences, while also carrying the prototypical meaning of the construction, for example the verb *give* is the most frequently occurring verb in VOO, *put* in VOL, etc.; see also Chapter 3 for a more in-depth discussion of prototypicality). Ellis and Ferreira-Junior (2009a, 2009b) found a strong relationship between the frequency of the verbs in the input the learners received and the frequencies of the verbs they used.

Many of the studies on constructions in L2 language in larger data were collected as part of experimental methodology, investigating constructional knowledge in advanced level learners only and do not investigate language development (Ellis *et al.* 2014; Römer *et al.* 2014; 2018). Two studies (Römer and Gardner 2019; Römer 2019) seek to address these limitations. The first investigates five VACs constructions, previously found to occur frequently in L2 spoken data (Römer *et al.* 2014, 2018). Using an L1 Italian and Spanish subcorpora (c. 1 million words) from the Trinity Lancaster Corpus Sample (TLCS), a cross-sectional spoken corpus from Trinity exams, Römer and Garner (2019) analysed use of the five constructions, to gain an insight into development of verb construction knowledge, comparing the findings with L1 usage using the BNC as a benchmark. They investigated usage across three tasks, and, dependent on the task, across two or three proficiency levels, with a predominance of data from the higher end of the proficiency scale. They observed strong consistency in the choice of lead verbs for each VAC, suggesting that learners at all levels are sensitive to frequency of usage and have an awareness of appropriate candidate verbs for the verb slots. They found that as proficiency increased, it aligned more with the L1 data; the distribution of usage in the C1/C2 data, compared to distribution in the B1 data, was found to be closer to the BNC and the variety of verb forms for each VAC in the higher level learners B2 to C2 was seen to be closer to the L1 data than in the lower level learners. Overall, there was evidence of development of VACs usage from a small set of fixed patterns

to a larger set of more varied patterning, with VAC usage become more predictable and more Zipfian as proficiency increased (Römer and Garner, 2019).

In the second of these studies, Römer (2019) takes a more comprehensive view of VACs usage. Using the L1 German learner data (c. 6 million words) from the pseudo-longitudinal EFCAMDAT (Alexopoulou *et al.* 2015), she explores all VACs used as they emerge, from A1 to C1 levels of proficiency, rather than investigate a preselected subset of VACs. In this study using the COCA as a proxy for L1 usage, she observes that the verbs associated with particular VACs move closer to L1 usage as proficiency increases. Aligning with previous studies, from both individual learners and bigger groups with the same L1 background (Eskilden, 2012; Eskilden and Cadierno, 2007, Römer and Garner 2019) Römer (2019) finds that lower level learners make use of a more restricted range of fixed verb associations which give way to a wider variety of associations at the higher levels of proficiency. In Römer's study, opportunities for future research are identified, including both qualitative and quantitative analysis beyond one L1 subset, and involving an investigation into how the VACs develop in relation to dominant verb associations and possible development of their functional characteristics.

While making great strides towards our enhanced understanding of emerging knowledge, the focus on the verb in VACs studies does not account for any development beyond the verb clause. Given the importance of the noun phrase in relation to proficiency and development identified in 2.3 above, it would seem the net for capturing structural development beyond the verb clause needs to be cast wider. Continuing along the lexical-structural continuum, in the following section I review studies that have made use of POS tag sequences in learner corpora as a means to capture a wider range of recurrent structural patterns.

2.6 L2 developmental studies: using POS tags

Using a POS tagged corpus, it is possible to search for POS n-grams (sequences of POS tags), from which the high frequency of the sequences can inform “expressions of syntactic patterning” (Kennedy, 1996), phraseological patterns (Granger and Bestgen 2014) and potentially identify constructions (Capelle and Grabar, 2016). Although it should be recognized that not all POS tag sequences are constructions (Gilquin 2018) in the traditional form-meaning mapping sense, POS n-grams offer a holistic approach in exploring the most commonly used syntactic patterns without having a preselected set of constructions as a starting point.

Two early corpus-based studies used POS tags to investigate learner language (Granger and Rayson, 1998; Aarts and Granger 1998). Both compared POS tag use in the ICLE corpus (comprising academic essays) with the L1 LOCNESS (Louvain Corpus of Native English Essays), a collection of similar register essays from American L1 English writers. Using an automatic profiling technique, Granger and Rayson (1998) compared the use of single grammatical tags across the LOCNESS and L1 French learner data in the ICLE, and found that the writing in the L1 French data displayed characteristics of spoken language. Register studies had previously focused on frequencies of single grammatical categories (Biber 1988), and while analysis of sequences of categories had been seen in authorship studies and literary text analysis, it had not yet been applied to L2 studies. With a relatively small sample of 150,000 from the ICLE corpus, Aarts and Granger (1998) compared the essay writing of Dutch, Finnish and French advanced learners of English with the LOCNESS (Louvain Corpus of Native English Essays) by looking at POS tag sequences. Trigrams were extracted and the ranking and the frequency counts of the top 20 in each of the four subcorpora were compared. The lexical sequences underlying two of the patterns were examined. Their findings showed both evidence of language patterns that were convergent in all three of the L2 subcorpora as well as L1 specific patterns, characterising the language of learners from different L1s. They also found universal deviation between the 3 L2 datasets and the L1 LOCNESS data. Most relevant to this present study, they point to tag sequence extraction as means of gaining new insights into L2 grammar, though their cross-sectional design has limitations when looking at development, and points to a need for longitudinal data. Granger and Rayson's 1998 study proposed that through automatic profiling their POS tag approach would "help researchers form a quick picture of the interlanguage of a given learner population and that it opens up interesting avenues for future research". (1998, p.131)

Increasing sophistication in corpus tools has meant that corpora can typically be tagged and parsed automatically, allowing for simple extraction of POS tag sequences. This offers the opportunity to take a multi-level, bottom-up data-driven approach to longitudinal data to investigate first the sequences of POS categories learners are putting together, at different levels of proficiency, and subsequently the lexical selections that are being made for these sequences.

Studies of constructions have already offered insights into this co-selection of parts of speech and lexis (see 2.5 above) at varying levels, however, as has been discussed, the constructions or sequences in focus are preselected by the researcher. In contrast, Gilquin (2018), following

Aarts and Granger (1998), took a corpus-driven approach, with no *a priori* selection, to explore the use of tag sequences as means to identify constructions in spoken L2 data, and to establish construction that “are likely to be entrenched in the EFL constructicon” (Gilquin 2018, p.2). Using raw text versions of LINDSEI and its L1 counterpart (LOCNEC), Gilquin extracted the most frequent POS tag sequences of any length from both data sets. Several insights emerged. The study revealed that the top four sequences were the same in both data, and that the top 30 (25 of which were shared in both data), perhaps unsurprisingly, were basic structures (e.g. determiner + noun, pronoun + verb, preposition + determiner, pronoun + verb) of 2- and 3-grams. These sequences did not always correspond to constructions in a form-meaning mapping sense, (e.g. *on the, was very*) and required another part of speech to ‘complete’ them. She noted the prevalence of the noun phrase both in incomplete (noun + preposition, e.g. *exam about*), and complete (determiner + noun *a bird*) sequences containing nouns as well as in sequences not containing nouns but which prime nouns for completion (preposition + determiner, e.g. *on the*). Overall the findings pointed to a similar picture of POS tag distribution in both the L1 and L2 data and Gilquin maintains that, approaching constructions through a POS tag lens and grouping sequences at this level of syntactic abstraction allows for generalisations about entrenchment in the learner ‘constructicon’. However, the study stops short of exploring the lexical instantiations of the tag sequences, of which careful examination, Gilquin concedes, is essential to their interpretation. In fact, Gilquin suggests a combination of POS extraction with examination of their lexical exponents to fully understand their usage. Additionally, the approach is restricted to a cross-sectional view of the learner data and does not address how the distribution and frequency of these POS tag sequences might emerge and evolve across proficiency levels.

2.7 Summarising: identifying the gaps

UB models of language acquisition have gained ground in recent years, particularly in the field of first language acquisition (FLA) (Tomasello 2003), and, even more recently, are gaining traction in second language acquisition (SLA) studies (Ellis *et al.* 2016). Within this context, Ellis *et al.* (2015) highlight the role that learner corpus data must continue to play in providing important insights into SLA, particularly in the use of constructions (see Chapter 3). They discuss the need for a range of types of methods to triangulate with learner corpus research (LCR) in the investigation of SLA and FLA and point to a dearth of substantial learner data. Learner corpora, they say are “essential in showing the evidence of learner

formulaic use, and dense longitudinal corpora allow the charting of the growth of learner use” (2015, p.358).

Key drivers include the *push* of the availability of more dense learner corpora of different types via online platforms and the *pull* of the demand from SLA researchers to expand beyond experimental evidence of acquisition by using large scale samples of learner language from corpora, especially in relation to usage-based approaches to second language acquisition.

However, there is a danger that some key conceptual and methodological considerations are being glossed over in the rush to the data. This research considers that, in the emerging body of corpus SLA work, there is an emphasis on the analysis of phraseology and formulaicity in cross-sectional research designs, alongside lesser attention to syntactic *development* in learner language. Studies have often been driven by usage-based constructions which are defined in a way that methodologically presupposes a top-down theory-driven approach, characteristic of SLA, and forecloses on the traditional data-driven bottom-up approach associated with corpus linguistics. Within this context, we find the emergence of substantive and convincing work on Verb-Argument Constructions (VACs) which prescribes findable items in a top-down generic manner. At the same time, research on L2 phraseology in the contrastive research tradition has systematically found that L1 and L2 speakers use formulaic sequences differently, L2 use being characterised by a "mixture of underuse, overuse and misuse" (Paquot and Granger 2012, p.136). Because of the nature of the data available until this point, conclusions are often drawn from small datasets (by contemporary corpus standards) and from a limited number of L1 backgrounds. Results are usually viewed contrastively while taking an unquestioning stance on the assumption that a First Language Acquisition (FLA) research paradigm transfers neatly to SLA and thus ignoring the fact that L2 users already have an established L1 system and prior knowledge of an inventory of constructions from their L1.

Ellis (2019, p.51) called for “big-data corpus investigation of representative language use in different sociocultural institutions and communities of practices and how these change over time”. Clearly there have been huge strides made in understanding both L1 and L2 usage from a VACs perspective, and insights from the Contrastive Interlanguage Analysis (CIA) tradition on phraseology and lexis in L2, but there is an absence of the analysis of the emergence of syntax as language proficiency develops. As a result, L2 researchers may be ignoring some of the patterns of language competence highlighted by Ellis (2019),

particularly the acquisition of syntactic patterning in formal L2 instruction. A lack of longitudinal data has led to an over-reliance on cross-sectional designs, capturing points-in-time usage rather than development over time. As described, the dearth of longitudinal data is now being mitigated by the release of corpora that are calibrated across scales of proficiency, namely the TLC, The Open Cambridge Learner Corpus and the Englishtown corpus (EFCAMDAT). It is timely therefore to address some of conceptual gaps and to explore a methodology to identify candidates for analysis with a large quasi-longitudinal corpus and through this, analyse the *development* of syntactic sequences in learner language, over time, and thus contribute to our understanding of which sequences are used by learners and how, and inform usage-inspired L2 teaching (Ortega *et al.* 2016).

A shift in the size and nature of the data brings with it a re-evaluation of approaches to investigation. Within the UB approach, for example, a top-down, hypothesis-driven analysis has been used to explore constructions, with the exception of Römer 2019, who extracts all VACs from a subset of L1 German data. Studies looking at development across proficiency levels from A1 to C2 in large-scale data, without recourse to L1 data, from a range of L1 backgrounds are lacking. Those that are truly longitudinal focus on a small number of participants, from one or two L1 backgrounds. Research on structural development concentrates around the intermediate to advanced levels, meaning that developmental territory from lower to higher levels across all levels has remained uncharted.

This study benefits hugely from the sum of the findings in the research described above and seeks to address some of the gaps highlighted. It aims to give a globalised view of structural development taking both a quantitative and qualitative approach. Previous studies have been limited in giving a generalised picture of structural development, for a host of reasons, from limitations concerning data size, L1 backgrounds, proficiency levels and objects of focus. This study has privileged access to a large-scale longitudinal corpus, drawing from a large pool of learners, from a diverse range of L1 background, across all levels of proficiency benchmarked to the CEFR. Taking a corpus-driven approach, and mapping the use of POS tag sequences, it trawls the L2 data in its entirety. It explores the emergence of structural generalisations from low to high proficiency level learners, in an attempt to help shift the balance from a description of learner output at points in time, to a developmental view of learner language usage.

“Frequency is a key determinant of acquisition because “rules” of language, at all levels of analysis from phonological, through syntax to discourse, are structural regularities which emerge from learners’ lifetime analysis of the distributional characteristics of language input”.

Ellis (2013, p.89)

Chapter 3 Foundations and concepts

In Chapter 1, I set out the background to this study and touched on how, according to a usage-based (UB) theory, learning a language happens through frequency and relevance of experience. I sketched out the notion that our minds are sensitive to patterns in our language experience, and I put forward questions for investigation about whether language development in an L2 was observable through a methodological approach investigating POS tag sequences across different levels of proficiency. Chapter 2 followed with a description of the main studies in the field focusing on L2 language proficiency and development, and identified an area of potential research that this study hopes to fill.

In this chapter I return to theoretical and methodological considerations arising from chapter 1 and explore these in more detail. I first consider the role that frequency and distribution play in language use, from a UB perspective. The chapter explores the notion of structural regularity in more detail and discusses concepts and terminology relating to theoretical and operational descriptions of language use including Emergent grammar (Hopper 1987), the idiom principle (Sinclair 1991), Pattern Grammar (Hunston and Francis 2000) and Construction grammar (Bybee, 2010; Goldberg 1995, 2006). It then considers definitions and descriptions of development in writing, and concludes by exploring the role of corpus linguistics as a method for exploration. It looks at how language usage and analysis is operationalised through different units of analysis, including patterns, sequences of words, bundles, chunks, phrases, n-gram, p-frames, constructions and POS tag sequences.

3.1 Language, frequency, structure and regularity

In his Emergent Grammar theory, Hopper proposes that our task as linguists is “to study a whole range of repetition in discourse, and in doing so to seek out those regularities which promise interest as incipient sub-systems. Structure, then, in this view is not an overarching set of abstract principles but more a question of a spreading of systematicity from individual words, phrases and small sets” (1987, p.143), both shaped by usage and shaping usage. Crucially it is emergent because it is constantly being negotiated. It changes and allows for change. This dynamic notion fits within a context of usage-based learning, where language development is seen as a continuous shuffling and reshuffling of the frequencies of encountered patterns. This line of thinking marked a departure from theories of language acquisition which propose abstract rules from which to create structures, and transcends the distinction of competence and performance found in traditional language acquisition studies (Ortega 2013). Frequency in language is a natural phenomenon. Some morphemes, words, phrases, chunks, sequences occur more frequently than others. Some are more useful and therefore more used than others. Frequency of language encounter plays a crucial role in learning. The first time we experience or notice a piece of language, it is an isolated event which “can result in a unitary representation in memory that binds all its properties (i.e. phonological make-up, spelling, etc.) together” (Wulff and Ellis 2018, p. 40). Subsequent encounters activate our pattern-finding mechanisms and strengthen form-function mappings, while at the same time our perceptual mechanisms are attending to the frequency and distribution of the sequence in the input. In UB terms, as seen in previous chapters, these form-function mappings are ‘constructions’ (See 3.1.2 below). We have stronger memories for the constructions that we experience more frequently, and our ability to access them from the ‘construction warehouse’ is easier as their form-meaning relationship becomes entrenched.

3.1.1 Zipf’s law and frequency

An analysis of frequency distribution in natural language can be described in terms of Zipf’s law (Zipf 1935). This law describes a relationship between the frequency of units of language and their frequency rank (Piantadosi 2014); for example, in naturally occurring language the first, most frequently occurring word occurs twice as often as the second most frequent word and three times as often as the third most frequent word, etc. When represented graphically this kind of distribution is characterised by the kind of ‘long tail’ illustrated in the example in Figure 3.1. Table 3.1 gives the top 10 ranked base verb forms (i.e. all infinitive form use of

verbs) in the 90-million-word BNC written corpus. Figure 3.1 shows how the frequency of these words decreases in relation to their frequency rank and how the highest ranking *types* (i.e. the different occurrences of base verb forms) constitute a large percentage of all *tokens* (i.e. the total occurrences of all base verb forms). One of the characterising features of this distribution is the high token frequencies of the highest ranking exemplars or types (Table 3.1) when compared with token occurrences of exemplars or types further down the ranking. For example, ‘be’ ranked #1 occurs approximately 10 times more frequently than ‘see’ ranked #4.

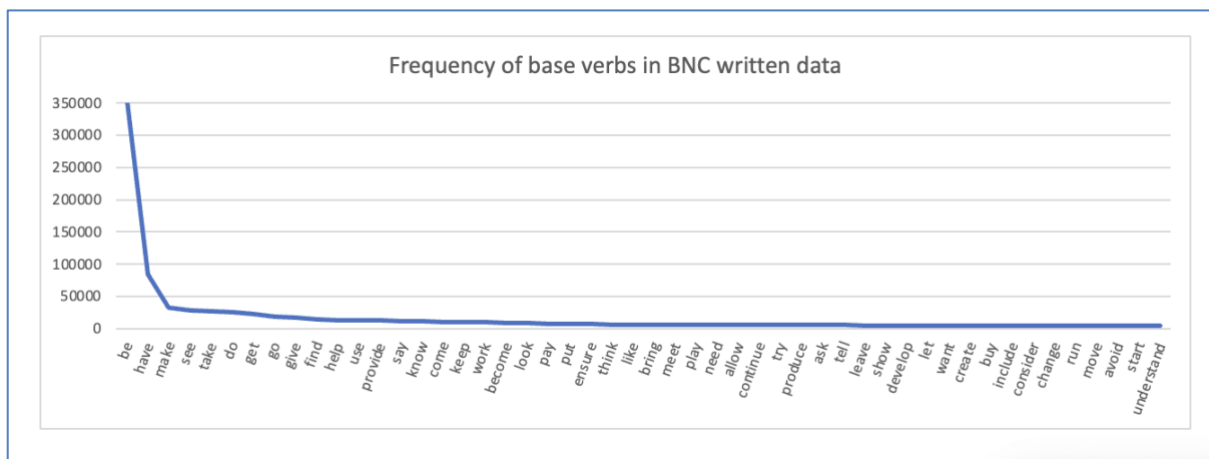


Figure 3.1 Frequency distribution of the top 50 base verb forms in the BNC written corpus

rank	verb base form types	raw token frequencies
1	be	391209
2	have	167438
3	make	159710
4	see	46867
5	take	36475
6	do	17373
7	get	15847
7	go	15209
8	give	14929
10	find	13269

Table 3.1 Top 10 ranked base verb form types and their raw token frequencies in the BNC written corpus

Research illustrating distributions of frequencies in natural language shows how Zipf's law is observable as a universal phenomenon across language, and can be seen not only across individual words (Evert 2005), but also, and most relevant to this study, across *other units of language*, other than words, for example constructions (Ninio 2005, Ellis *et al.* 2016).

Williams *et al.* (2015) hold that Zipf's law works better for phrases than for words, while Piantadosi makes the point that words 'may not be a precisely defined psychological class'; they happen to be a vehicle to demonstrate how the distribution plays out (2014, p. 112-130). The power law can be observed for example in the frequency ranking of patterns in natural language, and then in turn in the combinations of the lexical exponents of a given pattern (Ellis *et al.*, 2016). Usage-based theorists (Bybee 2008; Ellis 2008; Ellis *et al.* 2013; Ellis *et al.* 2016; Lieven and Tomasello 2008) assert that "it is the coming together of Zipfian-like distributions across linguistic form and linguistic function that promotes robust language learning despite learners' idiosyncratic experience." (Ellis *et al.* 2016, p.57-58).

Within this theory, structural conventions in language emerge as we subconsciously tune into and count the form-meaning regularities that we experience. UB studies have shown that language users are sensitive to the statistics of both types and tokens (see 3.1.3 below), of repeated patterns in language "tallying them implicitly during each processing episode", and that structural regularities emerge from the "conspiracy" of these encounters. (Ellis *et al.* 2015, p.36, *inter alia*). This illustrates subconscious statistical learning, not a conscious process but one which happens implicitly through language experience. Learners do not have explicit access to this tallying (Ellis *et al.* 2015).

3.1.2 Association, pattern-finding and schema

Tomasello (2013) maintains that in child language acquisition, children use two types of cognitive skills when first learning and developing language: they begin by associating the intentions of their adult caregivers with the linguistic conventions they use (intention-reading), and, secondly, they look out for repeated patterns of utterances to create abstract schema (pattern-finding). Pérez-Paredes *et al.* (2021) note that to do this, they develop skills of schematisation and analogy. They begin by understanding an entire communicative act rather than individual words or structures. Then, as they experience the same communicative act with the same words and structures, (and, indeed, the same communicative act with different words and structures), they map the words and structures onto their understanding of the function. They first become aware of meanings in context and begin by engaging in

imperative, declarative and informative communication using pointing. These gestures involve people, places and things (agents, locatives and objects), indicating understanding of entire acts. Children then begin to map utterances onto these agents, locatives and objects. For the child developing language, the utterance is the smallest unit of linguistic communication. They hear utterances as holophrases (Pit Corder 1973) or formulas and then learn individual words by figuring out and extracting their meaning in context when repeated in a range of different utterances and shared contexts. From this, by attending to the patterns in the utterances and using skills of analogy, they abstract meaningful grammatical constructions.

When it comes to *producing* sequences, children first put together two single words (word combinations) or holophrases that fit a context; for example, to refer to a bird in a tree they may say '*bird tree*'. To begin with, the word combinations like *bird tree* may not be words that have been experienced together in the input, but they have been abstracted individually from the input, as meaningful referents for objects (e.g. from *it's a bird*, *Look at the bird*, *See the bird in the garden*, etc and the same for tree). In UB terms they then begin to demonstrate early signs of abstracting grammatical patterns and evidence of 'slot-filling' and generalising by combining a word like *more* with a range of objects (*more banana*, *more milk*). This signals the second stage of a usage-based theory of development where learners have moved from a formula or holophrase to an abstraction of a low-scope pattern like *more + object* (Ellis 2003). From this they progress to abstracting more and more patterns (e.g. *where's + noun + gone*, *I want + noun*), figuring out more variable slots, and moving to verb islands (Tomasello 1992). This is where they first learn about verb complementation and collocational patterning on an individual verb by verb basis, before moving to a wider generalised understanding of all colligational patterns when they have a larger dataset of evidence to abstract from. As they move through these stages they figure out which words can go into which slots and an understanding of the degrees of fixedness of the relationship between words in slots. This completes the three steps of usage-based learning from formula/holophrase to a slot and frame system and through a continuous process of abstraction towards a fully abstracted system of mappings of form and meaning, otherwise referred to as 'constructions' (Pine *et al.* 1997; Lieven 2016).

It is these language constructions which constructionists, construction grammars and cognitive psychologists situate at the centre of the language learning experience (Bybee, 2008; Goldberg 1995, 2006). As we have already seen, constructions can range from

morphemes to words, phrases, and syntactic frames) and therefore carry varying levels of complexity:

simple morphemes such as *-aholic* (meaning ‘being addicted to something’) are constructions in the same way as simple words like *nut* (meaning ‘a fruit consisting of a hard or tough shell around an edible kernel’), idioms like *It is driving me nuts* (meaning ‘It is greatly frustrating me’), and abstract syntactic frames like Subject-Verb-Object-Object (meaning that something is being transferred) (Wulff and Ellis 2018, p.38).

This fully abstracted system is stored on a continuum of formulaicity, from heavily entrenched chunks such as the ‘It’s driving me nuts’ to the syntactically connected strings such as the verb argument construction (VAC) ‘VOO’ (verb + object + object), (*She gave him the book*). (Ellis *et al.* 2015). Wulff and Ellis describe this language knowledge as ‘a huge warehouse of constructions that vary in their degree of complexity and abstraction.’ (2018, p.39) which exist as mental constructs in the user’s mind.

3.1.3 Categorisation and prototypicality

The statistical tallying that we do as we encounter language old and new involves categorisation. By analogy, we figure out which words and sequences of words belong to the same categories. Not only do we categorise for example, labradors, poodles, spaniels, otterhounds as ‘dogs’, we also have a sense of that otterhounds are rarer than spaniels. The notion of ‘dog’ is a prototype, the breeds of dog are numerous, they share characteristics and a Zipf-like frequency of use. Extensive research into VACs has shown that constructions also have prototypes, that the types of verbs occupying the verb slot of any construction, share characteristics of the prototypical meaning and also have a Zipfian distribution (Bybee 2008; Goldberg 2006; Ellis *et al.*, 2016). For each VAC, there is one verb, which Ninio terms ‘pathbreaking’ (1999), which takes the largest share of the distribution and which is prototypical of the meaning of the construction. For example, in the VL construction, movement to place, *go* is the prototype verb, followed by *come*; in the VOO construction (verb + object + object), *give* is the prototype, followed by *send*. When learners come across subsequent verbs found in the same syntactic contexts, or slots, in the input, they already draw on the prototype from which to infer meaning. These prototypes are ‘the hubs in the construction’s semantic network’ (Ellis and Ogden 2017). As we learn these form-meaning mappings we learn to categorise. We learn to match what we come across for the first time against what we have already encountered and categorised.

3.1.4 Formulaicity and sequential probabilities

In UB terms, learning a language necessitates learning thousands upon thousands of constructions, and alongside this is the assertion that a high percentage of these patterns or constructions are stored as memorized wholes or formulae, clauses and clause structures that humans can retrieve as “automatic chains from the long-term memory” (Pawley and Syder 1983, p.192). It is predicted that learners of a language subconsciously acquire first the constructions that they come across most frequently in the input that they receive. Studies of the sequential probabilities within the language system have illustrated how mastery of the language system involves not only knowing these constructions but also about knowing the strength of association within and around these sequences. (Bybee, 1998; Elman, 2009; Ellis *et al.* 2015; Arnon and Christiansen 2017). It involves a subtle understanding of the strength of fixedness of elements in a sequence, for example knowing that *a wide variety of* is more frequently occurring than *a big variety of*. Studies of formulae in L1 English have shown that up to 50% of language produced is formulaic (De Cock *et al.* 1998; Erman and Warren 2000). These formulaic sequences are learnt because they are both frequent and prototypical. This assertion is also supported in CL studies which, since the 1990s, through analysis of large bodies of text, have been providing evidence of the existence of recurrent patterns of words and construction within language research. In pioneering CL studies, resulting in the notion of ‘the idiom principle’, Sinclair asserted that users of language have at their disposal “a large number of semi-preconstructed phrases that constitute single choices, even though they may be analyzable into segments.” (1991, p.110).

When it comes to the relevance for L2 learning, research in both L1 and learner data has illustrated that, on one hand, the ability to use formulaic language has been shown to be a distinguishing feature of L1 fluency, while struggling with such patterns, on the other hand, has become a marker of learner language (Pawley and Syder 1983; De Cock *et al.* 1998; Wray 2000). It appears that formulaic language in L2 data demonstrates marked usage, with a reliance on a limited set of expressions over others (De Cock *et al.* 1998; Durrant and Schmitt 2009).

The three stage usage-based framework from “formulaic phrase to limited scope slot-and-frame pattern to fully productive schematic pattern” has been shown to be applicable to both L1 and L2 learners (Ellis 2012, p.18). Formulaic language sits at the fully productive schematic end of the process. In order to get to the point where they subconsciously select *a huge amount of* over *a great amount of*, learners need to have experienced enough examples

of language usage “that their accidental and finite experience is truly representative of the total population of language of the speech community in terms of its overall content, the relative frequencies of that content, and the mappings of form to functional interpretation.” (Ellis 2002, p.167). Given the enormity of the L1 lexicon and breadth of possible constructions, it is therefore not surprising that L2 users might be distinguished by their ability or inability to use formulaic language in a fully productive way. Since all language learning is ‘sampling’ the L2 user needs access to a very large sample (Ellis 2009).

3.1.5 Contexts of L1 and L2 learning

Of obvious relevance are the differences between the *contexts* of L1 and L2 language acquisition. In First Language Acquisition (FLA), and therefore child language acquisition, the child is discovering the world while simultaneously discovering the language used to describe that world while moving through social, cognitive, emotional and physical developmental stages in general. Adult learners are building not only on pre-existing knowledge of the world but on pre-existing knowledge of language. Input in FLA comes from naturalistic exposure (Tomasello and Brooks 1999) whereas much L2 input takes place in classrooms, and is controlled and selected. Second language learners will range in age groups, from school-going children and adolescents to adults or, at the very least, people who already use a first language successfully, and in the 21st century, where multi-lingualism is the norm, the likelihood is that people are learning more than one additional language simultaneously (Douglas Fir Group 2016). Within a usage-based perspective, language learning for the multi-lingual learner both converges with and diverges from the first language learning context. In contrast to the child, the second language learner already has a well-developed schematised repertoire for at least one other language. As they learn additional languages, they are not discovering the world for the first time nor are they developing social and conceptual understanding of the world. Additionally, they have typically developed problem-solving and explicit learning skills. The learning context is typically different. The child, in the FLA process, begins with a blank sheet and abstracts syntactic categories from usage, while the second language learner begins with an L1 and builds on pre-existing knowledge of slots and frames, along with knowledge of how to combine them and what to put in them. Despite all this, there is evidence to suggest that the process of additional language learning still involves intention reading and pattern finding and that it develops along a similar cline from formula to low scope patterns to fully abstracted constructions (Ellis 2003). This has obvious implications for language teaching in

relation to transition along the cline of a multi-layered process. Second language learners move from a holophrase (e.g. something like *I'd like to ...* which at low levels learners seem to use as a whole formula) to a low scope slot and frame system (e.g. *I went/walked to the cinema/shop/restaurant* where learners are able to substitute different elements into slots) to a fully abstracted formulaic chunk (where the items are fixed and the meaning is specific e.g. *He came to the conclusion that ...*).

In summary, according to the UB model, the acquisition of 'constructions', the target of learning for both L1 and L2, is input-driven and depends upon exposure to meaningful form-function relations resulting in a language system which "emerges from the statistical abstraction of patterns latent within and across form and function use" (Ellis 2012, p.17). The language learner (first or additional) attends to these frequently used form-meaning pairings and they become "entrenched as grammatical knowledge in the speaker's mind" (Ellis and Ferreira-Junior 2009a, p.188). The degree of entrenchment, according to Ellis and Ferreira-Junior (2009a) is proportional to the frequency of usage. The question here is whether development of these form-meaning mappings is observable.

3.2 What is development?

3.2.1 L1 and L2 learning contexts for writing

In the case of first language learners, form-meanings mappings typically begin to emerge in speaking before writing. Children learn to speak before they learn to write. Specifically, they experience and use language while they are constructing and making sense of the world around them. This is a stark contrast with the landscape of language usage for the L2 learner, for both speaking and writing. Typically most L2 learners are in instructed classroom settings and they are not children; they have already learnt to speak and write in at least one other language and have already made sense of the world around them. According to the EF proficiency index which gives a global account of English language proficiency

<https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2021/ef-epi-2021-english.pdf> in 2021 the average age of English learners worldwide is 26 years.

When they are learning a new L2 language they are mapping new forms on to known meanings (even though the conceptualisation of these meanings may not always be the same in one language or culture as another) (Constantinou 2019; Pérez-Paredes *et al.* 2021). L2 learners typically learn written forms of the target language alongside spoken forms. L2 written language production is characterised by the context of language teaching and a genre

of conventional formats, found over and over in exam-type tasks and reflected in coursebooks, working towards a syllabus, often with an exam in mind. There is inevitably a circularity in this. Exam type tasks are identifiable as a genre in its own right which does not reflect the breadth of usage contexts typical of everyday writing in the world. This has obvious implications if taking an L1:L2 comparative approach and raises questions of whether we can ever adequately make comparisons between learner data and benchmark data such as the BNC.

When writing in L1 language emerges it is often as a result of classroom tasks, of school-aged children, which are often not so dissimilar from the exam style tasks of EFL. This is illustrated in Figure 3.2, which shows the writing produced by a 5-year-old L1 English speaker, a response to the task ‘What did you do on holiday?’. *I went to Scotland we nearly got flooded in a big flood 72 people had to be lifted up in a helicopter*

Though the child is at an early stage of letter and word formation, their language use is sophisticated, with use of modal meanings, passive constructions, a mid-clause adverb modifying a clause, etc.

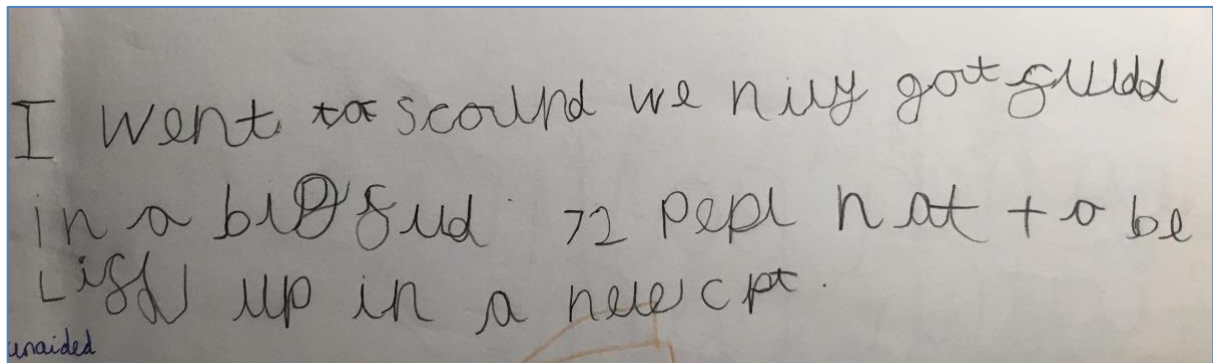


Figure 3.2 Example of writing from a 5-year-old L1 English user

We met a lot of friends of ours. I had much pleasure althought at the end I felt a bit tired. So at 2 o'clock we decided to got back home. It was the best idea. Next day, we both **had to** work early in the morning. [A2, Portuguese, PET, 2002]

Figure 3.3 Example of writing from an A2 L2 English learner

Figure 3.3 shows an example of writing from the L2 beginner language, in response to the following question from the PET exam, 2002.

‘You recently spent the day with a friend who you had not seen for a long time. Now you are writing a letter to your English-speaking pen-friend. Describe how your friend has changed, say how you spent your day together, and talk about your plans to meet the friend again in the future’

Both extracts describe a past event, both are characterised by use of the past simple, both use a *had to* sequence (as in *we had to be lifted up* (L1), *we both had to work* (L2)). The L2 user doesn't yet have the noun + *had to be* + past participle sequence (expressing necessity). According to the findings in the *English Grammar Profile* (see Chapter 1) structures containing the sequences *had to* + *be* + past participle are not seen in the learner data until C2, the highest level of learner proficiency. Passive constructions are perceived to be ‘hard’ grammar. The chances that the child knows this explicitly as ‘a passive construction’ are very low. Looking at the writing from a UB perspective, we might suggest that the partially schematic ‘noun + *had to be* + past participle sequence to express necessity’ is an entrenched form-meaning mapping in the child's warehouse of language. Taking another example *we nearly got flooded*, not only do we see the use of *get* + past participle, but also the use of *nearly*, all of which comes together to mark a meaning of ‘a lucky escape from a negative situation’. We might assume that the child is not consciously drawing on this structure and yet they have it at their disposal as a good fit for the context and meaning they wish to convey. First language learners typically have had at least a year of language experience, observation and sound making before they even start to put sounds together to make words and sequences. And then another two years or so before they put marks on the page to represent these words and sequences. In instructed settings L2 learners are generally expected to start using language productively both in speaking *and* writing as soon as they start learning (Durrant *et al.* 2021). On a surface level we can observe some similarities in the use of structures, but we cannot observe the mechanisms of retrieval and use of these structures from isolated examples. We do not know from this evidence whether L2 learners draw on the same cognitive mechanisms to retrieve language that L1 writers are using. When L2 learners begin to put words together what are the words and sequences that they first use and how does this usage develop as proficiency increases? How do we characterise development?

3.2.2 Defining and describing development

In first language development research, longitudinal studies following children from their first utterances to a point of stable acquisition take place over a relatively predictable time frame with widely agreed developmental milestones (Clark and Clark 1977; Slobin 1978) . These markers of development in child language are well-documented and measured in productive and receptive terms of increased lexical and phrasal repertoire of form to meaning mappings alongside the creation of a conceptual framework for the world around (Tomasello 2006) . First language milestones relate to speaking, with a writing system emerging typically in instructed settings, in early years education, from the age of three upwards, depending on cultural norms and expectations. Learner language does not have the equivalent in formalised developmental milestones; there are widely accepted levels of proficiency, though established linguistic milestones and criteria for these are fuzzy and nebulous. In order to describe development in both L1 and L2 writing Durrant *et al.* (2021) conceptualise writing development broadly along the axis of both time and quality. They note that proficiency in L2 writing is typically described in terms of what learners ‘can do’ but point out that movement along a proficiency scale does not necessarily constitute improvement or change in quality. Descriptions of proficiency vary considerably in relation to the “emphasis, description and scope” of criteria set out for different proficiency levels in different proficiency frameworks and that this likely results in variation in the linguistic features associated with particular levels of proficiency. (Durrant *et al.* 2021, p. 25). They argue that an analysis of development must involve analysis of linguistic features and offer following four premises which they argue help provide a framework for corpus analysis:

1. Writing development is at least partly a matter of linguistic development
2. This linguistic development is reflected in specific and consistent differences in the use of specific language features.
3. These features can be reliably identified by analysts.
4. Patterns of linguistic development can be identified at a valid level of description.

(Durrant *et al.* 2021, p. 28)

As they point out analysis of a large-scale data allows us to begin to draw generalisations about changes in linguistic features and patterns of linguistic development. This leads on to a consideration of what linguistic features to use as a departure for analysis. Section 3.3 is concerned with identifying suitable units of analysis to use to describe change in linguistic features.

3.3 Units of analysis

3.3.1 *Constructions: suitability and findability*

From a UB perspective, the obvious candidate for analysis of development would be the construction. However, when it comes to L2 data, analysis of constructions brings up issues of ‘findability’. Studies centring on constructions have traditionally used L1 norms to explore known constructions (Römer *et al.* 2014). Given that constructions represent a third and final stage in a process of abstraction, applying their usage to a L2 data encompassing a range of levels of proficiency gives a view on a series of ‘finished’ entrenched products rather than developing ones. Their presence in learner data depends on frequency of encounter.

Constructions will not have had the opportunity to become entrenched if they have not been encountered. Analysis of constructions driven largely by L1 usage may eventually perpetuate the predominance of the deficit hypothesis in SLA studies (Tyler and Ortega 2018) and may be to the detriment of studies of L2 language acquisition that embrace the notion of development as the focal point of analysis (Larsen-Freeman 2015).

Tomasello (2003) distinguishes between constructions and traditional linguistic units. The former can be abstract (at different levels of abstractions), item-based (lexical items) and mixed (V + for). Traditional linguistic units are seen as general patterns that, among others, include linguistic categories such as nouns, verbs or articles. Tomasello goes on to argue that UB approaches include all kinds of “usage patterns, even those of only limited generality” (2003, p.89). Tomasello does not exclude traditional linguistic units from the potential repertoire of constructions. On the contrary, he maintains that meaningful linguistic units from a constructionist perspective should look at competence, “not in terms of the possession of a formal grammar of semantically empty rules, but in terms of the mastery of a structured inventory of meaningful linguistic constructions” (2003, p.99).

Originally constructionist approaches arose from analysis of idioms, formulaic language and constructions of limited productivity (Fillmore 1988, Goldberg 1995) particularly those non-compositional sequences whose meaning was not predictable from their form (e.g. the X-er the Y-er). A current maximal definition of the construction (Goldberg 2006) maintains that they exist at all levels of frequency, from highly frequent to infrequent, and at all levels of description from morpheme to clause. Buerki (2018) points out that analysis of constructions at the more substantive level of productivity, particularly those that are compositional (whose meaning can be inferred from their component parts) or non-idiomatic, are often overlooked.

Those that have been of research interest represent a small number of the many thousands of constructions in use and they are not necessarily those that are most frequent (e.g. verb across noun, verb towards noun). This is corroborated by Hunston who states that a limitation of construction grammar studies is that they ‘offer detailed descriptions of a relatively small number of constructions only’ (Hunston 2019, p.324). This has clear implications when looking at their use in L2 data. The more infrequent the construction in L1 usage the lower the chance of finding it in developing L2 usage. This has both theoretical and methodological implications, involving questions of how to define and describe constructions in development and, most relevant to the present study, how to go about finding them in L2 data.

3.3.2 *Constructions and patterns*

As seen in Chapter 2, the study of learner language within corpus linguistics has used a variety of different units of analysis over the last 30 years, ranging from word classes to discourse markers, and has been dominated by a focus on lexis over structure. This is evident in the first iteration of Contrastive Interlanguage Analysis (CIA) research (Granger 1998). The evolution of the units of analysis throughout the years reflects a continuum from an interest in individual items to an interest in phraseological units and patterns but the choice of which units to use for analysis has not been underdiscussed. Römer *et al.* (2014) used 50 constructions from the COBUILD Grammar Patterns 1: Verbs (Francis, et al., 1996) to initiate a systematic analysis of VACs in the 100-million-word British National Corpus (BNC), and Römer *et al.* (2015) discussed the methodology used to extract these VACs. However, the authors did not address the motivation to choose these VACs in particular, nor the exact nature of the constructions investigated (i.e. if V against n, for example, is a construction, how does this construction integrate the different meanings identified in Francis *et al.* (1996)? In this instance in the COBUILD Grammar Patterns, there are six meanings / groups for one construction, that is, each of the *compete*, *campaign*, *preach*, *bump*, *insure* and *offend* group of verbs all occupy the verb slot in V *against* n, with a different meaning for each group).

The COBUILD patterns came from ‘Pattern grammar’ which in turn evolved from a need to regard lexis and grammar as a unified system, following Sinclair’s insight that “grammatical generalizations do not rest on a rigid foundation, but are the accumulations of the patterns of hundreds of individual words and phrases” (1991, p.100). Pattern grammar aimed to achieve a “systematised approach to the grammar of the lexicon” (Hunston 2019, p. 328) and is based

on lexically-dependent descriptions of the local grammar of a word which are then generalised to other words. For example the verb *find* occurs in the pattern ‘verb + noun phase + -ing clause’ (or *V n ing*). Other verbs with this behaviour include *catch, watch, see, hear*, but also *remember, forget, fear, tolerate* and *begin, end, start* (Hunston 2019).

Through pattern grammar, Hunston (2019) offers a solution for the paucity of coverage of constructions by proposing an alignment between pattern grammar (Francis 1993; Hunston and Francis 2000) and construction grammar (Goldberg 2006). She shows the potential of pattern grammar to develop “a taxonomy of forms (patterns) that can be used eventually in the understanding of learner L2 development and the identification of potential constructions” (Hunston, 2019, p. 330) and other morpho-syntactic units. What she offers through pattern grammar is “a systematic means of specifying the full range of mid-level constructions in English” (2019, p. 325). Unlike constructions, patterns do not necessarily exist as mental constructs with entrenched form-meaning mappings, neither do they carry any information about their frequency of occurrence. They are entirely driven by the form of the local grammar, and while criticised for not employing functional categorisation, they are loosely grouped according to meaning. It is these loose meaning groups that Hunston argues can be used for identifying constructions at a consistent level of specificity in which the construction can be used to “refer to a sub-set of instances of a grammar pattern [...] identified by the occurrence of a limited set of node words” (Hunston 2019, p.324).

3.3.2 Units of analysis: *n*-grams, bundles, *p*-frames, POS tags

As Chapter 2 outlines, many studies have used automatic extraction of *n*-grams (typically sequences of *n*-words) to extract lexical bundles (e.g. Biber *et al.* 2004; Allen 2009; Chen and Baker 2010; Juknevičiennė 2009; Ping 2009), collocations (e.g. Groom 2009; Granger and Bestgen 2014), formulaic sequences (e.g. Götz and Schilk 2011), *p*-frames (Garner 2016), and clausal sequences (e.g. De Cock 2007). This has contributed to an overreliance on the lexical which ties in with the expediencies of ‘findability’ of structural sequences in the absence of POS tagged data (e.g. examining closed word classes). Gilquin and Granger (2015, p.420) in their overview of learner corpus research (LCR) note under-representation of work on ‘grammatical features’ and they cite lack of POS tagging and parsed learner data in the past. As described in Chapter 2, two early exceptions are Aarts and Granger (1998) and Grayson and Granger (1998), and a more recent study, Gilquin (2018), who uses POS tags to compare sequences of parts of speech in L1 and L2 high proficiency level data. Gilquin

offers this approach as a way to find constructions in learner data but points out that not all POS tag sequences are constructions (Gilquin 2018). Within a construction context, and taking a pedagogical view, Cappelle and Grabar (2016) propose using what they call ‘POS n-grams’ as a way to develop a new type of learner grammar, based on n-gram frequency lists compiled from the 45-million-word COCA (Corpus of Contemporary American English) (www.ngrams.info). They point out that the teaching of vocabulary is littered with frequency-based word lists, but maintain that there is very little attention to frequency in grammar teaching, and that when grammar is explicitly taught “the sequence and selection of grammar patterns is mostly a matter of convention or convenience” (2016, p.272). They recount that materials development, and in particular pedagogical grammars, have paid detailed attention to frequency of usage found in corpora (see Biber *et al.* 1999, Biber *et al.* 2002, Carter and McCarthy 2006, Carter *et al.* 2016) but argue that grammar teaching is not informed by frequency *ranking* of grammatical sequences (their emphasis 2016, p.273). They demonstrate an approach to frequency ranking of grammatical sequences which uses POS n-grams as a proxy for constructions, noting like Gilquin (2018) that not all POS n-grams are constructions. They begin at a lexical level by using the 100 most frequent lexical 5-grams in the COCA frequency lists (e.g. *the rest of the world, at the end of the*) and then categorise them based on the most frequent POS 5-gram types, those lexical 5-grams that have the same structure (e.g. *the X (noun) of the Y (noun), in/at/at the X(noun)*). While they are termed as POS 5-grams by Cappelle and Grabar, this way of categorisation is more akin to a phrase frame or pattern grammar approach, where some elements are lexically specified and others are maximally general. They point out that their approach is hybrid in nature, neither wholly syntactic nor lexical, and justify that on the basis it is useful to the learner. In this regard a POS approach has much to offer as a methodology since it offers opportunities to explore the theoretical inseparability of grammar and lexis proposed by a UB theory. As this study hopes to illustrate, POS n-grams offer a holistic approach to explore the most commonly used syntactic patterns and their underlying lexical exponents without having a preselected set of constructions as a starting point.

3.4 Summary

In this chapter I have attempted to explore some of the theoretical and methodological concepts related to analysis of development in learner language at all levels of proficiency in large-scale data. I considered the central role that frequency and distribution play in language use, from a UB perspective, and briefly charted the affordances and limitations of various

units of analysis that have been employed so far in analysis of development. I concluded by explaining why taking a POS tag sequence approach offers an open bottom-up truly data driven way to capture language change in the use of linguistic features across proficiency levels. In the next chapter I describe in detail the data and methodological approach adopted in this study.

Chapter 4 From the bottom up: Data, tools and methods

4.0 Introduction: basic requirements

Chapter 3 was concerned with some of the theoretical and methodological considerations when approaching a study of generalised global development of sequences in written L2 English. This chapter answers the general question ‘What data and tools are used in this study to explore development in learner language and how are they used?’

The basic prerequisite components for such an investigation are access to large-scale data representative of development, and a means to identify and explore sequences within it that mark development. In this study, the first requirement is met by the Cambridge Learner Corpus (CLC), a 52-million word pseudo-longitudinal systematic collection of written exams, from six proficiency levels. The second is met by a combination of corpus tools and a bottom-up data-driven mixed methods approach.

More specifically this study is steered by these research questions:

RQ1 Is development in L2 writing observable through the frequency and distribution of POS sequences across proficiency levels?

RQ2 How does POS sequence usage develop across proficiency levels?

RQ3 Can existing frameworks for classification of language patterning account for a description of development in L2 writing?

RQ1 presupposes access to POS tagged data that has been reliably bench-marked for different stages of proficiency, as well as an understanding of development (see Chapter 3). As already noted, the data needs to be large-scale and requires a degree of homogeneity across levels in order to tell a generalisable comparative developmental story. To achieve this, the study takes an exploratory approach, using changes in POS tag sequence usage as a window into development (see section 3.3). Retrieval of frequency and distribution of POS tags across levels requires corpus tools, and a quantitative approach. It requires a method to identify where usage might converge in the data, both in quantitative distributional terms and at a fine-grained manual analysis of usage. Integral to RQ2 is a framework on which development can be plotted and compared, both quantitatively and qualitatively. Finally RQ3 considers whether existing frameworks can be used to describe development.

The chapter begins by discussing in 4.1 the type of data needed for a study on development, and the importance of bench-marked stages of proficiency. In 4.2, I describe the CLC and the CLC sub-corpus used in this study, its appropriateness for this study. In 4.3, I explain the methodological approaches used in the study, followed by an example of how the mixed methods approach is applied (4.4), how it relates to other approaches (4.5). I conclude with a summary (4.6) illustrating how the study sits at a cross-roads between SLA and LCR.

4.1 Largescale, longitudinal, levelled, homogeneous, tagged learner corpora

As noted in Chapter 2, large-scale longitudinal data of learner language is scarce. The task of collecting language from the same group of learners over time tracking all stages of their linguistic development is monumental, and on a large, representative scale, almost impossible. Factors such as participant retention, funding, short-term urgency taking priority over protracted long-term research are all cited as obstacles to successful curation of truly longitudinal learner data (See Myles 2005, 2015, Meunier 2015). Most longitudinal and developmental studies, few in number to date, involve small cohort studies, following the journey of a handful of learners; even then they are over a limited time duration, and rarely across all stages of proficiency (Myles *et al.* 1999). This present study needs a large-scale multi-lingual, multi-levelled corpus to explore evidence for a generalised view of development, across learner language as an entity.

In Chapter 3 it was pointed out that while L1 development is frequently plotted along a timeline of developmental milestones, time is a poor mechanism for describing development in L2 writing. Learner language does not have the equivalent developmental profile in formalised terms; though there are widely accepted levels of proficiency, established linguistic milestones and criteria for these are fuzzy and nebulous. Until such time as we might have truly longitudinal L2 data at our disposal, we adapt. The way that most SLA researchers wanting to use corpora to explore development have adapted to this is by using pseudo-longitudinal (Johnson and Johnson 1999) or quasi-longitudinal (Granger 2002; Thewissen 2013; Römer and Garner 2021) data, collections of cross-sectional data, that track learners over time, albeit not the same learners.

Most LCR in general, exploring development or otherwise, has used data which is calibrated by year of study (e.g. second year undergraduate university students). The challenges of this approach have been well documented (Callies *et al.* 2014; Pendar and Chapelle 2008; Díez-Bedmar 2012; Tono and Díez-Bedmar 2014; Meunier 2015; Myles 2015; Durrant *et al.*

2021), the most relevant of these for studies of development being that students at the same institutional level may not have a uniform proficiency level. The calibration entails that year of study acts as an equivalence for the amount of time a learner has spent learning the language, and does not account for individual differences within a cohort in proficiency, aptitude or experience and therefore cannot guarantee validity (Gablasova *et al.* 2017).

More recent developments see the investigation of learner language using data which is calibrated by level to the CEFR (Common European Framework of Reference for Languages), devised and published by the Council of Europe (Díez-Bedmar 2017). Such calibration with reference to the CEFR is of particular relevance and importance to this study. According to the Council of Europe, the overall purpose behind the CEFR was to create a standard “a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe”, and to provide “objective criteria for describing language proficiency” (Council of Europe 2001, p.1) by establishing a range of comparable levels describing learner language use. While not without its critics, over the past twenty years it has become a visible presence throughout the language teaching industry, its greatest impact seen within assessment (Little 2007; Díez-Bedmar 2018). More recently, in a companion volume to the original CEFR publication the framework is described, with particular reference to assessment, as a means “to provide transparency and clear reference points”, and “to provide a sound basis for the mutual recognition of language qualifications” (Council of Europe 2018, p.25).

The proficiency levels defined by the CEFR, evolved from a “wide, though by no means universal, consensus on the number and nature of levels appropriate to the organisation of language learning and the public recognition of achievement” (Council of Europe 2001, p.22) relating to stages broadly agreed by the language teaching community of beginner, intermediate and advanced. In the documentation explaining the CEFR, these three stages are reframed as Basic User (A), Independent User (B) and Proficient User (C) and further subdivided into 6 levels (A1, A2, B1, B2, C1, C2), shown in Figure 4.1, perceived to give adequate coverage of the (European) language learning space. Each stage is given a title intended to reflect the nature of the stage.

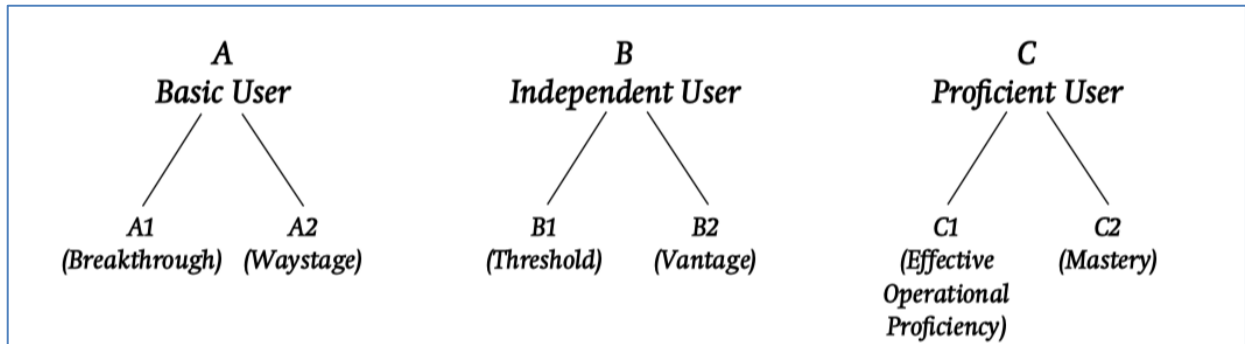


Figure 4.1 Original 'hyper-text' branching framework of the CEFR (Council of Europe 2001, p.2)

It is worth noting here that the architects of this framework are keen to acknowledge the continuous and individual nature of the language learning process, pointing out that the developmental pathway of no two users of a language is the same, and that the attempt to 'establish 'levels' of proficiency is to some extent arbitrary, as it is in any area of knowledge or skill (Council of Europe 2001, p. 17). Subsequent representations of these 6 levels reflect different purposes, e.g. a vertical scale as in Figure 4.3 below, comparing different qualifications along the scale. The originators emphasise that the levels should not be interpreted as linear, and point out that the appearance of equidistance between levels on a visual scale does not equate to equal learning time taken to move between levels. In the more recent companion volume to the CEFR the levels are represented as in Figure 4.2.

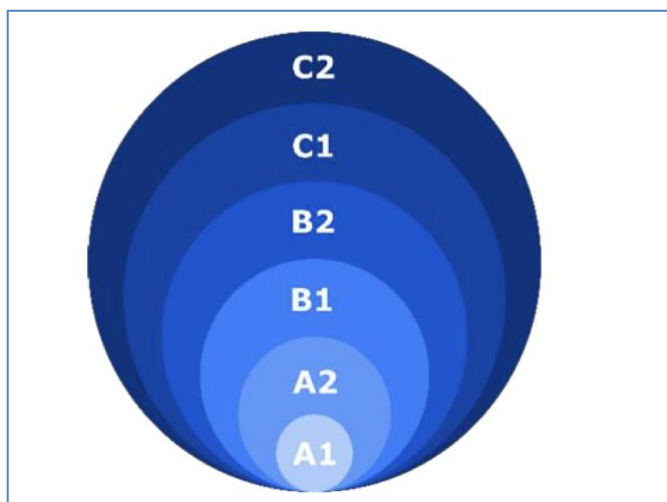


Figure 4.2 CEFR reference levels (Council of Europe 2018, p.34)

This description of the levels is designed to represent development in a progressive, illustrating language proficiency on a continuum. This visual representation of broadening development has particular relevance in this study.

For analysis of learner writing, Díez-Bedmar (2018) stresses the need for use of larger corpora and for the importance of using data which is aligned to CEFR. Use of the CLC in this study brings a unique opportunity to investigate L2 writing in the largest dataset of its kind. To date there is no other written dataset on this scale from an examination board, which has data derived from ‘reliable’ and ‘validated’ international English language exams. The data is benchmarked to the CEFR both in relation to the exams taken and the results achieved. The value of this alignment with proficiency levels is not to be underestimated, placing this research within a new wave of work that is using a more reliable means of calibrating learner data. Data of this size and configuration offers an opportunity to profile the developmental nature of sequences and chart their most frequent abstractions and uses. At 52 million words, the CLC is so large as a learner corpus and unique in a number of ways and this facilitates the research design across a quasi-longitudinal dataset. Römer *et al.* warn of a “scarcity of occurrences” (2014, p.132) in the use of learner data. Although the CLC data does not track individual learner development it allows for comparisons of cohorts of learners across a period of 17 years, and as such presents a quasi-longitudinal view. Given the size of CLC, this study will not suffer from paucity of data and allows for a broader approach to the investigation of learner language, across a wide range of L1 backgrounds and proficiency levels. Added to this, the data is not publicly available¹, affording this study a unique and timely opportunity to share insights from such a dense source.

4.2 The CLC and the CLC sub-corpus

The CLC is a corpus of written English compiled by Cambridge University Press and Cambridge Assessment English. The instance of the CLC used in this study is the ‘uncoded’ version which distinguishes it from a parallel, smaller scale version of the CLC which has been tagged with error codes. The uncoded CLC is a 52+-million word dataset of written learner language, aligned to the CEFR, encompassing the Cambridge Assessment English

¹ While the CLC is not publicly available, the researcher has received agreement from CUP to have access to it for this PhD research by licensed agreement

'main suite' of exams as well as other language exams under the Cambridge Assessment English umbrella including BULATS (Business Language Testing Service), assessment for young learners, for Business English and main suite exams adapted for Schools, as illustrated in Figure 4.3. It comprises 266,600 exam documents, spanning 148 different first language backgrounds, from a 20-year period (1993 - 2012). Only open-ended student writing is included in the CLC (i.e. it does not include gap-filling tasks). Each document is tagged with metadata providing candidate information (first language, nationality, level of education, age, gender), general exam information (exam taken, CEFR level, year of exam, exam performance) and task specific information (question number, task style, task format, task register).

In the CLC, all learner data is tagged both by exam taken *and* by actual performance achieved *at, above or below* the level of the exam taken. For example, assessment data from a candidate who performs at the top percentile of a C1 exam is tagged as C1 exam data but also as C2 level performance data (and hence, if a researcher opts to search for all data at C2 performance level, these data will be included); alternatively, a candidate performing at the lower margin in a C1 exam, e.g. a fail, will be tagged as C1 exam data (fail) but will also be tagged at the lower performance level of B2, and so on. Likewise, while there is no exam aligned at A1, candidates who take the A2 aligned exam (Key) but who perform lower than A2 level are tagged as performing at A1.

For the purposes of this research a sub-corpus of the CLC is used, which includes only the main suite of general exams. Other exams in the CLC were excluded to avoid discipline-specific use, such as is found in the academic, legal and business exams. The sub-corpus data, which makes up 63% of the entire CLC is typical of EFL mainstream language assessment and is characterised by a mix of task styles (e.g. advice, argument/opinion, complaint, criticism, descriptive, news), formats (e.g. article, essay, letter, email, report, proposal, speech, story), and of formal, informal and neutral registers (See Appendix 1 for a breakdown of tasks at each level).

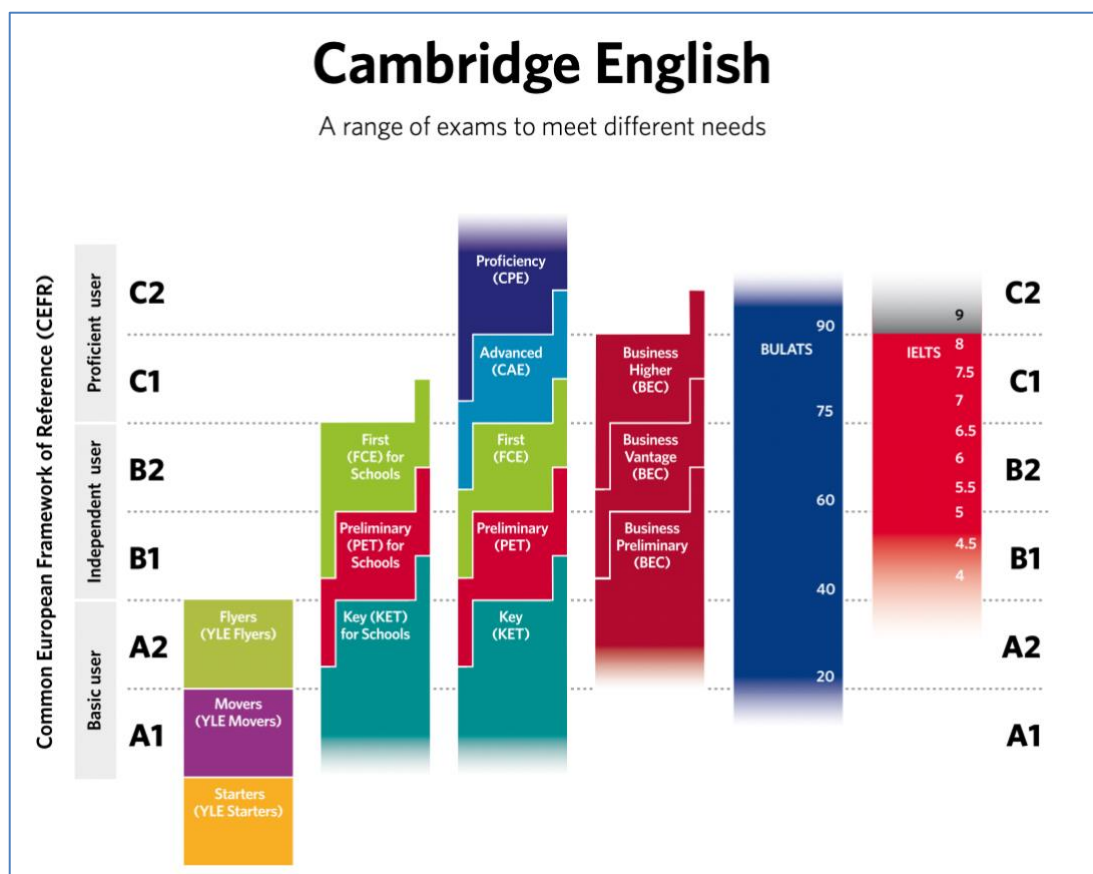


Figure 4.3 Range of Cambridge English qualifications benchmarked to the CEFR
<https://www.cambridgeenglish.org/Images/22695-principles-of-good-practice.pdf>

The main suite (illustrated in the third column from the left in Figure 4.3) encompasses five general English qualifications (KET, PET, FCE, CAE, CPE, aimed at attaining proficiency levels A2, B1, B2, C1, C2) which work together as a set so that ‘each exam builds on the skills [you] develop at the previous level.’ <https://www.cambridgeenglish.org/exams-and-tests/qualifications/general/>. Cambridge Assessment English make clear the importance of claims of alignment to the CEFR and highlights its endeavours to ensure empirical validation of its test calibration methodology and item-banking process. It attests a rigorous test developmental cycle, in which it describes the integration of test alignment with the CEFR (Taylor and Jones 2006; <https://www.cambridgeenglish.org/research-and-validation/validity-and-validation/>) As noted, all learner data is tagged both by exam taken *and* actual performance achieved. This proficiency level alignment and benchmarking is of particular relevance in this study for three reasons: firstly, as stated, overt alignment between levels in these exams allows for a quasi-longitudinal approach to comparing actual attainment and development between one level and another; secondly, using exams from the same suite

ensures a degree of comparability in the language content from one proficiency level to the next, and thirdly using performance-levelled data rather than exam-levelled data dilutes, somewhat, the effect of task, since performance data at a given level relates to more than one exam (e.g. performance level data at B1 constitutes data from three exams KET, PET and FCE).

Table 4.1 gives a breakdown of how the tokens per level are distributed across performance levels in the CLC sub-corpus (given in the columns A1 to C2) and across exams (in the rows 1 to 5). Total performance level tokens are given in row 6, shaded in green, and percentage distribution by performance level in row 7. A vertical reading of the table gives a breakdown of tokens by performance level achieved and the exams taken. A horizontal reading shows a breakdown of exam taken and the performance levels achieved in those exams. Figures in red indicate performance achieved *below* exam level, figures in black show performance level achieved *on a par* with the exam level taken and in blue show performance *above* exam level. For example, as there are no A1 exams, the tokens in the A1 performance column (2,456,971) reflect writing from candidates who took the A2 exam but did not meet the pass criteria for the exam and are awarded A, the level below. The shaded figures in green are the totals by performance level that are relevant for this study.

			PERFORMANCE LEVEL						Total
			A1	A2	B1	B2	C1	C2	
1	EXAM LEVEL	KET (A2)	2,456,971	713,366	89,709				3,260,046
2		PET (B1)		4,989,851	1,926,578	252,912			7,169,341
3		FCE (B2)			1,245,186	3,665,843	1,093,046		6,004,075
4		CAE (C1)				1,345,224	3,241,358	1,116,227	5,702,809
5		CPE (C2)					2,377,164	6,582,468	8,959,632
6		Total	2,456,971	5,703,217	3,261,473	5,263,979	6,711,568	7,698,695	31,095,903
7		%	7.9	18.34	10.49	16.93	21.58	24.76	

Table 4.1. Distribution of tokens across performance levels achieved and exams taken in the CLC main suite exam sub-corpus*

(*red indicates performance *below* exam level, black indicates performance *on a par* with the exam level and blue indicates performance *above* exam level)

In the mainsuite sub-corpus there are 140 first language backgrounds represented in the mainsuite data, in over 214,084 documents which vary in length from under 100 tokens at the lowest levels to just over 1000 at the C levels. The top 20 L1s overall distribution are illustrated in Figure 4.4. Clearly some language backgrounds contribute a larger proportion of the data than others.

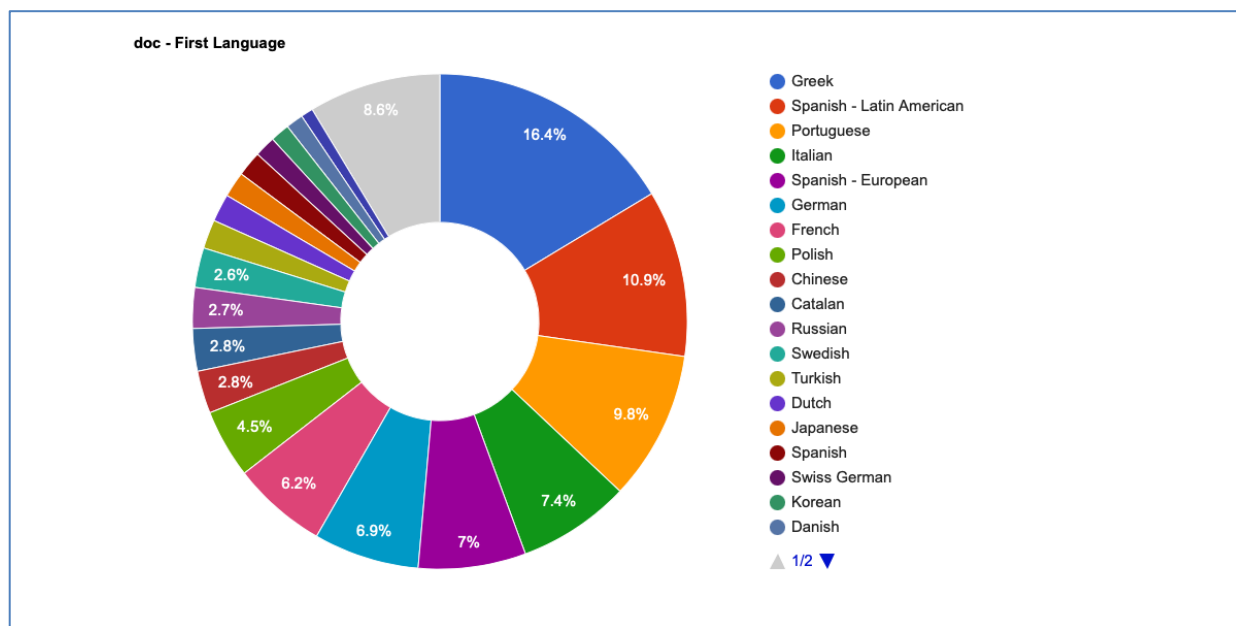
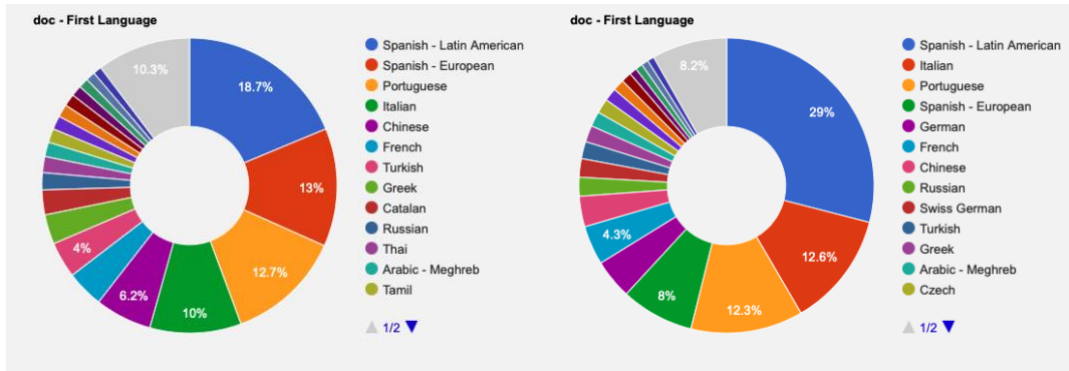


Figure 4.4 Breakdown of languages represented in the CLC mainsuite data across all levels

The detailed metadata and breakdown across levels allows for a breakdown of the data into 6 performance level subcorpora. This affords a comparative analysis across the whole sub-corpus, and or across any number of individual performance levels subcorpora. For example, the number of L1 backgrounds represented by level is given in Table 4.2 and the percentage contributions by level of the 20 most frequent are given in Figure 4.5.

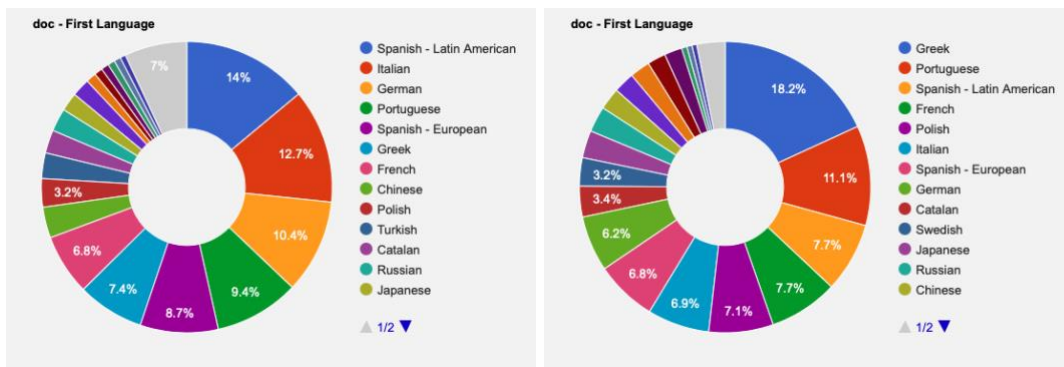
performance level	A1	A2	B1	B2	C1	C2
L1 backgrounds	116	108	94	75	73	67

Table 4.2 Number of L1 backgrounds by level



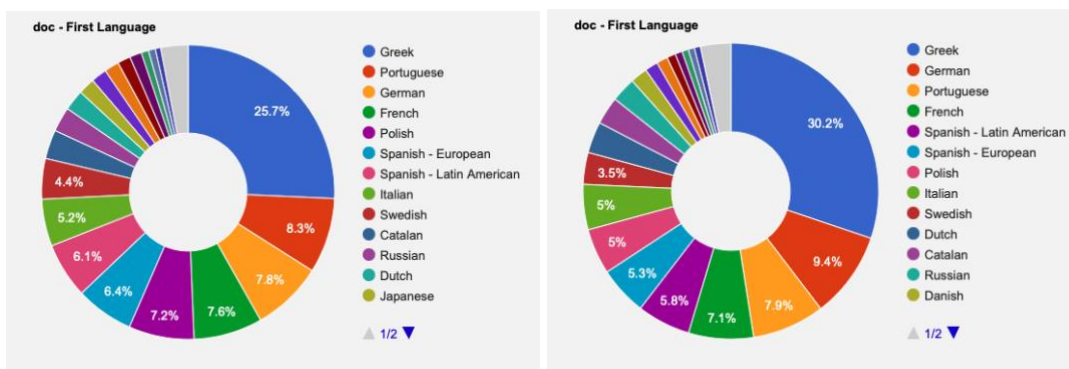
A1

A2



B1

B2



C1

C2

Figure 4.5 Distribution by L1 background across levels

As illustrated their distribution varies from level to level and the number of L1 backgrounds decreases as proficiency increases. The implications and limitations of this are discussed in Chapter 9.

The data sits on a bespoke platform of the Sketch Engine suite of tools, a customised interface for analyzing CLC data. A standard feature of the Sketch Engine tools is automatic POS tagging and lemmatization using the Treetagger tool set and The English Penn Treebank tagset. A summary of the tags can be found at Appendix 2. Each token has a POS label assigned to it, allowing for retrieval of n-gram sequences at varying levels of abstraction on a word, lemma, or POS level and/or a combination of these, using the Corpus Query Language (CQL) tool in the interface.

I now turn to the tools and approach used to retrieve POS tag sequences and describe the mixed methods approach applied to the analysis.

4.3 Mining the data: The tools and approach

How do you tackle development in large scale learner data? Many of the studies in development have taken a single item approach, be it verb argument constructions (VACs), or noun phrases or formulaic language, for example. The rationale for this is discussed in chapters 2 and 3. Targets of analysis have often been selected from an L1 usage perspective which runs the risk of situating L2 output in deficit terms and overlooking important features of the developmental journey. I have argued in previous chapters for the value of looking at learner language for its own sake and for this exploration to begin from *within* the L2 data.

Without abandoning focus on the verb, this research seeks to identify additional points of departure. Other studies (Diez Bedmar and Pérez-Paredes 2010; Biber and Gray 2016; Capelle and Grabar 2016; Kyle and Crossley 2017) have illustrated that the prevalence of the noun phrase and surrounding complexity is highly instrumental in development in learner language. In this study, I keep an open mind as to where the points of departure might emerge, with no preselection as the focus of analysis. I approach the data with an open view beginning with the learner data in its entirety, exploring what emerges from this haul, potentially capturing what might otherwise have escaped the spotlight.

I adopt a mixed methods approach and carry out the analysis in two phases, first to identify what syntactic sequences learners put together, and second to scrutinise sequence change and development in their frequency, distribution and use. In the first phase I take an exploratory approach and make use of the uppermost level of abstraction available in the toolset - POS tags - to trawl the data and quantitatively explore the frequency and distribution of sequences. I use a novel bottom-up approach and begin the retrieval with a series of open token slots, an n-gram approach aligning with Cappelle and Grabar (2016) and Gilquin

(2018). The CQL tool in the Sketch Engine interface allows for retrieval of open token slots of any number to produce lists of n-grams, abstracted on both a POS tag (node tag) level, or word level (node form) or a hybrid of both. This first phase seeks to find what individual parts of speech are put together, which sequences are most frequently used at each level and how this converges or diverges between levels.

Where the first phase involves trawling the data in its entirety, the second phase makes use of the frequency and distribution results from phase 1 and investigates individual POS tag sequences. In this phase the lexical exponents of individual sequences are retrieved, in this case using the Sketch Engine node form frequency tool, and type-token frequencies of lexical sequences are calculated across levels. The lexical and functional repertoires of both convergent and divergent POS tag sequences across levels are then investigated. This allows for (1) observation of how the frequency and distribution of POS tag sequences changes across proficiency levels (2) identification of any core POS tag sequences across all levels and (3) further investigation of how divergent and convergent sequences develop across levels. This comparative approach can be replicated across any learner data which is tagged by level and applied iteratively across any number of proficiency levels. It can be applied to L1 data to compare how development and usage in L2 data relates to L1 usage. Additionally it provides a mechanism to triangulate with other learner data sets at any point of proficiency, and/or L1 data.

This approach seeks to identify generalisations in how learners put together parts of speech to form meaningful strings of words. In doing so it finds sequences but does not claim that the sequences are constructions in a form-meaning mapping sense (Goldberg 2006). Some of them will be complete constructions, others will not. If language learning is a process of abstraction and generalisation, this methodology is trying to find a way into that process.

The two phase step by step process of the approach is illustrated in Figure 4.6. It shows a generic comparative approach which can be applied to a range of datasets.

PHASE 1 Retrieving POS sequences and identifying candidate for analysis	
STEP 1	Establish search query for n-gram POS tag sequence and retrieve all sequences across all data by variable (e.g. performance level)
STEP 2	Examine frequencies of occurrence and distribution of the n-gram sequence across selected variable (e.g. performance level), across a selected number of sequences (e.g. 10/20/50/100/1000).
STEP 3	Rank and number the results. Compare the rankings across selected variable (e.g. performance level). Trace convergence and divergence from front view and rear view perspectives.
PHASE 2 Analysing individual sequences identified in phase 1	
STEP 4	Select L2 proficiency levels for analysis. Establish sequences (of both convergent and divergent structures) for further investigation, using the ranking system of Phase 1.
STEP 5	Extract lexical realisations of sequences across selected L2 proficiency performance levels
STEP 6	Sort sequences into patterns
STEP 7	Apply functional descriptions to patterns
STEP 8	Identify lexical and functional growth across levels in relation to patterns
STEP 9	Compare across proficiency levels to identify developmental pathways

Figure 4.6 Applying a generic bottom-up iterative approach for retrieving and analysing POS tag sequences

The first phase addresses RQ1 by retrieving the frequency and distribution of POS tag sequences at each proficiency level in the data to discover, in quantitative terms, if there are POS tag sequences that mark both divergence and convergence across levels. In doing so, I am seeking to identify POS tag sequences that merit further investigation at varying levels of abstraction, e.g. at POS level, at a lexico-grammatical level, both for their potential role in transition between levels on the POS level of and analysis of what is core. This approach can be used with any n-gram POS sequences number. The second phase addresses RQ2 and RQ3, examining divergent and convergent sequences qualitatively to identify what they might reveal about development across proficiency levels, in terms of structural, lexical and functional growth and repertoire, and whether existing frameworks for classification of patterning account for development observed.

The phases can be applied iteratively to compare any number of L2 proficiency level datasets.

4.3.1 Phase 1

Establish search query and examine frequency and distribution across proficiency levels

First the number of items in the sequence is selected. Then using an open slot sequence of (e.g. 3-gram, 4, gram or 5-gram), the n-grams are retrieved. In the case of this study the data is searched to identify all occurring 4-gram POS tag sequences at each proficiency level and their distribution. At each level, all 4-gram sequences are retrieved, the most frequently occurring POS tag sequences identified, the total number of tokens at each level, and the occurrences per million word (PMW) are established. All sequences at all levels are collated. The next step is to look at how each sequence is distributed within each proficiency level and to trace the journey of these sequences.

Rank and number the results and trace convergent and divergent sequences

Having established a ‘master’ cohort of sequences, they are ranked and numbered in order from the most frequently-occurring to the least frequently-occurring, using a simple numbering and sorting function in Excel (For a sample, see Appendix 3). Here ranking is used as a proxy for distribution, making the assumption, after Ellis (2017), from a usage-based perspective, that learners have an implicit understanding of the frequency of items in usage (as described in Chapter 3). The frequency rankings are then compared across the L2 proficiency levels. This involves first comparing the rankings from both front view and rear view perspectives (Figure 4.7) and calculating the difference in ranking, using simple

subtraction to establish a rank difference; the front view begins with the most frequent sequences at the lowest level and compares how they are ranked at higher levels; in the rear view the highest frequency ranking sequences at the highest proficiency level are compared with how these rank at lower levels.

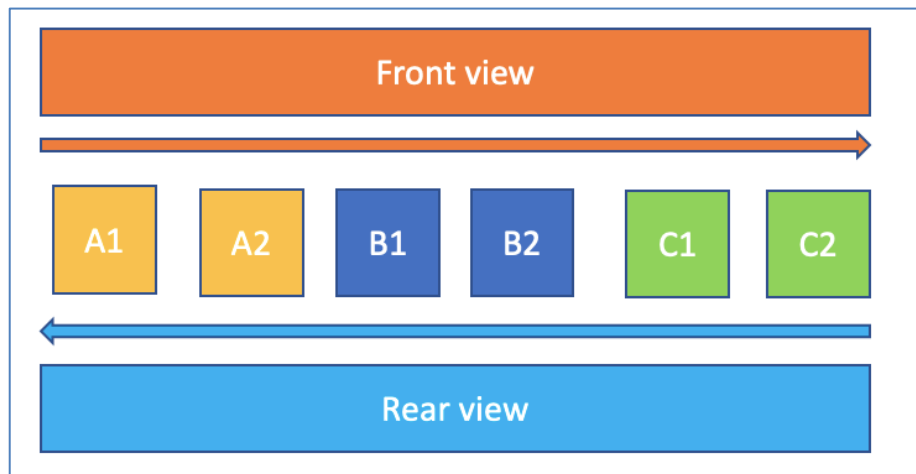


Figure 4.7 Front view from A1 to C2 and rear view from C2 to A1

Table 4.3 shows an example of how the rank difference is simply calculated. It shows the top 5 4-gram POS tag sequences at A2 level in the CLC and their rankings at each of the other five levels. It also gives the difference in their ranking. For example, sequence #1 ranks at #10 at A1 so it has a rank difference of -9, whereas it also ranks at #1 at B1, B2, C1, C2 giving it a rank difference of 0. This rank difference figure gives an indication of how closely a sequence ranks across the different levels. It gives an indication of convergence of ranking between levels. We can see for example that the highest ranking sequence at A2 remains 'core' across levels A2 to C2, i.e. it ranks consistently highly at all five levels.

A2 rank	Top 5 A2 POS tag sequences	ranking at other levels					rank difference				
		A1	B1	B2	C1	C2	A2 - A1	A2 - B1	A2 - B2	A2 - C1	A2 - C2
1	NN IN DT NN noun+prep+det+noun	10	1	1	1	1	-9	0	0	0	0
2	IN DT JJ NN prep+det+adj+noun	62	3	2	2	2	-60	-1	0	0	0
3	IN DT NN SENT prep+det+noun+.	3	2	3	6	6	0	1	0	-3	-3
4	IN DT NN IN prep+det+noun+prep	15	5	4	4	3	-11	-1	0	0	1
5	IN PPZ NN SENT prep+poss_pronoun+noun+.	4	6	8	17	18	1	-1	-3	-12	-13

Table 4.3 Top 5 POS tag sequences at A2 with frequency rankings at all other levels

At this point any number of sequences can be selected (e.g. top 10/20/100/1000 from each level). This approach can be applied to a comparison of any combination of the subcorpora as illustrated in Figure 4.8 to allow for a view on development at any point in the developmental process. For example two adjacent levels can be compared, or all levels, or any or all of the subcorpora with another corpus, using the same process.

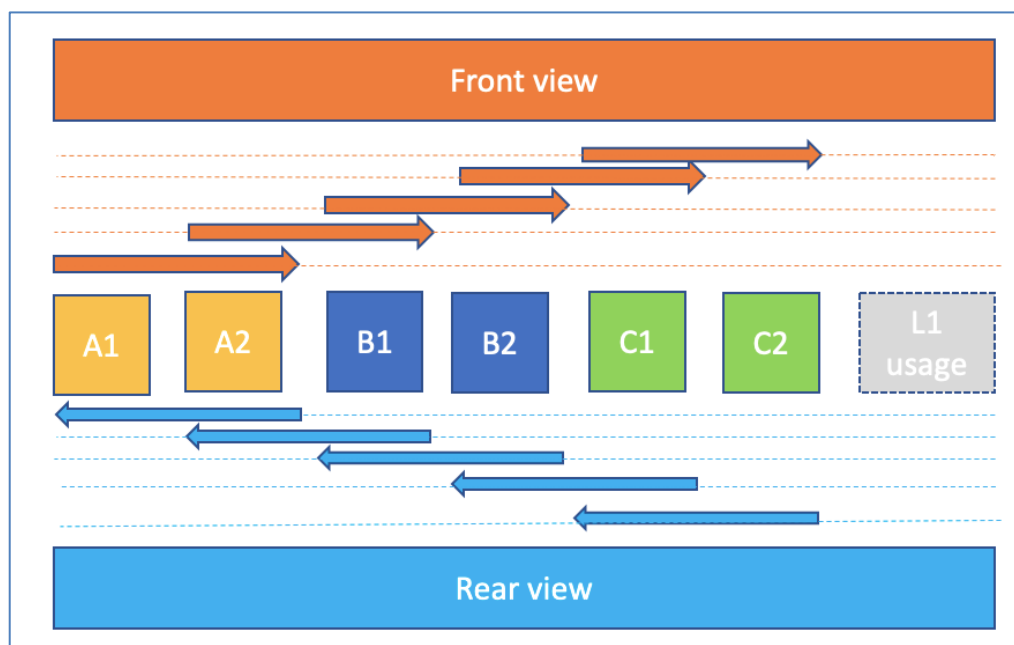


Figure 4.8. Representation of front and rear view comparison on individual subcorpora

There are two main reasons for this combination of views: firstly to capture those sequences which are most frequently used at lower levels, and investigate whether their rankings change, and to what extent, as learners gain competence with the language; and secondly to track back from the most frequently used sequences at the highest level, to identify which emerge as more frequently used, which sequences disappear as they become less frequently used and at what point these sequences emerge or disappear. It allows a view on what changes syntactically from level to level and what remains the same and offers a perspective on competence as it emerges from lowest to highest level and as it evolves from highest to lowest. It allows both a window into the syntactic profile at any one point along the proficiency scale as well a developmental view of what is transitional, what has gone before, what is to come, what is consistently ‘core’ to all levels and differences in convergence. It also validates the status of learner language as a variety worthy of investigation without the need to contrast it necessarily against L1 varieties using the deficit hypothesis.

4.3.2 Phase 2

In the second phase, I explore in more depth those sequences identified in phase 1 as being of interest for their potential to (1) play a transitional role in the development (2) illustrate what is consistently core in development. Individual sequences are selected for these two reasons and further examined.

For each of the sequences selected, lexical exponents are retrieved for each level under investigation. The type token frequencies for the lexical sequences per level are calculated.

The next steps (6 to 9) involve manual investigation, identifying patterns of use and comparing the lexical and functional use of the lexical components across levels, with a view to observing a developmental pathway for any one given POS sequence. For this analysis I use a combination of existing frameworks for classification, drawing on pattern grammar (Francis *et al.* 2000), lexical bundles (Biber *et al.* 2014) and p-frames (Garner 2016) and through this approach I observe whether existing categorisation frameworks can be used to account for development. For example in order to apply a pattern grammar approach the sequences are sorted into patterns, driven by a generalisation of the pattern, from a hybrid of word and POS (N *of the* N, *end of the* N, *Vpast to the* N, *went to the* N) and then aligned with the meaning groups of pattern grammar (Francis *et al.* 2000). (Subsequent chapters offer for further exemplification.)

PHASE 2 Analysing individual sequences	
STEP 5	Extract lexical realisations of sequences across selected L2 proficiency performance levels
STEP 6	Sort sequences into patterns
STEP 7	Apply functional descriptions to patterns
STEP 8	Identify lexical and functional growth across levels in relation to patterns
STEP 9	Compare across proficiency levels to identify developmental pathways

4.4 Methodology discussed

N-gram size

The decision of the size of the n-gram to retrieve is, to a degree, arbitrary. The nature of this study is exploratory and to this end the aim is to work with a number which captures sequence usage and lexico-grammatical generalisations as best as possible. I have chosen to go with 4-gram sequences. For this I take inspiration from other exploratory work: Biber *et al.*'s (1999) approach to lexical bundles and Cappelle and Grabar's quest to build an n-gram of English, using POS tags to identify frequency structural strings in COCA (2016). Biber *et al.*'s findings that lexical bundles of 3-grams are highly frequent, and often have an extended collocational association, whereas 4-grams and more are more phrasal in nature suggest a minimum of 4. They also point out that 4-word bundles can also co-occur to form 5- or 6-gram bundles and characteristically do not represent a structural unit. Cappelle and Grabar (2016) use 5-gram sequences and justify their choice by maintaining that 5-gram sequences can both harbour shorter sequences and can be extended 'manually'. The subject of their focus is intermediate to advanced level learners. They propose that the longer, more complex sequences generated by 5-grams would suit their data but that a shorter string would be more suitable for lower level data. Since the present study encompasses all levels on the proficiency scale, except for absolute beginners, a 4-gram sequence is used as a starting

point. A 4-gram approach allows for exploration of nested n-grams (e.g. (IN) DT (JJ) NN (preposition +) determiner (+ adjective) + noun), as well as extension of the 4-gram through colligational and collocational analysis of items preceding or following the retrieved sequence. By way of example, following the Phase 1 steps, a simple pilot study using a 4-gram search was applied to the CLC sub-corpus. A 4-gram POS tag sequence was used to search on each of the 5 performance levels. A one-million line random sample was then used for each level and all 4 POS-gram sequences retrieved.

Table 4.4 (next page) shows the top ten most frequent 4-gram POS tag sequence results for performance levels A1, A2, B1 numbered 1-10, and their raw and relative frequencies in the CLC sub-corpus. The first row for each level indicates the number of types of 4-gram POS sequences for each level. Initial observations show evidence that

- (1) There is overlap between levels.
- (2) There is overlap within sequences. For example, compare **NN IN DT NN noun + prep + det + noun** and **NN IN DT JJ noun + prep + det + adjective**, where noun + prep + det **NN IN DT** are common to both.
- (3) at face value, some of the sequences do not look structurally ‘complete’, in that a fifth item, collocating with the four in the sequence is predictable. For example **DT JJ NN IN** (e.g. *the yellow door in*) where a following noun phrase complement is likely or **SENT PP MD VV** (e.g. *. I would like*) where a following verb phrase or noun phrase is necessary.

Many of these sequences show the tag SENT, which indicates a full stop. The POS sequence search process includes the punctuation tags SENT and [,] as a POS tag. Being able to observe sentence boundary use may prove to be enlightening, especially in the early stages of development. As observable in Table 4.4, SENT appears as a tag in six of the top 10 in A1, five of the top ten in A2, decreasing to three at B1. To begin with, while it might be tempting to exclude the punctuation tags, I have elected to leave them in when looking at an overview of the POS sequences distribution and development, since they are frequently represented in recurring patterns, particularly at the lower levels (Chapter 5). As Gilquin points out (2018) relying on a specific POS tag set might be seen to compromise a corpus-driven approach in its purest sense since by their nature they impose a top down system of classification. In this regard I align with Gilquin’s justification that by first allowing the automatic retrieval of patterns with no intervention in the first phase of this analysis “suggests that the initial stages of the analysis are sufficiently atheoretical to qualify as a corpus-driven study” (2018, p.6).

However when drilling down to identify the POS tag sequences at each level (Chapters 6 to 8), and when looking at lexis and function, since the scope of this study does not allow for detailed analysis of every sequence, sequences that include a SENT tag are excluded.

A1	raw	%	A2	raw	%	B1	raw	%
SENT PP MD VV .+pronoun+modal+verb <i>. I would like</i>	14540	0.61	NN IN DT NN noun+preposition+determiner+noun <i>centre of the town</i>	24907	0.44	NN IN DT NN noun+preposition+determiner+noun <i>centre of the town</i>	14267	0.44
PP MD VV IN pronoun+modal+verb+preposition <i>You can come to</i>	12916	0.55	IN DT JJ NN preposition+determiner+adjective+noun <i>to a new shop</i>	23240	0.41	IN DT NN SENT preposition+determiner+noun+. <i>to the cinema.</i>	12606	0.39
IN DT NN SENT preposition+determiner+noun+ <i>in the morning.</i>	10676	0.45	IN DT NN SENT preposition+determiner+noun+. <i>in the morning.</i>	22602	0.40	IN DT JJ NN preposition+determiner+adjective+noun <i>on the other hand</i>	11001	0.34
IN PPZ NN SENT preposition+posspronoun+noun+. <i>to my house.</i>	9369	0.40	IN DT NN IN preposition+determiner+noun+preposition <i>in the centre of</i>	19831	0.35	NP NP NP NP proper noun x 4 <i>(any text with capital letters)</i>	10788	0.33
NN SENT PP VVP noun+pronoun+presentsimple <i>phone. I like</i>	9302	0.40	IN PPZ NN SENT preposition+posspronoun+noun+ <i>in my town.</i>	19663	0.35	IN DT NN IN preposition+determiner+noun+preposition <i>in the centre of</i>	9851	0.31
NN SENT PP MD noun+pronoun+modal <i>music. I can</i>	8854	0.37	PP MD VV IN pronoun+modal+verb-base+preposition <i>you should go to</i>	18026	0.32	IN PPZ NN SENT preposition+posspronoun+noun+. <i>in my room.</i>	9555	0.30
DT NN SENT PP determiner+noun+pronoun <i>another country. I</i>	8452	0.36	SENT PP MD VV .+pronoun+modal+verb <i>. I would like</i>	16556	0.30	DT NN IN DT determiner+noun+preposition+determiner <i>the end of the</i>	9413	0.29
NP NP , PP noun+noun+,pronoun <i>Dear Sam, I</i>	8357	0.35	NN SENT PP VVP noun+pronoun+presentsimple <i>phone. I like</i>	15737	0.28	DT NN SENT PP determiner+noun+pronoun <i>the music. I</i>	9396	0.29
PP VVP TO VV noun+preposition+determiner+noun <i>you want to come</i>	7696	0.32	DT JJ NN IN determiner+adjective+noun+preposition <i>the other side of</i>	15402	0.27	DT JJ NN IN determiner+adjective+noun+preposition <i>a great time with</i>	8379	0.26
NN IN DT NN noun+preposition+determiner+noun <i>concert in the morning</i>	7592	0.32	DT JJ NN SENT determiner+adjective+noun+. <i>the yellow door.</i>	15338	0.27	TO VV DT NN to-infinitive+verb+determiner+noun <i>to make a film</i>	7914	0.25

Table 4.4 Top 10 4-gram POS sequences for each proficiency level in the CLC sub-corpus, by raw and per million word (PMW) frequencies

Sequence ranking

The approach taken in this study allows for any number of sequences to be ranked. Taking into account Zipf (1935) and frequency effects (see Chapters 1 and 3), using the most frequently occurring sequences accelerates an efficient zoning in on what is most used. This is illustrated in Table 4.5 using a data sample. In the CLC sub-corpus, the top 100 4-gram sequences represent between only 0.09% (at A level) and 0.03 % (at C2 level and in the BNC) of the possible 4-gram types. However, in terms of distribution of occurrence, the top 100 4-grams represent between 19.14% (A1) and 12.67 % (C2) of all 4-gram tokens/occurrences in the sample, a not insignificant proportion of all occurrences. Even an analysis of the top 20 sequences covers between 5 and 6% of all occurrences while only representing between 0.02 and 0.03 of all 4-gram types.

Level	No. of 4-gram sequence types	Top 100 types as % of all types	Top 100 type occurrences as % of total occurrences	Top 20 types as % of all types	Top 20 type occurrences as % of total occurrences
A1	110703	0.09	19.14	0.02	6.90
A2	200384	0.05	15.99	0.02	6.03
B1	164828	0.06	14.55	0.02	5.38
B2	230464	0.04	13.16	0.02	5.15
C1	278605	0.04	12.68	0.02	5.34
C2	299916	0.03	12.67	0.02	5.56

Table 4.5 relative distribution of Top 100 4-grams as types and tokens

In line with Zipf (1935) we are observing that the most frequent 4-gram types account for a large number of occurrences. The same phenomena is evident when we look at the lexical realisations of each sequence. Looking at 1 POS n-gram affords a perspective on thousands of lexical n-grams. Looking at 100 or 1000 increases the perspective exponentially.

Whereas studies on recurrences in learner language typically look at generalisations at the word level, this study begins at the structural level as a way to understand structural generalisation. Typically, as seen in Chapter 2, analysis of lexical bundles moves from lexis to structural and functional generalisation, whereas a POS-grams approach offers first a perspective on structure and then on lexis and function. Looking at POS n-grams allows for generalisations beyond task and context. For example, Table 4.6. shows, by way of contrasting an n-gram approach with a POS- sequence approach, that we can immediately see regularities in terms of structure through POS tags in a way that lexical n-grams does not afford. Table 4.6 shows both the top 20 most frequent POS 4-grams alongside the most frequent lexical 4-grams in a 1 million concordance line sample of A2 CLC sub-corpus. From Table 4.6 we can see that NN IN DT NN (noun + preposition + determiner + noun) is the most frequently occurring sequence in the A2 data and yet there are no lexical exponents in the 4-gram lexical sequences which have this pattern. In contrast the first 4-gram lexical sequence (*How are you?*) is made of up WRB VBP PP SENT which does not appear in the top 20 A2 sequences.

A2	4-gram POS tag sequences	4-gram lexical sequences
1	NN IN DT NN	<i>How are you ?</i>
2	IN DT JJ NN	<i>are you ? I</i>
3	IN DT NN SENT	<i>I 'm going to</i>
4	IN DT NN IN	<i>, How are you</i>
5	IN PPZ NN SENT	<i>. See you soon</i>
6	PP MD VV IN	<i>to go to the</i>
7	SENT PP MD VV	<i>I would like to</i>
8	NN SENT PP VVP	<i>Would you like to</i>
9	DT JJ NN IN	<i>you ? I 'm</i>
10	DT JJ NN SENT	<i>in the centre of</i>
11	NN IN PPZ NN	<i>! How are you</i>
12	DT NN SENT PP	<i>. Would you like</i>
13	NP NP NP NP	<i>. I think that</i>
14	VV IN DT NN	<i>to a new shop</i>
15	PP VVP TO VV	<i>. I hope you</i>

16	DT NN IN DT	<i>I think you should</i>
17	NP NP , PP	<i>. I do n't</i>
18	PP MD VV PP	<i>are a lot of</i>
19	SENT PP VVP PP	<i>for your letter .</i>
20	JJ NN SENT PP	<i>I went to the</i>

Table 4.6 Top 20 4-gram POS tag sequences and lexical 4-grams in a sample of A2 data

Through taking a POS tag approach not only is it possible to observe the syntactic generalisations made by learners at each level, but we are able to take one of those generalisations and explore it in greater detail at the lexical and functional level. This filtering process is illustrated in Figure 4.9.

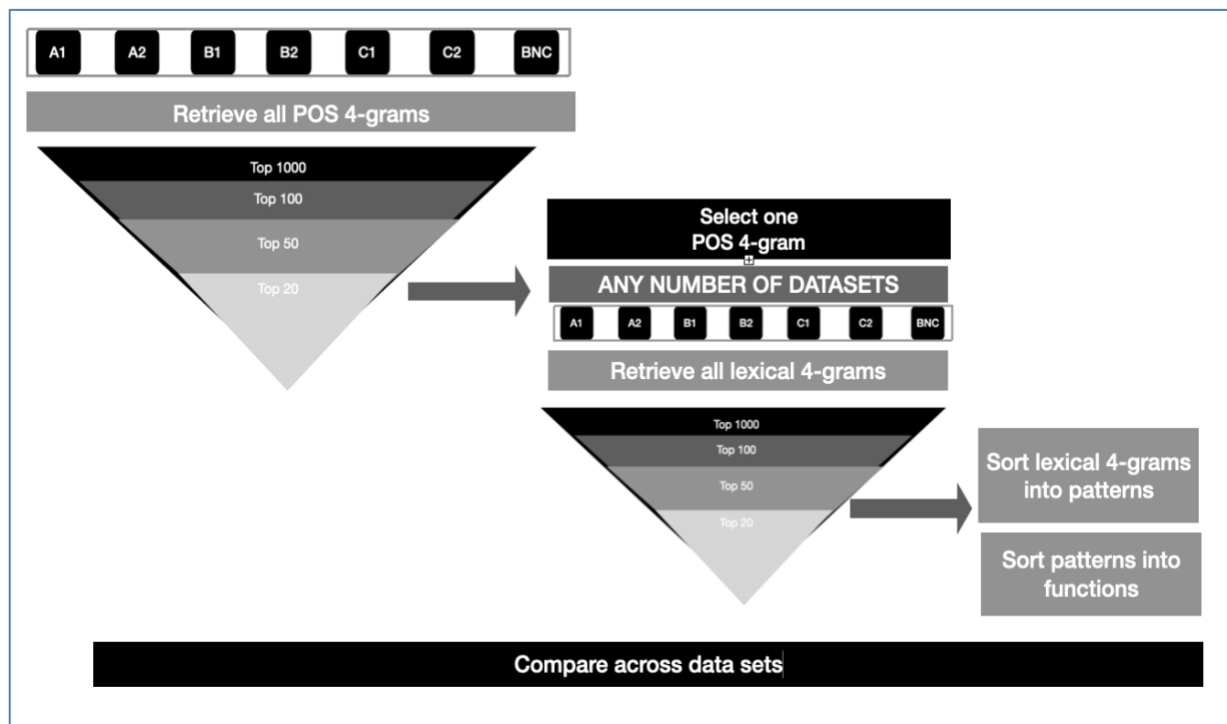


Figure 4.9 Retrieval and filtering process for investigation of POS tag sequences

4.5 Towards a methodology for bottom-up lexical and functional analysis

There is not a one size fits all approach to applying this to L2 data. Existing frameworks for categorising form and function (e.g. Pattern Grammar) (Francis *et al.* 1996, 1998; Hunston and Francis 2000), lexical bundles (Biber *et al.* 1999, 2004), Constructions (Goldberg 1995, 2006) and VACs (Ellis *et al.* 2016) work with the frequencies that are found in L1 data. Part of the motivation behind this study is to generalise about those elements which are put

together right from early proficiency levels. However, as already seen in Chapter 3, very often these lexical bundles, grammar patterns, constructions that are frequent in L1 data are scarce in learner data, particularly at the lower levels of proficiency. This may present a problem when categorising both form and functional usage in L2 data. While Sinclair (2004) argues that words do not ‘constitute independent selections’, there may indeed be evidence in learner language of a kind of independence of selection of highly frequent concrete words that are selected because they fill a grammatical slot. This would be in line with the slot and frame developmental process of a usage-based approach (Ellis 2012), in which learners begin to understand the relationship between words and the syntagmatic slots they can fill, but not yet the extent of the relationship between multiple sequences of words in multiple slots. Alongside these independently selected items there is a need to describe the emergence, in the learner data, of collocational knowledge - understanding of ‘co-selection’ (Sinclair, 2004), sequences of words that go together (e.g. compare the relatively independent selection of *a + new + shop + in* and *a wide range of*, both instances of the same POS-gram sequence), and the movement from slot and frame to fully abstracted system. This points to a need for a functional taxonomy which incorporates both combinations of high frequency words and ones where the co-selection factor is stronger and which does not exclude any lexical instantiations of a POS sequence because it does not meet the definition of a bundle or a construction or a pattern. For this reason I opt to categorise the sequences on a case by case basis, applying and adapting existing frameworks as relevant. In summary as illustrated in Figure 4.10, taking a POS-gram approach allows for an expansive bottom-up approach at the outset that potentially allows for analysis of all linguistic features.

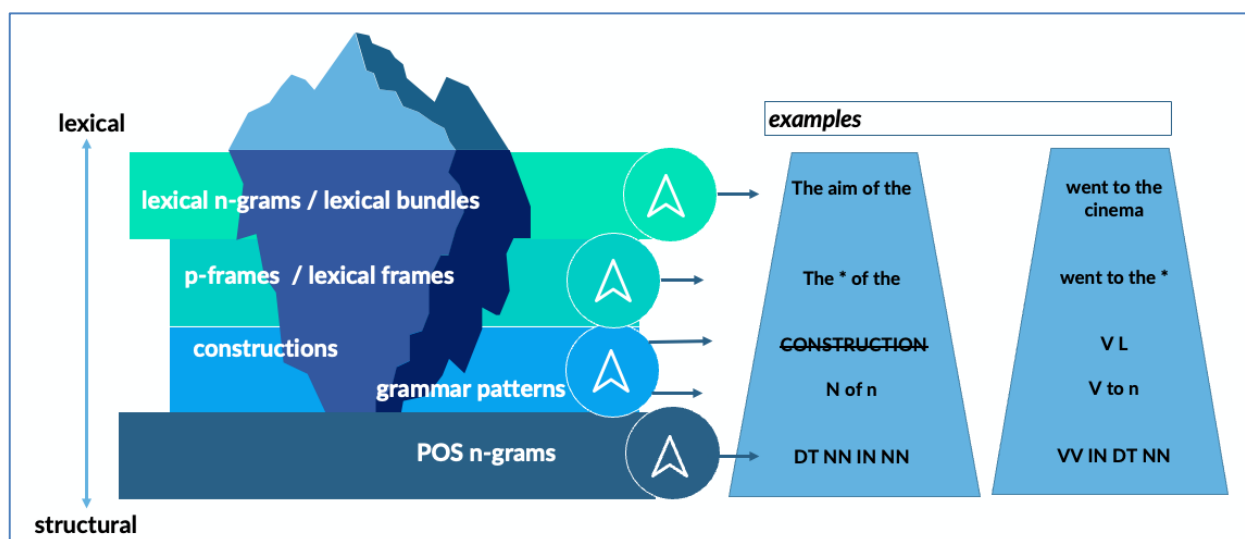


Figure 4.10 Description of bottom-up POS tag sequence approach

4.6 Summary

This chapter began by considering the type of data needed for a study on development, and the importance of bench-marked stages of proficiency. In 4.2, the Cambridge Learner Corpus (CLC) and the CLC sub-corpus used in this study was described and its appropriateness for this study. In 4.3, the methodological approaches used in the study were outlined, followed by an example of how the mixed methods approach is applied (4.4). In summary this study comes at a cross-roads between SLA and LCR taking elements of both to address the research questions identified at the beginning of this chapter. Figure 4.11 situates this study in relation to the two fields.

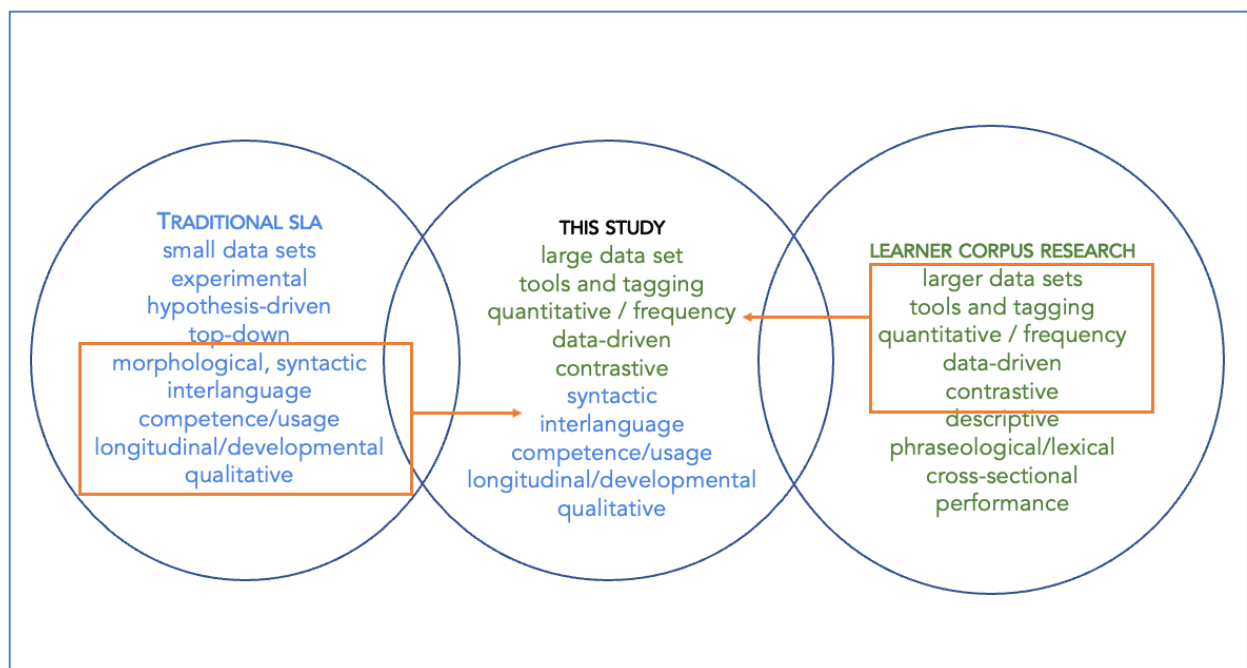


Figure 4.11 Illustration of the common ground between this study, traditional SLA and LCR

Chapter 5 follows with a summary view of development from A1 to C2. It provides an illustration of the methodology in practice, demonstrating the approach and describing initial global findings.

Chapter 5 Scanning the landscape: looking forward and looking back

5.0 Introduction

This baseline chapter presents a summary view of development across all six CEFR levels of proficiency, from both a front view (low to high proficiency) and rear view (high to low proficiency) perspectives. It looks at the developmental landscape on a global level before picking up on the detail which is scrutinised in Chapters 6 to 8 across different levels and stages of the learner developmental journey.

The investigations outlined in the chapter begin to address all research questions:

RQ1 Is development in L2 writing observable through the frequency and distribution of POS sequences across proficiency levels?

RQ2 How does POS sequence usage develop across proficiency levels?

RQ3 Can existing frameworks for classification of language patterning account for a description of development in L2 writing?

Usage-based models of language learning have frequency at their core and so we begin with an assumption, after Ellis (2017), that learners have an implicit understanding of the distributions of items in usage and how these items are put together. Observing the statistical properties of language across different levels of proficiency can help understand how learners make sense of their language experience. As already described in chapters 2 and 3, research illustrating distributions of frequencies in language show how Zipf's law (1935) (which describes the relationship between the frequency of an item and its frequency rank) is observable as a universal phenomenon across language, and can be seen, not only across words, but, relevant to this study, across sequences of words (Ellis *et al.* 2016) and phrases (Ryland Williams *et al.* 2015). Here I make the assumption that this power law can be observed in the frequency ranking of POS tag sequences occurring in learner language, and to this end I approach the data using ranking of sequences as a proxy for distribution. Among the thousands of possible POS tag sequences that learners might put together, it is possible to observe which are the most frequently used and whether this changes as language proficiency increases.

I begin with an overall view of the distribution of patterns across all proficiency levels to illustrate the power law in action. Frequency rankings and distribution of A1 4-gram POS sequences are then compared, using their rank difference (see chapter 4), in terms of their

convergence and divergence with other proficiency levels. This comparison is repeated first with C2 sequences looking back at development and then with all other intermediary levels (A2, B1, B2, C1) looking forward and back. This is followed by a simple distributional picture of the types and frequencies of POS sequences and their structural characteristics across all levels. I conclude the chapter with case studies of two sequences which exemplify development in relation to their lexical exponents and functional profiles.

5.1 Frequency ranking and distribution: overall view

Initial observations of the data indicate a Zipfian-like distribution of 4-gram POS sequences across all levels. This is illustrated by plotting the normalised frequencies for each level in a distribution graph. Figure 5.1 illustrates how the top 100 POS tag sequences (x axis) for each level are distributed in normalised frequencies per 1 million words (y axis). It shows that the highest ranking sequences account for a higher proportion of occurrences, followed by a decrease and levelling out of frequencies, characterised by the long tail pattern.

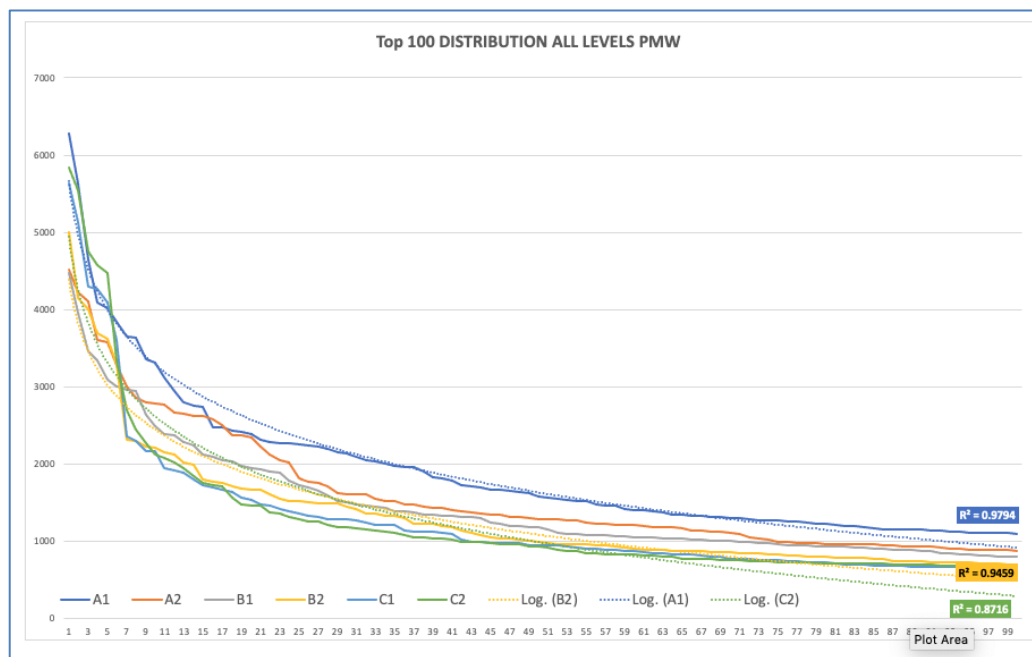


Figure 5.1 Distribution of the top 100 sequences in normalised frequencies across all proficiency levels

Logarithmic trendlines for the A1, B2 and C2 frequencies are shown, with values of 0.98, 0.95, and 0.87 respectively, indicating a good fit of the line to the data. (It is noticeable that the best fit of the line to the data is at A1 and may be an indicator of development and the

internalisation of the statistical properties of language. The distributional profile for each level is subsequently described in detail in chapters 6 to 8.)

5.1.1 All 4-gram POS tag occurrences by level

4-gram POS tag sequences were extracted from all levels of the data in the CLC main suite exams sub-corpora (see chapter 4) using a bespoke *Sketch Engine* platform, and databases of all sequences were compiled (For a small sample of the database, see Appendix 3). Table 5.1 gives (1) results for total occurrences of 4-gram POS sequences per level retrieved from each subcorpus (2) total individual 4-gram POS sequence types, and (3) subcorpus size in tokens per level. I note that at this stage in order to preserve a truly data-driven, bottom-up approach this includes sequences containing punctuation. I also note that there is a discrepancy between the subcorpus token sizes and the expected number of 4-gram occurrences. Since the discrepancy diminishes as proficiency increases (7.1% at A1, 3.6% at A2, 2.5% at B1, 1.4% at B2, 0.9% at C1, 0.8% at C2) it is suspected that the lower levels present greater difficulties in tag assignment. I will come back to general issues of tagging and punctuation, however a detailed investigation of tag assignment is beyond the scope of this study because of limited access to the raw data.

	4-gram POS sequences	A1	A2	B1	B2	C1	C2
1	occurrences	2293600	5496831	3183197	5190020	6648802	7640531
2	types	110703	200384	164828	230464	278605	299916
3	subcorpus size (tokens)	2456971	5703217	3261473	5263979	6711568	7698695

Table 5.1 Types and occurrences of POS 4-gram sequences per level

The POS tag sequences were ranked in order of frequency at all levels of proficiency and rankings compared across all levels. Frequencies were normalised by a factor of a million.

5.1.2 Top 100 4-gram POS tag occurrences by level

To gain an overall snapshot of development in this chapter I focus in on the top 100 at each level. 100 sequence types represent between 0.09% (at A level) and 0.03 % (at C2 level) of the possible 4-gram types, however, in terms of distribution of occurrence, they represent

between 19.14% (A1) and 12.67 % (C2) of all 4-gram tokens/occurrences in the sample, a not insignificant proportion of all occurrences, as shown in Table 5.2:

4-gram POS tag sequences	A1	A2	B1	B2	C1	C2
Total occurrences Top 100	439012	879148	463240	682966	842976	967812
Top 100 sequences as % of all sequence types	0.09	0.05	0.06	0.04	0.04	0.03
Top 100 as % of all sequence occurrences	19.14	15.99	14.55	13.16	12.68	12.67

Table 5.2 Top 100 4-gram POS tag sequences as percentage of all 4-gram POS sequences

The top 100 are the highest ranked and as such they are the most frequently occurring which means it is possible to observe a lot from relatively little. I next look at the difference in rankings (of sequences) across levels.

5.2 Overall view: a picture of convergence and divergence

Looking at change in sequence rankings is designed to gain an overall view of development. This is done through both a front view and rear view perspective on the data (Figure 5.2), first comparing the ranking and distribution of sequences at A1 level and their rank differences across subsequent higher levels of proficiency. The same is then done for C2 sequences across lower levels of proficiency. (See Chapter 4 for methodology and rationale.)

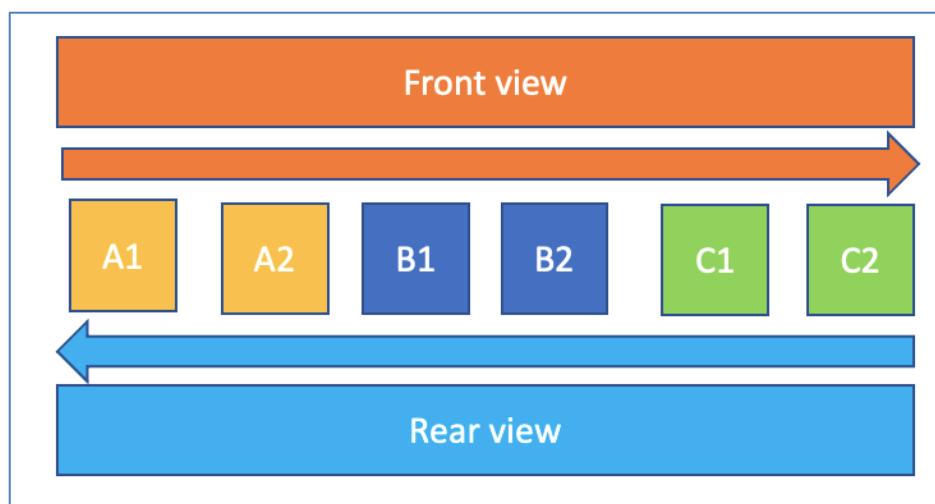


Figure 5.2 Front view from A1 to C2 and rear view from C2 to A1

The purpose of these two perspectives is to give a snapshot of usage from lower levels up *and* from higher proficiency levels back, and a sense of whether development is observable through frequency and distribution of POS tag sequences. It affords a view of the convergence and divergence of sequence distribution across levels in two directions. It allows us to see:

- (1) which of the sequences most frequently used at A1 level continue to be relied upon as proficiency increases (5.2.1 below) and
- (2) which of the sequences most frequently used at C2 level were also relied upon at lower levels (5.2.2 below).

5.2.1 Overall perspectives: front view, A1 to C2

We first look at the front view, the sequences which are most frequently used at A1. Figure 5.3 shows the percentage of divergence and convergence between the top 100 A1 sequences when comparing their frequency of occurrence and distribution at other levels (A2 to C2), based on the rank difference categorisation. Colours within each bar indicate bands of differences in ranking (shown in the table legend). They range from those that are closest in rank across levels by a rank difference of +/-5 (dark green sections) to those that are not found anywhere in the data at subsequent levels (red sections). Numbers within each bar show the percentage of sequences occurring within each band.

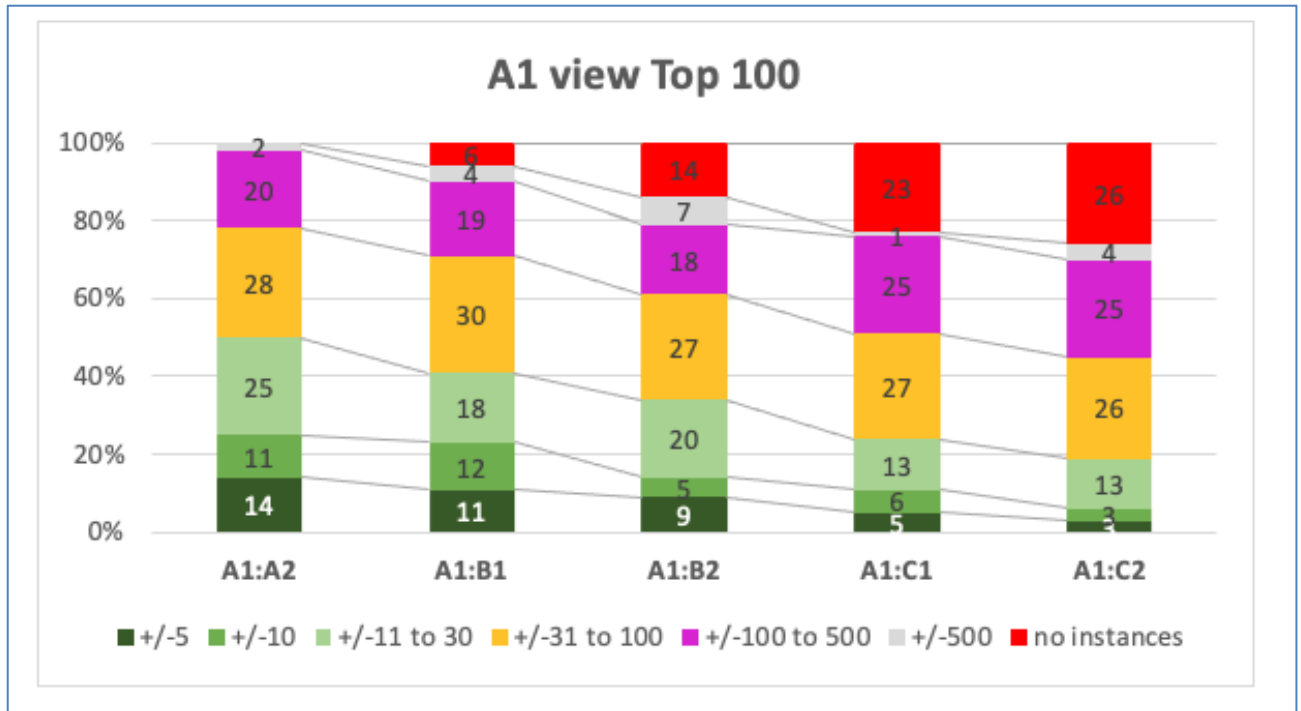


Figure 5.3 Top 100 A1 and C2 sequences: convergence and divergence across all levels (by percentage)

Taking each categorisation in turn, they illustrate the closeness in ranking of the A1 sequences in relation to where they are ranked at other levels. For example, 14% of the top 100 A1 sequences are also found within a ranking of +/-5 of all A2 sequences (indicated in dark green), whereas only 3% of the top 100 A1 sequences are found within a ranking of +/-5 of all C2 sequences. When comparing adjacent levels, all of the top 100 A1 sequences are also found in the A2 data, albeit to varying degrees of convergence. However, as proficiency increases so does the percentage of A1 sequences which are used with decreasing frequency at C2 or not at all. There is increasing divergence of sequence usage between A1 and C2. Across all levels the number of sequences not found anywhere in each data set increases as proficiency increases (indicated in red). 6% of the sequences that are found in the top 100 A1 data are not found anywhere in the B1 data, this rises to 26% which are not found at all in the whole of the C2 data. Over a quarter of the sequences that feature in the top 100 in A1 learner use are no longer used by C2 learners. 45% of all A1 sequences are found within +/-100 ranking at C2.

5.2.2 Overall perspectives: rear view, C2 to A1

Turning now to the rear view perspective, from the C2 data, Figure 5.4 shows the percentage of divergence and convergence between the top 100 C2 sequences and other levels, based on the rank difference categorisation.

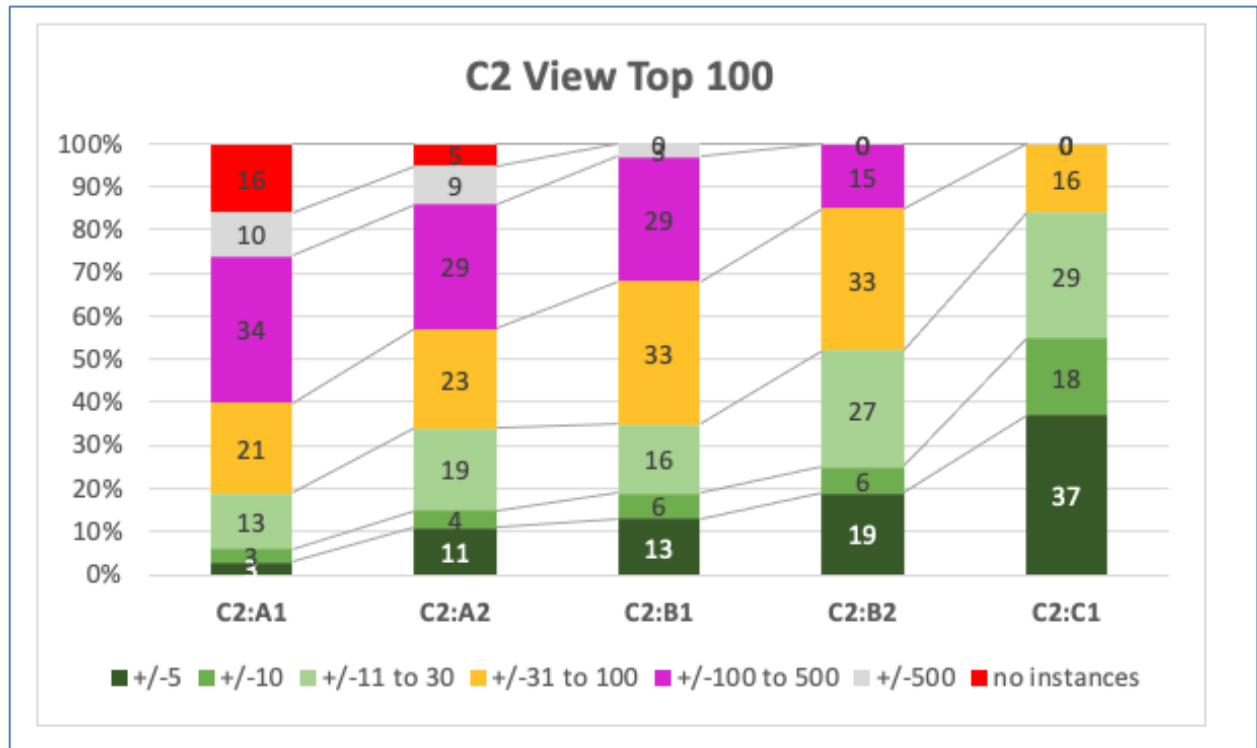


Figure 5.4 Top 100 C2 sequences: convergence and divergence across all levels

Taking each categorisation in turn, they illustrate the closeness in ranking of the C2 sequences in relation to their rankings at other levels. For example, 37 of the top C2 sequences are also found within a ranking of +/-5 of all C1 sequences, whereas only 3 of the sequences are also found within a ranking of +/-5 of all A1 sequences. The percentage of convergent sequences increases and the percentage of divergent sequences decreases as proficiency increases. A growing core of consistently ranked sequences is observable.

100% of the top 100 C2 sequences are also found to be ranked within +/-100 in the C1 data, albeit to varying degrees of convergence, 84% of which are found within +/-30 ranks at C1.

5.2.3 Overall perspectives: A2 to C1

Having taken a view from both ends of this proficiency scale (A1 and C2), in this section I describe an overall view from the perspective of each of the levels in between: A2, B1, B2 and C1. The purpose is not to look at the detail of specific sequences but to observe

tendencies of how sequence distribution develops from any point within the developmental process (illustrated in Figure 5.5). Adjacent levels, all levels, any or all of the subcorpora, can be compared using the same approach.

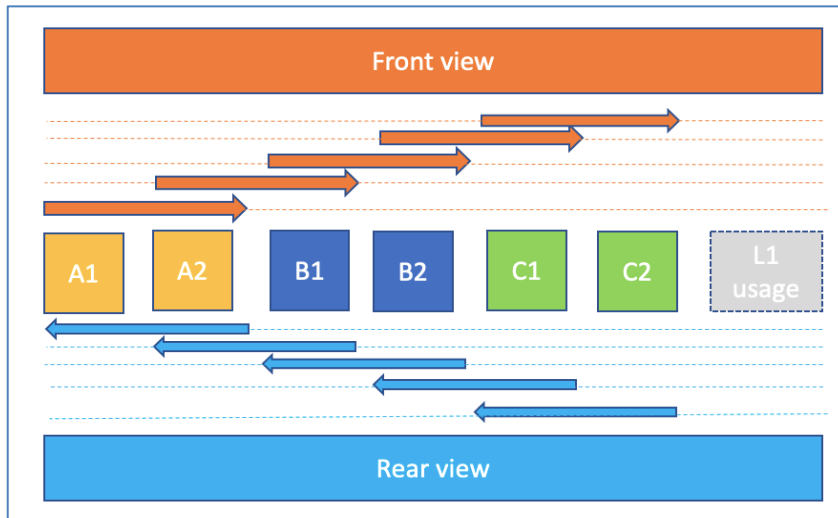


Figure 5.5. Representation of front and rear view comparison on individual subcorpora

A2 view

Figure 5.6 illustrates the A2 data perspective, a comparison of the top 100 A2 sequences and their rankings at all other levels. 76% of sequences found in the top 100 at A2 are ranked within +/- 30 at B1 (all green sections), indicating strong convergence of sequence use between these two adjacent levels. 28% of these are core to both, ranking within +/-5 of each other (dark green sections). 84% of sequences are found within +/- 100 ranking at B2, indicating strong convergence. 92% and 90% of all A2 sequences are found within +/- 100 ranking at C1 and C2, respectively.

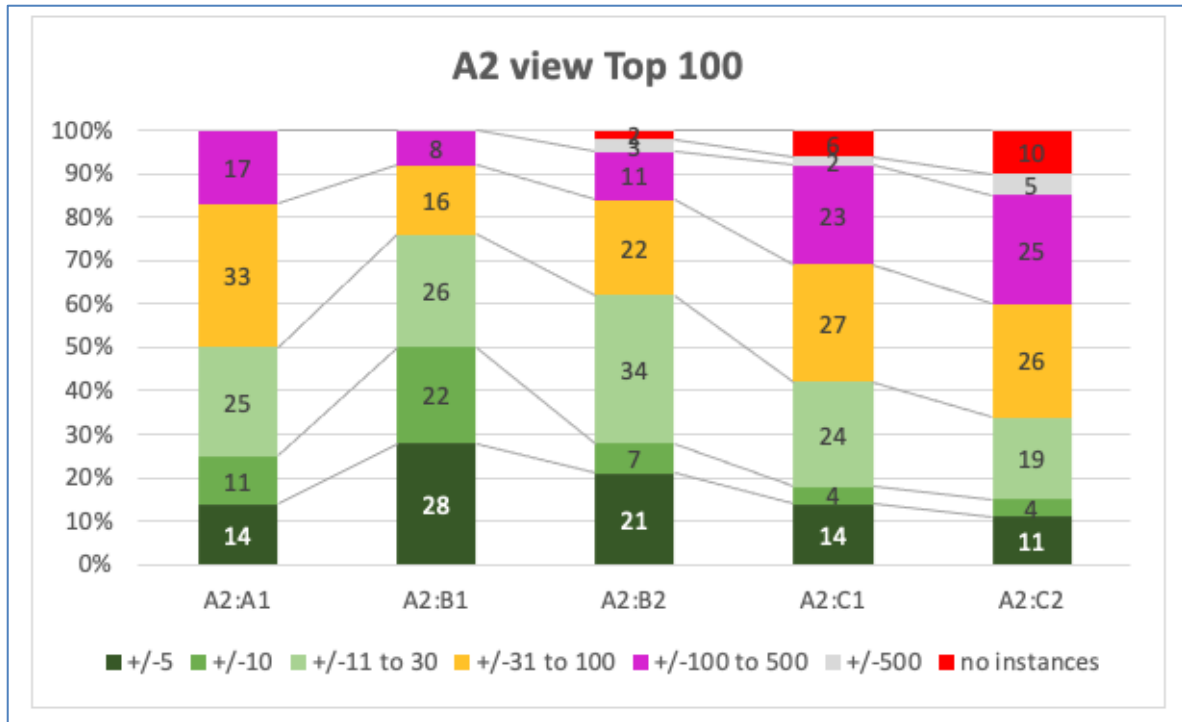


Figure 5.6 Top 100 A2 sequences: convergence and divergence across all levels

There is greater convergence in core sequences and closer ranking generally between A2 and B1, the adjacent higher level, than with A1 the adjacent lower level, which may indicate a growing sensitivity to the statistical patternings in increased language input, and a restructuring of the frequencies in which sequences observed in one level move closer to the distribution of sequences observed in the next highest. This is investigated in more detail in Chapter 6.

B1 view

Figure 5.7 illustrates the B1 data perspective, a comparison of the top 100 B1 sequences and their rankings at all other levels. When compared with the core sequences found in A1 (Figure 5.3) and A2 (Figure 5.6) we see an overall increase in closely ranked core sequences across all levels.

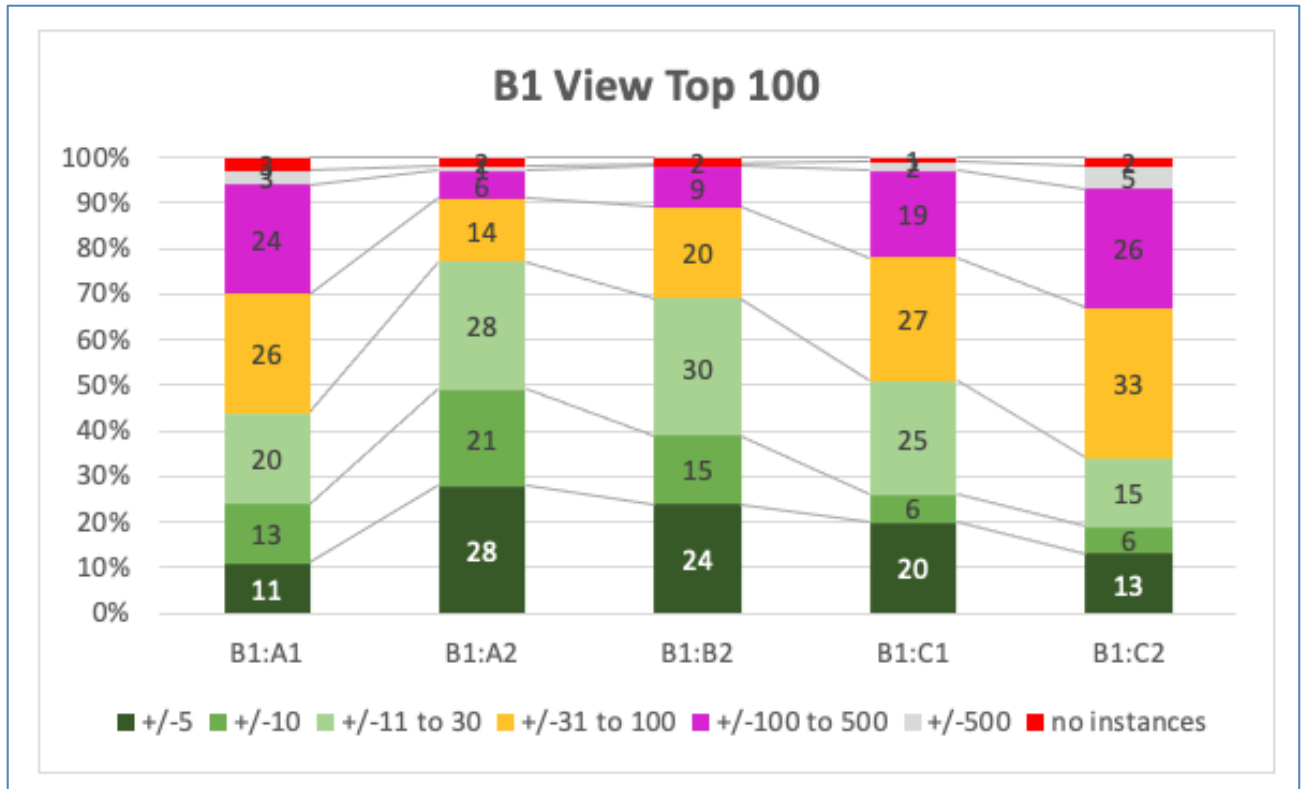


Figure 5.7 Top 100 B1 sequences: convergence and divergence across all levels

When looking at adjacent levels we see 91% of the top 100 sequences at B1 are within ranking of +/-100 at A2 (investigated in more detail in Chapter 6). Between B1 and B2, there is an increase overall in those sequences found within +/- 100. 89% of these sequences are ranked within +/-100 at B2. However it is notable that there is no increase in the consistently closely ranked (+/-5) core sequences between B1 and B2. This may coincide with a period of syntactic stabilisation and increased experimentation with a lexical and functional repertoire. This is investigated in detail in chapter 7.

B2 view

In the overall picture at B2, an increasing convergence in ranking is observed when comparing the top 100 sequences at B2 with their distribution from A1 to C1 (Figure 5.8):

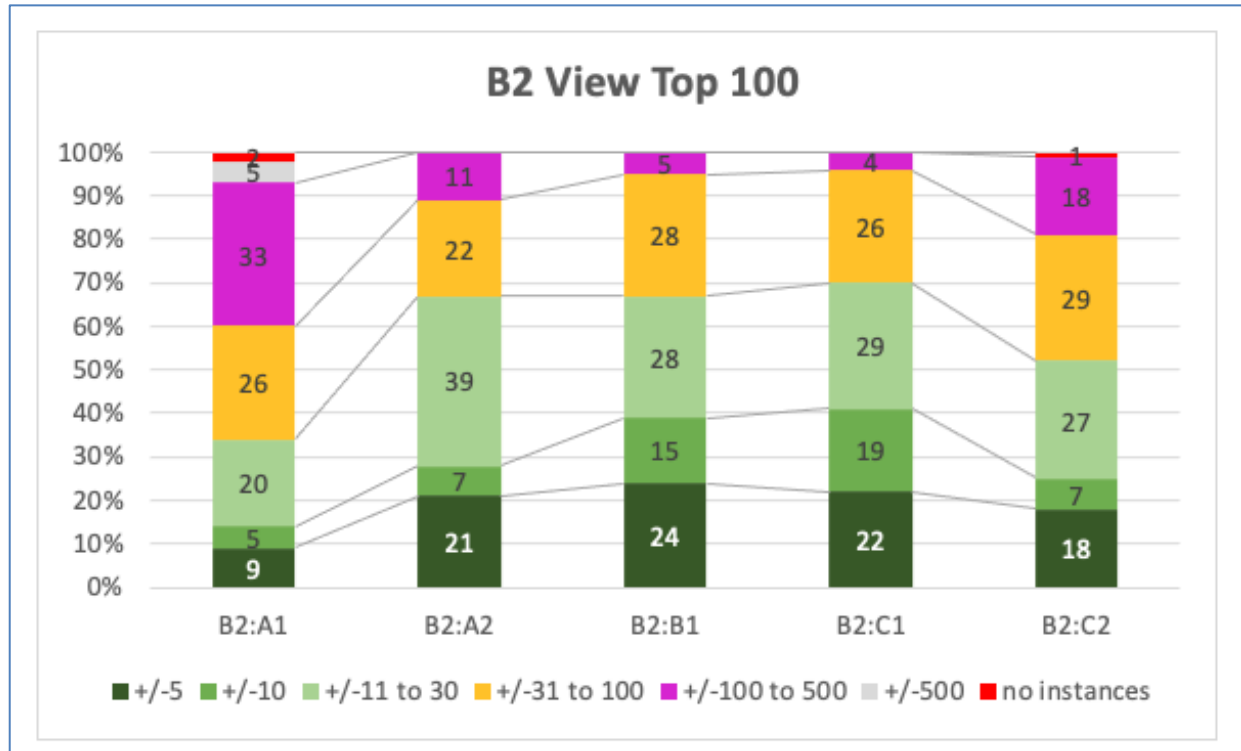


Figure 5.8 Top 100 B2 sequences: convergence and divergence across all levels

89% of the top 100 B2 sequences are found within a rank of +/- 100 at A2, 95% at B1, and 96% at C1, showing convergence particularly with adjacent levels. 81% of the B2 sequences are also found within a rank of +/-100 at C2. Between 67% and 70% of the B2 sequences rank within +/-30 at A2, B1 and C1, indicating increased stabilisation of sequence usage, once again investigated in chapter 7.

C1 view

There is a clear picture of increasing convergence when looking at the top 100 sequences used at C1 level (Figure 5.9).

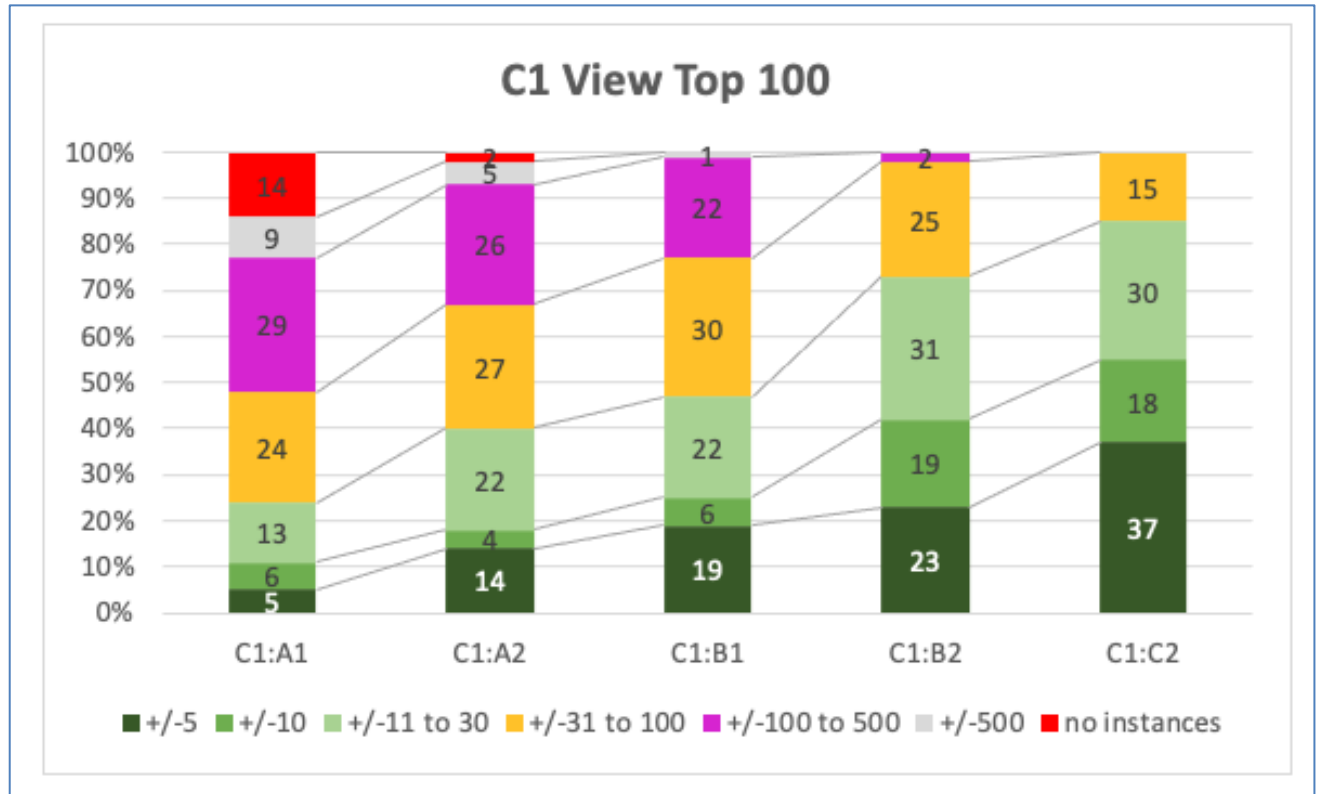


Figure 5.9 Top 100 C1 sequences: convergence and divergence across all levels

85% of the C1 sequences are shared with C2 within a rank difference of +/- 30, 37% of which are within a rank of +/-5. At other levels the number of sequences within a rank difference of +/-5 increase as proficiency increases, giving evidence to a growing core of sequences. Only 2% and 14% of the C1 sequences are not found in the A2 and A1 data, respectively.

5.3 Sequence types: A1 and C2

The previous sections have explored the overall frequency and distribution of the top 100 and revealed a tendency for three types of sequence that are observable across all the data:

- (1) core sequences, those that rank consistently closely across levels
- (2) emerging sequences, those that increase in rank across levels
- (3) decreasing sequences, those that decrease in rank across levels

In this section I look at examples of these types at both ends of the frequency ranking scale, first the top 10 sequences from each of levels A1 and C2, i.e. the most frequent, and then sequences which rank highly in one level but are not found in the other.

5.3.1 Sequence types: Top 10 A1

The top 10 sequences at A1 show the three different types of sequence (Table 5.3).

The highest ranked POS tag sequence at A1 is #1 SENT PP MD VV (which begins with a punctuation marker) .+pronoun+modal+verb, (e.g. . *I would like*). This is an example of a decreasing sequence. Across other levels this sequence ranks at #7 at A2 (-6), #17 at B1 (-16), #21 at B2 (-20), #36 at C1 (-35), #70 at C2 (-69), with rank differences in brackets. The decrease in ranking indicates that this sequence becomes increasingly less important in the learner repertoire as proficiency increases.

In contrast, sequence #3 IN DT NN SENT preposition+determiner+noun+. (e.g. *in the morning*.), is an example of a sequence with a high degree of convergence in ranks across all levels, all ranking within +/- 5 of each other: #3 at A2 (0), #2 at B1 (1), #3 at B2 (0), #6 at C1 (-3), #6 at C2 (-3), with rank differences in brackets. This sequence is an example of a core sequence, one that is consistently closely ranked and frequently used across levels.

Sequence #10 NN IN DT NN noun+preposition+determiner+noun (e.g. *concert in the morning*) ranks at #1 across all other levels. This is an example of an emerging sequence at A1. While still ranking within the top 10 at A1, it increases in rank at A2 and becomes consistently important in the repertoire of A2 to C2 learners. It becomes part of the core sequence group from A2 onwards.

Of note here is that seven of the top 10 at A1 contain punctuation. Some of these straddle sentences boundaries and others are representative of the short basic clause patterns, building blocks which are characteristic of this level. In order to retain the integrity of the methodology, all sequences containing punctuation tags are retained for this initial analysis. I return to this decision in subsequent chapters in which sequences containing punctuation are removed.

Rank difference between top A1 sequences and other levels						
	A1 4-gram POS tag sequences and <i>examples</i>	A2	B1	B2	C1	C2
1	.+pronoun+modal+verb SENT PP MD VV . <i>I would like</i>	-6	-16	-20	-35	-69
2	pronoun+modal+verb+preposition PP MD VV IN <i>You can come to</i>	-4	-10	-34	-58	-93
3	preposition+determiner+noun+. IN DT NN SENT <i>in the morning.</i>	0	1	0	-3	-3
4	preposition+posspronoun+noun+. IN PPZ NN SENT <i>to my house.</i>	-1	-2	-4	-13	-14
5	noun+.+pronoun+presentsimpleV NN SENT PP VVP <i>phone. I like</i>	-3	-10	-19	-38	-85
6	noun+.+pronoun+modal NN SENT PP MD <i>music. I can</i>	-	20	-31	-32	-47
7	determiner+noun+.+pronoun DT NN SENT PP <i>another country. I</i>	-5	-1	-4	-14	-23
8	propernoun+propernoun+,+pronoun NP NP SENT PP <i>Dear Sam, I</i>	-9	-18	-60	-	-
9	pronoun+presentsimpleV+to+verb PP VVP TO VV <i>you want to come</i>	-6	-13	-25	-54	-85
10	noun+preposition+determiner+noun NN IN DT NN <i>concert in the morning</i>	9	9	9	9	9

Table 5.3 Example of Top 10 sequences at A1 and their rank difference across all levels

Categorisation key:

variance of	5	10	11-	31-	101-	501	not
rank +/-			30	100	500	+	found

5.3.2 Top 10 C2

A visual snapshot of the rear view, from the C2 perspective (Table 5.4), indicates strong convergence between the top 10 at C2 and the top 10 at A2 to C1 levels though a highly

divergent picture between C2 and A1. This is distinct from the front view which indicates overall a picture of increased divergence as proficiency increases from A1, shown by the variation in colour and rank difference, for most of the sequences.

The most frequent sequence at C2 is **noun+preposition+determiner+noun** (e.g. *aim of this report*). Across other levels this sequence also ranks at #1, apart from A1, where it is ranked #10. Sequence #6 **preposition+determiner+noun+**. (e.g. *on the desk.*) is core to all levels. Sequences #1, 2, 3, 4, 9 are all core to all levels other than A1. Sequences #5 and #10 are core to all levels other than A1 and A2, i.e. they are core to the repertoire of B1 level upwards. Sequences #7 and #8 become increasingly used as proficiency increases.

Rank difference between top C2 sequences and other levels						
	Top C2 4-gram POS tag sequences and examples	A1	A2	B1	B2	C1
1	noun+preposition+determiner+noun NN IN DT NN <i>aim of this report</i>	-9	0	0	0	0
2	preposition+determiner+adjective+noun IN DT JJ NN <i>on the other hand</i>	-60	0	-1	0	0
3	preposition+determiner+noun+preposition IN DT NN IN <i>in the middle of</i>	-12	-1	-2	-1	-1
4	determiner+adjective+noun+preposition DT JJ NN IN <i>a great deal of</i>	-105	-5	-5	-2	-1
5	determiner+noun+preposition+determiner DT NN IN DT <i>the aim of this</i>	-27	-11	-2	0	2
6	preposition+determiner+noun+. IN DT NN SENT <i>on the desk.</i>	3	3	4	3	0
7	determiner+noun+preposition+noun DT NN IN NN <i>a matter of fact</i>	-77	-24	-25	-11	-3
8	preposition+determiner+noun+, IN DT NN , <i>As a result,</i>	-81	-15	-13	-2	0
9	determiner+adjective+noun+. DT JJ NN SENT <i>the same time.</i>	-25	-1	-2	2	2
10	To+verb+determiner+noun TO VV DT NN <i>to find a job</i>	-33	-18	0	1	1

Table 5.4 Example of Top 10 sequences at C2 and their rank difference across all levels

The ranking and distribution of sequences at A1 level, from this perspective, looks less stable in comparison with other levels. All of the sequences that appear in the top 10 at C2 are also highly ranked, albeit to varying degrees, across all of the lower levels, apart from A1. The implication here is that learners at A2 are already sensitive to what is most frequently used in the input they are experiencing. I note also here that of the top 10 C2 sequences, only one contains a verb form, all others contain a noun phrase or prepositional phrase. In contrast five of the top 10 A1 sequences contain a verb form. I return to this observation in section 5.4.

5.3.3 A1 sequences not found in C2

Overall, we have seen a tendency for convergence in the use of sequences as proficiency increases. All of the top 100 C2 sequences are also found within the B1, B2 and C1 data and only 16% and 5% of the C2 sequences are not found anywhere in the A1 and A2 data respectively. However the number of top 100 A1 sequences not found anywhere in each of the higher levels increases as proficiency increases (illustrated in red in Figure 5.3). These sequences that appear in the A1 top 100 but nowhere in the C2 data are shown in Table 5.5.

Sequences / TAGS	Typical examples
preposition-number+noun+"" IN CD NN ""	<i>at 2 o'</i>
number+noun+""+noun CD NN "" NN	<i>Five o'clock</i>
preposition+number+:+noun IN CD : CD	<i>At 6 : 00</i>
.+verb+pronoun+adverb SENT VV PP RB	<i>. Thank you very</i>
noun+.+verb+pronoun NN SENT VV PP	<i>house. See you</i>
.+pronoun+verb_be+verb_ing SENT PP VBP VVG	<i>. I'm going to</i>
noun+preposition+number+noun NN IN CD NN	<i>airport at 8 o</i>
wh+verb_be+pronoun+. WRB VBP PP SENT	<i>How are you?</i>
Modal+verb+preposition+possessive pronoun	<i>can write to my</i>

MD VV IN PPZ	
determiner+presentsimple_V+ preposition DT NN VBZ IN	<i>a bicycle is from</i>
possessive pronoun+noun+preposition+number PPZ NN IN CD	<i>your house at 7</i>
propernoun+propernoun+.+pronoun NP SENT PP	<i>National Park. It</i>
+.modal+pronoun+verb SENT MD PP VV	<i>. Could you help</i>
number+:+noun CD : CD NN	<i>6:00 am</i>
modal+verb+preposition+number MD VV IN CD	<i>can write for 2</i>
+.pronoun+have+TO SENT PP VHP TO	<i>. I have to</i>
+.pronoun+verb_be+preposition SENT PP VBZ IN	<i>. He is from</i>
verb+pronoun+adverb+. VV PP RB SENT	<i>See you soon.</i>

Table 5.5 Examples of A1 POS tag sequences not occurring at C2

Initial observation of these sequences and examples of their lexical exponents throw up issues of tagging and task and topic effect and require further investigation:

- The lexical instantiations seen in table 5.5 may indicate task effect. Tagging of punctuation relating to time (5 o'clock, 6:00) may have skewed the non-occurrence of these sequences at higher levels.
- Taking a bottom-up, truly corpus-driven approach will inevitably unearth limitations relating to the tagging system (discussed in Chapter 4), e.g. in this case relating to punctuation.
- All POS tag sequences are seen to be of relevance and are dealt with in the detailed analysis in chapters 6 to 8, even though many 4-gram POS sequences do not produce 'complete' phrasal units, and need to be either completed by another element or reduced.

- These examples may be indicative of formula-oriented holistic production at A1, where one syntactic sequence corresponds to one holistic example and one function, e.g. where the only lexical realisation for the frequently occurring sequence VV PP RB SENT at A1 is See you soon., indicating that no syntactic generalisation is taking place for this sequence at A1 level. This is investigated further in Chapter 6.

5.3.4 C2 sequences not found in A1

The C2 sequences that do not occur at all in the A1 data are shown in table 5.6.

Sequences / TAGS	Typical examples
past participle+preposition+determiner+noun VVN IN DT NN	<i>created by this situation</i>
preposition+determiner+adjective+plural noun IN DT JJ NNS	<i>of those exotic places</i>
adjective+plural noun+preposition+determiner JJ NNS IN DT	<i>personal experiences of the</i>
preposition+noun+TO+verb IN NN TO VV	<i>in order to check</i>
noun+TO+verb+determiner NN TO VV DT	<i>opportunity to enjoy the</i>
preposition+determiner +plural noun+preposition IN DT NNS IN	<i>from the mistakes of</i>
past participle+preposition+determiner+adjective VVN IN DT JJ	<i>organised by the international</i>
plural noun+ .+adverb+ , NNS SENT RB ,	<i>products. Therefore,</i>
.+adverb+ ,+determiner SENT RB , DT	<i>. Finally, the</i>
determiner+noun+preposition+ing DT NN IN VVG	<i>The idea of improving</i>
modal+be+pastparticiple+preposition MD VB VVN IN	<i>should be taken into</i>
.+preposition+noun+ , SENT IN NN ,	<i>. In addition,</i>

plural noun+preposition+determiner+adjective NNS IN DT JJ	<i>children in the same</i>
.+preposition+determiner+adjective SENT IN DT JJ	<i>. On the other</i>
determiner+adjective+plural noun+preposition DT JJ NNS IN	<i>the negative aspects of</i>
adjective+noun+preposition+plural noun JJ NN IN NNS	<i>considerable number of people</i>

Table 5.6 C2 POS tag sequences not occurring at A1, with lexical examples

Both the syntactic tags and examples of the lexical realisations might point to evidence that

- C2 learners are reliant on a wider range of syntactic forms in a variety of contexts (VVN past participle, VVG -ing form), prepositional and noun phrases (e.g. *of those exotic places, The idea of improving*)
- C2 learners display sensitivity to the discourse management, orientation and signposting needs of writing (e.g. *. Finally, the*).

This is investigated in Chapter 8.

5.4 Overall sequence types: qualitative analysis of phrasal categorisation

As seen in previous research on lexical bundles and p-frames structural characteristics found in learner data have been investigated at different levels of proficiency (Chen and Baker 2010, 2016; Gray and Biber 2013; Staples et al. 2013; Garner 2016, among others). Put crudely, this has revealed a tendency for verb-based bundle use at lower levels and noun-based use at higher levels. A snapshot of the top 10 sequences from both A1 and C2 in sections 5.3.1 and 5.3.2 above gives a crude indication for a similar tendency.

A broad approach to classifying the 4-gram POS tag sequences across all six levels was applied in this study to identify whether there were any observable trends in relation to structural type across levels. A classification system adapted from Gray and Biber (2013) is used here to group the sequences as follows (labels in brackets):

- (1) Noun-based sequences which contain one or more nouns, and no verbs (NP). With a sub-category of
 - (a) prepositional phrases containing a noun phrase (prep NP)

- (2) Verb-based sequences which contain one or more modal, auxiliary or main verb (V), with subcategories of
 - (a) Verb-based sequences which contain a verb followed by a noun (V NP)
 - (b) Verb-based sequences contain a verb followed by a prepositional phrase (V prep NP)
- (3) Sequences with punctuation in medial position (SENT)
- (4) Sequences with initial adverb (RB)
- (5) Miscellaneous sequences (MISC)

Using this taxonomy, the top 100 sequences across all levels were categorised and the percentage distribution of types was calculated, as illustrated in Figure 5.10.

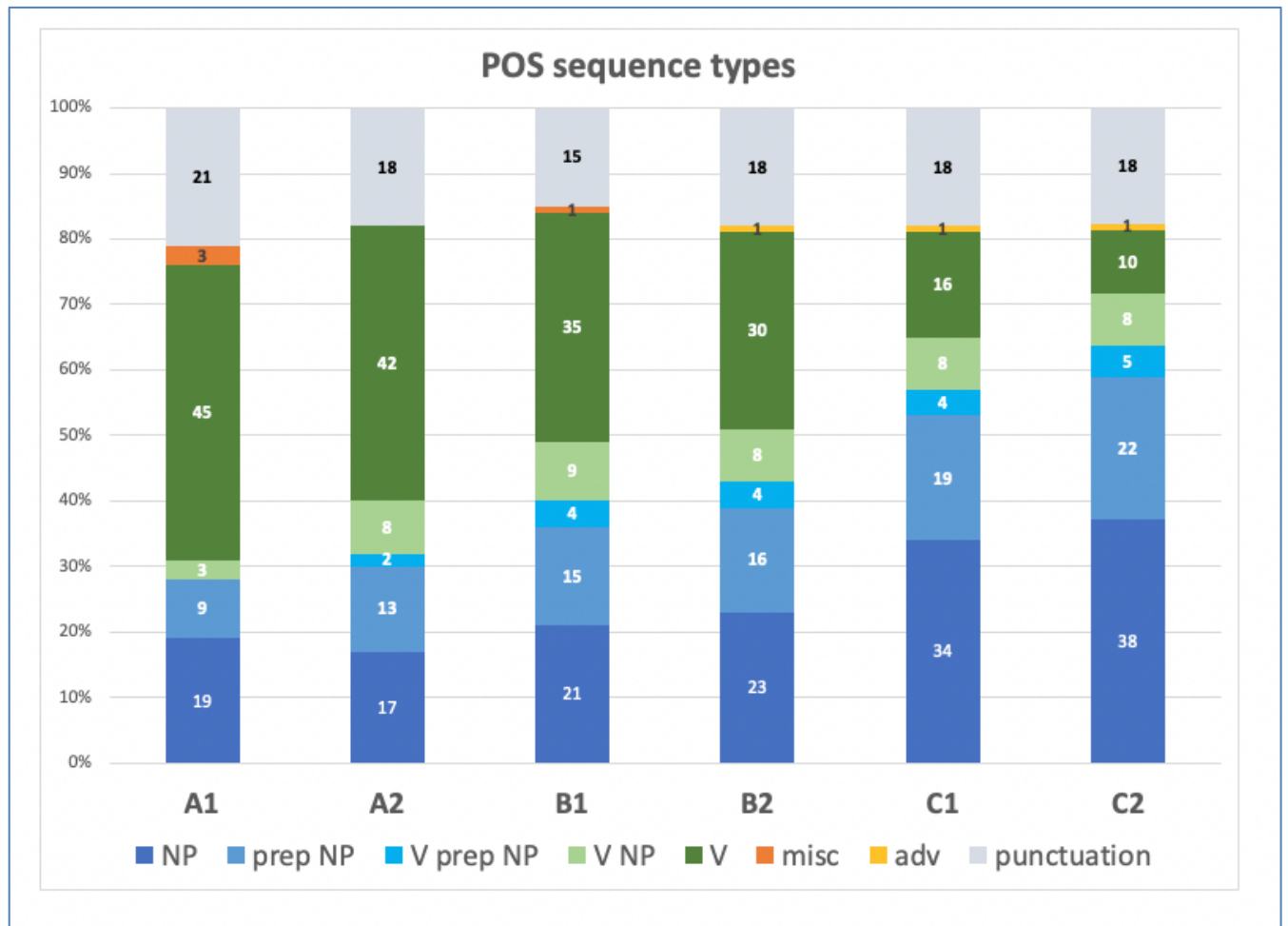


Figure 5.10 Distribution of sequence types: structural categorisation

In broad brush terms, a picture of increased noun phrase use emerges as proficiency increases while verb-based sequence usage decreases. Normed frequencies (per million words) of types in the classification system were calculated to compare usage between groups (Table 5.7) and an overall view is illustrated in Figure 5.11.

Categories	A1	A2	B1	B2	C1	C2
Noun based PMW (1)	56693	60728	62452	64143	79331	88129
Verb based PMW (2)	86667	72742	60667	46415	29497	23113
SENT (3) Other (4) and (5)	42168	26467	22408	20322	16744	14238

Table 5.7 Structural classification of the top 100 4-gram POS sequences across levels: normalised occurrences (PMW)

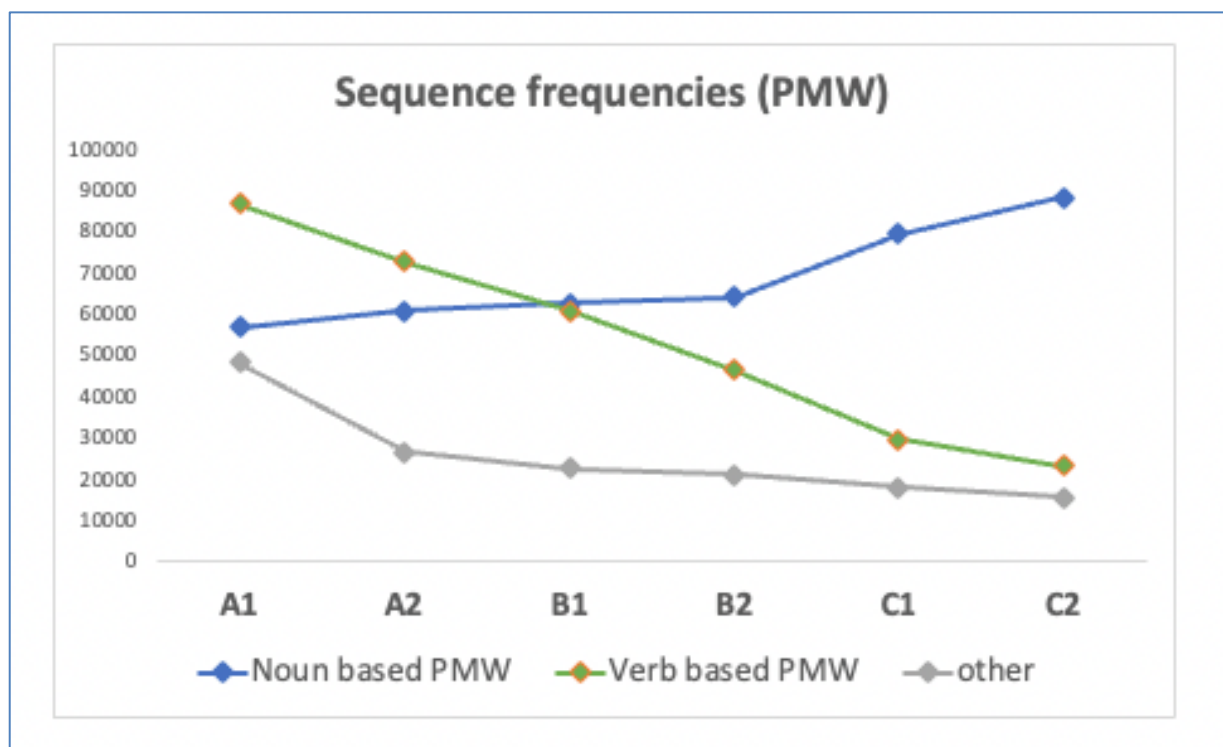


Figure 5.11 Overall occurrences (PMW) of noun-based and verb-based sequences in the top 100 sequences at all levels

The picture of development that emerges over the top 100 sequences at each level shows a clear preference for noun-based sequence use at C2 level alongside a comparative steady decrease in verb-based sequence usage from A1 to C2, and this is coherent with word class distributions in Biber *et al.* (1999) across registers. There is a steady rise in noun-based usage from A1 to B1 where the intersection between verb- and noun-based usage indicates a comparatively equal distribution of both. The steady rise in noun-based usage continues from B1 to B2 where there is a noticeable increase to C1. This finding, that is, the movement from verb-based to noun-based sequences offers an important contribution in answering to RQ2 which focuses on how POS tag sequence usage develop across proficiency levels (further elaboration on this can be found in Chapters 6 to 8).

5.5 Individual sequences: case study analysis of A1 and C2 #1 sequences

Through taking a POS-gram approach not only is it possible to observe the syntactic generalisations made by learners at each level, but it also allows greater exploration of one of these generalisations at the lexical and functional level. Here I take two sequences, the #1 ranked sequence from each of the A1 and C2 levels and explore their usage and development across all levels, by looking at actual occurrences. The first is an example of the highest ranking sequence at A1, a verb-based sequence, on which, as we have seen in terms of overall distribution, learners become less reliant as proficiency increases; the second is a noun-based sequence which is consistently core across all levels.

5.5.1 Case study 1: SENT PP MD VV .+pronoun+modal+verb

The highest ranked POS tag sequence at A1 is #1 **+.pronoun+modal+verb**, (e.g. *I would like*). Across other levels this sequence ranks at #7 at A2, #17 at B1, #21 at B2, #36 at C1, #70 at C2, and becomes increasingly less important in the learner repertoire as proficiency increases. Three of the elements in this sequence are relatively fixed: two belong to closed word classes (pronoun, modal), the punctuation (.) stands only for full point, question mark or exclamation mark. It is only the final element verb which belongs to an open word class, with thousands of candidates for this lexical slot.

Overall frequencies: 1000 types

The first thousand lexical exponents were extracted from the subcorpora, using the Sketch Engine corpus query language (CQL) and the following CQL string:

[tag="SENT"][tag="PP"][tag="MD"][tag="VV"]. Table 5.8 gives a breakdown of raw and

relative occurrences of this sequence by level. Relative (per 1 million) occurrences are normalised using each level subcorpus size.

	subcorpus size	raw occurrences	*relative occurrences	total occurrences 1000 types	1000 types as % occurrences
A1	2456971	14599	5942	14024	96
A2	5703217	16610	2912	14852	89
B1	3261473	6559	2011	5684	87
B2	5263979	8685	1650	7136	82
C1	6711568	7665	1142	5993	78
C2	7698695	5817	756	4405	76

* relative to subcorpus size per 1 million

Table 5.8 Breakdown of occurrences by level of .+pronoun+modal+verb

The Sketch Engine platform has a routine download limit of maximum of 1000 items, and for this reason percentage amounts are also given to show the proportion of all occurrences that 1000 types constitutes for each level and to give an indication of type-token ratio. These findings confirm decreasing use of this sequence as proficiency increases. The sequence is used twice as frequently in the A1 than the A2 data and eight times as frequently in the A1 data than in the C2 data. However, as proficiency increases so does the range of lexical exponents, indicated by the total occurrences of 1000 types as a percentage of all occurrences. The implication in simple, formal terms is that A1 learners repeatedly use the same restricted range of forms often and C2 learners, while using this sequence far less frequently, do so with a wider range of forms.

Top 100 types: lexical exponents

To investigate this range of lexical exponents further, the top 100 exponents for each level were categorised first by using the modal verb forms, illustrated in Figure 5.12. Initial observations show a decrease in the use of *can* and *will* as proficiency increases, and an increase in the use of *would*. *Must* decreases in use from A1 to A2 and then increases. The range of modal verbs used increases proficiency increases, with all eight core modal verbs being used at C2 (*would, can, will, must, should, could, may, might*) in comparison with five at A1 (*can, will, would, must, should*).

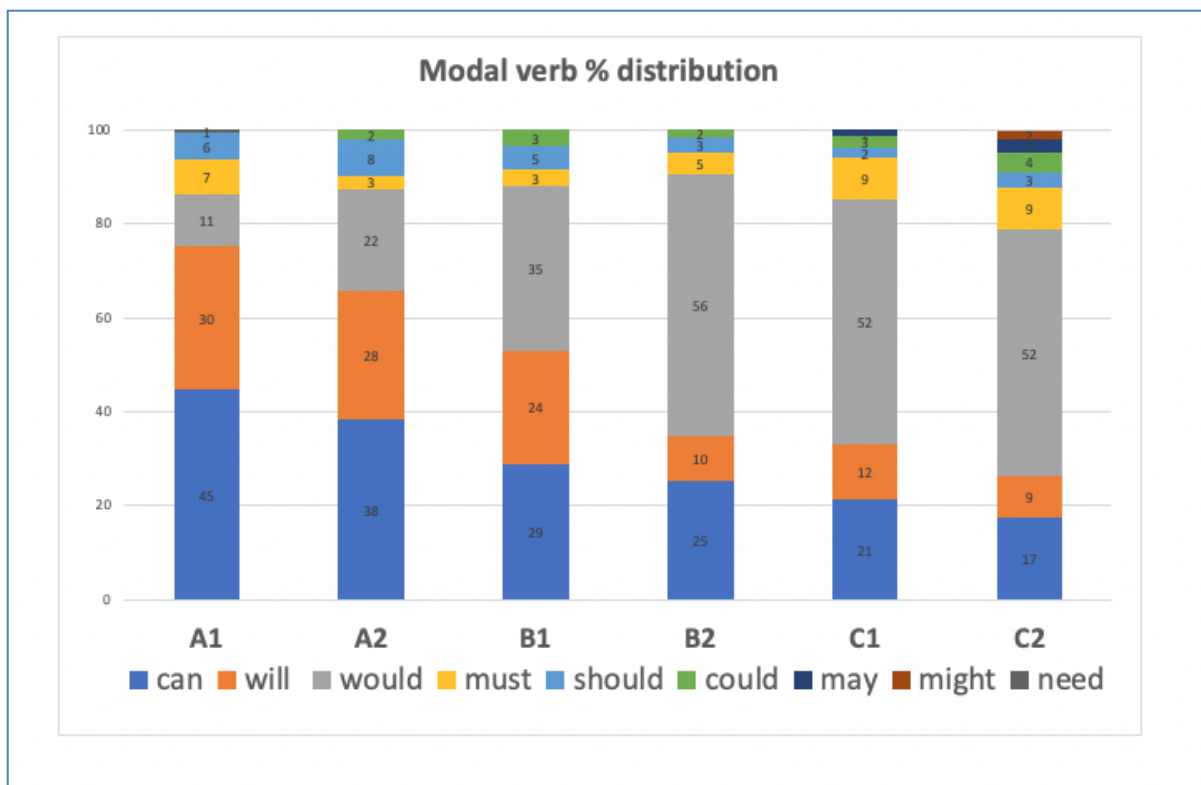


Figure 5.12 Percentage distribution of modal verbs in top 100 lexical exponents of .+pronoun+modal+verb all levels

Clearly an analysis looking only at the occurrence of the modal verb in this sequence is oversimplistic, and hides a complexity of lexical and functional patterning, along with possible task effect. To explore this further the top 30 lexical patterns and frequencies from the A1 and the C2 data are exemplified in Table 5.9.

A1	Relative frequency	%		C2	Relative frequency	%
. You can come	483	4.8		. I would like	1313	13.13
. I would like	438	4.4		. I must say	155	1.55
. You can bring	416	4.1		. I must admit	143	1.43
. I will start	310	3.1		. I would suggest	119	1.19
. I can write	296	3.0		. I would say	101	1.01
. You must bring	275	2.7		. I'd like	98	0.98
. You can get	266	2.7		. I would recommend	96	0.96
. I'd like	266	2.7		. I will try	83	0.83

. <i>You can wear</i>	229	2.3		. <i>I would try</i>	79	0.79
. <i>You should bring</i>	183	1.8		. <i>You will find</i>	79	0.79
. <i>You can go</i>	182	1.8		. <i>We would like</i>	79	0.79
. <i>We can go</i>	175	1.7		. <i>You can find</i>	70	0.70
. <i>You should wear</i>	159	1.6		. <i>You can see</i>	67	0.67
! <i>You can come</i>	151	1.5		. <i>We can see</i>	52	0.52
. <i>I will arrive</i>	132	1.3		. <i>I would appreciate</i>	48	0.48
. <i>We will start</i>	119	1.2		. <i>I can say</i>	45	0.45
. <i>We will go</i>	112	1.1		. <i>It may sound</i>	43	0.43
. <i>I 'll start</i>	105	1.0		. <i>I must confess</i>	40	0.40
. <i>I will go</i>	97	1.0		. <i>It may seem</i>	40	0.40
. <i>We can meet</i>	81	0.8		. <i>I can see</i>	36	0.36
. <i>You can take</i>	78	0.8		. <i>I can assure</i>	36	0.36
. <i>I will travel</i>	78	0.8		. <i>I would love</i>	36	0.36
. <i>I can help</i>	78	0.8		. <i>You can go</i>	34	0.34
? <i>You can come</i>	76	0.8		. <i>I could see</i>	33	0.33
. <i>You must wear</i>	76	0.8		. <i>I can understand</i>	33	0.33
. <i>I 'll wait</i>	58	0.6		. <i>You can imagine</i>	31	0.31
. <i>We shall meet</i>	54	0.5		. <i>You can get</i>	31	0.31
. <i>I 'll arrive</i>	52	0.5		. <i>I would give</i>	29	0.29
. <i>We must bring</i>	51	0.5		. <i>You can do</i>	29	0.29
. <i>I will come</i>	51	0.5		. <i>I can imagine</i>	28	0.28

Table 5.9 Top 30 lexical exponents of the **+.pronoun+modal+verb** from A1 and C2.

I note here that a detailed analysis of the lexical and functional usage of sequences with modal forms across proficiency levels is a study in itself, and I will restrict it here to some general observations. However, what is of immediate interest is the distribution of this sequence across different lexical realisations. In the A1 data it is more evenly distributed

across a range of lexical exponents, whereas in the C2 data (as well as in the other levels) the predominant use tends towards the formulaic *I would like*. Top A1 examples are directive (*you can/must/should bring, we shall meet*), transactional and topic-oriented relating to events and arrangements (indicated by the main verb forms, e.g. *wear, bring, come, arrive, meet, travel*). C2 users appear to be sensitive to specialised pragmatic meanings, e.g. sequences with *must* have moved from intrinsic, directive functions characteristic of A1, with *you* and a following dynamic verb (after Biber *et al.* 1999) e.g. *. You must bring*, to extrinsic functions, with a following stative verb *I must say/admit/confess*, employing verbs with a declarative function, expressing stance and concession. Likewise sequences with *I would like* move from expressing preference at A1 (e.g. *I would like to go*) to routinised semi-fixed strings expressing stance, foregrounding thanks, and elaborating (e.g. *I would like to thank/mention/comment*) There is evidence of a greater degree of fixedness and formulaicity between each of these elements in the sequence, indicated by high LogDice and MI scores in the C2 data, indicating strength of collocation with *I must*, as illustrated in Figure 5.13, in comparison with *you must* at A2.

	Word	Cooccurrences ?	Candidates ?	T-score	MI	LogDice ↓
1	<input type="checkbox"/> admit	210	788	14.49	13.98	11.92 ...
2	<input type="checkbox"/> confess	61	109	7.81	15.05	10.83 ...
3	<input type="checkbox"/> say	313	5,714	17.69	11.70	10.58 ...

Figure 5.13 Strength of collocation of *I must* and following verb in C2 data.

Instances of *You+must* are infrequent in the C2 data (compared to A1), the first of which is *. You must try* which is ranked #222 of all of the lexical exponents, with only 4 occurrences. This may be revealing of pedagogical description of *must* often found in lower level resources which give a broad and underspecified account as the modal verb for obligation, and/or an indication of a feature of spoken register, characteristic in the writing of lower levels.

The C2 top 30 examples also show a range of additional specialised functions in sequences with *I would* from hedged suggestion (*I would recommend/suggest/try*) to hedged request / preference (*I would love/appreciate*). Lexical exponents containing *may* and *might* also rank highly in the top C2 examples. Typical examples are with pronoun *It* e.g. *. It may seem/sound., It might sound*, performing both a focussing and impersonal stance function,

frequently found in the context of following clause initial *but* signalling an opinion or fact about something that is potentially contradictory, controversial or surprising:

Extract 5.1

It may sound complicated, but it really isn't. (C2, L1 Korean, CAE, 1998 Q4)

Extract 5.2

It may sound a little awkward, but music plays such an important role in my life that it defines my acts and my future. (C2, L1 Greek, CPE, 2007 Q4)

Extract 5.3

It might seem natural to be kind to your friends **but** not all people treat their friends in a correct way. (C2, L1 Swedish, CPE, 1999 Q1)

Extract 5.4

It might sound a bit idealistic and naive, but I think this concept of communication will make the world a better place (C2, L1 Danish, CAE, 1999, Q2)

There is evidence from a range of L1 backgrounds and exam tasks that the form *It may/might sound/seem* + adjective phrase (optional comma) + *but* sequence with this highly specialised function has become entrenched at C2 level. I emphasise here that this meaning is found not in the individual elements but in the patterning. There is also evidence of further formulaic patterning and hedging in the adjective phrase (with the addition of *a bit / a little*) exemplified in #2 and 4.

Overall, a summary of the differences of the sequence SENT PP MD VV .+pronoun+modal+verb shows a broader distribution of exponents at A1, with no one single exponent taking the share of occurrences. However there is a reliance on one single exponent in C2 (*I would like*) but also evidence of use of a wider range of modal verbs and a wider range of lexical exponents. There is also evidence of a more fixed patterning between items in the sequence in C2 and a wider range of functions: some of which are pragmatically specialised.

5.5.2 Case study 2: NN IN DT NN noun+preposition+determiner+noun

The next sequence to be investigated is the highest ranking sequence at C2: NN IN DT NN noun+preposition+determiner+noun (e.g. *aim of this report*). This is also consistently the highest ranking sequence at all other levels, other than A1 where it ranks at #10. Unlike the

modal verb sequences in 5.5.1, two of the elements belong to closed word classes (preposition, determiner), and both the initial and final element (noun in both cases) belongs to an open word class, with thousands of candidates for each of these lexical slots.

Overall frequencies: Top 1000 types

The first thousand lexical exponents were extracted from the subcorpora using the CQL string [tag="NN"][tag="IN"][tag="DT"][tag="NN"]. Table 5.10 gives a breakdown of raw and relative occurrences of this sequence by level. Relative (per 1 million) occurrences are normalised using each level subcorpus size.

	subcorpus size	raw occurrences	*relative occurrences	total occurrences	1000 types as % occurrences
				1000 types	
A1	2456971	7601	3094	4472	58.83
A2	5703217	24919	4369	11884	47.69
B1	3261473	14268	4375	5485	38.44
B2	5263979	25994	4938	6913	26.59
C1	6711568	37415	5575	9191	24.57
C2	7698695	44670	5802	11303	25.30

Table 5.10 Breakdown of occurrences by level of **noun+preposition+determiner+noun**

As previously mentioned, the Sketch Engine platform has a download limit of maximum of 1000 items, and for this reason percentage amounts are also given to show the proportion of all occurrences that 1000 types constitutes for each level and to give an indication of type-token ratio. These findings confirm increasing use of this sequence as proficiency increases. In relative terms the sequence is used a third more frequently in the A2 than the A1 data, remaining stable from A2 to B1 and increasing in usage steadily from B1 to C2. The relative occurrences at A2 and B1 are almost identical however, as indicated by the total occurrences of 1000 types as a percentage of all occurrences, the range of lexical exponents increases. A higher percentage reflects a lower range of types. For example, the first 1000 types make up 58.83% of all occurrences at A1, decreasing to 47.69% at A2, 38.44% at B1 and 26.59% at B2. From B2 to C2 the range of lexical range, stabilises. The implication in simple, formal

terms is that B2, C1 and C2 learners use a similar range of lexical exponents. The functional range is explored next.

Top 100 types

In order to get a better understanding of any recurrent generalisations first in terms of lexical choices and functional use, the top 100 most frequent lexical realisations for all levels were examined and categorised using a pattern grammar approach set out in Hunston and Francis. (2000). (See also <https://grammar.collinsdictionary.com/grammar-pattern>). This involves first identifying form groupings or ‘grammar patterns’ (see Chapters 3 and 4), e.g. N of n (noun of noun), N to n (noun to noun) and secondly the meaning groupings for each pattern (e.g. era/fraction/site, access/response). By way of example, Table 5.11 illustrates the top 20 from A1 and C2, with their form groupings.

	A1	Form groupings	C2	Form groupings
1	<i>clock in the morning</i>	N in n	<i>photo from the drawer*</i>	N from n
2	<i>concert in the town*</i>	N in n	<i>aim of this report</i>	N of n
3	<i>day of the class*</i>	N of n	<i>library with an internet*</i>	N with n
4	<i>pen-friend in another country*</i>	N in n	<i>aim of this proposal</i>	N of n
5	<i>clock in the evening</i>	N in n	<i>purpose of this proposal</i>	N of n
6	<i>meeting about the concert*</i>	N about n	<i>purpose of this report</i>	N of n
7	<i>clock in the afternoon</i>	N in n	<i>response to the article</i>	N to n
8	<i>centre of the city</i>	N of n	<i>end of the day</i>	N of n
9	<i>day of the art</i>	N of the n	<i>use of the land</i>	N of the n
10	<i>front of the cinema</i>	N of n	<i>understanding of the world</i>	N of n
11	<i>m in the morning</i>	N in n	<i>centre of the town</i>	N of n
12	<i>information about the art</i>	N about n	<i>centre of the city</i>	N of n
13	<i>price of the ticket</i>	N of the n	<i>solution to this problem</i>	N to this n

14	<i>front of the supermarket</i>	N of n	<i>response to the campaign</i>	N to n
15	<i>a.m. in the morning</i>	N in the n	<i>part of the world</i>	N of the n
16	<i>name of the music</i>	N of n	<i>rest of the world</i>	N of n
17	<i>center of the city</i>	N of n	<i>solution to the problem</i>	N to n
18	<i>centre of the town</i>	N of n	<i>area of the hotel</i>	N of n
19	<i>kind of the concert</i>	N of n	<i>clock in the morning</i>	N in n
20	<i>front of the shopping</i>	N of n	<i>side of the coin</i>	N of n

(The lexical sequences marked * are also found in exam rubrics.)

Table 5.11 Top 20 most frequent lexical realisations of noun+preposition+determiner+noun at A1 and C2, categorised using Pattern grammar taxonomy (Hunston and Francis 2000)

Figure 5.14 shows how the distribution of the top 100 of these form groupings change across levels. At A1 the N of n pattern is used as frequently as the N in n pattern. From A1 there is an observable increase of the share that the N of n pattern occupies, as it becomes dominant and a gradual decrease in the share N in n pattern occupies, from A2 as proficiency increases. The N to n pattern also increases from B1 to C2.

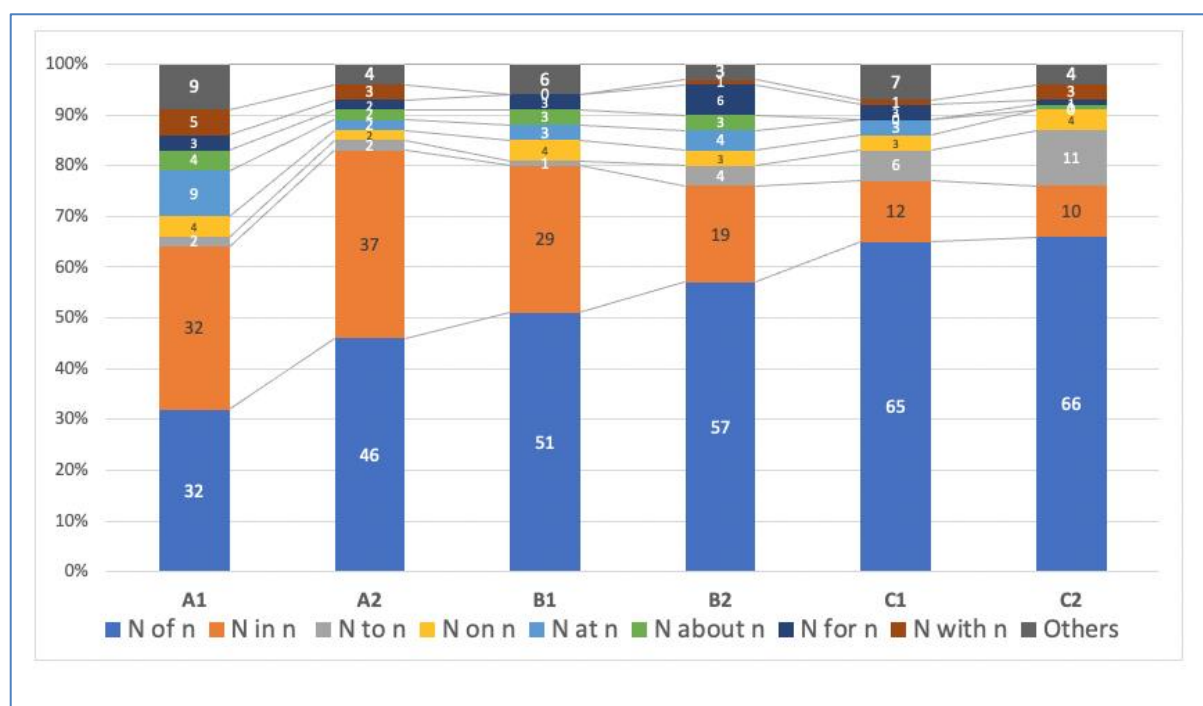


Figure 5.14 Distribution of noun form groupings of noun+preposition+determiner+noun sequence across the top 100 lexical realisations

An investigation into meaning groupings revealed that a decrease in distribution does not correspond to a decrease in functions. For example, at A1 the **N in n** pattern, accounting for 32% of the top 100, was consistently used with both a time function and a place function (*clock in the morning, table in the kitchen*), neither of which, incidentally, are categorised as meaning groups in ‘pattern grammar’ (I highlight further the challenges of the assignment of meaning groups below). At C2, despite the decrease in distribution (10% of the top 100), the **N in n** form performs three functions: an ‘increase’ function (*increase in the number*), after Hunston and Francis 2000, as well as the time and place functions seen at A1; additionally these last two functions are used at C2 with increased lexical range (*development in the city, tourism in the world*). This form group alone demonstrates a growing lexical and functional repertoire, as proficiency increases, even though its frequency of usage decreases.

The **N of n** form grouping constitutes 32% of the top 100 lexical realisations in the A1 data and 66% in the C2 data. Table 5.12 shows the meaning groups for the **N of n** occurrences in the top 100 of both data sets, with examples, categorised according to the Pattern grammar taxonomy. There are many cases where no corresponding group was found in Hunston and Francis 2000 but where there is a clear meaning. These are labelled uncat+MEANING, with a relevant meaning group specified (e.g. uncat+TIME). Asterisks indicate cases where the meaning is not found in a **N of n** grouping but is found elsewhere in the Pattern grammar groupings. For example *centre of the city, front of the television* are categorised with a ‘front’ meaning under the **in N** category. This raises a problem of categorisation of meaning which has methodological implications and which is discussed in Chapters 4 and 9.

Meaning group	Top A1 examples	Meaning group	Top C2 examples
front*	<i>centre of the city</i>	aim	<i>aim of this proposal</i>
	<i>centre of the town</i>		<i>aim of this report</i>
	<i>front of the bank</i>		<i>purpose of this letter</i>
	<i>front of the bus</i>		<i>purpose of this proposal</i>
	<i>front of the church</i>		<i>purpose of this report</i>
	<i>front of the cinema</i>	announcement	<i>importance of the choice</i>
	<i>front of the hospital</i>		<i>knowledge of the language</i>

rise and fall	<i>end of the party</i>		<i>knowledge of the world</i>
	<i>end of the street</i>		<i>nature of the problem</i>
uncat_MONEY	<i>cost of the ticket</i>		<i>understanding of the world</i>
	<i>price of the ticket</i>	construction	<i>construction of a park</i>
uncat_TIME	<i>date of the class</i>		<i>creation of a park</i>
	<i>day of the art</i>		<i>use of the car</i>
uncat_NAME	<i>name of the band</i>		<i>use of the land</i>
	<i>name of the club</i>		<i>destruction of the environment</i>
	<i>name of the music</i>	front*	<i>centre of the city</i>
uncat_PLACE	<i>place of the concert</i>		<i>centre of the town</i>
uncategorised	<i>colour of the mobile</i>		<i>front of a computer</i>
			<i>front of the television</i>
			<i>middle of the night</i>
		issue	<i>culture of the country</i>
		percentage*	<i>majority of the population</i>
			<i>part of the country</i>
			<i>part of the population</i>
			<i>part of the world</i>
			<i>rest of the day</i>
			<i>rest of the world</i>
		rim	<i>area of the hotel</i>
		rise and fall	<i>beginning of the novel</i>
			<i>beginning of this century</i>
			<i>end of the book</i>
			<i>end of the day</i>
			<i>end of the world</i>
			<i>end of the year</i>
		uncat_PERSON	<i>member of the family</i>

		uncat_TIME	<i>time of the day</i>
			<i>time of the year</i>
		uncat_PLACE	<i>top of the mountain</i>
			<i>view of the world</i>

Table 5.12 Noun of noun pattern grammar meaning groups and examples from A1 and C2.

As well as these meaning groupings there are also examples in the top 100 at C2 which are form part of fixed or semi-fixed phrases and which do not correspond to meaning groups (e.g. without a *shadow of a doubt*, have (lemma) a *whale of a time*). They have specific form-meaning mappings with specialised functions (Table 5.13).

<i>shadow of a doubt</i>	<i>spite of the fact</i>
<i>side of the coin</i>	<i>view of the fact</i>
<i>whale of a time</i>	

Table 5.13 (Semi)-fixed phraseological examples from the top 100 lexical exponents at C2

It is worth noting here that the frequency of this type of formulaic example increases as proficiency increases from B2 upwards. (For example at B2 we start to see exponents such as *opportunity of a lifetime*, and at C1 *(in) case of an emergency*. This last example is noteworthy as an illustration of a sequence which is displaying understanding of the fixedness of the formula, *in case of emergency*, though possibly with evidence of the slot and frame mechanism with the insertion of *an* before *emergency*. I return to this in Chapter 9.)

In terms of the development of form-meaning pairings within this core pattern common to both A1 and C2, and highly ranking at all levels, results from the top 100 examples suggest that:

- A1 learners rely predominantly on two patterns across a limited range of meaning categories and a limited range of examples, producing holistic strings such as *front of the cinema*, *centre of the town*, *clock in the morning*, *table in the kitchen*.
- By C2, even when there are fewer pattern realizations there is an increase in meaning groupings. This suggests some kind of honing within the same pattern.
- As proficiency increases so does lexical and functional range.

- At C2, there appears to be some movement towards fixedness of patterning, which suggests a type of a sensitivity to item co-selection and more formulaic abstractions. Whereas at earlier levels there is a predominance for few literal references, possibly driven by tasks in the exam by C2, we see the emergence of more formulaic use. For instance, we see more shell nouns (Hunston and Francis, 2000) followed by a post modifier (*understanding of the world, majority of the population*). This seems to suggest that the learners at C2 are engaging in a selection process that is sensitive to the collocational choices in the entire sequence and the wider textual context in which the sequence is used.
- The groupings described in Pattern Grammar for categorising form-meaning relationships do not account for all forms and associated meanings, nor for the changes in form-meaning relationships across levels.

5.6 Scanning the landscape: general tendencies in POS tag sequence use

This chapter has set out a global approach to development in L2 writing and begun to explore whether development is observable through the frequency and distribution of POS tag sequences across proficiency levels, whether there are core POS tag sequences in L2 writing and how POS tag sequence usage changes across levels. The findings have been illustrated at different levels of abstraction: first through a bird's eye view of the top 100 sequences by level, and next by the top 10 POS tag sequences per level, then filtering down to the top sequence at both ends of the proficiency scale, and by investigating these on a lexical and functional level. This has revealed some general tendencies in development across levels:

- There are sequences that are consistently highly ranked, and therefore frequently used across all proficiency levels, categorised as core to the learner repertoire.
- This core of consistently used sequences grows as proficiency increases.
- There are sequences that decrease in rank, and therefore are less used than other sequences, as proficiency increases.
- There are sequences that increase in rank and therefore become more useful as proficiency increases.
- There is greater convergence between the highly ranked sequences at C2 and other levels, than between the highly ranked sequences at A1 and other levels, i.e. other levels make more use of the C2 top 100 than the A1 top 100.
- Adjacent proficiency levels show overall greater convergence than non-adjacent levels.

- Verb-based sequences decrease as proficiency increases.
- Noun-based sequences increase as proficiency increases.
- The B1 level is a turning point where verb-based and noun-based frequencies come together.
- There is a settling of usage between A1 and A2.
- Sequences containing punctuation remain consistent as proficiency levels increase.
- As proficiency increases learners include a wider range of syntactic forms in their repertoire, including past participles, prepositional and noun phrases (e.g. *of those exotic place, The idea of improving*), and display sensitivity to the register, the discourse management, orientation and signposting needs of writing (e.g. . *Finally, the*). This can only be explored through the lexical and functional usage of sequences across levels.
- There is inherent development in terms of the range, type and nature of the form-meaning pairings.
- Task effect on sequence usage must be considered.

Having taken an overall view of development in this chapter, the next three chapters explore each of the three proficiency level groupings in more detail. Continuing with the journey metaphor, Chapter 6 begins with the A level, the starting point in the language learning journey, and takes a front view and rear view perspective on the change in POS tag distribution and usage characterising change between the A and B levels.

Chapter 6 Setting out

Chapter 5 offered a broad global overview of development, from A1 to C2. This chapter is the first of a series of three in which I take a detailed look at the three broad CEFR levels, A, B and C. In this chapter I address the developmental journey from the perspective of the A2 learner, looking forward to B1 and back to A1, using the main suite subcorpus of the CLC (defined in Chapter 4). In the Common European Framework of Reference (CEFR), A1 and A2 levels are defined as belonging to the category ‘Basic User’. At A2, learners are described as having moved from the ‘Breakthrough’ level (A1) to ‘Waystage’ (A2), a stage defined as marking ‘the conclusion of the first significant phase for learners on their way to Threshold’ (B1) (Van Ek and Trim 1990).

There is strong evidence that global proficiency levels in English are forever on the increase (the EF proficiency index gives an annual global account of English language proficiency <https://www.ef.com/assetscdn/WIBIwq6RdJvcD9bc8RMd/cefcom-epi-site/reports/2021/ef-epi-2021-english.pdf>). Against this backdrop of evidence, A2 level is rarely a stopping point in the learner journey and is more likely to occupy a transitional place along a dynamic path. Here I examine what the data along this pathway illustrates, identifying the sequences that the A2 user relies on, what they have gathered and taken forward from A1, what they have left behind, and how this shapes the accumulation of language they take to B1.

In Table 6.1 the A1, A2 and B1 levels are described in general terms on the CEFR global scale. In these terms, L2 users are not expected to produce connected text until B1. However, the expectation in this present study is that A2 users do what is deemed to be characteristic of B1, that is, ‘produce simple connected text on topics which are familiar or of personal interest’.

INDEPENDENT USER	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe
---------------------	----	---

		experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
BASIC USER	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

<https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale>

Table 6.1 Global scale descriptors for A1 A2 and B1 as defined by the Council of Europe

The data in this study shows A2 users repeatedly demonstrating the ability to generalise how to put words together in a coherent way. For example one of the most frequent sequences of POS tags found in the A2 data is determiner+adjective+noun+preposition (e.g. *a large school in*). A2 learners are consistently and frequently putting a determiner before an adjective and following it with a noun and then a preposition. In doing so, potentially they are showing evidence of abstracting structural generalisations from the language they experience. In this chapter I describe what the A2 learners do with these generalisations and what marks differences in use between their usage and that of the levels above and the level below.

In the first part of the chapter (6.1 to 6.3) I first explore if development is observable through the frequency and distribution of POS sequences across proficiency levels A1, A2 to B1 (RQ1). I then take a case study approach (6.4 to 6.6) to address how POS sequences develop across proficiency levels A1, A2, B1 (RQ2) and explore if existing frameworks for classification of language patterning account for a description of development (RQ3).

6.1 Focusing in: overall distribution A1, A2 and B1

I begin with initial observations about the 4-gram POS tag sequences across A1, A2 and B1 levels. All 4-gram POS tag sequences were extracted from all three levels. The total number of POS 4-gram sequence occurrences and total POS 4-gram sequence types per level are shown in Table 6.2:

	A1	A2	B1
total 4-gram raw occurrences: all types	2293600	5496831	3183197
Total 4-gram types	110703	200384	164828

Table 6.2 Occurrences of POS 4-gram sequences across levels A1, A2 and B1

All sequences were ranked and the rankings of the top 50 types from each level selected for further analysis and comparison. The total number of POS 4-gram sequence occurrences in the top 50 types can be seen in Table 6.3.

	A1	A2	B1
total 4-gram raw occurrences: top 50	292781	590731	310365
50 types as % of all types	0.05	0.02	0.03
Total occurrences in top 50 as % of all	12.77	10.75	9.75

Table 6.3 Distribution of top 50 types across levels

These top 50 types constitute between 0.05% and 0.02% of types of POS 4-grams in the A1, A2 and B1 data (Table 6.3). However they account for 12.77%, 10.75% and 9.75% of all POS 4-gram occurrences. In short, even though they represent a small fraction of types, their token occurrences are so high that they make up 10% and more of all occurrences at each level. The highest ranked are the most frequently occurring which means it is possible to observe a lot from relatively little (See also the overall picture of distribution in Chapter 5, Figure 5.1).

6.1.1 Overall distribution: top 50 A2 sequences

Here I focus on change from the perspective of the top 50 A2 sequences. A snapshot of this change can be seen in Figure 6.1, with specific focus within the red box. The colour coding gives a visual overview of the convergence in ranking between levels, from dark green (highly convergent) to pink (highly divergent) (see key). The difference in ranking shows how the top 50 A2 POS tag sequences are distributed across other levels and allows us to observe how their distribution changes.

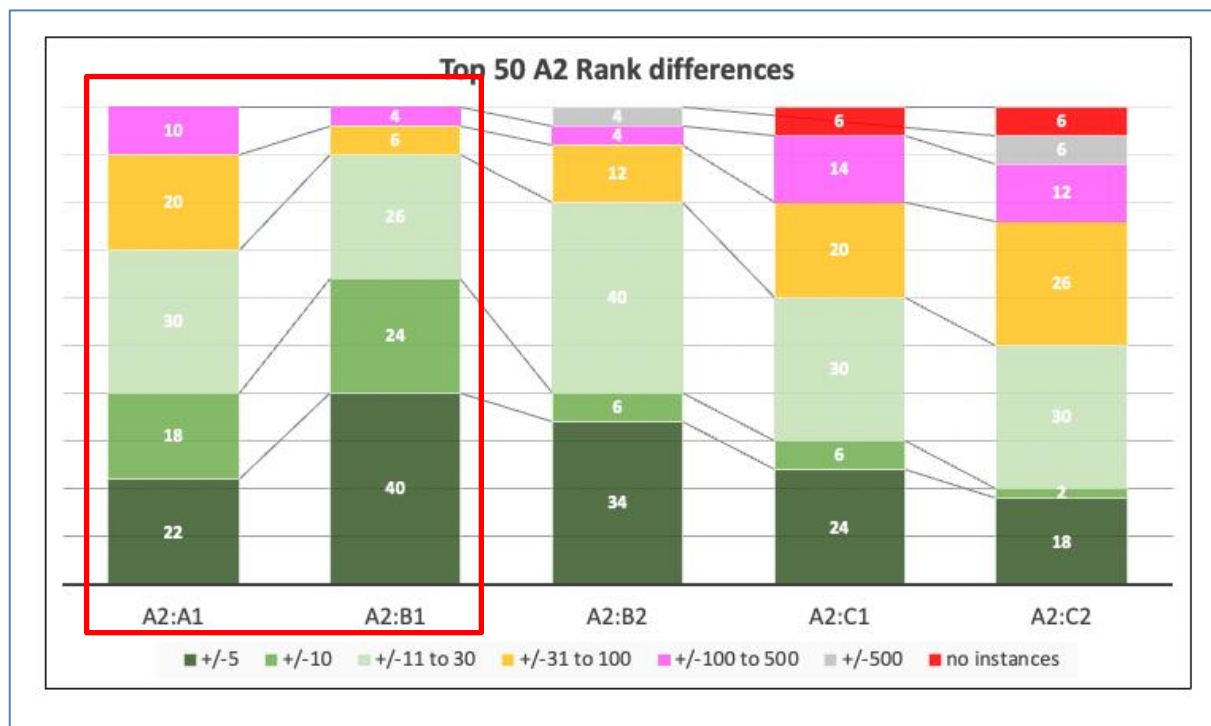


Figure 6.1 Percentage convergence of the top 50 sequences at A2 with their rankings at all other levels

Initial observations indicate that, in terms of POS tag sequence usage, A2 learner writing looks more like B1 writing than A1 writing. There is a clear convergence between A2 and B1 sequence usage. 90% (45) of the top 50 at A2 are also within a range of +/- 30 ranks at B1 (indicated by all of the green sections in the second column). 40% (20) of these are within a range of +/-5 (dark green), indicating these sequences are closely and highly ranked at both A2 and B1 levels.

There is less convergence between the A2 and A1 rankings, illustrated in the first column to the left (Figure 6.1). 70% (35) of the A2 top 50 are ranked within a range of +/- 30 at A1.

22% (11) are closely and highly ranked, within a range of +/-5 ranks. The orange section in this left most column (beyond the range of +/- 30 to 100) and the pink section (with a difference in rank of +/-100 to 500) indicate sequences that are used much less at A1 than at A2.

It is important to consider exam and task effect here. As detailed in chapter 4 (Table 4.1) 100% of the A1 performance data comprises scripts from the A2 (KET) exam, that is students who took their exam at A2 level but whose performance places them at A1. 12.5% of the A2 performance data consists of scripts which met the 'at grade' criteria for the A2 exam, whereas 87.5% comes from those which achieved A2 performance level in the B1 (PET) level exam (i.e. they performed 'below level'). In the B1 performance data 2.8% comes from A2 exam data, 59% from B1 exam data and 38.2% from B2 exam data. Given the high contribution in the A2 performance data from the B1 exam it might therefore be unsurprising that overall there are more similarities between A2 and B1 writing, since the data may be coming from similar tasks. However, the fact that almost 40% of the B1 performance data is from B2 exam (and 100% of the A1 performance data comes from the A2 exam) suggests that there is more than exam or task effect at play here. This is explored further in section 6.6).

Looking beyond the B1 level, the sequences used most frequently at A2, become less and less used as proficiency increases (See Figure 6.1). This is illustrated by a gradual decrease in the highly convergent ranked sequences (dark green) and a gradual increase in the divergence in sequences (illustrate by the orange, pink, grey and red sections). The sequences which are important to A2 usage become less and less important as proficiency increases.

6.2 A2 sequences: looking back and looking forward

The rankings of the top 50 sequences at A2 were compared with their ranks at A1 and B1, and their relative rank variance calculated using the simple rank difference calculation described in chapter 4.

As alluded to in Chapter 5, many of the top A2 50 sequences which rank higher at A1 than at A2 also contain punctuation. This seems to be characteristic of lower levels and may be indicative of the fact that writing at A1 level is made up of short syntactic units. Punctuation may be of interest developmentally in relation to length and complexity of structures but falls beyond the scope of this study since many of them are composed of two separate fragments

separated by punctuation. From this point onwards, sequences with punctuation are removed.

The 31 remaining sequences are listed in Table 6.4:

COLOUR KEY	
rank variance of +/-	
5	
10	
11 to 30	
31 to 100	
101 to 500	
501+	
#VALUE!	= not found

A2 rank		Rank difference	
	POS tag sequences and <i>examples</i>	A1	B1
1	noun prep det noun NN IN DT NN <i>centre of the town</i>	-9	0
2	prep det adj noun IN DT JJ NN <i>to a new shop</i>	-60	-1
4	prep det noun prep IN DT NN IN <i>in the centre of</i>	-11	-1
6	pronoun modal verb-base prep PP MD VV IN <i>you should go to</i>	4	-6
9	det adj noun prep DT JJ NN IN <i>the other side of</i>	-100	0
11	noun prep poss-pronoun noun NN IN PPZ NN <i>poster for my room</i>	0	-2
13	proper-noun proper-noun proper-noun proper-noun NP NP NP NP A <i>VERY LARGE GARDEN</i>	-27	9
14	verb-base prep det noun VV IN DT NN <i>go to the cinema</i>	-5	-15
15	pronoun pres-simpleV verb-base PP VVP TO VV <i>I want to see</i>	6	-7
16	det noun prep det DT NN IN DT <i>a desk in the</i>	-16	9
18	pronoun modal verb-base pronoun	-28	0

	PP MD VV PP <i>you would like it</i>		
	pronoun modal verb-base det		
24	PP MD VV DT <i>you can visit a</i>	-2	-4
	prep det noun noun		
25	IN DT NN NN <i>with a tennis ball</i>	-11	-18
	to-inf verb-base det noun		
28	TO VV DT NN <i>to buy a sofa</i>	-15	18
	pronoun modal adverb verb-base		
29	PP MD RB VV <i>I can't think</i>	-177	10
	to-inf verb-base prep det		
30	TO VV IN DT <i>to go in the</i>	-31	-15
	det noun prep noun		
31	DT NN IN NN <i>a large school in</i>	-53	-1
	pronoun pres-simpleV pronoun modal		
33	PP VVP PP MD <i>I think you can</i>	-62	2
	pronoun pres-simple-be -ing-form to		
34	PP VBP VVG TO <i>I'm going to</i>	5	-16
	pronoun pres-simpleV adverb verb-base		
35	PP VVP RB VV <i>I don't like</i>	-184	-25
	verb-base det noun prep		
36	VV DT NN IN <i>meet a lot of</i>	-60	11
	adj noun prep det		
37	JJ NN IN DT <i>new shop in the</i>	-337	-10
	det noun prep plural noun		
38	DT NN IN NNS <i>a lot of things</i>	-43	-17
	det noun prep pronoun		
39	DT NN IN PP <i>the cinema with me</i>	-27	-15
	pronoun pres-simple-have to-inf verb-base		
40	PP VHP TO VV <i>I have to take</i>	22	6
	pres-simple-be -ing-form to-inf verb-base		
41	VBP VVG TO VV <i>'m going to buy</i>	4	-15
	prep det noun conj		
42	IN DT NN CC <i>in the corner and</i>	-137	2

44	det noun prep poss-pronoun DT NN IN PPZ <i>a shop near my</i>	-23	9
45	to-inf verb-base prep pronoun TO VV IN PP <i>to hear from you</i>	-8	-40
46	pronoun past-simpleV det noun PP VVD DT NN <i>I bought a dress</i>	16	19
49	modal verb-base prep det MD VV IN DT <i>can go to the</i>	-61	-76

Table 6.4 Top 50 4-gram POS sequences at A2, and their rank differences at A1 and B1

This list is then used to identify changes in sequence usage and provide a starting point for further exploration of lexical and functional characteristics. The changes in sequence usage are identified by their change in ranking, which is shown in the rank difference columns to the right of the table (Table 6.4). Negative rank difference figures indicate a lower ranking at the other levels, and positive figures indicate a higher ranking. For example item #2 at A2 (preposition+determiner+adjective+noun e.g. *to a new shop*) is ranked at #62 in the A1 data (with a rank variance of -60) and at #3 in the B1 data (with a rank variance of -1). This indicates a jump in the increase in usage in the A2 and B1 repertoires in comparison with the A1 repertoire. The colour coding gives a visual overview of the degrees of convergence in ranking between levels, from dark green (highly convergent) to pink (highly divergent) (see key).

The overall results across all proficiency levels, described in Chapter 5, pointed to three types of sequences: (1) core sequences (2) emerging sequences and (3) decreasing sequences (see section 5.3).

In the next two sections I look first at these types and how they change between the A1, A2 and B1 levels, before exploring examples of their lexical and functional characteristics.

6.3 A2 sequences: looking ahead to B1

6.3.1 Core sequences: A2 and B1

There are 10 core sequences that are ranked closely (within +/-5) at both A2 and B1 (Table 6.5). They are dominated by noun phrases, two of which contain adjectives (#2, #9), and verb sequences containing modal verbs (#18 #24 #33).

A2 rank	POS tag sequences and examples	Rank difference B1
1	noun prep det noun NN IN DT NN <i>centre of the town</i>	0
2	prep det adj noun IN DT JJ NN <i>to a new shop</i>	-1
4	prep det noun prep IN DT NN IN <i>in the centre of</i>	-1
9	det adj noun prep DT JJ NN IN <i>the other side of</i>	0
11	noun prep poss pronoun noun NN IN PPZ NN <i>poster for my room</i>	-2
18	pronoun modal verb-base pronoun PP MD VV PP <i>you would like it</i>	0
24	pronoun modal verb-base det PP MD VV DT <i>you can visit a</i>	-4
31	det noun prep noun DT NN IN NN <i>a large school in</i>	-1
33	pronoun pres-simpleV pronoun modal PP VVP PP MD <i>I think you can</i>	2
42	prep det noun conj IN DT NN CC <i>in the corner and</i>	2

Table 6.5 Core sequences: A2 sequences which are closely ranked at both A2 and B1.

6.3.2 Emerging sequences: A2 and B1

The emerging sequences are those which rank higher at B1 than A2 and therefore become increasingly more important for B1 learners (Shown by the increase in rank difference at B1 in Table 6.6). Noticeable here is the increasing number of sequences containing verb forms.

A2 rank	POS tag sequences and examples	Rank difference B1
13	proper-noun proper-noun proper-noun proper-noun NP NP NP NP A <i>VERY LARGE GARDEN</i>	9

16	det noun prep det DT NN IN DT <i>a desk in the</i>	9
28	to-inf verb-base det noun TO VV DT NN <i>to buy a sofa</i>	18
29	pronoun modal adverb verb-base PP MD RB VV <i>I can't think</i>	10
36	verb-base det noun prep VV DT NN IN <i>meet a lot of</i>	11
40	pronoun pres-simple-have to-inf verb-base PP VHP TO VV <i>I have to take</i>	6
44	det noun prep poss pronoun DT NN IN PPZ <i>a shop near my</i>	9
46	pronoun past-simpleV det noun PP VVD DT NN <i>I bought a dress</i>	19

Table 6.6 Emerging sequences: A2 sequences which are higher ranked at B1 than A2 (with rank difference).

6.3.3 Decreasing sequences: A2 and B1

Decreasing sequences are those that are lower ranking at B1 than at A2 (shown by the decrease in rank difference in Table 6.7). At the top of the table are those that are the least used at B1 in relation to other sequences, becoming less relevant in the B1 repertoire.

A2 rank	POS tag sequences and <i>examples</i>	Rank difference B1
49	modal verb-base prep det MD VV IN DT <i>can go to the</i>	-76
45	to-inf verb-base prep pronoun TO VV IN PP <i>to hear from you</i>	-40
35	pronoun pres-simpleV adverb verb-base PP VVP RB VV <i>I don't like</i>	-25
25	prep det noun noun IN DT NN NN <i>with a tennis ball</i>	-18
38	det noun prep nounS	-17

	DT NN IN NNS <i>a lot of things</i>	
34	pronoun pres-simple-be -ing-form to-inf PP VBP VVG TO <i>I'm going to</i>	-16
14	verb-base prep det noun VV IN DT NN <i>go to the cinema</i>	-15
30	to-inf verb-base prep det TO VV IN DT <i>to go in the</i>	-15
39	det noun prep pronoun DT NN IN PP <i>the cinema with me</i>	-15
41	pres-simple-be -ing-form to-inf verb-base VBP VVG TO VV <i>'m going to buy</i>	-15
37	adj noun prep det JJ NN IN DT <i>new shop in the</i>	-10
15	pronoun pres-simpleV to-inf verb-base PP VVP TO VV <i>I want to see</i>	-7
6	pronoun modal verb-base prep PP MD VV IN <i>you should go to</i>	-6

Table 6.7 A2 sequences decreasing in ranking at B1 (with rank difference).

These appear to be a mixed bag of structures. Of particular interest here are the ones that have dropped dramatically in ranking, e.g. #49 modal verb-base prep det (*can go to the*) and #45 to-inf prep pronoun (*to hear from you*). An initial look at the lexical instances of these suggests that there is a task effect at play, with a small number of tasks generating a large percentage of these instances. For example at A2, there are 7181 instances of the #49 modal+verb-base+prep+det (*can go to the*) sequence, 63% of which are generated by three questions, asking for advice about what to do in a town, a typical type of exam task at these levels. However, the three questions come from three different years of the PET B1 exam, and the range of lexical instances suggest some structural abstraction (e.g. *should go to the / can go to the / can meet at the / will go to the*). The A2 performance data for this sequence is dominated by a B1 exam task. The B1 performance data suggests that those who attain a B1 level are using this sequence less frequently than those attaining a A2 level in a B1 exam.

6.4 A2 sequences: looking back to A1

As with the A2 and B1 comparison the shift in ranking between the top A2 50 sequences and their relative rankings at A1 shows a picture of convergence and divergence.

6.4.1 Core sequences: A2 and A1

The sequences that are closely ranked at both A1 and A2 levels (within +/-5) (Table 6.8) are dominated by verb sequences containing modal verbs and present progressives with one noun phrase sequence (noun prep poss pronoun noun. *theatre with your aunt*) unlike the core sequences between A2 and B1 which are characterised by noun phrases. Two of these core sequences also remain core in the B1 repertoire, **noun+prep+poss pronoun+noun** (*theatre with your aunt*) and **pronoun+modal+verb-base+det** (*you can visit a*).

POS tag sequences and examples		
A2 rank		Rank difference A1
14	verb-base prep det noun VV IN DT NN <i>go to the beach</i>	-5
24	pronoun modal verb-base det PP MD VV DT <i>you can visit a</i>	-2
11	noun prep poss pronoun noun NN IN PPZ NN <i>theatre with your aunt</i>	0
6	pronoun modal verb-base prep PP MD VV IN <i>you can go to</i>	4
41	pres-simple-be -ing-form to-inf verb-base VBP VVG TO VV <i>'m going to buy</i>	4
34	pronoun pres-simple-be -ing-form to PP VBP VVG TO <i>I'm going to</i>	5

Table 6.8 Core sequences: A2 sequences which are closely ranked at both A2 and A1.

6.4.2 Emerging sequences: A2 and A1

The emerging sequences are noticeable for their divergence in the ranking (Table 6.9). The higher up the table the more divergent their use. 55% of these sequences are used far more frequently at A2 than A1, some having a rank frequency difference of tens or hundreds.

Amongst these sequences are noun phrases with adjectives and conjunctions, e.g. #37, 42, 9,

2 and two verb sequences containing adverbs e.g. #35, 29. In these two sequences the RB adverb tag refers to the negative *n't*, indicating the emergence of verb sequences with negative forms at A2.

A2 rank	POS tag sequences and <i>examples</i>	
		Rank difference A1
37	adj noun prep det JJ NN IN DT <i>large desk with a</i>	-337
35	pronoun pres-simpleV adverb verb-base PP VVP RB VV <i>I don't like</i>	-184
29	pronoun modal adverb verb-base PP MD RB VV <i>I can't see</i>	-177
42	prep det noun conj IN DT NN CC <i>in the corner and</i>	-137
9	det adj noun prep DT JJ NN IN <i>the other side of</i>	-100
33	pronoun pres-simpleV pronoun modal PP VVP PP MD <i>I think you can</i>	-62
49	modal verb-base prep det MD VV IN DT <i>can go to the</i>	-61
2	prep det adj noun IN DT JJ NN <i>to a new shop</i>	-60
36	verb-base det noun prep VV DT NN IN <i>visit a castle in</i>	-60
31	det noun prep noun DT NN IN NN <i>a lot of money</i>	-53
38	det noun prep plural noun DT NN IN NNS <i>a lot of things</i>	-43
30	to-inf verb-base prep det TO VV IN DT <i>to go in the</i>	-31
18	pronoun modal verb-base pronoun PP MD VV PP <i>you can give me a</i>	-28
13	proper-noun proper-noun proper-noun proper-noun noun	-27

	NP NP NP NP <i>A VERY LARGE GARDEN</i>	
39	det noun prep pronoun DT NN IN PP <i>the shop with me</i>	-27
44	det noun prep poss pronoun DT NN IN PPZ <i>a shop near me</i>	-23
16	det noun prep det DT NN IN DT <i>a desk in the</i>	-16
28	to-inf verb-base det noun TO VV DT NN <i>to buy a sofa</i>	-15
4	prep det noun prep IN DT NN IN <i>in the middle of</i>	-11
25	prep det noun noun IN DT NN NN <i>with a tennis ball</i>	-11
1	noun prep det noun NN IN DT NN <i>centre of the town</i>	-9
45	to-inf verb-base prep pronoun TO VV IN PP <i>to hear from you</i>	-8

Table 6.9 Emerging sequences used more at A2 than A1

Overall there is a lack of stabilisation between sequences used at A1 and A2 in comparison with those whose usage settles between A2 and B1. I note that these sequences contain among other elements, adjectives and adverbs, non-finite verb forms, and a plural noun form. These sequences may be revealing about the transition from A1 to A2.

6.4.3 Decreasing sequences: A2 and A1

Decreasing sequences are those that are ranked lower in the A2 data than the A1 data, however they are all within a +/- rank of 30, as shown in Table 6.10. still relatively highly ranked. A2 users however are less reliant on them than A1 users. It is noticeable that they all contain verbs phrases. Given that the A1 data is taken from A2 exams it may indicate that those who do not attain a sufficiently high mark to reach the A2 level may not be relying on verb phrases and not demonstrating a wide enough range of sequence use.

A2 rank	POS tag sequences and <i>examples</i>	Rank difference A2-A1
15	pronoun pres-simpleV to-inf verb-base PP VVP TO VV <i>I want to buy</i>	6
46	pronoun past-simpleV det noun PP VVD DT NN <i>I bought a dress</i>	16
40	pronoun pres-simple-have to-inf verb base PP VHP TO VV <i>I have to go</i>	22

Table 6.10 Decreasing sequences used less at A2 than A1

6.4.4 A developmental picture: summary

Overall, there is evidence of more convergence between A2 and B1 than A2 and A1. This may indicate the beginning of a settling of the usage of the high ranking sequences at A2, which then continues to B1, and as a result of task effect or a combination of both. A general summary to inform the next phase of analysis is that:

- The sequences that are core to adjacent levels increase as proficiency increases.
- There is greater stabilisation of usage between A2 and B1 than between A1 and A2.
- Sequences with nouns and noun phrases in high ranking positions are dominant, particularly at A2 and B1, and less prevalent at A1, pointing to an increase in noun phrase development from A1 to A2.
- Some sequences with modal verbs and present progressive forms are core to A1 and A2 and become less central to B1 repertoire.
- Sequences with modal verbs are the most frequent and consistently highly ranked sequences with verbs at all levels.
- Sequences containing other verb phrases increase at B1.

The first phase of the analysis has shown the changes in distribution of POS tag sequences. Clearly, in order to give a comprehensive view of emerging development any number of these need to be investigated further. Since it is not possible to discuss every sequence some filtering is needed. Representative sequences for core and emerging types are illustrated for how they play out lexically and functionally. The following sections explore two emerging sequences, identified on the basis of these initial observations, and selected to illustrate some

of the questions, limitations and methodological challenges of approaching emerging development in large-scale exam data.

The sequences represent a noun phrase sequence, with adjectives DT JJ NN IN (*a new shop in*) and a sequence with a modal verb, PP MD RB VV (*I can't think*).

The following sections (6.5 and 6.6) explore research questions RQ2 and RQ3 firstly to examine how representative sequences develop across proficiency levels and in doing so explore whether existing frameworks for classification of language patterning account for this development.

6.5 Case study 1: Determiner + adjective + noun + preposition (DT JJ NN IN)

In this first case study I look at the sequence determiner+adjective+noun+preposition (DT JJ NN IN) (e.g. *a big concert in, an essential part for*), ranked #9 at A2. It is an example of an emerging sequence, not highly ranking at A1 (#109) but which jumps to the top 10 at A2 and becomes consistently important at B1, where it is also ranked #9 and continues to rise in ranking and consistently highly ranked at B2 (#6), C1(#5) and C2 (#4) levels. It has been selected to illustrate the challenges of applying existing functional frameworks to low proficiency level data.

6.5.1 Determiner + adjective + noun + preposition: occurrences by level

Table 6.11 shows the breakdown of the raw and relative occurrences of this sequence by level and shows the proportion of occurrences that the top 1000 types constitutes for each level and give an indication of type-token ratio.

	subcorpus size	raw occurrences	relative PMW occurrences	total occurrences	1000 types as % occurrences
				1000 types	
A1	2456971	2352	957	1684	71.6
A2	5703217	15404	2701	9382	60.9
B1	3261473	8379	2569	3977	47.5
B2	5263979	17135	3255	6741	39.3
C1	6711568	27197	4052	9217	33.9
C2	7698695	34971	4542	10813	30.9

Table 6.11 Breakdown of occurrences by level of determiner+adjective+noun+preposition

These findings confirm increasing use of this sequence as proficiency increases. In relative terms the sequence is used three times more frequently in the A2 than the A1 data, dips slightly in relative frequency from A2 to B1 and increases in usage steadily from B1 to C2. The relative occurrences at A2 and B1 are almost identical however, as indicated by the total occurrences of 1000 types as a percentage of all occurrences, the range of lexical exponents increases. A higher percentage reflects a lower range of types. For example the first 1000 types make up 71.96% of all occurrences at A1, decreasing to 60.9% at A2, 47.5% at B1, 39.3% at B2, 33.9% at C1 and 30.9% at C2. The implication in simple, formal terms is that the range of lexical exponents used increases as proficiency level increases, with a slight dip at B1. The sequence is an example of one where two open word class slots (adjective+noun) sit together, (between two closed word class slots (determiner, preposition), and provides the opportunity for a greater range of patterning in terms of both independent selection of high frequency (e.g. *a/the + new/black + door/shop/shirt/house + with/near*) items and the emergence of fixed co-selected patterns (e.g. *a wide range/variety of, a huge/large amount/number of, a free copy of*)

6.5.2 Determiner + adjective + noun + preposition: Top 50 lexical exponents: applying frameworks for structural and functional characteristics

To look at the structural and functional characteristics at each level, the top 50 most frequent lexical exponents of the DT JJ NN IN sequence were extracted from the A1, A2 and B1 data, using the word forms in KWIC function in Sketch Engine. The top 20 are shown in Table 6.12. Any lexical sequences which appear in exam rubrics were first identified and marked with an asterisk. They are not excluded from the analysis since they may play a role in abstracting structural generalisation.

A1	A2	B1
<i>a new pair of</i>	<i>*a large school in</i>	<i>*a large school in</i>
<i>this mobile phone because</i>	<i>a new shop in</i>	<i>*a small school in</i>
<i>a new job in</i>	<i>*a small school in</i>	<i>*the large school in</i>
<i>a big concert in</i>	<i>*the large school in</i>	<i>a long time since</i>
<i>a good time with</i>	<i>a new shop near</i>	<i>a special day in</i>
<i>a pop concert in</i>	<i>*the small school in</i>	<i>a great time with</i>
<i>a pop concert on</i>	<i>the new shop in</i>	<i>a good time with</i>
<i>a new house in</i>	<i>a long time since</i>	<i>the new class because</i>

<i>a pop concert with</i>	<i>a new bed for</i>	<i>*the small school in</i>
<i>a mobile phone for</i>	<i>a good time with</i>	<i>the other side of</i>
<i>a new dress for</i>	<i>a new pair of</i>	<i>a good idea because</i>
<i>the mobile phone because</i>	<i>a new shop at</i>	<i>an essential part of</i>
<i>a big party with</i>	<i>a new shop of</i>	<i>a good time in</i>
<i>a great concert on</i>	<i>a new lamp for</i>	<i>a new collection of</i>
<i>a great time in</i>	<i>a great time with</i>	<i>the same problem as</i>
<i>a new concert in</i>	<i>a new desk for</i>	<i>a free copy of</i>
<i>the first day of</i>	<i>a big house with</i>	<i>a new bed for</i>
<i>the next week on</i>	<i>the new shop near</i>	<i>the first time in</i>
<i>a beautiful pair of</i>	<i>a new shop on</i>	<i>the other half with</i>
<i>a great time at</i>	<i>a new computer for</i>	<i>a great time in</i>

*indicate sequences that appear in exam question rubrics

Table 6.12 Top 20 most frequent lexical realisations of determiner + adjective + noun + preposition at A1, A2 and B1

Existing frameworks for categorisation were then applied according to the exploratory methodology set out in Chapter 4 to see how the sequences change from level to level. This process and the findings are described in sections 6.5.3 to 6.5.6 below.

6.5.3 *Determiner + adjective + noun + preposition: applying Pattern Grammar*

Applying a pattern grammar classification, set out by Hunston and Francis (2000) (see also <https://grammar.collinsdictionary.com/grammar-pattern>), involves identifying firstly form groupings or ‘grammar patterns’ (see also chapter 5), e.g. adj N (adjective + noun) or N to n (noun phrase to noun) and secondly the meaning groupings for each pattern (e.g. manner, power, era). While this approach was successfully applied to some 4-gram POS sequences, see chapter 5 (section 5.5.2), it proved difficult to apply in this case study, for the following reasons:

Firstly 4-gram POS tag sequences are often not structurally complete. Pattern grammar does not accommodate fragments of (noun) phrases that 4-gram sequences often produce. For example, a 4-gram sequence like determiner+noun+preposition+noun (NN IN DT NN, e.g. *centre of the city*), as seen in chapter 5, fits neatly under some of the groupings, e.g. N of n, N in N, where N stands for noun phrase and the determiner is understood as part of the pattern

((*the*) *centre of the city*) ; however the sequence (and noun phrase fragment) determiner+adjective+noun+preposition (DT JJ NN IN e.g. *a large school in*) under examination here is not accounted for by a grammar pattern. One approach would be to assume a following complementing noun, and categorise this sequence under noun phrase patterns in pattern grammar (e.g. N of N, N in N). This moves into the territory of longer POS tag n-grams (5-/6-grams) and is beyond the scope of this study. Added to that it would not accommodate the internal elements of the sequence, e.g. adjective + noun, determiner + adjective + noun, and in this case it is the increased use of the adjective in this sequence which contributes to and marks the transition from A1 to A2 (cf 6.3.1).

Secondly where a fragment or part of a fragment is categorised, both structurally and semantically (e.g. adj N), many of the lexical items found in the learner data are not specified under any of the meaning categories in pattern grammar, or are categorised under a general heading, ‘Nouns with other meanings’ and subcategorised as descriptive (e.g. *great time, good condition, interesting view*) or classifying (e.g. *black community, musical prodigy*). It does not account for subtle differences in compositionality. For example it does not account for the differences in compositional strength between, e.g. *a new pair of* and *a new job in*, nor for any potential emerging structural generalisations, e.g. *a new lamp/bed/computer for*

Thirdly, as acknowledged by Hunston and Francis (2000), it is the occurrence of repeated forms that drive the pattern grammar categorisation. Meanings are arrived at intuitively and subjectively and are of secondary importance to form in this framework. Added to this, the relative frequency of one pattern over another is not central, which means that there is no indication in this framework whether one pattern or meaning group occurs more frequently than another. One result of this is that some of the most frequently occurring lexical realisations of the patterns (e.g. *a great/good time with, a good friend of, a wide range/variety of*) are not accounted for in the pattern grammar meaning groups.

In summary pattern grammar provides a descriptive framework for some of the structural and functional elements in some 4-gram sequences but does not accommodate all, nor does it account for emerging generalisations.

The B1 sequence ‘a free copy of’ is one such example of the limitations of applying this classification. The word ‘copy’ appears under the ‘diagram’ group in the (grammatically complete) N of n pattern, along with the explanation illustrated in Figure 6.2.

Grammar Patterns > Nouns > N of n

N of n

The 'diagram' group

The noun refers to a picture or representation of some kind. The noun group after *of* indicates what it represents.

*The policeman showed a **diagram of the position of the body**.*
*I did what I thought was a competent **drawing of the nude model**.*
*The colour **illustration of Alice** by Sir John Tenniel for 'Alice in Wonderland' was first published in 1911.*
*Every schoolroom and every office had a **portrait or bust of Lenin**.*

• bust	• illustration	• picture	• representation
• copy	• imitation	• portrait	• sketch
• diagram	• map	• portrayal	• statue
• drawing	• model	• prototype	
	• photograph	• replica	

Figure 6.2 Extract from the 'diagram' group from Pattern Grammar

https://grammar.collinsdictionary.com/grammar-pattern/n-of-n_5

However in the example from the B1 data it is the semi-fixedness of all elements of *a+free+copy+of+noun* which holds meaning rather than *a+copy+of+noun* of the picture/representation type in the 'diagram' group. The meaning of 'copy' in this B1 example does not fit with the definition of the diagram group.

Following the limitations of a pattern grammar approach, a lexical bundle approach was applied.

6.5.4 *Determiner + adjective + noun + preposition: applying a lexical bundle approach*

According to Biber *et al.*'s operational definition (1999, p. 993), a lexical sequence is counted as recurrent only if it occurs at least ten times per million words in a register, and across five different texts in a register. Since many of the top 50 instances of this sequence fall below this threshold, I begin by looking at all instances of the bundles, following Gray and Biber's observations concerning continuous and discontinuous sequences (2013). In their analysis of lexical bundles, Chen and Baker operationalise this further by filtering context-dependent combinations, e.g. those that occur because of the context of an essay topic, since they are not considered to constitute 'building blocks' of language (2010, p. 855). For this reason the bundles were first categorised functionally broadly to identify if they are topic or content-driven and if not, to then look at their distribution following the functional taxonomy in Biber *et al.* (2004), of referential, stance and discourse organising bundles (See Appendix 4 for the categorisation used). All levels (A1 to C2) were analysed in order to explore any developmental changes at higher proficiency. Further analysis showed a split

between topic combinations and referential use. Referential uses were then subcategorised into quantifying (Ref_Quantity), framing (Ref_Frame) and those specifying attributes of place (Ref_Place) and time (Ref_Time), illustrated in Figure 6.3:

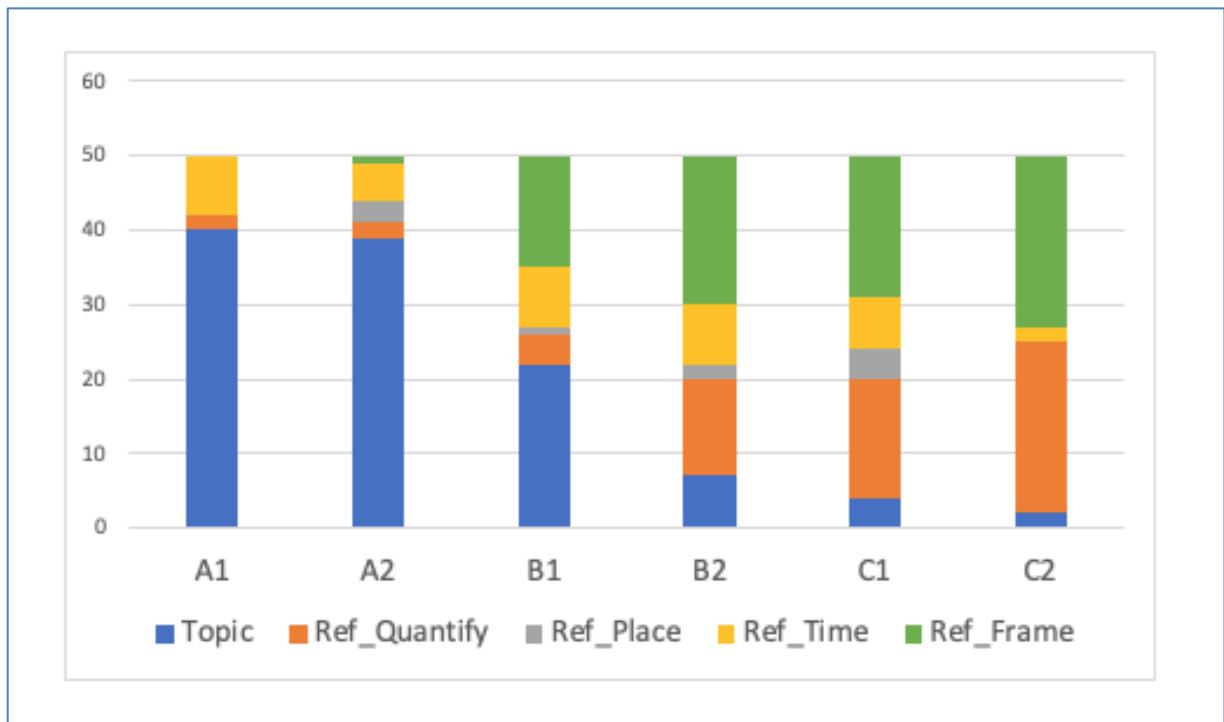


Figure 6.3 Functional categorisation of lexical sequences of determiner + adjective + noun + preposition DT JJ NN IN

Overall, the picture that emerges is that lexical sequences at the A levels are driven by the topic and context of the exam task (e.g. *a big concert in, a pop concert in, a new job in, a new shop near*). At B1 level, once more a pivotal point in development (see chapter 5), there is an equitable distribution between sequences driven by topic (e.g. *a new bed for, a big house with*) and referential lexical-bundle type sequences (e.g. *a long time since, a good place for, the other side of*) which are semi-fixed and formulaic in nature. By C2, the top 50 sequences are dominated by quantifying expressions (e.g. *a wide range of, a great deal of, a wide variety of, a great number of*) and framing expressions (e.g. *a great opportunity for, the main reason for, a major role in*), with only 4% driven by the topic or context. There is a clear shift from context-dependent sequences to referential sequences, as proficiency increases, alongside, more specifically, a growth in quantifying and framing use. If, following a traditional lexical bundle taxonomy approach, context-dependent sequences are removed this would exclude 80%, 78% and 22% of the A1, A2 and B1 DT JJ NN IN sequences. In this instance, while a lexical bundle approach appears to be appropriate and revealing for

development of proficiency from B1 above, it does not provide an adequate means for investigating development at the lower end of the proficiency scale, up to B1, the focus of this chapter. It accounts for the sequences of the referential type, for example with a quantifying function (*a wide range of, a little bit of, a large number of*) or a framing function (*an essential part of, an important role in*) but not for the recurrent frequently occurring sequences found in the A level data (*a new job in, a new shop in, a new desk for, a new lamp for, a big house with*).

However there are some important avenues for further investigation from this. What it broadly suggests is that learners at A1 and A2 levels rely heavily on topic to put together sequences, and in this case it is the concrete adjectives and nouns relating to the topic or task which are the building blocks for sequences. This may also lend evidence for the early slot and frame stage of the developmental sequence proposed by a usage-based theory of language learning (Ellis 2002; Lieven 2016) . Further exploration of the lexical and structural elements of this sequence (DT JJ NN IN) is needed to support this and to do so I turn to a p-frame approach.

6.5.5 Determiner + adjective + noun + preposition: applying a p-frame or lexical frame approach

Approaching the data using p-frames (otherwise known as phrase frames, lexical frames or collocational frames) investigates recurrent word sequences that differ only by one word (See Chapter 2). Previous studies (Chapter 3) have suggested that lower level learners rely on fixed and predictable p-frames with little variance within each slot and that these reflect the topic of the communicative context. However, as noted above, the sequence DT JJ NN IN is an example of one where two open word class slots (adjective+noun) sit together, which offers the opportunity for a greater range of independently selected items of high frequency (e.g. *a/the + new/black + door/shop/shirt/house + with/near*) items as well as fixed co-selected patterns (e.g. *a wide range/variety of, a huge/large amount/number of, a free copy of*). The lower level lexical sequences at A1 and A2 are characterised by two independently selected different items in the adjective and noun slots and therefore a p-frame approach of sequences identical apart from *one* element is not appropriate.

6.5.6 *Determiner + adjective + noun + preposition: structural and functional generalisations from A to B*

Increasing structural and functional generalisation at the A1 to B1 levels is observable, despite the fact that none of the approaches taken so far provide an adequate description. To dig deeper into these generalisations, using the top 50 lexical sequences at each level, I looked at the first and last closed class elements of the sequence DET (determiner) and IN (preposition) surrounding two open word slots JJ (adjective) and NN (noun), revealing the results shown in Figure 6.4.

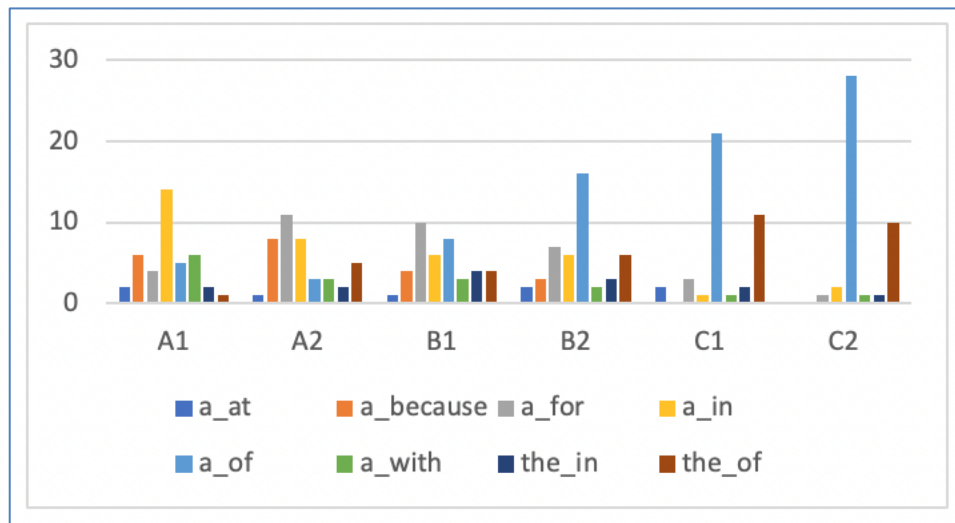


Figure 6.4 Distribution of forms for determiner and preposition positions in DT JJ NN IN across all levels

The A1 and A2 data show a mixed bag of usage, though they exhibit both structural and functional generalisation in the recurrent use of *a+adj+noun+preposition*, e.g. *a+adj+noun+in* to refer to place (*a new shop/house in*), *a+adj+noun+for* to refer to purpose (*a new dress for*, *a mobile phone for*), etc. What is noticeable is that at the A1 and A2 levels these functions are largely driven by the prepositional meaning and is also evidenced in the lower usage of *a+adj+noun+of*, which as attested by Sinclair (1991, p. 109) is a difficult item to categorise semantically. As discussed above, this may be evidence of the slot-filling phase as proposed by a usage-based theory, akin to early signs of grammatical abstraction seen in first language acquisition, when children map combinations of words onto agents, locatives and objects. They demonstrate understanding of concrete procedural frames which signal completion with people, places and things. Further evidence for this is seen in the consistent rise in usage of *a+adj+noun+of* and *the +adj+noun+of* usage as proficiency increases, (see Figure 6.4)

corresponding to semi-fixed referential expressions which quantify and frame (discussed below), where the semantic categorisation is driven by the sequence in its entirety.

A close look at candidates for each slot shows that the top 50 lexical sequences at A1 (See Table 6.12 for the top 20) are characterised by the use of descriptive adjectives and concrete nouns (*a black shirt for, a big shop in*), which are firmly tied to task topic, alongside early, though less frequent, signs of frames used for evaluation and quantity (*a great time with, a little bit of*). By A2 there is growth in the use of descriptive adjectives (the yellow door in), alongside an increase in evaluative use (*a good idea for, an interesting place because*). The descriptive topic-derived sequences are dominant in the top sequences at A1 and A2. Beyond the top 20, evaluative and referential sequences begin to appear. The increase in frames expressing evaluative and quantifying functions continues into B1 (*a great opportunity for, a large number of*), rising higher in the ranks, with context-dependent sequences decreasing steadily.

Table 6.13 shows how this distribution from topic-derived sequences to evaluative and quantifying now allows for a lexical bundle framework to be applied (Biber *et al.* 2004). This is possible once the usage begins to move away from independent slot and frame selection towards semi-fixedness. Table 6.13 shows the steady increase, from A1 to A2 to B1, of referential-type sequences with framing and quantifying functions, also illustrated in Figure 6.4.

A1					A2						B1					
the				topic	a	new	shop	at	topic	a	short	film	about	topic		
a	mobile	phone	because		a	new	bed	because			new	home	computer		because	
this						home						house				
an	interesting	place	because			house						bed				
a	black	shirt	because/with		a/the	large	school	because				desk			for	
a	new	T-shirt	because		a	small						lamp				
a	new	dress	for		a	new	bed	for			the		class		because	
	pink						chair									
	good	news	for				computer									
	mobile	phone	for/with				desk									
	beautiful	concert	in	dress												
	big			table												
	classical			television												
	musical			TV												
	new															
	pop						a		new	house	in	a	special	day	in	
great	concert	on	a	big				a	big	school	in					
musical			a/the	large	school	in	a/the	large								
pop	concert	of/in/on/with	a/the	small					a	small	town	near				
a	new	house	in	the	small	school	because	a	short	holiday	with					
a/the		job		the	yellow	door	in	a	good	idea	because					
a	big	party	in/with	a/the	new	shop	in/near/of/on	a	great	opportunity	if					
		shop		a	new	shop	of/on	a	good	idea	if					
		shop		a	big	house	with	a	good/great	experience	for					
		week	on	a	big	city	like	a		opportunity		idea				
		monday	at	the	last	film	of	a		place		opportunity				
		weekend	in	the	new			the	first	prize	in					
the/The	first	day	of	the	second	floor	of	the	first	time	in					
the	second	house	on	the	other	side	of	the	last	day	of					
	second	street	after	the				the	daily	life	of					
the	first	street	on	the				a	long	time	since					
half	past	ten	in	the				the	first	part	of					
the	first	turning	on	the				the	other	side	of					
a	good	time	at/in/with	the				the	same	place	as					
	great			the				the	same	problem	as/with					
a	beautiful	pair	of	ref_frame	ref_quant	a	good	time	in	ref_frame	a	good	way	of		
	new					a	good	idea	for		a	good	time	in/with		
	nice					a	good	idea	because		a	great	friend	of		
	an	interesting	place			a	good	copy			of					
a	little	bit				a	good	time	with		a	free				
						a	great				an	essential	part			
						a	good	time	in		a	little	bit	of	ref_quant	
						a	long	time	since		a	new	collection			
						a	new	pair			a	large	number			
						a	big	variety	of		a	new	pair			

Table 6.13 Lexical breakdown of DT JJ NN IN across A1, A2 and B1 categorised according to a lexical bundle framework (Biber *et al.* 2004)

6.5.7 *Determiner + adjective + noun + preposition: structural and functional generalisations beyond B*

Projecting forward and taking the longer front view perspective beyond the B levels, illustrated in Figure 6.4, an increasing reliance on *a/the + adjective+noun+of*, can be seen, with a corresponding function expressing quantity or group rising to the highest ranking sequences by C2.

The top lexical items at all levels, show the increase in quantity referring expressions. 9 of 10 of the top 10 lexical sequences at C2 are *a+of* expressions, 7 of which express quantity, illustrated in green in Table 6.14, with highly frequent use of semi-fixed (e.g. *a wide variety/range of*, *a great/large number of*) and fixed sequences (e.g. *a great deal of*, *the vast majority of*.)

rank	A1	A2	B1	B2	C1	C2
1	a new pair of	a large school in	a large school in	a wide range of	a wide range of	a wide range of
2	this mobile phone because	a new shop in	a small school in	a new shop in	a great deal of	a great deal of
3	a new job in	a small school in	the large school in	a new collection of	a wide variety of	a great number of
4	a big concert in	the large school in	a long time since	a special day in	a great number of	a large number of
5	a good time with	a new shop near	a special day in	a free copy of	a large number of	a wide variety of
6	a pop concert in	the small school in	a great time with	a great deal of	an important role in	an important role in
7	a pop concert on	the new shop in	a good time with	the other side of	the other side of	the other side of
8	a new house in	a long time since	the new class because	a long time since	a great variety of	a great amount of

9	a pop concert with	a new bed for	the small school in	an essential part of	a great opportunity for	the vast majority of
10	a mobile phone for	a good time with	the other side of	a great number of	the back row in	an important part of

Table 6.14 lexical breakdown of DT JJ NN IN across the top 10 lexical instances at all levels

6.5.8 Summarising development: case study 1

This case study explores one of the emerging sequences at A2, increasingly used as proficiency increases. Initial analysis has shown that learners at A1 and A2 levels rely heavily on topic to put together sequences, and in this case it is the concrete adjectives and nouns relating to the topic or task which are the building blocks for sequences. This may also lend evidence for the early slot and frame stage of the developmental sequence proposed by a usage-based theory of language learning. Existing frameworks for structural and functional classification do not adequately account for early output at the A1/A2 levels and the growth in the lexical and functional diversity of the sequence as it increases with proficiency.

Looking forward beyond B1, a dominant form and function combination for this sequence starts to emerge at B2 and dominates the most highly ranking lexical sequences by C2, namely a +adjective+noun+ of to express quantity (e.g. *a wide range of, a huge amount of, a great deal of*). Although we see a variety of candidates continuing to ‘fill’ the POS tag slots at B2 and C1, there is increasing distillation of ‘slot candidates’ so that by C2 level there is evidence, on the one hand, of an increasingly specialised function (quantity) alongside increasing fixedness and constraint on the selection and combination of lexical items (see chapter 8). At A2 and B1 we see independent paradigmatic choices at a POS item level.

6.6 Case study 2: PP MD RB VV pronoun modal adverb verb-base

The second case study explores another example of an emerging sequence, one with a modal verb pronoun modal adverb verb-base (PP MD RB VV) (e.g. *I can't think, I couldn't believe, I'll never forget, you shouldn't bring*). It is not highly ranked at A1 (#206) but it jumps to the top 30 at A2 (#29) and continues to rise in ranking at B1 (#19), after which level the ranking stabilises B2 (#13), C1 (#12), C2 (#15). As well as tracing the development of usage, this

sequence has been selected to illustrate two important methodological challenges: POS tagging and task effect.

6.6.1 Occurrences by level

Table 6.15 shows the breakdown of the raw and relative occurrences of this sequence by level and shows the proportion of occurrences that constitute the top 1000 types for each level, giving an indication of type-token ratio (apart from occurrences at A1 for which there are 495 types in total of this sequence in the A1 data). In relative terms the sequence is used three times more frequently in the A2 than the A1 data, rising in relative frequency from A2 to B1, remaining consistently frequent from B1 to C1, and dipping slightly at C2. While the relative occurrences at B2, B1 and C1 are almost identical, the range of lexical exponents increases - indicated by the total occurrences of 1000 types as a percentage of all occurrences. At A1 495 types constitute 100% of all occurrences; at A2, 1000 types constitute 80.6% of all occurrences decreasing to 71.1% at B1, 57.1% at B2, 45.6% at C1 and 39.2% at C2, where even though the relative frequencies are lower than the previous three levels, the range of types is greater. The implication in simple, formal terms is that the lexical diversity increases as proficiency levels rise. As proficiency increases, learners do more with the same.

	subcorpus size	raw occurrences	relative PMW occurrences	total occurrences	1000 types as % occurrences
				1000 types	
A1	2456971	1,412	575		
A2	5703217	8,935	1567	7201	80.6
B1	3261473	6,305	1933	4485	71.1
B2	5263979	10,471	1989	5975	57.1
C1	6711568	12,789	1906	5836	45.6
C2	7698695	13,436	1745	5262	39.2

Table 6.15 Breakdown of occurrences by level of pronoun+modal+adverb+verb-base

6.6.2 Distribution of top 50 lexical occurrences by level: task effect

The top 50 most frequent lexical realisations at A1, A2 and B1 levels were extracted. Table 6.16 shows the top 20 and the percentage of the entire occurrences for each level.

A1	%	A2	%	B1	%
<i>you couldn't come</i>	22.0	<i>I can't go</i>	7.22	<i>I couldn't believe</i>	2.59
<i>you can't come</i>	4.9	<i>I can't meet</i>	5.19	<i>I can't wait</i>	1.86
<i>you can't find</i>	4.5	<i>I can't come</i>	2.43	<i>I can't meet</i>	1.73
<i>I can't wait</i>	2.1	<i>I couldn't believe</i>	1.79	<i>I will never forget</i>	1.44
<i>I can't go</i>	1.9	<i>I can't wait</i>	1.41	<i>I can't go</i>	1.38
<i>I can't find</i>	1.2	<i>you can't go</i>	1.34	<i>I couldn't find</i>	1.08
<i>I can't see</i>	1.2	<i>I can't see</i>	1.15	<i>You won't believe</i>	1.05
<i>You can't find</i>	1.1	<i>you couldn't come</i>	1.12	<i>you can't do</i>	0.94
<i>I would also like</i>	1.1	<i>we can't meet</i>	0.92	<i>I'll never forget</i>	0.84
<i>I can't come</i>	1.0	<i>I won't go</i>	0.84	<i>I can't come</i>	0.79
<i>I can't remember</i>	1.0	<i>I can't believe</i>	0.75	<i>you can't go</i>	0.71
<i>I couldn't come</i>	1.0	<i>I couldn't see</i>	0.73	<i>You can't imagine</i>	0.68
<i>I can't do</i>	0.8	<i>I can't sleep</i>	0.67	<i>I can't believe</i>	0.59
<i>you can't go</i>	0.8	<i>you can't do</i>	0.66	<i>I couldn't see</i>	0.56
<i>you could not come</i>	0.7	<i>I couldn't find</i>	0.66	<i>I would also like</i>	0.52
<i>You shouldn't bring</i>	0.7	<i>I cannot go</i>	0.65	<i>you won't go</i>	0.52
<i>You mustn't bring</i>	0.6	<i>I can't do</i>	0.57	<i>you can also go</i>	0.51
<i>You couldn't come</i>	0.6	<i>I can't miss</i>	0.54	<i>I can't forget</i>	0.49
<i>I can't call</i>	0.6	<i>You can also go</i>	0.53	<i>You can also go</i>	0.49
<i>You can't bring</i>	0.6	<i>I couldn't go</i>	0.53	<i>I can't stay</i>	0.44

Table 6.16 lexical breakdown of PP MD RB VV across the top 20 lexical instances at A1, A2, B1

Initial observations show a dominance of one form in the A1 data, and an increased levelling out of distribution of forms from A1 to A2 to B1, illustrated by the percentage distribution figure for each lexical form. There is a strong indication of a task effect in the A1 data, with the most frequent lexical exponent *you couldn't come* making up 22% of all A1 occurrences. At A2 *you couldn't come* is the 8th most frequent lexical exponent and makes up 1.12% of the occurrences, whereas it does not feature at all in the top 50 B1 for this POS tag sequence. Further analysis of this in the Cambridge Learner Question Papers corpus results in two

occurrences of *couldn't come*, with one direct match (see (1) below) *you couldn't come* in question paper rubrics:

(1) PET (B1), 1998:

Part 3 Question 16 You recently had a birthday party, which an English-speaking friend was unable to attend. Your friend has just sent you a birthday present. Now you are writing a letter to your friend. Thank your friend for the present, tell your friend all about the party, and suggest when you could next meet. Finish the letter on your answer sheet, using about 100 words. You may use this page for any rough work. *Dear I'm really sorry **you couldn't come** to my party*

(2) KET (A2), 2004:

Question 56 Read this note from your friend, Ally. *Sorry **I couldn't come** to your birthday party. What did you do at the party? Who was there? What presents did you get? Ally* Write a note to Ally and answer her questions. Write 25-35 words. Write the note on your answer sheet

In the A1 data, however, there are no instances of *you couldn't come* resulting from exam question (1) (above). The corpus visualisation tool in Sketch Engine shows that all instances of *you couldn't come* are concentrated around one part of the corpus, the KET 2004 exam, (2) above, as illustrated in Figure 6.5.

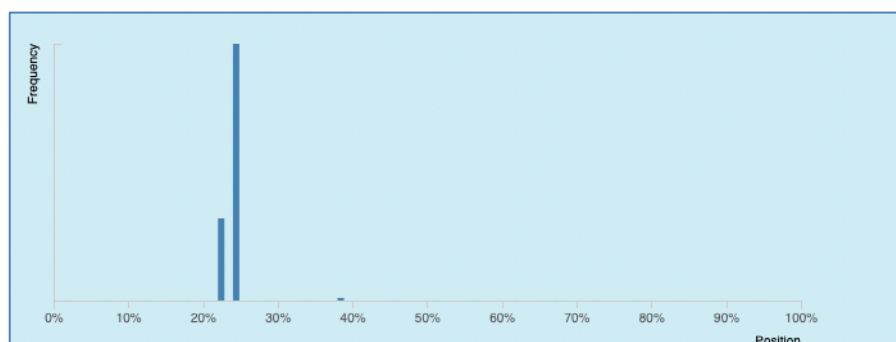


Figure 6.5 Distribution of hits of *you couldn't come* in the A1 data.

Analysis of the concordance lines (Figure 6.6) shows a tendency for the phrase *to my (birthday) party* after *you couldn't come*, a direct transformation of the rubric (2) above. It is noticeable that there is greater diversity preceding *you couldn't come* (see also Figure 6.6 for

a snapshot), ranging from *It's a pity / I'm sorry / I'm/was sad (that) / I'm sad/sorry because / it doesn't matter if / Don't worry if / No problem if*.

1	<input type="checkbox"/>	①	KET • 2004	being you next year	</s></s>	Dear ALLY	</s></s>	It 's ok about	</s></s>	you could n't come	</s></s>	to my party	</s></s>	I did many things ,
2	<input type="checkbox"/>	①	KET • 2004	Yours Sincerely	</s></s>	Dear Ally	</s></s>	It 's a pity	</s></s>	you could n't come	</s></s>	I miss you a lot	</s></s>	A
3	<input type="checkbox"/>	①	KET • 2004	See you soon	</s></s>	Hallo Ally ,	</s></s>	it ' was not very nice ,	</s></s>	you could n't come	</s></s>	to my birthday party ,	</s></s>	but we had also fun
4	<input type="checkbox"/>	①	KET • 2004	I am and Rocky	</s></s>	Dear Ally ,	</s></s>	I regret that	</s></s>	you could n't come	</s></s>	to my birthday party	</s></s>	It was so co
5	<input type="checkbox"/>	①	KET • 2004	of cloth	</s></s>	See you soon	</s></s>	Ally : No problem if	</s></s>	you could n't come	</s></s>	to my party	</s></s>	Well , I danced , ate
6	<input type="checkbox"/>	①	KET • 2004	Hello Ally : How are you ?	</s></s>	I 'm very sad because	</s></s>	you could n't come	</s></s>	to my birthday	</s></s>	My party was funny and inte		
7	<input type="checkbox"/>	①	KET • 2004	I have a new car	</s></s>	Bye !	</s></s>	It 's alright if	</s></s>	you could n't come	</s></s>	to my birthday party	</s></s>	Everybody wa
8	<input type="checkbox"/>	①	KET • 2004	See you soon	</s></s>	Dear Ally I 'm sorry that	</s></s>	you could n't come	</s></s>	to my birthday Party	</s></s>	At my party v		
9	<input type="checkbox"/>	①	KET • 2004	WITH LOVE	</s></s>	Dear Ally ,	</s></s>	it is n't bad that	</s></s>	you could n't come	</s></s>	to my party	</s></s>	We heard music and d
10	<input type="checkbox"/>	①	KET • 2004	How you are very busy	</s></s>	So it does n't matter that	</s></s>	you could n't come	</s></s>	to my birthday party	</s></s>	First , I play		
11	<input type="checkbox"/>	①	KET • 2004	At the	</s></s>	My party was very funny	</s></s>	What	</s></s>	you could n't come	</s></s>	DEAR DAISY ,	</s></s>	YESTERD
12	<input type="checkbox"/>	①	KET • 2004	At my home	</s></s>	From	</s></s>	It 's ok because	</s></s>	you could n't come	</s></s>	to my birthday party	</s></s>	I cut my b
13	<input type="checkbox"/>	①	KET • 2004	With love	</s></s>	Dear Ally	</s></s>	That 's a pity , that	</s></s>	you could n't come	</s></s>	but I danced and drank very much	</s></s>	
14	<input type="checkbox"/>	①	KET • 2004	KISSES	</s></s>	Dear Ally ,	</s></s>	That was a pity	</s></s>	you could n't come	</s></s>	to my birthday party !	</s></s>	I danced a

Figure 6.6 Sample of the concordance lines with *you couldn't come* in the A1 data.

Quite clearly and inevitably there is some element of task effect at play here. However there are some important points to consider:

- While 22% of all A1 occurrences are *you couldn't come*, there are still 78% other forms to account for. Figure 6.7 shows the same distribution as in Figure 6.5, circled in red, but within the context of all other lexical exponents. The remaining 78% range in form and distribution, e.g. *you can't come* (4.9%), *you can't find* (4.5%), *I can't believe* (0.4%), *you shouldn't bring* (0.4%) (See Table 6.16).
- *Couldn't come to my birthday party* can be found in the question rubric for one question in KET A2 exam but the A1 data exhibits a variety of usage contexts, suggesting an ability to extract this lexical string from the input and modify it appropriately. This is not unlike the type of holistic patterning found in first language acquisition and may be indicative of the effect of recency.

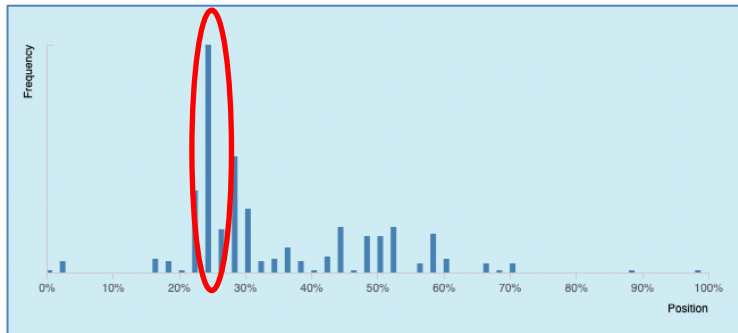


Figure 6.7 Distribution of hits of all PP MD RB VV occurrences in the A1 data.

In summary, looking the A1 data, on the one hand there is evidence of a kind of holistic extraction of a formula ‘*couldn’t come to my birthday party*’, while also demonstration of some generalised abstraction of the form (e.g. *you can’t come*) as well as an ability to vary the contexts of use, prompted by the one-word ‘*Sorry*’ in the rubric (e.g. *it’s a pity that, I’m sad that*).

6.5.3 A2 and B1: Task effect or no task effect?

A2 data

If we look at the A2 distribution of the sequences PP MD RB VV using the Sketch Engine visualisation tool, results show a spike, as illustrated in Figure 6.8.

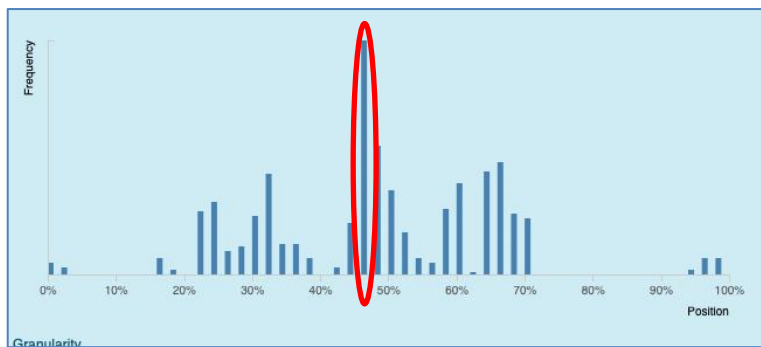


Figure 6.8 Distribution of hits of all PP MD RB VV occurrences in the A2 data.

The expectation here might be that, like the A1 level spike, there is a high frequency of one form. Concordance lines show that while one form (*I can’t meet*) is dominant (See Table 6.17 for the frequency breakdown of this spike) the data comes from 4 different tasks, over two exam levels, from ten different years. So even though one task might be accounting for the lion’s share of this form, there is evidence of an ability to select a variety of forms for the same structure, e.g. *I/we can’t meet/go/come*. A move from the holistic patterning seen at A1

to a filling of independent slots. Some more fixed patterning *I couldn't/can't believe (it)* also begins to emerge at A1.

	Word	Frequency ↓	Relative ?	% Of conc.
1	<input type="checkbox"/> I ca n't meet	344	5.71	25.79 %
2	<input type="checkbox"/> I ca n't go	149	2.47	11.17 %
3	<input type="checkbox"/> I ca n't come	65	1.08	4.87 %
4	<input type="checkbox"/> we ca n't meet	47	0.78	3.52 %
5	<input type="checkbox"/> I ca n't see	34	0.56	2.55 %
6	<input type="checkbox"/> I can not meet	23	0.38	1.72 %
7	<input type="checkbox"/> she could n't find	20	0.33	1.50 %
8	<input type="checkbox"/> you ca n't do	15	0.25	1.12 %
9	<input type="checkbox"/> you ca n't go	15	0.25	1.12 %
10	<input type="checkbox"/> I can not go	14	0.23	1.05 %
11	<input type="checkbox"/> I could n't meet	11	0.18	0.82 %

Table 6.17 Lexical breakdown of PP MD RB VV across the top 11 lexical instances at A2

B1 distribution

At B1, the distribution of the PP MD RB VV sequence also has a strong spike (Figure 6.9) Unlike the A2 data, this all comes from one exam, PET, and one year, 2008, from three questions in the same exam, 68% comes from one question, examples of which can be seen in Figure 6.10.

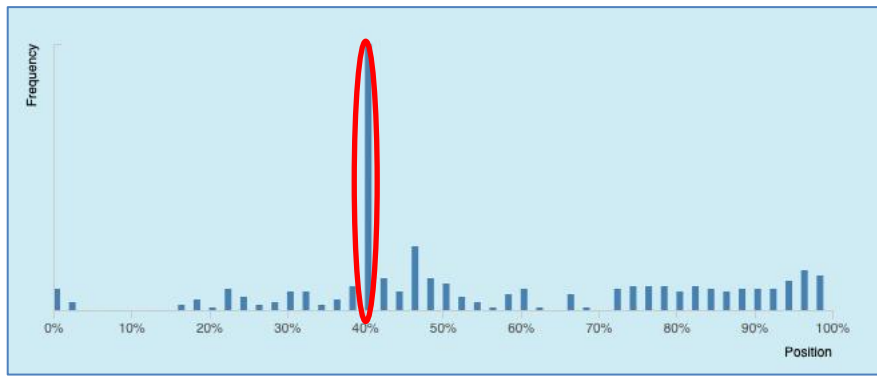


Figure 6.9 Distribution of hits of all PP MD RB VV occurrences in the B1 data.

1	<input type="checkbox"/> ① PET • 2008 day a TV company came to our school to make a film .	<input type="checkbox"/> I could n't believe it .	<input type="checkbox"/> They chose our school because they th
2	<input type="checkbox"/> ① PET • 2008 , you 're not arguing with them .	<input type="checkbox"/> I ca n't decide	<input type="checkbox"/> for you . It 's not my holiday .
3	<input type="checkbox"/> ① PET • 2008 ing fine .	<input type="checkbox"/> I will definitely go	<input type="checkbox"/> on holiday with my parents . There are som
4	<input type="checkbox"/> ① PET • 2008 o with your friends , because this is your last year before university and	<input type="checkbox"/> you may not see	<input type="checkbox"/> them again . I believe that your parents wi
5	<input type="checkbox"/> ① PET • 2008 ne ?	<input type="checkbox"/> I can really understand	<input type="checkbox"/> you because I have the same problem However , I kr
6	<input type="checkbox"/> ① PET • 2008 somewhere with your family , you have to stay next to your parents and	<input type="checkbox"/> you ca n't do	<input type="checkbox"/> anything ; but if you go on holiday with your friends you
7	<input type="checkbox"/> ① PET • 2008 nake this decision .	<input type="checkbox"/> you would rather go	<input type="checkbox"/> on vacation with your friends . Maybe you ca
8	<input type="checkbox"/> ① PET • 2008 very boring for you .	<input type="checkbox"/> you can alone decide	<input type="checkbox"/> what you want to do . I 'm going to fly
9	<input type="checkbox"/> ① PET • 2008 y and I swear you I 've never got bored .	<input type="checkbox"/> you ca n't choose	<input type="checkbox"/> , you could go on holiday with your family in July , and wi
10	<input type="checkbox"/> ① PET • 2008 / with your friends , just tell your parents your idea .	<input type="checkbox"/> they wo n't blame	<input type="checkbox"/> you for this . It 's quite hard to choose b
11	<input type="checkbox"/> ① PET • 2008 /our problem .	<input type="checkbox"/> they will never know	<input type="checkbox"/> what you really want . Hello Alice !
12	<input type="checkbox"/> ① PET • 2008 have a good time more with them than with your parents .	<input type="checkbox"/> I ca n't say	<input type="checkbox"/> to you anything enough . You have to decid

Figure 6.10 Concordance lines of B1 occurrences of PP MD RB VV from PET 2008

What is noticeable about the frequency of the different lexical exponents from this spike (Table 6.18) is that (1) they are distributed more evenly (e.g. *you can't do* makes up 3.62% of the occurrences, *you won't go* is 2.79%), (2) there is more variability in the selection of items for each POS tag slot and adverbs such as *also* and *rather* begin appear alongside a wider lexical repertoire in the final VV slot (e.g. *you won't enjoy*, *they won't let*, *I'd rather go*, *you shouldn't eat*). The semi-fixed *I couldn't/can't believe (it)* rise up the frequency list. Further down the frequency list other semi-fixed sequences appear (e.g. *I'll never forget*, *I can't remember*).

	Word	Frequency ↓	Relative ?	% Of conc.		Word	Frequency ↓	Relative ?	% Of conc.
1	<input type="checkbox"/> you ca n't do	39	0.65	3.62 %	41	<input type="checkbox"/> I will never forget	30	0.50	0.34 %
2	<input type="checkbox"/> you wo n't go	30	0.50	2.79 %	42	<input type="checkbox"/> I would really like	30	0.50	0.34 %
3	<input type="checkbox"/> you ca n't go	29	0.48	2.69 %	43	<input type="checkbox"/> you ca n't find	30	0.50	0.34 %
4	<input type="checkbox"/> you can also go	16	0.27	1.49 %	44	<input type="checkbox"/> I 'll always live	29	0.48	0.32 %
5	<input type="checkbox"/> You can also go	15	0.25	1.39 %	45	<input type="checkbox"/> I ca n't tell	28	0.46	0.31 %
6	<input type="checkbox"/> I would n't go	14	0.23	1.30 %	46	<input type="checkbox"/> I could n't answer	28	0.46	0.31 %
7	<input type="checkbox"/> you wo n't enjoy	11	0.18	1.02 %	47	<input type="checkbox"/> she could n't find	28	0.46	0.31 %
8	<input type="checkbox"/> I could n't answer	10	0.17	0.93 %	48	<input type="checkbox"/> I ca n't live	28	0.46	0.31 %
9	<input type="checkbox"/> you 'd rather go	10	0.17	0.93 %	49	<input type="checkbox"/> I ca n't attend	26	0.43	0.29 %
10	<input type="checkbox"/> they wo n't let	9	0.15	0.84 %	50	<input type="checkbox"/> you ca n't come	26	0.43	0.29 %
11	<input type="checkbox"/> I 'd rather go	9	0.15	0.84 %	51	<input type="checkbox"/> I will not go	26	0.43	0.29 %
12	<input type="checkbox"/> you could n't do	9	0.15	0.84 %	52	<input type="checkbox"/> you ca n't imagine	25	0.41	0.28 %
13	<input type="checkbox"/> you wo n't get	9	0.15	0.84 %	53	<input type="checkbox"/> You ca n't imagine	24	0.40	0.27 %
14	<input type="checkbox"/> you could also go	8	0.13	0.74 %	54	<input type="checkbox"/> you should n't disappoint	24	0.40	0.27 %
15	<input type="checkbox"/> I would n't know	8	0.13	0.74 %	55	<input type="checkbox"/> you wo n't disappoint	24	0.40	0.27 %
16	<input type="checkbox"/> you wo n't do	7	0.12	0.65 %	56	<input type="checkbox"/> I could n't say	22	0.37	0.25 %
17	<input type="checkbox"/> I ca n't choose	7	0.12	0.65 %	57	<input type="checkbox"/> You should n't eat	21	0.35	0.24 %
18	<input type="checkbox"/> You can also say	7	0.12	0.65 %	58	<input type="checkbox"/> I ca n't write	20	0.33	0.22 %
19	<input type="checkbox"/> I ca n't wait	7	0.12	0.65 %	59	<input type="checkbox"/> you can not go	20	0.33	0.22 %
20	<input type="checkbox"/> you can also ask	6	0.10	0.56 %	60	<input type="checkbox"/> you should n't go	20	0.33	0.22 %

Table 6.18 lexical breakdown of PP MD RB VV across the top 20 and top 40-60 lexical instances at B1

6.7 A1 to B1: Setting out

In the first part of the chapter (6.1 to 6.3) I first explored if development is observable through the frequency and distribution of POS sequences across proficiency levels looking back from A2 to A1 and forward to B1, in an attempt to address RQ1. Overall there was evidence of increasing convergence of POS tag sequence usage as proficiency increased. Noun phrase usage increased between A1 and A2, and verb phrase usage increased at B1. I then took a case study approach (6.4 to 6.6) to address how POS sequences developed across proficiency levels A1, A2, B1 (RQ2) and noted an increase in lexical and functional repertoire. Learners at A1 and A2 levels showed a reliance on topic-driven concrete adjectives and nouns as the building blocks to put together sequences. There was evidence for early slot and frame development, as well as some indication of the emergence of pioneer or path-breaking type form-function mappings. I explored if existing frameworks for classification of language patterning account for a description of development (RQ3) and applied a lexical bundle approach, underlining the movement from topic-driven sequences to more abstract frames for reference purposes.

Chapter 6 considers the next stage of the learner journey, moving through the proficiency levels to B1 and B2. As already observed in the global view of development in chapter 5, there is evidence pointing to the B1 level as a turning point in POS tag sequence usage.

Chapter 7 On the road, gathering pace from B1 to B2

According to the Common European Framework (CEFR) categorisation, the B level learner has moved from a ‘basic user’ of language to an ‘independent user’ (see also Chapter 6). In Table 7.1 the B1 and B2 levels are described in very general terms on the CEFR global scale.

B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics, which are familiar, or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.

<https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale>

Table 7.1 Global scale descriptors for B1 and B2 as defined by the Council of Europe

According to the CEFR self-assessment guidelines for writing

(<https://www.coe.int/en/web/portfolio/self-assessment-grid>) learners at B1 level are expected to demonstrate that they “can write simple connected text on topics which are familiar or of personal interest’ and “write personal letters describing experiences and impressions.”

B2 level users are expected to “write clear, detailed text on a wide range of subjects related to my interests ... write an essay or report, passing on information or giving reasons in support of or against a particular point of view ... write letters highlighting the personal significance of events and experiences”. In simplistic terms they are expected to move from simple topic-related descriptions of personal experience to more complex evaluative descriptions.

Attaining a B1 level is increasingly a universal job requirement globally and is often a target for learning. As already discussed in Chapters 5 and 6, the B level is seen as a transitional phase. The CEFR terms ‘Waystage’ and ‘Threshold’, used to describe the B1 and B2 levels respectively implying a point at which the language experience enters a new phase. This chapter seeks to define and describe change at the B levels. In the first part of the chapter (7.1 to 7.3) I first explore if development is observable through the frequency and distribution of the highest-ranking POS sequences across proficiency levels B1 and B2 (RQ1). I then take a case study approach in sections 7.4 to 7.7. Aligning with RQ2, these will address how POS sequences develop across proficiency levels. The case studies will also attend to RQ3 and explore if existing frameworks for classification of language patterning account for a description of development.

7.1 Focusing in: overall distribution B1 and B2

I begin with initial observations about the POS 4-gram distribution across B1 and B2 levels. All 4-gram POS tag sequences were extracted from the B1 and B2 data. The total number of POS 4-gram sequence occurrences and total 4-gram POS tag sequence types per level are shown in Table 7.2:

	B1	B2
Total 4-gram raw occurrences: all types	3183197	5190020
Total 4-gram types	164828	230464

Table 7.2 Occurrences of 4-gram POS tag sequences across levels B1 and B2

All sequences were ranked and the rankings of the top 50 types from each level selected for further analysis and comparison. The total number of 4-gram POS tag sequence occurrences in the top 50 types can be seen in Table 7.3. These top 50 types constitute 0.03% and 0.02% of types of POS 4-grams in the B1 and B2 data. However, they account for 9.75% and 9.04% of all POS 4-gram token occurrences at each level (Table 7.3). In short, even though they represent a small fraction of a percentage of all types, they make up almost 10% of all occurrences at each level (see also the overall picture of distribution in Chapter 5 Figure 5.1), because they are the highest ranking and therefore the most frequently used.

	B1	B2
Total 4-gram raw occurrences: top 50	310365	469106
50 types as % of all types	0.03	0.02
Total occurrences in top 50 as % of all occurrences	9.75	9.04

Table 7.3 Occurrences and percentage distribution of the top 50 types across B1 and B2

Next, in sections 7.1.1 and 7.1.2, I examine changes in the distribution of sequence ranking of the top 50 sequences across B1 and B2 levels and their adjacent levels.

7.1.1 Overall distribution: top 50 B1 sequences

Here I focus on change from the perspective of the top 50 B1 sequences. A snapshot of this change can be seen in Figure 7.1, with specific focus within the red box. The colour coding gives a visual overview of the convergence in ranking between levels, from dark green (highly convergent) to pink (highly divergent) (see key).

As already observed in chapter 6, 90% of the top 50 B1 sequences are also found within a rank of +/-30 at A2 (all three green sections in the B1:A2 column), with 40% ranked closely within a range of +/-5 (dark green section only).

When the B1 sequences are compared with their ranking at B2, there is less convergence: 80% of the top 50 B1 sequences are found within a rank difference of +/- 30 at B2 (all three green sections in the B1:B2 bar column), and those closely ranked (i.e. core or highly convergent) (within a range of +/-5) decrease to 34% (dark green section only).

There is a visible decrease in those top 50 B1 sequences which remain core at both B1 and B2 (see Figure 7.1, dark green). This divergence in ranking continues to increase as proficiency increases beyond the B2 level. 66% of the top B1 sequences are also found within a rank of +/-30 at C1, and 52% at C2 (represented by the three green bands in each bar column). Sequences that are core and high ranking at B1 are less and less likely to be core as proficiency increases. However a core of high ranking sequences remains. By C2, 24% of the top 50 B1 sequences also ranked within +/-5 of the B1 rankings. This represents a picture of both divergence and convergence. On the one hand there is stable usage of a quarter of the B1 sequences from A2 to C2 alongside a shift in ranking of the remainder of the sequences. This points to evidence of a statistical restructuring of frequency and distribution of these sequences as proficiency increases.

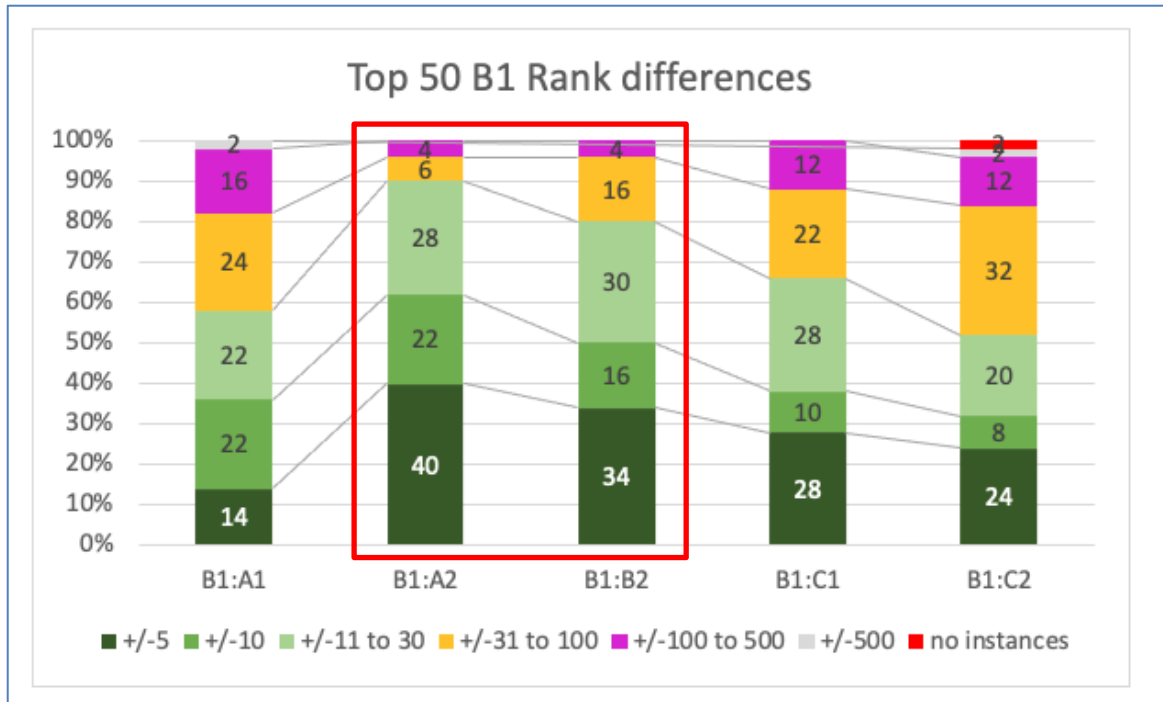


Figure 7.1 Percentage convergence/divergence of the top 50 sequences at B1 with their rankings at all other levels

Through this 4 POS tag sequence lens, B1 writing looks closer in usage to the adjacent lower proficiency level A2 writing than to B2 writing. Exam level effect was checked and ruled out as an influence on these results: of the 3.2 million+ words in the B1 performance data, 59.05% comes from the B1 level exam (PET), 38.27% from a B2 level exam (FCE) and 2.68% from the A2 level exam (KET) (see Chapter 4).

7.1.2 Overall distribution: top 50 B2 sequences

A picture of accumulating stabilisation and convergence emerges when looking at the top 50 sequences at B2 and comparing their rankings at adjacent levels (Figure 7.2). Of the top 50 B2 sequences, 94% are found within a rank difference of +/- 30 at C1, rising from 88% convergence between B2 and B1 (indicated by the three green bands of shading in each bar column). 34% of these top 50 are found to be consistently and closely ranked (i.e. those with a rank difference of +/-5) across four levels: A2, B1, B2 and C1, dropping slightly at C2 (32%) (indicated by the dark green bands). As proficiency increases so does the similarity of the distribution of core POS tag sequences. When compared to the B1 profile, there is increasing consistency between B2 sequences and adjacent lower and higher proficiency levels.

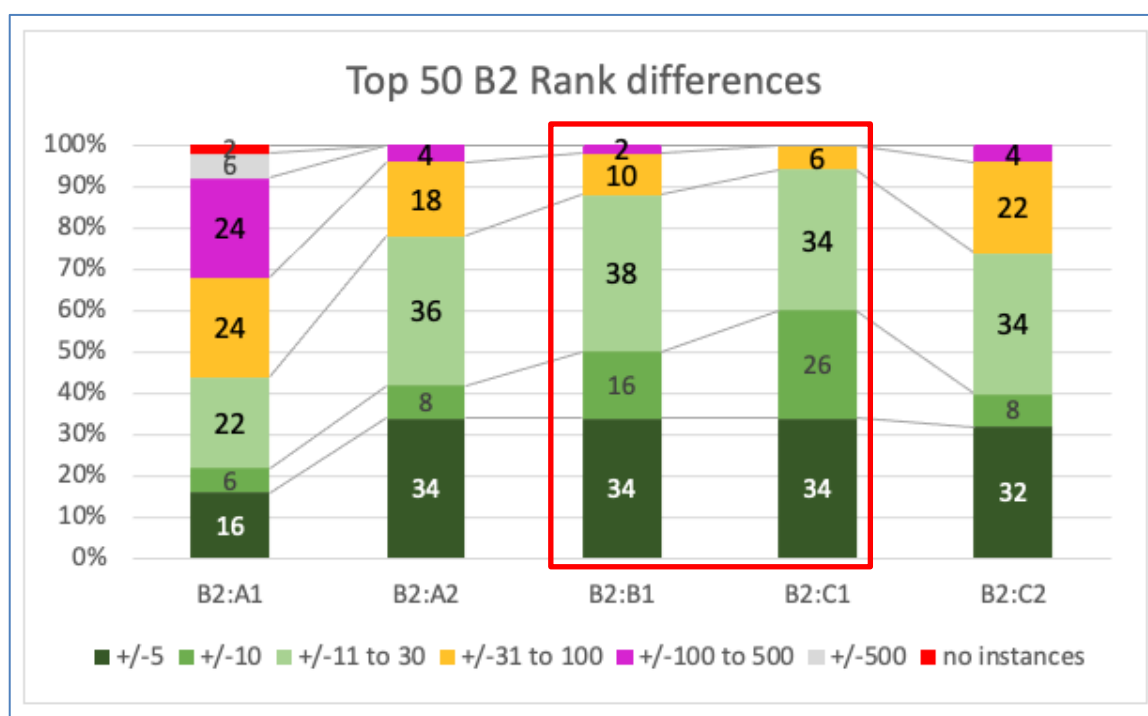


Figure 7.2 Percentage convergence of the top 50 sequences at B2 with their rankings at all other levels

In overall terms, initial observations from both B1 and B2 perspectives indicate a potential leap in development spanning the B level, in relation to sequence distribution: B1 writing has more in common with the adjacent lower level (A2) and the B2 writing has more in common with the adjacent higher level (C1). This picture appears to show adjustment of frequency of use up to B1 and between B1 and B2 which stabilises at B2 and beyond.

In the following two sections, 7.2 and 7.3, I look at the constituents in the sequences at B1 and B2 and identify changes in their usage. This provides a starting point for further exploration of lexical and functional characteristics.

7.2 B1 sequences

As described above, the rankings of the top 50 sequences at B1 were compared with their ranks at A2 and B2 using a simple rank difference calculation to calculate their relative rank variance (as described in chapter 4). Sequences with punctuation were then removed (see also Chapters 4 and 5), leaving 31 sequences (Table 7.4). Of the 31 sequences under investigation, 17 contain verbs. The remaining 14 contain part or whole noun phrases. Six of these sequences (marked in blue font) are new to the top 50 at B1 and do not occur in the top

50 at the adjacent lower level (A2). Four (marked in red font) do not occur in the top 50 of the adjacent higher level (B2).

COLOUR KEY	
rank variance of +/-	
5	
10	
11 to 30	
31 to 100	
101 to 500	
501+	
#VALUE! = not found	

B1 rank	POS tag sequences and examples	B1- A2	B1- B2
1	noun prep det noun NN IN DT NN <i>centre of the town</i>	0	0
3	prep det adj noun IN DT JJ NN <i>on the other hand</i>	1	1
4	proper-noun proper-noun proper-noun proper-noun NP NP NP NP (this generates any text in capitals)	-9	-8
5	prep det noun prep IN DT NN IN <i>in the centre of</i>	1	1
7	det noun prep det DT NN IN DT <i>the end of the</i>	-9	2
9	det adj noun prep DT JJ NN IN <i>a great time with</i>	0	3
10	to-inf verb-base det noun TO VV DT NN <i>to make a film</i>	-18	1
12	pronoun modal verb-base prep PP MD VV IN <i>you should go to</i>	6	-24
13	noun prep poss-pronoun noun. NN IN PPZ NN <i>holiday with your family</i>	2	-1
18	pronoun modal verb-base pronoun PP MD VV PP <i>I must tell you</i>	0	-7
19	pronoun modal adverb verb-base PP MD RB VV <i>I will never forget</i>	-10	6
22	pronoun pres-simple to-inf verb-base PP VVP TO VV <i>I want to tell</i>	7	-12

	verb-base det noun prep		
25	VV DT NN IN <i>do a lot of</i>	-11	5
27	pronoun past-simple det noun PP VVD DT NN <i>I opened the door</i>	-19	-33
28	pronoun modal verb-base det PP MD VV DT <i>you can see the</i>	4	1
29	verb-base prep det noun VV IN DT NN <i>apply for the job</i>	15	7
31	pronoun pres-simple pronoun modal PP VVP PP MD <i>I think you should</i>	-2	-35
32	det noun prep noun DT NN IN NN <i>a lot of money</i>	1	14
34	pronoun pres-simple-have to-inf verb-base PP VHP TO VV <i>I have to get</i>	-6	-38
35	det noun prep possessive pronoun DT NN IN PPZ <i>the rest of my</i>	-9	-7
36	pronoun past-simpleV to-inf verb-base PP VVD TO VV <i>I decided to go</i>	-47	-10
38	noun prep poss-pronoun plural-noun NN IN PPZ NNS <i>holiday with your friends</i>	-346	-197
39	prep det noun pronoun IN DT NN PP <i>In the morning I</i>	-23	-2
40	prep det noun conjunction IN DT NN CC <i>on the television and</i>	-2	12
41	pronoun modal verb-base to-inf PP MD VV TO <i>I would like to</i>	-19	25
42	modal verb-base to-inf verb-base MD VV TO VV <i>would like to know</i>	-19	27
43	prep det noun IN DT NN <i>in the city centre</i>	18	13
45	to-inf verb-base prep det TO VV IN DT <i>to go to the</i>	15	14
47	adj noun prep det	10	24

	JJ NN IN DT <i>other side of the</i>		
48	pronoun pres-simple-be ing-form prep. PP VVP VVG IN <i>I'm working in</i>	-38	-40
50	pronoun pres-simple-be ing-form to-inf PP VVP VVG TO <i>I'm going to</i>	16	-73

Table 7.4 Top 50 4-gram POS sequences at B1, and their rank differences at A2 and B2

As before, the rank difference figures and colours show degrees of difference, and give an indication of the shift in distribution across the levels. The colour coding gives a visual overview of the convergence in ranking between levels, from dark green (highly convergent) to pink (highly divergent) (see Colour key above). Negative rank difference figures indicate a lower ranking at the other levels, and positive figures indicate a higher ranking. For example item #1 at B1 (noun+preposition+determiner+noun e.g. *centre of the town*) is also ranked at #1 in the A2 and B2 data (with a rank variance of 0), whereas item #10 (to-infinitive+verb-base+determiner+noun e.g. *to make a film*) is ranked at #28 in the A2 (with a rank variance of -18) and #9 in the B2 data (with a rank variance of +1). This drop in ranking at A2 indicates that this sequence is less used in the A2 repertoire than in the B1 and B2 repertoires, where it is consistently used.

Overall results described in Chapter 5 pointed to three types of sequences: (1) core sequences (2) emerging sequences and (3) decreasing sequences (see section 5.3). In the following sections I examine how the B1 and B2 data are also characterised by these sequence types, before exploring examples of their lexical and functional characteristics.

7.2.1 Core sequences at B1

There are 10 core sequences that are highly convergent in ranking (within +/-5) at both B1 and B2 (Table 7.5). Six of these also converge closely in rank at A2 (#1, 3, 5, 9, 13, 28). Three other sequences (#7, 10, 25), which were in the top 50 in the A2 data, have become more highly ranking at both B1 and B2 and therefore more used in the B1 and B2 repertoire than in the A2. At rank #39, we see IN DT NN PP (preposition+determiner+noun+pronoun e.g. *In the morning I*), which is a new sequence in the top 50 at B1, not found in the top 50 at A2 (rank difference -23), and remains consistently ranked at B2.

B1 rank	POS tag sequences and <i>examples</i>	Rank difference	
		A2	B2
1	noun prep det noun NN IN DT NN <i>centre of the town</i>	0	0
3	prep det adj noun IN DT JJ NN <i>on the other hand</i>	1	1
5	prep det noun prep IN DT NN IN <i>in the centre of</i>	1	1
7	det noun prep det DT NN IN DT <i>the end of the</i>	-9	2
9	det adj noun prep DT JJ NN IN <i>a great time with</i>	0	3
10	to-inf verb-base det noun TO VV DT NN <i>to make a film</i>	-18	1
13	noun prep poss-pronoun noun NN IN PPZ NN <i>holiday with your family</i>	2	-1
25	verb-base det noun prep VV DT NN IN <i>do a lot of</i>	-11	5
28	pronoun modal verb-base determiner PP MD VV DT <i>you can see the</i>	4	1
39	prep det noun pronoun IN DT NN PP <i>In the morning I</i>	-23	-2

Table 7.5 Core sequences: B1 sequences which are highly convergent in ranking at both B1 and B2.

When compared with the core sequences seen at A2 (Section 6.2.1), sequences containing noun phrases continue to dominate, with an increase in noun phrases containing adjectives (#3, 9). Alongside this there is a decrease in sequences containing modal verbs, and an increase in those containing verbs (#10, 25), though note here the continued absence of verb forms marked for tense in these highly convergent sequences.

7.2.2 Emerging sequences at B1

There are nine emerging sequences in the B1 top 50, those which rank higher at B2 than B1, and become increasingly more important for B2 learners (Table 7.6):

B1 rank	POS tag sequences and B1 examples	B1-B2
19	pronoun modal adverb verb-base PP MD RB VV <i>I will never forget</i>	6
29	verb-base prep det noun VV IN DT NN <i>apply for the job</i>	7
32	det noun prep noun DT NN IN NN <i>a lot of money</i>	14
40	prep det noun conjunction IN DT NN CC <i>on the television and</i>	12
41	pronoun modal verb-base to-inf PP MD VV TO <i>I would like to</i>	25
42	modal verb-base to-inf verb-base MD VV TO VV <i>would like to know</i>	27
43	prep det noun IN DT NN <i>in the city centre</i>	13
45	to-inf verb-base prep det TO VV IN DT <i>to go to the</i>	14
47	adj noun prep det JJ NN IN DT <i>other side of the</i>	24

Table 7.6 Emerging sequences: B1 sequences which are higher ranked at B2 than B1 (with rank difference).

Noticeable here is the increase in the range of sequences containing base verb forms, particularly with modal verb forms (#19, 41, 42), and in sequences with to-infinitive structures (#41, 42, 45) as well as a rise in the rank for extended noun phrase sequences (#32, 40, 43, 47). As discussed in Chapter 4, some of these may be extensions of the same sequence, e.g. #41 and 42 (pronoun+modal+verb-base+to-inf, e.g. *I would like to*) and (modal+verb-base+to-inf+verb-base, e.g. *would like to know*) may be part of the same 5-gram

sequence. These two, ranking at #16 and #15 at B2 are new to the top 50 in B1 (#41 and 42), having been ranked at #60 and #61 in the A2 data.

7.2.3 Decreasing sequences at B1

There are 12 decreasing sequences in the top 50 B1 sequences (i.e. those that rank lower at B2 than at B1) (Table 7.7). At the top of the table are those that are the least used at B2 in relation to other sequences, and less relevant in the B2 repertoire, (to varying points of difference, shown by the rank difference figure). Those in red are not carried forward into the top 50 at B2.

B1 rank	POS tag sequences and B1 examples	B1-B2
38	noun prep poss-pronoun plural-noun NN IN PPZ NNS <i>holiday with your friends</i>	-197
50	pronoun pres-simple-be ing-form to-inf PP VVP VVG TO <i>I'm going to</i>	-73
48	pronoun pres-simple-be ing-form prep PP VVP VVG IN <i>I'm working in</i>	-40
34	pronoun pres-simple-have to-inf verb-base PP VHP TO VV <i>I have to get</i>	-38
31	pronoun pres-simple pronoun modal PP VVP PP MD <i>I think you should</i>	-35
27	pronoun past-simple det noun PP VVD DT NN <i>I opened the door</i>	-33
12	pronoun modal verb-base prep PP MD VV IN <i>you should go to</i>	-24
22	pronoun pres-simple to-inf verb-base PP VVP TO VV <i>I want to tell</i>	-12
36	pronoun past-simpleV to-inf verb-base PP VVD TO VV <i>I decided to go</i>	-10
4	proper-noun proper-noun proper-noun proper-noun NP NP NP NP (this generates any text in capitals)	-8
35	det noun prep possessive pronoun DT NN IN PPZ <i>the rest of my</i>	-7

18	pronoun modal verb-base pronoun PP MD VV PP <u>I must tell you</u>	-7
----	---	----

Table 7.7 B1 sequences decreasing in ranking at B2 (with rank difference).

There is a prevalence of sequences with pronoun plus verbs *be* and *have* (#50, 48, 34) as well as four sequences with verbs marked for tense, two with present simple verbs (#31, 22) and two with past simple verbs (#27, 36). These sequences marked for tense appear to peak in ranking at B1. Also of note are two modal verb structures #12 pronoun+modal+verb-base+preposition (e.g. *you should go to*) and #18 pronoun+modal+ verb-base+pronoun which rank highly at the A levels, and continue to drop in rank at B2, and the noun phrase structure #38 noun_preposition+possessive-pronoun+plural-noun which drops in rank dramatically at B2.

From this initial analysis, overall characteristics of the B1 sequences include:

- an increase in noun phrases containing adjectives
- an increase in core sequences containing non-finite verb forms
- a decrease in sequences containing modal verbs
- a peak in the sequences containing pronoun + verbs marked for tense

7.3 B2 sequences

As described in 7.2 above, and in previous chapters, the rankings of the top 50 sequences at B2 were compared with their ranks at B1 and C1 and their relative rank variance calculated. Three sequence types, core, emerging and decreasing, continue to be observable in the B2 data.

Sequences with punctuation were removed (in line with the rationale outlined in previous chapters), leaving 33 sequences (Table 7.8). Of these 33, 16 contain verbs. The remainder contain part or whole noun phrases. Seven of these sequences (marked in blue) are new to the top 50 at B2, not occurring in the top 50 at the adjacent lower level (B1). Of the remaining 25, two (marked in red font) do not occur in the top 50 of the adjacent higher level (C1).

COLOUR KEY	
rank variance of +/-	
5	
10	
11 to 30	
31 to 100	
101 to 500	
501+	
#VALUE! = not found	

B2 rank	POS tag sequences and B2 examples	B2-B1	B2-C1
1	noun prep det noun NN IN DT NN <i>centre of the town</i>	0	0
2	prep det adj noun IN DT JJ NN <i>On the other hand</i>	-1	0
4	prep det noun prep IN DT NN IN <i>at the end of</i>	-1	0
5	det noun prep det DT NN IN DT <i>The aim of this</i>	-2	2
6	det adj noun prep DT JJ NN IN <i>a wide range of</i>	-3	1
9	to-inf det noun TO VV DT NN <i>to find a job</i>	-1	0
12	proper-noun proper-noun proper-noun proper-noun NP NP NP NP (this generates any text in capitals)	8	-6
13	pronoun modal adverb verb-base PP MD RB VV <i>I couldn't believe</i>	-6	1
14	noun prep poss-pronoun noun NN IN PPZ NN <i>rest of my holiday</i>	1	-1
15	modal verb-base to-inf verb-base MD VV TO VV <i>would like to know</i>	-27	-13
16	pronoun modal verb-base to-inf PP MD VV TO <i>I would like to</i>	-25	-17

18	det noun prep noun DT NN IN NN <i>this kind of job</i>	-14	8
20	verb-base det noun prep VV DT NN IN <i>see a lot of</i>	-5	4
22	verb-base prep det noun VV IN DT NN <i>go to the cinema</i>	-7	-8
23	adj noun prep det JJ NN IN DT <i>new shop in the</i>	-24	12
25	pronoun modal verb-base pronoun PP MD VV PP <i>you could send me</i>	7	-25
26	plural-noun prep det noun NNS IN DT NN <i>animals in a zoo</i>	-32	13
27	pronoun modal verb-base determiner PP MD VV DT <i>you can see the</i>	-1	-8
28	prep det noun conjunction IN DT NN CC <i>in the morning and</i>	-12	6
30	prep det noun IN DT NN <i>to the city centre</i>	-13	10
31	to-inf verb-base prep det TO VV IN DT <i>to go to the</i>	-14	-1
33	noun prep det adj NN IN DT JJ <i>house with the yellow</i>	-34	19
34	pronoun pres-simple to-inf verb-base PP VVP TO VV <i>I want to thank</i>	12	-29
36	pronoun modal verb-base prep PP MD VV IN <i>you can go to</i>	24	-24
39	det noun prep plural-noun DT NN IN NNS <i>a lot of people</i>	-16	8
41	prep det noun pronoun IN DT NN PP <i>at that moment he</i>	2	4
42	det noun prep possessive pronoun DT NN IN PPZ <i>the rest of my</i>	7	3
43	det noun prep to-inf verb-base -67	-67	14

	DT NN TO VV <i>the opportunity to see</i>		
45	adverb det noun prep RB IN DT NN <i>all over the world</i>	-57	11
46	pronoun past-simpleV to-inf verb-base PP VVD TO VV <i>he decided to go</i>	10	-36
48	pres-simple-be det adj noun VVZ DT JJ NN <i>is a good idea</i>	-25	1
49	modal verb-base det noun MD VV DT NN <i>can learn a lot</i>	-20	0
50	ed-form prep det noun VVN IN DT NN <i>returned from a trip</i>	-161	26

Table 7.8 Top 50 4-gram POS sequences at B2, and their rank differences at B1 and C1

Overall these sequences illustrate that a core body of sequences is growing and a picture of greater convergence as proficiency increases is emerging.

7.3.1 Core sequences at B2

There are 14 core sequences that are highly convergent in ranking (within +/-5) in the B2 and C1 data (Table 7.9). Nine of these also converge closely in rank with the B1 data (#1, 2, 4, 5, 6, 9, 14, 20, 41). Two other sequences (#13, 31), which were in the top 50 in the B1 data, have become more highly and closely ranked at both B2 and C1 and therefore more used in the B2 and C1 data than in the B1. Six of these core sequences contain verbs and two are new to the top 50 of B2; #48 (pres-simple-be+determiner+adj+noun) and #49 (modal+verb-base+determiner+noun) are not found in the top 50 at B1 (rank difference -25 and -20), and remain consistently highly ranked at C1.

B2 rank	POS sequences and <i>examples</i>	B2-B1	B2-C1
1	noun prep det noun NN IN DT NN <i>centre of the town</i>	0	0
2	prep det adj noun IN DT JJ NN <i>On the other hand</i>	-1	0
4	prep det noun prep IN DT NN IN <i>at the end of</i>	-1	0

5	det noun prep det DT NN IN DT <i>The aim of this</i>	-2	2
6	det adj noun prep DT JJ NN IN <i>a wide range of</i>	-3	1
9	to-inf verb-base det noun TO VV DT NN <i>to find a job</i>	-1	0
13	pronoun modal adverb verb-base PP MD RB VV <i>I couldn't believe</i>	-6	1
14	noun prep poss-pronoun noun NN IN PPZ NN <i>rest of my holiday</i>	1	-1
20	verb-base det noun prep VV DT NN IN <i>see a lot of</i>	-5	4
31	to-inf verb-base prep det TO VV IN DT <i>to go to the</i>	-14	-1
41	prep det noun pronoun IN DT NN PP <i>at that moment he</i>	2	4
42	det noun prep possessive pronoun DT NN IN PPZ <i>the rest of my</i>	7	3
48	pres-simple-be det adj noun VVZ DT JJ NN <i>is a good idea</i>	-25	1
49	modal verb-base det noun MD VV DT NN <i>can learn a lot</i>	-20	0

Table 7.9 Core sequences: B2 sequences which are highly convergent in ranking at both B2 and C1.

7.3.2 Emerging sequences at B2

Noun phrases dominate the ten emerging sequences in the B2 top 50 – those which rank higher at C1 than B2 and therefore become increasingly more important for C1 learners (Table 7.10). Five of these sequences are new to the top 50 at B2 #45, 26, 43, 33, 50, all of which contain noun phrases and jump markedly from a lower ranking at B1 to a higher ranking position at C1. As well as moving even further up the ranks at C1, and therefore becoming increasingly used, they are also much lower ranking at B1 and may mark the leap between B1 and C1. Also noticeable here is the first appearance of a past participle tag in #50

VVN IN DT NN (past-participle+preposition+determiner+noun, e.g. *returned from a trip*), a noun phrase containing a non-finite verb, both indicators of an increased range of syntactic complexity #43 DT NN TO VV (determiner+noun+to-infinitive+verb-base, and the appearance of an adverb sequence #45 RB IN DT NN *adverb+noun+determiner+verb-base*).

B2 rank	POS sequences and examples	B2-B1	B2-C1
28	prep det noun conjunction IN DT NN CC <i>in the morning and</i>	-12	6
18	det noun prep noun DT NN IN NN <i>this kind of job</i>	-14	8
39	det noun prep plural-noun DT NN IN NNS <i>a lot of people</i>	-16	8
30	prep det noun IN DT NN <i>to the city centre</i>	-13	10
45	adverb det noun prep RB IN DT NN <i>all over the world</i>	-57	11
23	adj noun prep det JJ NN IN DT <i>new shop in the</i>	-24	12
26	plural-noun prep det noun NNS IN DT NN <i>animals in a zoo</i>	-32	13
43	det noun prep to-inf verb-base DT NN TO VV <i>the opportunity to see</i>	-67	14
33	noun prep det adj NN IN DT JJ <i>house with the yellow</i>	-34	19
50	ed-form prep det noun VVN IN DT NN <i>returned from a trip</i>	-161	26

Table 7.10 Emerging sequences: B2 sequences which are higher ranked at C1 than B2 (with rank difference).

7.3.3 Decreasing sequences at B2

The nine decreasing sequences in the top 50 B2 are pivotal to the transition from B to C levels (Table 7.11). Like the decreasing sequences in the B1 data (7.2.3), they are dominated by verb phrases with an initial pronoun (with the exception of #12 – a sequence that is

affected by task). Together the decreasing sequences at B1 and B2 appear to mark the movement from verb-dominated sequences to the noun dominated sequences seen in the C levels.

B2 rank	POS sequences and examples	B2-B1	B2-C1
46	pronoun past-simpleV to-inf verb-base PP VVD TO VV <i>he decided to go</i>	10	-36
34	pronoun pres-simple to-inf verb-base PP VVP TO VV <i>I want to thank</i>	12	-29
25	pronoun modal verb-base pronoun PP MD VV PP <i>you could send me</i>	7	-25
36	pronoun modal verb-base prep PP MD VV IN <i>you can go to</i>	24	-24
16	pronoun modal verb-base to-inf PP MD VV TO <i>I would like to</i>	-25	-17
15	modal verb-base to-inf verb-base MD VV TO VV <i>would like to know</i>	-27	-13
22	verb-base prep det noun VV IN DT NN <i>go to the cinema</i>	-7	-8
27	pronoun modal verb-base determiner PP MD VV DT <i>you can see the</i>	-1	-8
12	proper-noun proper-noun proper-noun NP NP NP (this generates any text in capitals)	8	-6

Table 7.11 B2 sequences decreasing in ranking at C1 (with rank difference).

From this initial analysis, overall characteristics of the B2 sequences include:

- an increasing body of core sequences
- a picture of greater convergence with adjacent higher level
- some sequences indicating greater syntactic complexity (e.g. past participle, noun phrases with non-finite verb)
- a decrease in sequences containing a pronoun and tensed verb.

7.3.4 A developmental picture: summary from the B level perspective

In the previous chapter, when looking at the data from the perspective of the A2 user, we saw greater convergence in the distribution of POS tag sequences between A2 and B1 than between A2 and A1. So far in this chapter we have taken both B1 and B2 as starting points, looking back and forward. We have seen overall that:

- There is greater similarity in distribution of the POS tag sequences between A2 and B1 and between B2 and C1 than there is between B1 and B2, pointing to a leap in development at the B level.
- There is a growth of the body of core sequences between B1 and B2, which continues to build in C1.
- Noun phrases with adjectives increase at B1 and continue to increase to C levels.
- Where verb phrases at the A levels were dominated with modal verbs and present continuous forms, these decrease at the B level.
- Tensed verbs following pronouns peak at B1 but begin to decrease at B2.
- There is an increase in core sequences with greater syntactic complexity at B2.

7.3.5 Background to case study selection

Again in this first phase of analysis, comparing sequence ranking and distribution has revealed changes which warrant further investigation. In the following sections I explore the lexical and functional exponents of some representative core, emerging and decreasing sequences, beginning with two sequences with pronoun + tensed verb + to-inf+ verb-base:

PP VVD TO VV (*I started to cry, I decided to go*) and PP VVP TO VV (*I want to apply, you need to go*).

The following sections (7.4 to 7.6) explore research questions RQ2 and RQ3 firstly to examine how representative sequences develop across proficiency levels and to explore whether existing frameworks for classification of language patterning account for this development.

In these case studies the lexical and functional properties of two sequences PP VVD TO VV (*I started to cry, I decided to go*) PP VVP TO VV (*I want to apply, you need to go*) are investigated. These two sequences are of interest because (1) they represent the same sequence, but with different verb forms in the first verb slot (one containing a past tense verb form (VVD) and the other a present simple verb form (VVP)), and (2) as already observed,

sequences with tensed verbs peak in ranking at B1. Both sequences contain a tensed verb form and a non-finite verb form. The tag VVD generates all past simple forms of verbs apart from *be* and *have*. VVP represents the first and second person singular and plural present simple forms of all verbs apart from *be* and *have*. From a B level perspective, both are examples of decreasing sequences, ranking higher at B1, decreasing in rank at B2 and continuing to decrease at C1 (and C2) as shown in the shaded sections in Table 7.12.

sequence	A1	A2	B1	B2	C1	C2
PP VVP TO VV <i>I want to thank</i>	#9	#15	#22	#34	#63	#94
PP VVD TO VV <i>he decided to go</i>	#566	#83	#36	#46	#82	#80

Table 7.12 Ranking of PP VVP TO VV and PP VVD TO VV across all levels

However, taking a look back at the A level usage (see the unshaded sections in Table 7.12), we see that the **pronoun + pres-simple + to-inf + verb-base** (PP VVP TO VV *I want to thank*) ranks highest at A1, decreases steadily to B2 and drops in rank at C1 and C2. In contrast, the **pronoun + past-simple + to-inf + verb-base** (PP VVD TO VV *he decided to go*) is not a frequently used sequence at A1. Its ranking jumps from #566 at A1 to #83 at A2, it peaks at B1, then decreases and stabilises at the C levels (back to a ranking on a par with A2). Both sequences decrease in ranking from B1 to C2.

We also note here the differences in the rankings at the A1 level. The present simple sequence is in the top 10 at A1 (#9) but the past simple sequence ranks at #566. There may be a range of reasons for this, including, among others, possible task effect, input factors (past simple typically being introduced after the present simple in instructional syllabi) and a possible indication of structural generalisation taking place at B1 but not at the A levels. These factors are further explored below, first in relation to PP VVD TO VV, the past simple sequence (7.4) and then by exploring PP VVP TO VV (7.5), the present simple sequence.

7.4 Case study 1: pronoun + past-simple verb + to-inf + verb-base (PP VVD TO VV)

I begin by investigating the frequency and distribution of the top 1000 types of this sequence. I then consider the lexical and functional usage of these sequence types and applying a pattern grammar (Hunston and Francis 2000) approach as a framework for analysis

7.4.1 Occurrences by level: *pronoun + past-simple + to-inf + verb-base (PP VVP TO VV)*

Table 7.13 shows the breakdown of the raw and relative occurrences of this sequence over all six levels as well as the percentage of occurrences covered by the top 1000 types for each level. The change in the % figures in the right-hand ‘1000 types as % of occurrences’ column indicates that the number of types for this sequence increases with proficiency, i.e. the range of different lexical exponents for the sequences increases.

	subcorpus size	raw occurrences	relative PMW occurrences	total occurrences	1000 types as % occurrences
				1000 types	
A1	2456971	766	312	766	100.00
A2	5703217	7277	1276	5873	80.71
B1	3261473	5473	1678	4335	79.21
B2	5263979	6455	1226	4379	67.84
C1	6711568	5090	758	3166	62.20
C2	7698695	5583	725	3325	59.55

Table 7.13 Breakdown of occurrences by level of pronoun+past-simple+to-inf+verb-base

Initial observations indicate a peak in the frequency of this sequence at B1 (Figure 7.3). In relative terms the sequence is used 5.3 times more frequently at B1 than at A1, 1.3 times more than at A2, 1.4 times more at B2, and over twice more than at C1 and C2.

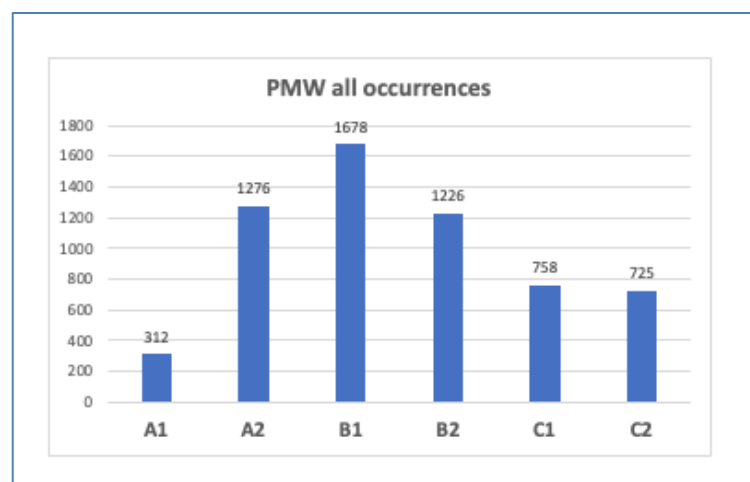


Figure 7.3 PMW frequency of all occurrences of PP VVD TO VV by level

However, frequency is just part of the picture. Let us consider how these frequencies are distributed across lexical exponents or types. Figure 7.4 shows the distribution of the frequency per level, in terms of the first 1000 and 100 types per level and Figure 7.5 shows the range of different types, indicated by the percentages that the first 1000 and 100 types constitute of all occurrences. For example, the first 1000 at A2 make up 80.71% of all occurrences, 79.21% at B1 and 67.84% at B2. At C2, even though the relative frequency of occurrences is a third of the number at B1, this first 1000 constitutes 59.55% of all types in comparison with the 79% of all types at B1.

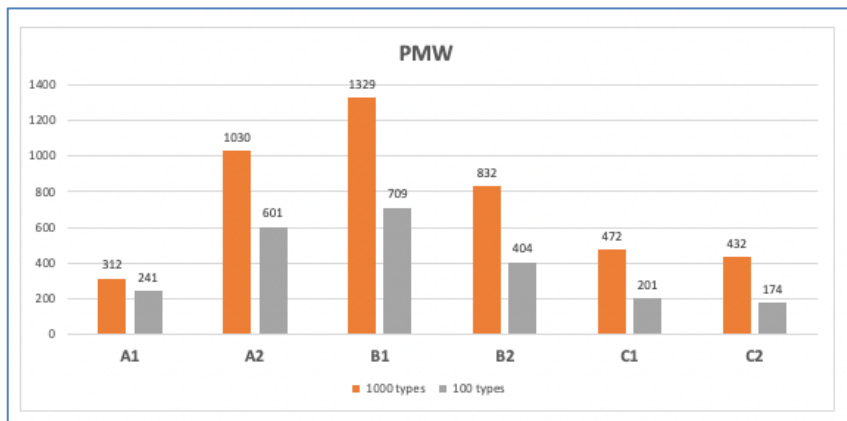


Figure 7.4 PMW frequency of the first 1000 and 100 types of PP VVD TO VV by level

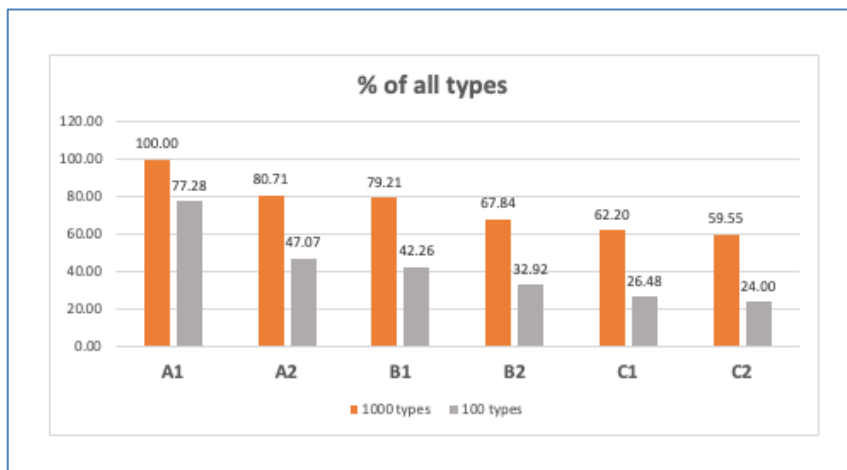


Figure 7.5 Percentage of all types of the first 1000 and 100 types of PP VVD TO VV by level

The implication in simple, formal terms, represented by Figures 7.4 and 7.5, is that even though the frequency of occurrence peaks at B1 and decreases as proficiency increases, the range of lexical exponents used for this POS tag sequence increases steadily. This is shown by the decrease in % types as proficiency increases, as learners at higher levels appear to do

more with less, lending further evidence to for a B1 to B2 leap in development identified above. First I look to see how this sequence plays out lexically and functionally across all levels.

7.4.2 Lexical and functional distribution by level: pronoun + past-simple + to-inf + verb-base (PP VVD TO VV)

The top 100 most frequent lexical exponents were extracted to investigate their structural and functional characteristics. The top 20 for each level, colour coded by verb, are shown in Table 7.14.

Colour key: went = brown; needed = pale orange; wanted = blue; used to = orange; started = pink; decided = green; arranged = lime green

A1	A2	B1	B2	C1	C2
I went to buy	I needed to buy	I decided to go	I decided to go	I decided to write	I decided to write
I went to see	I decided to go	I decided to join	I decided to write	I decided to go	I decided to go
I went to watch	I went to see	we decided to go	we decided to go	I used to go	I decided to take
I wanted to ask	I decided to buy	I wanted to go	I went to see	I wanted to go	I got to know
I decided to buy	we decided to go	I decided to buy	I used to go	I decided to take	I used to go
I wanted to tell	I wanted to tell	they wanted to film	I used to work	we decided to go	I used to play
I decided to paint	we arranged to meet	I wanted to tell	I wanted to go	I got to know	I used to enjoy
I wanted to know	I wanted to buy	They wanted to film	I decided to buy	I used to play	he decided to go
I started to work	I wanted to go	They wanted to make	I used to play	I went to see	I wanted to go

I wanted to buy	I went to buy	we arranged to meet	I decided to take	I decided to buy	I used to live
I needed to buy	We decided to go	I went to see	I wanted to tell	we went to see	I wanted to do
I went to go	I wanted to know	I needed to buy	I wanted to see	I decided to give	I used to think
I went to do	you wanted to go	they wanted to make	you wanted to know	I used to live	I used to spend
I went to saw	we started to talk	I decided to write	I wanted to know	I wanted to do	they used to do
I liked to watch	I forgot to tell	I started to cry	we used to go	I used to work	she decided to go
I went to shop	I wanted to ask	I wanted to thank	I wanted to buy	I managed to get	I used to visit
I decided to go	I used to go	I decided to change	I decided to visit	they used to do	I decided to send
I went to swim	I went to visit	I used to go	I wanted to visit	I tried to find	we used to go
I forgot to tell	I started to run	We decided to go	I wanted to do	I used to spend	you used to enjoy

Table 7.14 Top 20 most frequent lexical realisations of PP VVD TO VV at all levels.

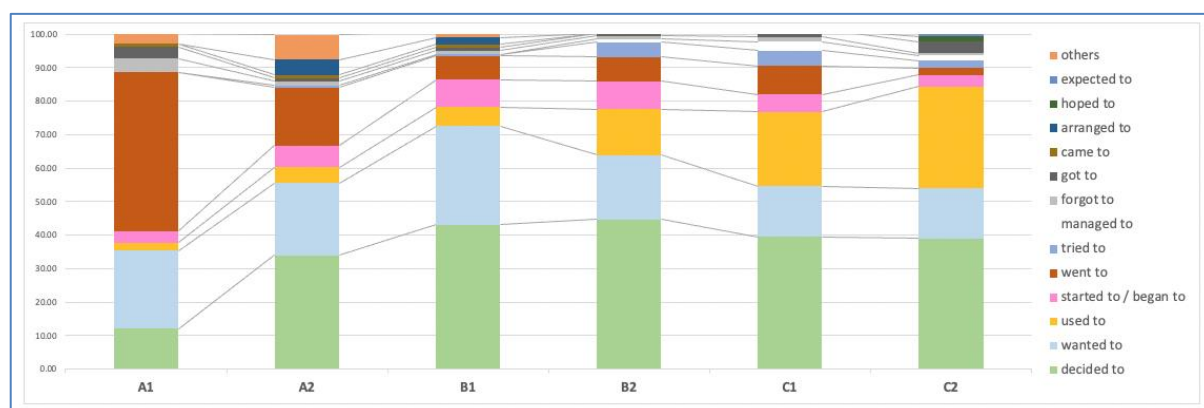


Figure 7.6 Overall breakdown of PP VVD TO VV in the top 100 types, across levels, by past simple verb form

An overall breakdown of verb forms in the top 100 across levels is illustrated in Figure 7.6.

Initial overall observations show that

- there is a preference at A1 level for *I went to* + activity verb (see brown section at A1).
- *decided to* rises in use from A1 to A2 and is consistently used across B and C levels (see green sections overall).
- the increase in the verb range as proficiency increases is limited, with *wanted to* (light blue) and *decided to* (green) dominating the mid levels, and *used to* (orange) increasing at C levels.
- *I* prevails in the pronoun slot and an activity verb in the VV slot across all levels.
- an overall narrative and recounting function across all levels, pointing to a task sensitivity.

7.4.3 Applying a pattern grammar categorisation

Having sorted the lexical sequences by past simple verb type they are categorised using pattern grammar (PG) set out by Hunston and Francis (2000) (see also

<https://grammar.collinsdictionary.com/grammar-pattern>) (See Table 7.15). This PG

categorisation first involves identifying form groupings and secondly meaning groupings.

The grammar pattern which is the closest fit for the sequence PP VVD TO VV comes under ‘Simple patterns V to-inf.’ This pattern is subdivided into three structures, I-III (Figure 7.7), each of which is subdivided into different meaning groups, exemplified in Table 7.15.

V to-inf	
V to-inf	
The verb is followed by a to-infinitive .	
This pattern has three structures:	
Structure I: Verbs in phase	<i>The number of victims continues to rise.</i>
Structure II: Verb with Object	<i>He expects to fly to Beijing soon.</i>
Structure III: Verb with Clause	<i>He hurried to catch up with his friend.</i>
Structure I: Verbs in phase	
Structure II: Verb with Object	
Structure III: Verb with Adjunct	
Productive uses	

Figure 7.7 Extract from the V to-ing grammar pattern

Structures	Example groups	Example verbs
Structure I	The ‘begin’ group	begin, cease, grow, come, start
Verb in phase	The ‘try’ group	attempt, battle, strive, try, strain, fight

	The 'fail' group The 'manage' group	decline, fail, forget, (not) need, refuse get, contrive, manage, serve
Structure II Verb with Object	The 'promise' group The 'hope' group The 'need' group The 'like' group	agree, arrange, decide, fix, plan, promise ache, aspire, desire, expect, hope, seek, want deserve, need prefer, like, love, hate
Structure III Verb with Adjunct	The 'collaborate' group The 'hurry' group The 'wait' group The 'qualify' group	collaborate, conspire, collude, gang up come, go, hurry, rush queue, wait, stand by qualify, register, train

Table 7.15 Pattern grammar classification of verb + to-inf

Table 7.16 summarises the top 100 types for each level classified first according to the past simple verb and then pattern grammar groupings. Each row shows the percentage breakdown for each past simple verb form by level, followed by its pattern grammar categorisation, according to structure type and meaning group. Cases where no corresponding group was found in Hunston and Francis 2000 but where there is a clear meaning are labelled uncat+MEANING, with a relevant meaning group specified (e.g. uncat+TIME).

	% of total						Pattern grammar	
	A1	A2	B1	B2	C1	C2	structure	group
<i>decided to</i>	12	34	43	45	39	39	II	promise
<i>arranged to</i>	0	5	2	0	0	0	II	promise
<i>wanted to</i>	23	22	29	19	15	15	II	hope
<i>hoped to</i>	0	0	0	0	0	2	II	hope
<i>expected to</i>	0	0	0	0	0	1	II	hope
<i>used to</i>	2	5	6	14	22	31	uncat_TIME	
<i>started to / began to</i>	4	7	8	8	5	3	I	begin
<i>got to</i>	3	1	1	1	2	4	I	begin
<i>went to</i>	47	17	7	7	8	2	III	hurry
<i>came to</i>	1	1	1	0	0	0	III	hurry
<i>tried to</i>	0	1	0	4	5	2	I	try
<i>managed to</i>	0	0	0	1	3	2	I	manage

<i>forgot to</i>	4	1	1	1	1	1	I	fail
<i>others</i>	3	7	1	0	0	0	uncat	

Table 7.16 Breakdown of the past simple verb form by level and grammar pattern

The PG categorisation approach is useful for matching forms with meaning groupings when looking at verb complementation patterns, and for identifying some clustering around meanings (for example, as evident in the promise group, the hope group, the begin group, the hurry group, where two or more forms belong to the same group). However, as a means to identify development of the 4-gram sequences in this study there are limitations, some of which are outlined here:

- the pattern grammar classification is restricted to only the VVD TO part of the POS-gram, and may miss nuances of meaning inherent in a longer sequence; for example it does not account for:
 - the effect and frequency of pronoun choice
 - the effect of tense on meaning and usage in the verb form
 - the collocational patterning in the following base verb form.
- its structural classification does not take into account frequency, distribution or meaning
- its ‘group’ classification does not take into account the frequency and distributional properties of the verbs, nor the possibility of a pioneering or pathbreaking superordinate form (which takes the largest share of the distribution and which is prototypical of the meaning of the sequence). For example *went to*, *decide to*, and *used to* are the forms that dominate at different levels. However *decide to* is classified under the ‘promise’ group and *went to* under ‘hurry’, though neither *promise* nor *hurry* are found in the data nor are their meanings implicitly or explicitly expressed in *decide to* or *went to*.
- It does not account for all occurrences, e.g. *used to*, one of the most frequently occurring forms at the C levels, is not categorised at all in this complementation pattern.
- It does not account for the effect that register and in this case task might have on verb use and meaning.

While there is evidence of a clustering of form-meaning groups in the data the PG approach of classifying offers a disparate collection of verbs in use in the data.

7.4.4 An alternative functional framework for an analysis of development

An alternative way of looking at this sequence involves a multi-layered approach, taking into account not just the form-meaning categorisation of the verb + complement, but also the surrounding co-selection of both the pronoun and the following verb-base form, all considered within the context of the task.

The sample of examples in Table 7.16 reflects a recounting of a series of volitional acts or experiences, typical of the narrative and descriptive task types found across the exam data in focus. This is exemplified from A1 *I went to see* to C2 *I used to go* with other sequences, e.g. *I decided to buy, I wanted to thank, I hoped to get* all fitting into this meaning. Below are extracts of these sequences in context, from the Cambridge mainsuite exam data:

Extract 7.1

I went to the cinema recently and we saw the film "the lord of the rings". I meet my new friend two weeks ago. I meet her in the tube station when she fall down stair, and **I went to help** her, then she invited me to a cup of coffe. We start to talk about us and is was like I knew her before (*A2 performance level PET exam 2002*)

Extract 7.2

My bedroom have two single beds on the either side of the room for my sister and me. There are two desks placed right in front of the big window. However, I still need to get a big carpet for decorating my bedroom. I decided to put it in the middle of my room so that it will look more bright. My dad bought us two closets for our clothes. It is placed in the corner, my sister and **I decided to get** a lot of posters of some famous actors, singers, etc for decorating our walls! (*B1 performance level PET exam 2003*)

Extract 7.3

Hi Mary, I'm really sorry because I haven't been in touch for so long time. I'm fine by the way. Yesterday I went to the cinema because for along time **I wanted to see** this film which director is Martin Voscopoulos (*B2 performance level FCE exam 2003*)

Extract 7.4

The doors were opened by four strong men. One of them told Paul to give him the suitcase. Paul refused and started to run so fast that he surprised the four men. **They tried to catch** him but failed. Hidden, Paul wanted to see why they wanted his suitcase (*B2 performance level FCE exam 1997*)

On a semantic or functional level, the dominant 'I decided to + verb' sequence seen at all levels performs a 'volitional' function, with speaker/writer agency to do with decisions, desires and the accomplishment of 'acts'. The pronoun *I* is dominant in this sequence at all levels. This same overall function is found in other verbs, with lower frequencies, *wanted*,

hope, expected, arranged, tried. As proficiency increases, a form-function pattern of **pronoun + verb of volition + to inf** appears to emerge, and the most frequently used verbs express this function at all levels. Within this series of volitional acts there is the additional expression of degrees of accomplishment, some more successful than the others, e.g. *I went to buy / I decided to buy / I started to learn / I wanted to visit / I tried to buy / I managed to buy / I failed to buy / I forgot to buy*.

Learners appear to show an increasing understanding of complexity of use of this sequence beyond the verb complementation pattern. It is their understanding of the linguistic choices offered by the context and the repertoire to fulfil function which develops. They seem to show an understanding that the **form** PP VVD TO VV performs a specialised **function** within a particular **context**, illustrated in Figure 7.8. When looking at this in simplistic terms the *I went to buy* sequence represents an act of total volition and accomplishment, the *I used to buy* a series of accomplishments, whereas the other options represent more nuanced colours of volition and achievement (*I wanted to buy, I managed to change*).

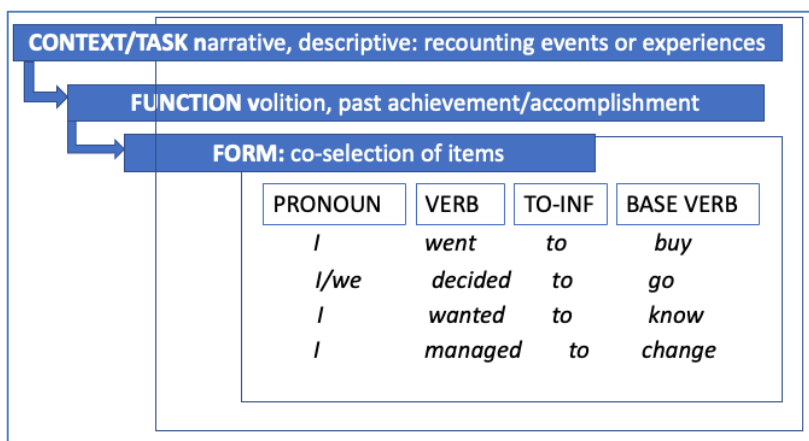


Figure 7.8 A context-function-form overview of the PP VVD TO VV sequence

Learners demonstrate a growing understanding that this highly specialised function is restricted to a limited range of verbs in the VVD slot. Returning to the B level focus of this chapter, there is subtle change between B1 and B2. This is evident in:

- a decrease in the *I/we went to* sequence
- an increase in the *I used to* sequence, moving from recounting one event (e.g. *I went to*) to several (e.g. *I used to go*)
- an increase in the degrees of accomplishment sequences (*tried to, managed to*)
- a stabilisation in the choice of verbs in the VVD slot.

It is also important to note that all of these trends continue as proficiency increases. *Decided, used, wanted* make up 78% of the top VVD forms at B1 and 77% at B2, rising to 84% of the top 100 types at C2. This is moving towards a fixedness of choice of verbs in the VVD slot at the C level. The increase in types at the higher levels comes with a wider range of pronoun use, in combination with a wider range of following verbs expressing acts.

Moving to the second case study (PP VVP TO VV), we might expect to find a similar picture of usage, the only difference between the two sequences being the use of a present simple verb after the pronoun.

7.5 Case study 2: pronoun + past-simple verb + to-inf + verb-base (PP VVP TO VV)

7.5.1 Occurrences by level: pronoun + present-simple + to-inf + verb-base (PP VVP TO VV)

Unlike the past simple sequence, which peaks in frequency at B1, this sequence decreases from A1 to C2. Figure 7.9 shows the relative occurrences of this sequence over all six levels and Figure 7.10 the percentage of occurrences covered by the top 1000 types for each.

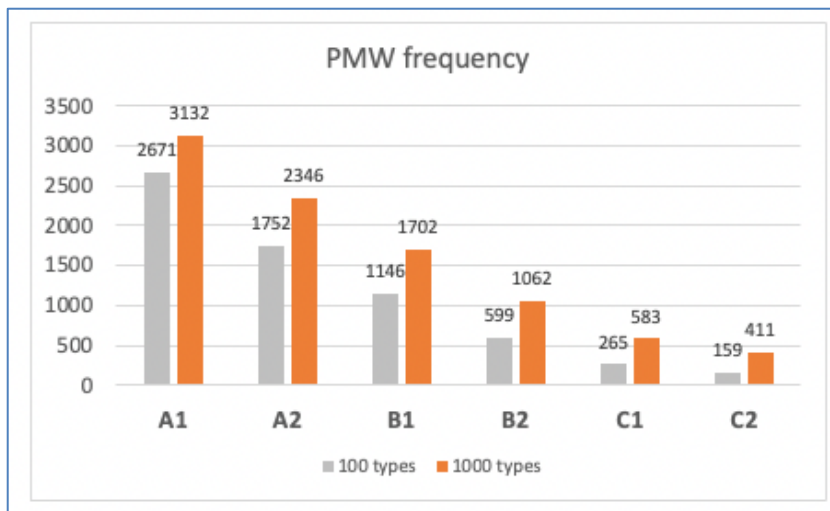


Figure 7.9 PMW frequency of the first 1000 and 100 types of PP VVP TO VV by level

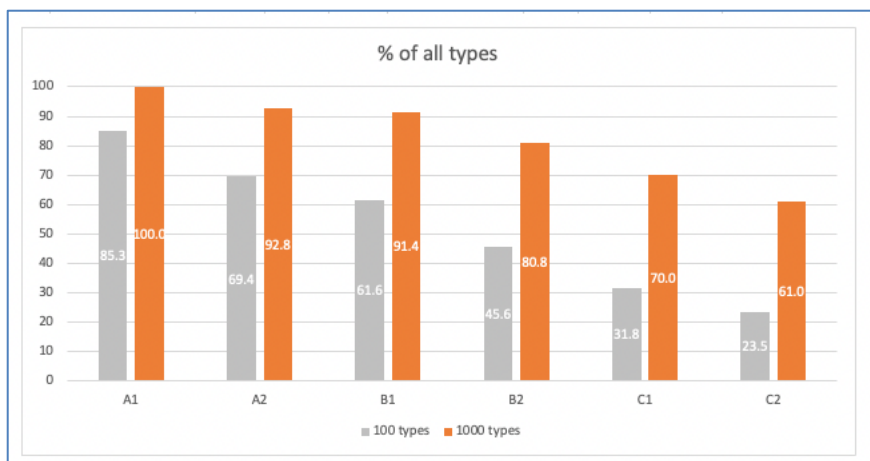


Figure 7.10 Percentage of all types of the first 1000 and 100 types of PP VVP TO VV by level

As with the previous case study, the implication in simple, formal terms is that even though the occurrences decrease as proficiency increases, the range of lexical exponents used for this POS tag sequence increases steadily. As proficiency increases learners appear to do more with less. Next we turn to the lexical exponents and their functional use, by level.

7.5.2 Lexical and functional distribution by level: *pronoun + present-simple + to-inf + verb-base (PP VVP TO VV)*

The top 100 most frequent lexical exponents were extracted to investigate their structural and functional characteristics. The top 20 for each level (colour coded by verb) are shown in Table 7.17.

Colour key: want to = blue; need to = green; hope to = purple; prefer to = yellow; like to = orange; get to = grey; tend to = dark grey

A1	A2	B1	B2	C1	C2
you want to come	you want to go	I want to tell	I want to thank	I hope to hear	you want to go
you want to go	you want to come	you want to go	I want to tell	I want to thank	they want to do
You need to wear	I want to tell	I hope to see	you want to go	you want to go	I hope to see
I want to go	I hope to see	you want to do	I want to know	I want to say	I want to say

I want to see	I want to buy	I hope to hear	I hope to see	I hope to see	you want to do
you need to wear	I want to sell	you prefer to go	I hope to hear	I want to tell	you get to know
I want to paint	I want to go	I want to thank	I want to join	you want to do	I want to give
I want to tell	I like to wear	I want to know	I want to say	you want to know	I hope to hear
I hope to see	you want to visit	you want to see	you want to come	I want to know	you want to get
I start to work	I need to change	I start to work	I want to ask	you need to know	they want to go
I want to help	I want to see	I want to go	I want to learn	you want to get	we want to do
I want to invite	you want to do	you want to come	you want to do	you want to take	I want to do
I want to come	I want to thank	I want to see	I want to go	I want to make	we get to know
I want to ask	I want to know	you want to know	you want to know	you get to know	you want to know
I want to know	I hope to hear	I want to say	I need to know	they want to do	I tend to believe
you like to come	you want to know	you want to visit	you want to see	I want to give	we want to make
You need to bring	You need to wear	they want to go	you want to take	you want to see	you want to see
I like to go	I need to buy	I want to ask	you want to buy	you want to learn	you need to do
I like to play	I want to say	I need to change	you want to learn	I need to know	I want to mention
you need to bring	I like to go	you want to spend	I want to give	I want to ask	they get to know

Table 7.17 Top 20 most frequent lexical realisations of PP VVD TO VV at all levels.

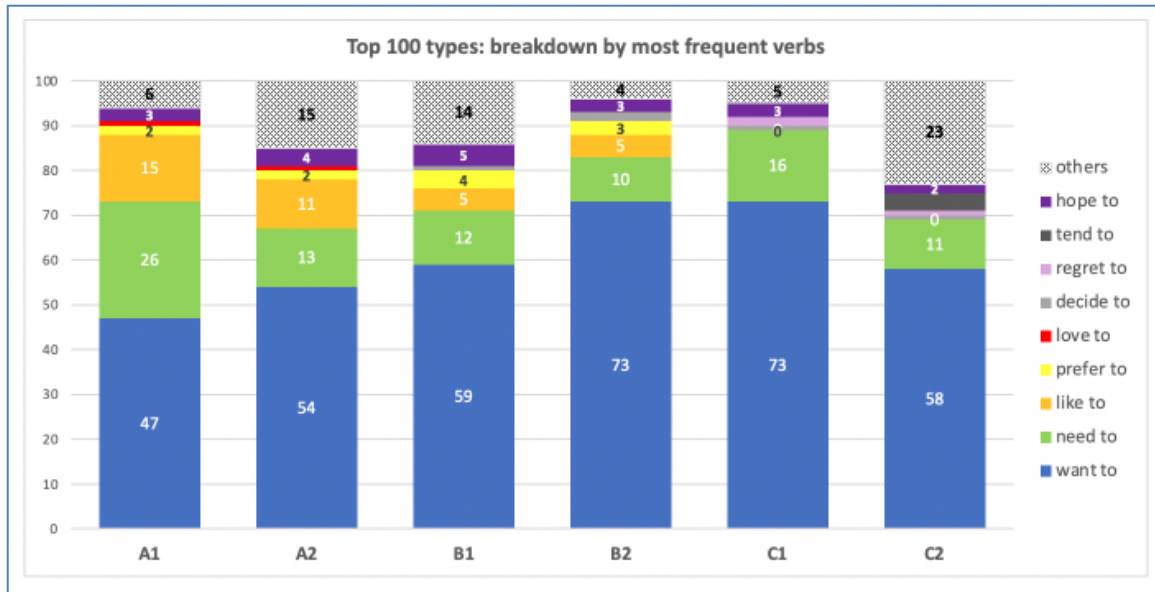


Figure 7.11 Overall breakdown of PP VVP TO VV in the top 100 types, across levels, by present simple verb forms

An overall breakdown of the most frequent verb forms in the VVP slot of the top 100 types is illustrated in Figure 7.11. Initial observations show that:

- *want to* is the dominant form across all levels, as indicated by its percentage use across the top 100 types (Figure 7.11). It increases in usage from A1 to B2, stabilises at C1 and decreases at C2.
- *need to* is the second most frequent, used across all levels.
- Verbs of preference other than *want* (*like to*, *prefer to*, *love to*) decrease in use from A1 to B2 and do not feature in the top 100 types at C1 and C2.
- At first sight the B2 level and C1 level have a more similar distribution in comparison with B1 and B2.
- *tend to* is used at C2 level.
- There is movement in the pronoun slot which may be indicative of formulaic use.
- There is an increase in other various verb forms and sequences from C1 to C2.
- Dominant forms in the PP VVD TO VV structure (*decide to*, *used to* and *go to*), occur either infrequently or not at all in the top 100 types of this VVP structure.

7.5.3 Applying a pattern grammar categorisation

Following the pattern grammar taxonomy described in 7.4 above (Table 7.15), the verbs in the top 100 types are categorised into structure and meaning groups. Table 7.18 summarises

the top 100 types classified first according to the present simple verb and then pattern grammar groupings. Each row shows the percentage breakdown for each present simple verb form by level, followed by its pattern grammar categorisation, according to structure type and meaning group. Cases where no corresponding group was found in Hunston and Francis (2000) but where there is a clear meaning are labelled 'uncat' and where relevant a meaning suggested, e.g. uncat+try.

	% of 100 types						Pattern grammar	
	A1	A2	B1	B2	C1	C2	structure	group
<i>want to</i>	47	57	59	73	73	58	II	hope
<i>need to</i>	26	13	12	10	16	11	II	need
<i>like to</i>	15	11	5	5	0	0	II	like
<i>prefer to</i>	2	2	4	3	0	0	II	like
<i>love to</i>	1	1	0	0	0	0	II	like
<i>hope to</i>	3	4	5	3	3	2	II	hope
<i>decide to</i>	0	3	1	2	1	1	II	promise
<i>regret to</i>	0	0	0	0	2	1	III	regret to say
<i>tend to</i>	0	0	0	0	0	4	III	tend
<i>go to</i>	1	1	3	1	0	0	III	hurry
<i>come to</i>	1	1	1	1	0	**1	III	hurry
<i>start to</i>	1	0	1	0	0	0	I	begin
<i>plan to</i>	0	2	1	0	0	0	II	promise
<i>learn to</i>	1	2	0	0	0	0	uncat+try?	
<i>choose to</i>	0	0	1	0	0	1	II	promise
<i>get to</i>	0	0	0	2	2	5	I	manage
<i>try to</i>	0	0	0	0	1	1	I	try
<i>wish to</i>	0	0	0	0	0	1	II	like
<i>dare to</i>	0	0	0	0	1	1	uncat	
<i>beg to</i>	0	0	0	0	0	1	uncat I beg to differ	
<i>dread to</i>	0	0	0	0	0	1	uncat I dread to think	
<i>mean to</i>	0	0	0	0	0	1	uncat I mean to say	
<i>*use to</i>	1	0	0	0	0	0	uncat	
<i>*write to</i>	1	1	1	0	0	0	uncat	

* <i>suggest to</i>	0	1	1	0	0	0	uncat
---------------------	---	---	---	---	---	---	-------

*indicates sequence which on closer inspection show complementation errors

** indicates a formulaic use *come to think

Table 7.18 Breakdown of the present simple verb form by level and grammar pattern

The PG categorisation approach is useful for matching forms with meaning groupings when looking at verb complementation patterns, and for identifying some clustering around meanings. It reveals a strong form-meaning correspondence between this verb pattern and functions of hope, promise, preference, and, at C2, tendency. However, some of the classifications do not take into account related meanings, e.g. *like to*, *prefer to*, *love to* are categorised under meaning group ‘like’, and *want to* under ‘hope’. An investigation of usage on a qualitative level shows that all of these verbs could fall under a ‘preference’ meaning. As a means to identify development of the 4-gram sequences in this study there are limitations, some of which are outlined here:

- the pattern grammar classification is restricted to only the VVP TO part of the POS-gram, and may miss nuances of meaning inherent in a longer sequence; for example it does not account for:
 - the effect and frequency of pronoun choice
 - the effect of tense on meaning and usage in the verb form, as revealed from the comparison with past simple usage in the same sequence.
 - the preceding and following collocational patterning.
- The classification does not take into account frequency and usage beyond the context of the pattern (e.g. see 7.5.4 below).
- It does not account for all occurrences, e.g. not all occurrences can be categorised using the taxonomy. Many of the occurrences at C2 are part of a fixed structure, e.g. *I dread to think*, *I beg to differ*, *you get to know*.

7.5.4 Pronoun use

An illustration of pronoun use of the nine most frequently occurring verbs in the top 100 types can be seen in Figure 7.12. The verbs are colour coded (e.g. *want* is blue, *need* is green) and the pronoun use of each verb is illustrated by a shade of colour (see shades of blue for *I/you/we/they want to*).

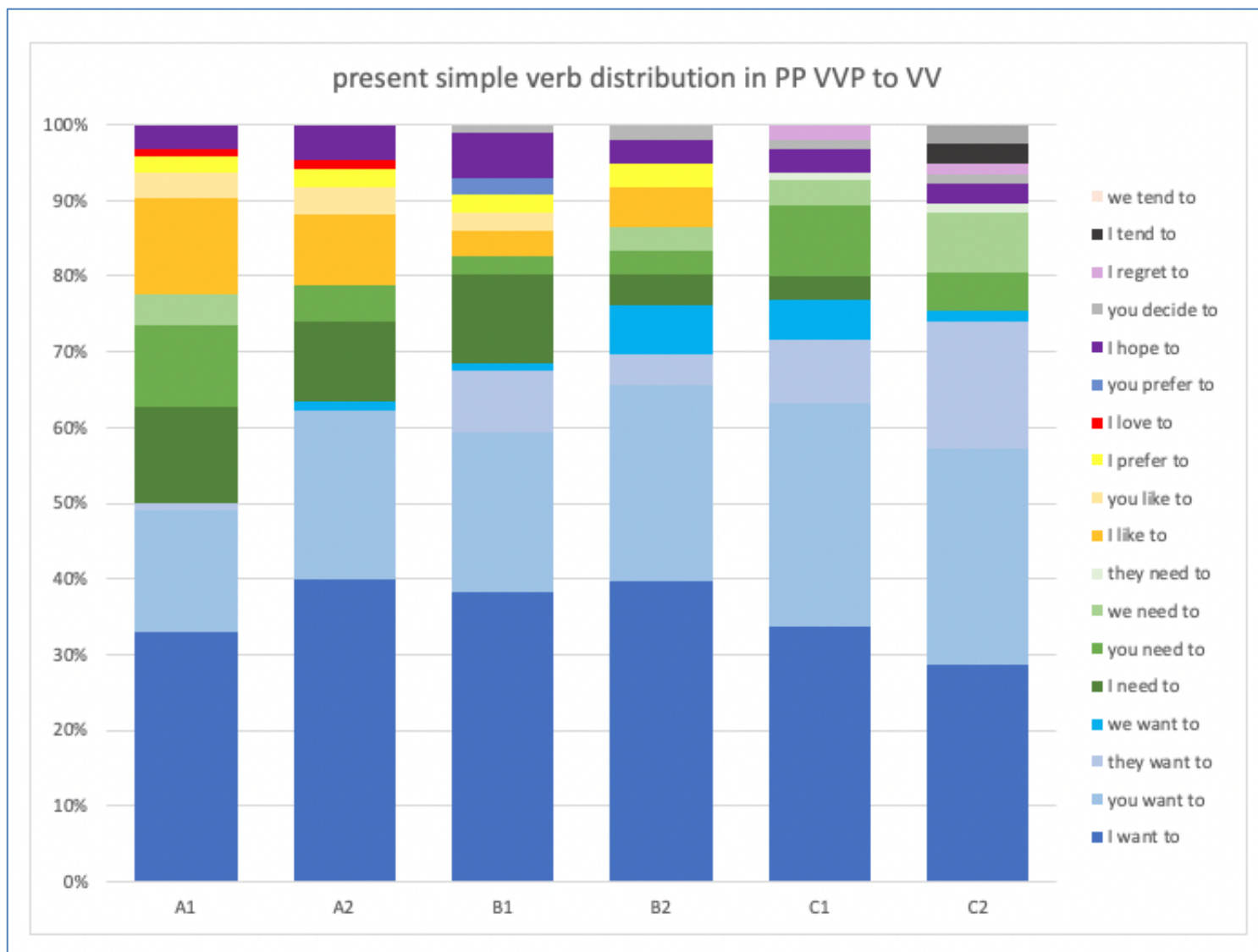


Figure 7.12 pronoun + present verb distribution of most frequently occurring verbs in top 100 types

Figure 7.12 shows change over the levels evident in an increased repertoire of pronouns before *want to*, and a redistribution of the present simple verb use. Candidates for the ‘like’ group decrease in frequency and disappear by C1 but *want to* persists. (See the disappearance of the orange and yellow and the increase in blue). This may be an indication that task is at play or that **pronoun + want to** represents a pioneering form (a form that takes the largest share of the distribution and which is prototypical of the meaning of the sequence) for a preference function, and by the C levels represents the most economical and efficient formula for expressing this function.

Pronoun + want to + verb-base: Looking for development beyond the 4-gram sequence

A closer analysis of the collocational patterns and concordance lines of *want to* shows further change across levels, pointing to a need to look at development beyond the 4-gram sequence in the surrounding cotext and context. (see further exploration in 7.6.1). The following four extracts exemplify a development from usage within a simple main clause expressing preference (extract 7.5) to repeated fixed patterning with e.g. *if, whenever, no matter what* in subordinate clauses exemplified in extracts 7.6-7.8

Extract 7.5

I have a small desk and perhaps I can put my PC screen on my desk table so that I can throw my PC desk table in place of it. **I want to buy** a larger desk but I still wonder if it does fit of it. Because I don't have much time I will stop now. (A2, PET, 2001)

Extract 7.6

The film starts at 9:00 pm, but we'll be there at 8:50pm. **If you want to go**, meet us there. Dear John, I think it's great that you will come here next Holiday! If you really come you can stay in my home in the city. There we can do many things like visiting the Zoo, go shopping (B1, PET, 2004)

Extract 7.7

the solution to the problem in not ending tourism but to make people realise that they should take care of the places they visit. If they do so it will always be there **whenever they want to go** and also allow future generations to see the world as they did. (B2, CAE, 2004)

Extract 7.8

Seriously, I believe we should study and maths and history and ancient greek no matter what we want to do in our life. Secondly, I would like to inform that I am totally against the idea that school education is a waste of time. (C2, CPE, 2003)

To explore this usage in context further beyond the ‘pronoun + want to + verb base’ instances, I take a look at the collocational patterning for all verbs in this case study 2 PP VVP TO VV (pronoun + present-simple + to-inf + verb-base) and case study 1 PP VVD TO VV (pronoun + past simple + to-inf + verb-base).

7.6 Beyond the 4-gram sequence: collocational patterning in case studies 1 and 2

In this section I compare the collocational patterning of the two sequences analysed in case studies 1 and 2 (7.4 and 7.5). PP VVP TO VV and PP VVD TO VV are similar sequences, but with a different tense use in the tensed verb slot. Analysis of these sequences so far suggests that the tense of the verb has an impact on lexical choice in the pronoun and verb slots, and that this varies across proficiency levels and contexts of usage. To look at this in more detail, I examine the wider textual patterning, and investigate what precedes each of the sequences at all levels. I use a combination of (1) the collocation tools available in Sketch Engine to identify the most frequent items preceding each sequence (N-1) and (2) concordance lines of the most frequently occurring collocations to qualitatively examine the contexts of usage.

7.6.1 Collocational patterning in case study 1: N-1 + pronoun + present simple verb + to-inf + verb

Table 7.19 shows the top 20 words that occur in position N-1 before the sequence PP VVP TO VV. They are ranked, not in terms of their frequency of occurrence but in terms of their strength of collocation score with the PP VVP TO VV sequence, all having a logDice score of >5. (A score of >5 or more indicates a strong collocation). Note that the collocates in Table 7.19 are uncorrected and so include misspellings as separate collocates (e.g. would instead of would)

	A1	A2	B1	B2	C1	C2
1	Do	Do	If	If	If	if
2	If	If	if	if	if	If
3	if	if	Do	Do	why	whether
4	Robbie	because	what	why	when	what
5	Did	?	why	because	case	whatever
6	Joe	So	because	case	what	when
7	p.m.	!	Now	when	whatever	why

8	Would	Now	Also	Also	whether	unless
9	do	do	where	Firstly	Do	What
10	because	so	whatever	what	whenever	wherever
11	Because	that	So	thing	thing	whenever
12	And	when	Would	whatever	unless	case
13	?	but	when	So	Finally	Do
14	.	.	whenever	Finally	because	makes
15	!	When	that	When	where	Whenever
16	Elena	what	?	Now	When	qualities
17	Nick	Would	unless	where	everything	whom
18	:	Because	Finally	reason	What	before
19	because	!!	And	What	makes	When
20	Wold	before	so	now	before	thing

Table 7.19 Top 20 collocations N-1 preceding PP VVP TO VV

An analysis of the concordance lines at each level shows that A levels are dominated by a *Do you want to + verb* frame, often in the context of invitations, as extracts 7.9 and 7.10:

Extract 7.9

Dan, Me, Lucas, Kehim and Dayana are going to the cinema tonight. **Do you want to come** with us? (A2, PET 2004)

Extract 7.10

Do you want to go to Karen's house? She has a tennis court! (A2, PET 2008)

In B1 level results, examples with *if* and *wh-* words prevail, often in the context of suggesting, offering alternatives (in the context of an *if you/they want + verb* frame) or giving reasons:

Extract 7.11

If they want to stay in a hotel, I would suggest the "Fiesta Americana Inn" which is comfortable and not expensive. (B1, FCE 2002)

Extract 7.12

I think you should speak with your parents, explaining them your **reasons why you want to go** with your friends! (B1, PET 2008)

The functions of giving reasons and offering alternatives continue into the B2 and the C levels, seen in the contexts of *because + I need/want to + verb* and *whether/whatever/whenever you/we want to + verb* frames:

Extract 7.13

In library you meet people who need quiet and peace, generally **because they want to learn** or read something interesting (C1, CAE 2002)

Extract 7.14

In this period of life teenagers make up their minds **whether they want to become** similar to their parents or change something in their behavior (C2, CPE 2007)

7.6.2 Collocational patterning in case study 1: N-1 + pronoun + past simple verb + to-inf + verb

Table 7.20 shows the top 20 words that occur in position N-1 before the sequence PP VVD TO VV. They are ranked, not in terms of their frequency of occurrence but in terms of their strength of collocation score with the PP VVD TO VV sequence, all having a logDice score of >5. (A score of >5 or more indicates a strong collocation).

	A1	A2	B1	B2	C1	C2
1	Yesterday	Yesterday	Then	Then	when	when
2	yesterday	So	why	So	why	where
3	Robbie	so	So	child	Then	Then
4	Elena	Then	so	why	child	child
5	So	then	then	so	When	When
6	so	night	when	when	where	So
7	Because	When	Remember	When	So	why
8	because	when	because	younger	so	what
9	then	Suddenly	finally	ago	then	so
10	!	why	When	then	ago	whenever
11	after	Remember	lesson	Finally	younger	Suddenly
12	When	because	end	Afterwards	since	Finally
13	week	child	Suddenly	suddenly	whenever	although
14	?	if	finished	where	what	ago
15	If	where	Later	night	nevertheless	finally

16	when	morning	Finally	if	if	Before
17	.	finally	child	Therefore	Before	if
18	first	remember	Therefore	finally	Thus	though
19	,	film	!	therefore	moment	afterwards
20	if	moment	said	because	later	how

Table 7.20 Top 20 collocations N-1 preceding PP VVD TO VV

An analysis of concordance lines shows at A2 that this sequence occurs within the context of recounting events with first person pronouns (e.g. *Yesterday I went to + verb*) and sequencing of events (*Then I/we decided to + verb*), while also providing background reasons for actions (*So I decided to + verb*):

Extract 7.15

Yesterday I went to buy some clothes. I bought a T-shirt, a sweater and jeans. I bought them because tomorrow I have a party. (A2, KET 2008)

Extract 7.16

After dinner we went to cinema. and we watched a horror film. It made me frightened. **Then we decided to go** my home. (A2, PET 2006)

Extract 7.17

As you know I like to watch Tv before I feel asleep. **So I decided to buy** a Tv. (A2, PET 2009)

By the B levels, there is more sequencing of events (*Then, Finally, Afterwards*), accounting for actions (*So, therefore*), and increased use of background reasons for actions following the semi-fixed phrasal patterning *That's the reason why / That's why I wanted/decided to + verb*):

Extract 7.18

I was completely terrified. **Finally we managed to land** on Milo's airport. (B2, FCE 1993)

Extract 7.19

Honestly I believe that you were given several misunderstanding informations about it, and **that is the reason why I decided to write** to you.(B2, CAE 2000)

Extract 7.20

Lots of them are treated very badly, but people don't realise it. **That is why I wanted to write** about this matter. (B2 FCE 2008)

By C2 level, the *when* prevails in the context of recounting events and time frames, and very frequently in the extended semi-fixed patterning, *It was + time frame + when + I decided/managed to*, showing evidence of a specialised focusing routine for storytelling:

Extract 7.21

It was a rainy day of December when I decided to get the car , against my mother's will , and have a go . (C2, CPE 2009)

Extract 7.22

It was on Friday when she started to hear rumors about a new boss coming to the department she was working in. (C2, CPE 1993)

Extract 7.23

It was then when she decided to take a risk and, disguising herself as a boy, she took up the audition for the role of Romeo in Shakespeare's perhaps most immortal tragedy: Romeo and Julliet (C2, CPE 2010)

At C levels *where* appears both in the context of an extended noun phrase (the noun + *where I/we used to / wanted to*) and in relative clauses (noun, *where I decided/managed to*), adding background information:

Extract 7.24

I was bought up in **an environment where I learnt to abide** by rules and principles in my life (C2, CPE 2003)

Extract 7.25

After the cafe "tour" I visited **the Art Gallery, where I managed to see** a real good exhibition (C2, CAE 2002)

7.6.3 Comparing case studies 1 and 2

A comparison between the two profiles, from an analysis of collocational patterning and concordance lines shows:

In the PP VVP TO VV sequence, there is

- An increase in *if/whether/wh*-words across levels, indicating both conditionality and fixedness of patterning in *if you want to, whether you want to, whatever I want to, unless you want to*
- Increased use in the expressions giving reasons in the B levels (*because, why, reason*)

- An increase in the use of discourse management and staging devices (*Firstly, Finally, So Also*) at the B levels.
- A movement from spoken-like discourse, involving questions at the A levels to more written-style register.

In the PP VVD TO VV sequence there is

- Evidence of characteristics of a narrative register, recounting past events, from Yesterday at A1 to *Suddenly/Finally/afterwards*
- Increased use of backgrounding and reasoning markers, *so, when, why* and semi-fixed patterning *That's why, that's the reason why I/we decided to + verb*
- Increased use as part of extended noun phrase at the C levels.

In both sequences:

- A movement from topic-based collocations (*Robbie, Joe, p.m.*) at the A1 level to discourse function marking collocations.
- A movement from simple clause use to more complex patterns of subordination: sequence use at lower levels characterised by short simple clauses (e.g. *yesterday I went to buy, then I decided to go*) and as proficiency increases the sequence is manipulated and woven into more complex subordinated patterns (e.g. *I visited the Art Gallery where I managed to see ...*)

7.7 Insights from comparing case studies 1 and 2: tense, context, register and theoretical alignment

Looking at two similar sequences has resulted in the following observations:

- The profile of lexical exponents differs depending on the choice of tense. The function of the sequence is affected by tense: 'preference' is the main function when used in a present simple sequence and 'volition' in the past simple sequence.
- As proficiency increases there is greater awareness of the effect of pronoun choice and tense, e.g. compare *I want to* which in certain contexts may be perceived as direct decreases at C levels, with *I wanted to* which takes on a pragmatic function of politeness.
- As proficiency increases, learners appear to do more with the same patterns, and demonstrate ability to use the same structure in different contexts with different meanings and collocational patterns.
- Learners appear to be sensitive to the register and task requirements.

These two sequences in this case study are examples where there are two open word class slots (tensed verb and verb-base) and two closed slots (pronouns and TO) (Table 7.21). In theory the verb slots (2 and 4) might be seen to provide the opportunity for a greater range of patterning of independent selection. In reality all slots are restricted by the syntagmatic context, and the wider context of the task. It appears, to varying degrees, from all levels in the data that there is a sensitivity to this.

tag position	1	2	3	4
tag	PP	VVD/VVP	TO	VV
slot	closed	open	closed	open
possible exponents	I you he she we they	all past simple / present simple verb forms	to	all verb-base forms

Table 7.21 Breakdown of occurrences by level of pronoun+past-simple+to-inf+verb-base

A brief analysis of these two sequences demonstrates the need to take a multi-faceted approach to development, looking beyond a form-meaning mapping taxonomy characteristic of pattern grammar (and VAC analysis).

It shows that development of structural generalisation is multi-layered: it is about understanding the nuts and bolts, the syntactic and lexical combinations and restrictions and the form-function mappings of a given sequence; it is also about understanding the requirements of the wider context, the time frames and task, selecting forms to fit functions, and fixedness of cotext.

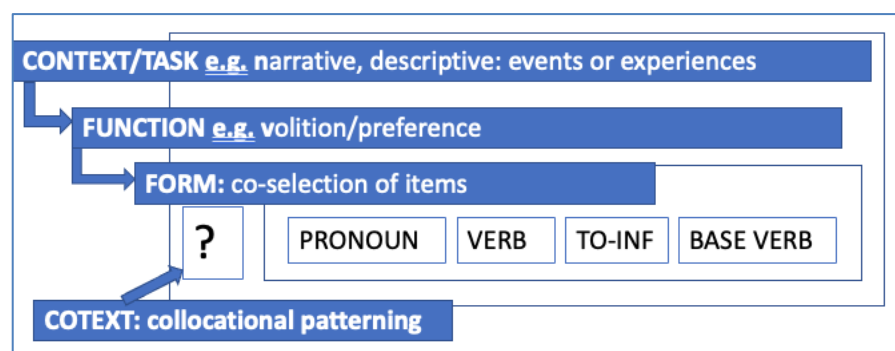


Figure 7.13 Context-Function-Form-Cotext description

In order to see what is happening we need to look beyond a simple n-gram sequence. The POS tag sequence offers a way in, a starting point to see where the patterns of form and function lie. An indepth qualitative analysis of the wider co-text and an understanding of the communicative context of the task is needed in order to gain a fuller picture of development. This is discussed further in Chapter 9.

7.8 B1 to B2: on the road

Overall, in this chapter, sequence change in the B levels has been observable in the convergence of core POS tags between B2 and higher levels and in a stabilisation of sequences at B2. The body of sequences that are high ranking and continue to be consistently highly ranked grows in number. There was evidence of a greater similarity in distribution of the POS tag sequences between A2 and B1 and between B2 and C1 than there is between B1 and B2, pointing to a leap in development at the B level, and a change in the types of sequence use; tensed verb sequences peak at B1 and noun phrase sequences increase through the B levels into the C levels. Through a case study analysis it became evident that as proficiency increased learners were able to do more with the same sequences and were sensitive to the fixedness of the patterning in the sequences as well as to the communicative demands of the task. Applying a pattern grammar categorisation approach proved to be useful for matching forms with meaning groupings for identifying some clustering around meanings. However, it was not able to provide a comprehensive framework for all instances of the sequences and aspects of the patterning as detailed above. The notion of the pathbreaking or pioneering sequence began to surface with sequences with *decided to* emerging as the sequence most associated with past volition, and sequences with *want to* carrying the strongest preference function. The appearance of sequences with *tend to* and *used to* at the C2 levels are also worthy of further investigation, as sequences carrying characteristics of habitual activities often associated with a present simple or past simple function.

At this stage, by the B2 level, learners have negotiated a great deal of linguistic territory. The next step, in Chapter 8, is to consider whether sequence change at C1 and C2 can throw additional light on development.

Chapter 8 Cruising: from C1 to C2

According to Cambridge Assessment website “a C2 Proficiency qualification shows the world that you have mastered English to an exceptional level”

<https://www.cambridgeenglish.org/exams-and-tests/proficiency/>. In the CEFR global scale learners at C1 and C2 levels are categorised as Proficient users with the following broad descriptions:

PROFICIENT USER	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.

Table 8.1 Global scale descriptors for C1 and C2 as defined by the Council of Europe

<https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale>

While it is not clear from these summaries where development or change might be observable between the two levels, there is a clear indication of a high degree of competence in the descriptions, such as ability to “produce clear, well-structured, detailed text on complex subjects”, to show “controlled use of organisational patterns” and to differentiate “finer shades of meaning even in more complex situations”.

This chapter seeks to investigate and describe development at this high level of proficiency. In the first part of the chapter, in line with the structure of chapters 6 and 7, I first explore if development is observable through the frequency and distribution of the highest-ranking POS sequences across proficiency levels C1 and C2 (RQ1). I then take a case study approach in sections 8.4 to 8.6. Aligning with RQ2, these will address how POS sequences develop across proficiency levels. The case studies will also attend to RQ3 and explore if existing frameworks for classification of language patterning account for a description of development.

8.1 Focusing in: overall distribution C1 and C2

I begin with initial observations about the POS 4-gram distribution across C1 and C2 levels. All POS 4-gram sequences were extracted from the C1 and C2 data. The total number of POS 4-gram sequence occurrences and total POS 4-gram sequence types per level are shown in Table 8.2:

	C1	C2
Total 4-gram raw occurrences: all types	6648802	7640531
Total 4-gram types	278605	299916

Table 8.2 Occurrences of POS 4-gram sequences across levels C1 and C2

All sequences were ranked and the rankings of the top 50 types from each level selected for further analysis and comparison. The total number of POS 4-gram sequence occurrences in the top 50 types can be seen in Table 8.3. These top 50 types constitute 0.02% of types of POS 4-grams in both the C1 and C2 data (Table 8.1), however because of their frequent usage, they account for 8.86% and 8.92% respectively of all POS 4-gram occurrences, making up almost 10% of the usage in each dataset (see the overall picture of distribution in chapter 5 Figure 5.1).

	C1	C2
Total 4-gram raw occurrences: top 50	589183	681367
50 types as % of all types	0.02	0.02
Total occurrences in top 50 as % of all	8.86	8.92

Table 8.3 Distribution of top 50 types across levels

Next, in sections 8.1.1 and 8.1.2, I examine changes in the distribution of sequence ranking of the top 50 sequences across C1 and C2 levels, taking both a retrospective and prospective view.

8.1.1 Overall distribution: top 50 C1 sequences

Here I focus on change from the perspective of the top 50 C1 sequences. A snapshot of this change can be seen in Figure 8.1, with specific focus within the red box. The colour coding gives a visual overview of the convergence in ranking between all levels, from dark green (highly convergent) to pink (highly divergent) (see key).

94% of the top 50 C1 sequences are also found within a rank of +/-30 at B2 (all three green sections in the C1:B2 column), with 34% ranked closely within a range of +/-5 (dark green section only).

When the C1 sequences are compared with their ranking at C2 we see that 90% are found within a rank difference of +/- 30 at C2 (all three green sections in the C1:C2 column), and there is even greater convergence when the C1 sequences are compared with those closely ranked, within a range of +/-5 (dark green only) which increases to 54%.

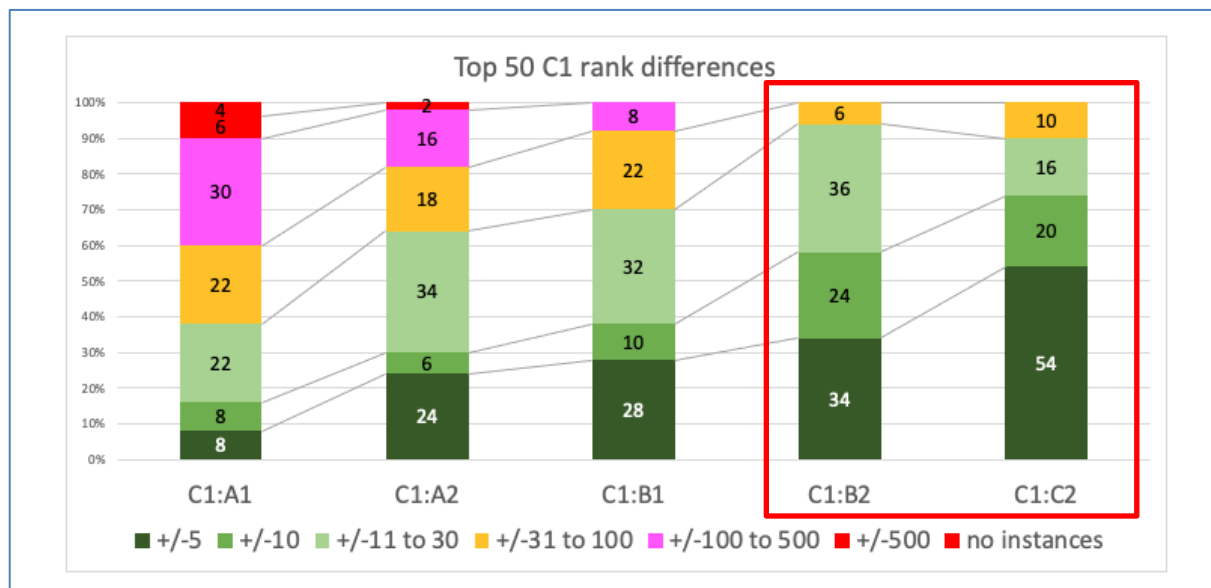


Figure 8.1 Percentage convergence of the top 50 sequences at C1 with their rankings at all other levels

There is a marked increase in the top 50 C1 sequences which are core to both C1 and C2. Through this 4-gram POS tag sequence lens, C1 writing looks closer in sequence use to the adjacent higher proficiency level C2 writing than to B2 writing. When looking at Figure 8.1

in its entirety it presents a strong visual picture of increasing convergence in the core sequences, and increasing stability of usage of all the sequences. Of the top 50 sequences at C1, 8% of them are within a ranking of +/-5 at A1, rising to 24% at A2, 28% at B1 34% at B2 and 54% at C2. Retrospectively, C1 data looks less and less like A1 data as proficiency increases and, prospectively, more and more like C2 data.

8.1.2 Overall distribution: top 50 C2 sequences

A picture of accumulating stabilisation and convergence continues when looking retrospectively at the top 50 sequences at C2 and comparing their rankings at previous lower proficiency levels (Figure 8.2). Of the top 50 C2 sequences, 100% are found within a rank difference of +/- 30 at C1, with 56% of these top 50 found to be consistently and closely ranked (with a difference of +/-5).

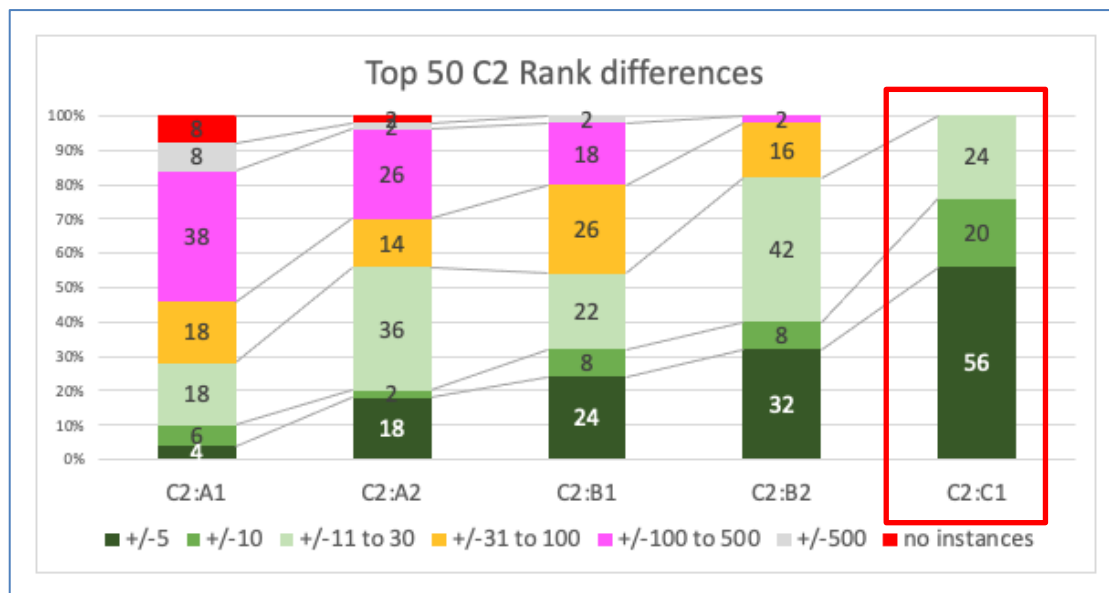


Figure 8.2 Percentage convergence of the top 50 sequences at C2 with their rankings at all other levels

Figure 8.2 illustrates that the highest ranking sequences converge as proficiency increases. This builds on the overall findings described in chapter 5. An increasing core of closely ranked sequences is evident. Retrospectively, from a POS tag sequence lens, C2 data looks less and less like A1 data as proficiency increases. In other chapters, I have shown each level (with the exception of A1) from both a rear and front view. It is, of course, possible to compare the C2 data with L1 data, if this is where the C2 level is 'headed', but the data would need to be comparable. To my knowledge there is no L1 data of L2 exam style tasks

such as the main suite Cambridge exams which would mean that any comparison would be flawed. I discuss this further in Chapter 9.

In the next two sections I look at the constituents in the sequences at C1 and C2 and identify changes in their usage. This provides a starting point for further exploration of lexical and functional characteristics.

8.2 C1 sequences

As described above, the rankings of the top 50 sequences at C1 were compared with their ranks at B2 and C2 and their relative rank variance calculated using a simple rank difference calculation.

Sequences with punctuation were removed, leaving 34 sequences (Table 8.4), of which 17 contain verbs. The remainder contain part or whole noun phrases. Three of these sequences (marked in blue) are new to the top 50 at C1, not occurring in the top 50 at the adjacent lower level (B2). 5 of the sequences (marked in red font) do not occur in the top 50 of the adjacent higher level (C2).

COLOUR KEY	
rank variance of +/-	
5	
10	
11 to 30	
31 to 100	
101 to 500	
501+	
#VALUE! = not found	

C1 rank	POS tag sequences and <i>examples</i>	C1-B2	C1-C2
1	noun prep det noun NN IN DT NN <i>aim of this report</i>	0	0
2	prep det adj noun IN DT JJ NN <i>on the other hand</i>	0	0
3	det noun prep det DT NN IN DT <i>the end of the</i>	-2	-2
4	prep det noun prep IN DT NN IN <i>at the end of</i>	0	1
5	det adj noun prep DT JJ NN IN <i>a wide range of</i>	-1	1

9	to-inf verb-base det noun TO VV DT NN <i>to find a job</i>	0	-1
10	det noun prep noun DT NN IN NN <i>a lot of money</i>	-8	3
11	adj noun prep det JJ NN IN DT <i>other side of the</i>	-12	0
12	pronoun modal adverb verb-base PP MD RB VV <i>I would also like</i>	-1	-3
13	plural-noun prep det noun NNS IN DT NN <i>parts of the world</i>	-13	0
14	noun prep det adj NN IN DT JJ <i>lunch in a typical</i>	-19	2
15	noun prep poss-pronoun noun NN IN PPZ NN <i>response to your letter</i>	1	1
16	verb-base det noun prep VV DT NN IN <i>spend a lot of</i>	-4	-1
18	proper-noun x4 NP NP NP (all results are capital letters or proper nouns)	6	-33
20	prep det noun prep IN DT NN <i>in the city centre</i>	-10	-6
22	prep det noun conj IN DT NN CC <i>in the morning and</i>	-6	3
24	-ed-form prep det noun VVN IN DT NN <i>given to the hospital</i>	-26	4
28	modal verb-base to-inf verb-base MD VV TO VV <i>would like to thank</i>	13	-27
29	det noun to-inf verb-base DT NN TO VV <i>the opportunity to meet</i>	-14	6
30	verb-base prep det noun VV IN DT NN <i>go for a walk</i>	8	-6
31	det noun prep plural-noun DT NN IN NNS <i>a lot of people</i>	-8	9

32	to-inf verb-base prep det TO VV IN DT <i>to go to the</i>	1	-3
33	pronoun modal verb-base to-inf PP MD VV TO <i>I would like to</i>	17	-38
34	adverb prep det noun RB IN DT NN <i>all over the world</i>	-11	5
35	pronoun modal verb-base det PP MD VV DT <i>you will find a</i>	8	-26
37	prep det noun pronoun IN DT NN PP <i>As a result I</i>	-4	-7
38	verb-base det adj noun VV DT JJ NN <i>play a musical instrument</i>	-15	6
39	det noun prep poss-pronoun DT NN IN PPZ <i>the rest of my</i>	-3	14
42	noun prep det plural-noun NN IN DT NNS <i>majority of the students</i>	-43	-1
45	adj noun prep noun JJ NN IN NN <i>sudden downpour of rain</i>	-68	17
46	prep det adj plural-noun IN DT JJ NNS <i>for a few days</i>	-58	12
47	pres-simple-verb det adj noun VBZ DT JJ NN <i>is a good idea</i>	-1	2
49	modal verb-base det noun MD VV DT NN <i>can learn a lot</i>	0	-13
50	pronoun modal verb-base pronoun PP MD VV PP <i>I would like you</i>	25	-86

Table 8.4 Top 50 4-gram POS tag sequences at C1, and their rank differences at B2 and C2

As in previous chapters the rank difference figures and colours indicate degrees of difference. The colour coding in the two right-hand columns gives a visual overview of the convergence in ranking between levels, from dark green (highly convergent) to pink (highly divergent) (see key). Negative rank difference figures indicate a lower ranking at the other levels, and positive figures indicate a higher ranking. The first coloured column shows the differences

between the C1 and B2 ranking and the second column shows the differences between the C1 and C2 ranking.

Overall results continue to point to three types of sequences: (1) core sequences (2) emerging sequences and (3) decreasing sequences. In the following sections I examine how the C1 and C2 data are characterised by these sequence types, before exploring examples of their lexical and functional characteristics.

8.2.1 Core sequences

There are 19 sequences that are highly convergent in ranking (within +/-5) at both C1 and C2 (Table 8.5). As with the core sequences seen in previous levels noun phrases continue to dominate and become consistently core to the C1 and C2 levels, with 18 of the 19 containing a noun phrase.

C1 rank	POS tag sequences and examples	C1-B2	C1-C2
1	noun prep det noun NN IN DT NN <i>aim of this report</i>	0	0
2	prep det adj noun IN DT JJ NN <i>on the other hand</i>	0	0
3	det noun prep det DT NN IN DT <i>the end of the</i>	-2	-2
4	prep det noun prep IN DT NN IN <i>at the end of the</i>	0	1
5	det adj noun prep DT JJ NN IN <i>a wide range of</i>	-1	1
9	to-inf verb-base det noun TO VV DT NN <i>to find a job</i>	0	-1
10	det noun prep noun DT NN IN NN <i>a lot of money</i>	-8	3
11	adj noun prep det JJ NN IN DT <i>other side of the</i>	-12	0
12	pronoun modal adverb verb-base PP MD RB VV <i>I would also like to</i>	-1	-3

13	plural-noun prep det noun NNS IN DT NN <i>parts of the world</i>	-13	0
14	noun prep det adj NN IN DT JJ <i>lunch in a typical</i>	-19	2
15	noun prep poss-pronoun noun NN IN PPZ NN <i>response to your letter</i>	1	1
16	verb-base det noun prep VV DT NN IN <i>spend a lot of</i>	-4	-1
22	prep det noun conj IN DT NN CC <i>in the morning and</i>	-6	3
24	-ed-form prep det noun VVN IN DT NN <i>given to the hospital</i>	-26	4
32	to-inf verb-base prep det TO VV IN DT <i>to go to the</i>	1	-3
34	adverb prep det noun RB IN DT NN <i>all over the world</i>	-11	5
42	noun prep det plural-noun NN IN DT NNS <i>majority of the students</i>	-43	-1
47	pres-simple-verb det adj noun VBZ DT JJ NN <i>is a good idea</i>	-1	2

Table 8.5 Core sequences: C1 sequences which are highly convergent in ranking at both C1 and C2.

The sequence ranked #42 NN IN DT NNS (noun+preposition+determiner+pluralnoun, e.g. *majority of the students*) is new in the top 50 at C1. It is not found in the top 50 at B2 (rank difference -43), and remains consistently ranked at C2.

The past participle -ed form sequence #24 VVN IN DT NN (-ed-form+prep+det+noun e.g. *given to the hospital*) which was first seen as an emerging sequence in the top 50 at B2 is also highly and consistently ranked at C1, as is one sequence with a tensed verb #47 VBZ DT JJ NN present-simple-verb+determiner+adjective+noun .

8.2.2 Emerging sequences

There are six emerging sequences in the C1 top 50 – those which rank higher at C2 than C1 and therefore become increasingly more important for C2 learners (Table 8.6). Noticeable here is the increase in the range of sequences containing noun phrases with adjectives, (#38, 46, 45) which are all new to the top 50 in C1.

C1 rank	POS tag sequences and <i>examples</i>	Rank difference	
		C1-B2	C1-C2
29	det noun to-inf verb-base DT NN TO VV <i>the opportunity to meet</i>	-14	6
38	verb-base det adj noun VV DT JJ NN <i>play a musical instrument</i>	-15	6
31	det noun prep plural-noun DT NN IN NNS <i>a lot of people</i>	-8	9
46	prep det adj plural-noun IN DT JJ NNS <i>for a few days</i>	-58	12
39	det noun prep poss-pronoun DT NN IN PPZ <i>the rest of my</i>	-3	14
45	adj noun prep noun JJ NN IN NN <i>sudden downpour of rain</i>	-68	17

Table 8.6 Emerging sequences: C1 sequences which are higher ranked at C2 than C1 (with rank difference).

8.2.3 Decreasing sequences

The sequences that indicate what C1 learners do more frequently in contrast to C2 learners are decreasing sequences, i.e. those that rank lower at C2 than at C1 (Table 8.7). There are 9 of these in the top 50 C1 sequences. At the top of the table are those that are the least used at C2 in relation to other sequences, and less relevant in the C2 repertoire, (to varying points of difference, shown by the rank difference figure). Those in red are not carried forward into the top 50 at C2. Noticeable here is the prevalence of sequences with modal verbs plus tensed verbs (#50, 33, 28, 35, 49). These sequences with modal verbs become less important in the

C2 data. *The sequence NP NP NP, #18 at C1, represents four capitalised nouns or proper nouns and is discounted here as the tagging is unreliable.

C1 rank	POS tag sequences and <i>examples</i>	rank difference	
		C1- B2	C1- C2
50	pronoun modal verb-base pronoun PP MD VV PP <i>I would like you</i>	25	-86
33	pronoun modal verb-base to-inf PP MD VV TO <i>I would like to</i>	17	-38
18	proper-noun x4 NP NP NP NP (sequences of capital letters)	6	-33
28	modal verb-base to-inf verb-base MD VV TO VV <i>would like to thank</i>	13	-27
35	pronoun modal verb-base det PP MD VV DT <i>you will find a</i>	8	-26
49	modal verb-base det noun MD VV DT NN <i>can learn a lot</i>	0	-13
37	prep det noun pronoun IN DT NN PP <i>as a result I</i>	-4	-7
20	prep det noun prep IN DT NN <i>in the city centre</i>	-10	-6
30	verb-base prep det noun VV IN DT NN <i>go for a walk</i>	8	-6

Table 8.7 C1 sequences decreasing in ranking at C2 (with rank difference).

From this initial analysis, overall characteristics of the C1 sequences include:

- a growing core of sequences convergent with the B2 and C2 levels indicating a stabilisation
- noun phrases containing adjectives, and plural nouns, are increasing in usage
- sequences containing modal verbs are decreasing in usage

8.3 C2 sequences

Moving on to the C2 data, the rankings of the top 50 sequences at C2 were compared with their ranks at C1 and their relative rank variance calculated using a simple rank difference calculation. As this is the highest proficiency level there is no front view, only a rear view perspective of development in relation to the previous level. The three sequence types, core, emerging and decreasing, continue to be observable in the C2 data.

Sequences with punctuation were removed, leaving 35 sequences (Table 8.8). All of the sequences contain noun phrases, and ten contain verbs some of which are constituents of the noun phrases. Seven of these sequences (marked in blue) are new to the top 50 at C2, not occurring in the top 50 at the adjacent lower level (C1), and five of these contain adjectives.

COLOUR KEY	
rank variance of +/-	
5	
10	
11 to 30	
31 to 100	
101 to 500	
501+	
#VALUE! = not found	

C2 rank	POS tag sequences and <i>examples</i>	Rank difference C2-C1
1	noun prep det noun NN IN DT NN <i>aim of this proposal</i>	0
2	prep det adj noun IN DT JJ NN <i>at the same time</i>	0
3	prep det noun prep IN DT NN IN <i>in the middle of</i>	-1
4	prep det adj noun DT JJ NN IN <i>a great deal of</i>	-1
5	det noun prep det DT NN IN DT <i>the aim of this</i>	2
7	det noun prep noun DT NN IN NN <i>a lot of time</i>	-3
10	to-inf verb base det noun TO VV DT NN <i>to get a job</i>	1
11	adj noun prep det	0

	JJ NN IN DT <i>other side of the</i>	
12	noun prep det adj NN IN DT JJ <i>cinema for a weekly</i>	-2
13	plural-noun prep det noun NNS IN DT NN <i>parts of the world</i>	0
14	noun prep poss-pronoun noun NN IN PPZ NN <i>response to your article</i>	-1
15	pronoun modal adverb verb-base PP MD RB VV <i>I will never forget</i>	3
17	verb-base det noun prep VV DT NN IN <i>improve the quality of</i>	1
19	prep det noun conj IN DT NN CC <i>of the world and</i>	-3
20	-ed-form prep det noun VVN IN DT NN <i>created by this situation</i>	-4
22	det noun prep plural-noun DT NN IN NNS <i>a lot of people</i>	-9
23	det noun to-inf verb-base DT NN TO VV <i>the opportunity to learn</i>	-6
25	det noun prep poss-pronoun DT NN IN PPZ <i>the rest of their</i>	-14
26	prep det noun noun IN DT NN NN <i>with an internet café</i>	6
28	adj noun prep noun JJ NN IN NN <i>short period of time</i>	-17
29	adverb prep det noun RB IN DT NN <i>all over the world</i>	-5
32	verb-base det adj noun VV DT JJ NN <i>play an important role</i>	-6
34	prep det adj plural-noun IN DT JJ NNS <i>of the main reasons</i>	-12
35	to-inf verb-base prep det TO VV IN DT <i>to go to the</i>	3

36	verb-base prep det noun VV IN DT NN <i>go for a walk</i>	6
38	noun prep adj plural-noun NN IN JJ NNS <i>number of old people</i>	-20
40	prep poss-pronoun adj noun IN PPZ JJ NN <i>in our everyday life</i>	-12
42	prep noun prep det IN NN IN DT <i>in response to the</i>	-14
43	noun prep det plural-noun NN IN DT NNS <i>majority of the people</i>	1
44	prep det noun pronoun IN DT NN PP <i>as a result they</i>	7
45	verb-is det adj noun VBZ DT JJ NN <i>is the only way</i>	-2
46	adj plural-noun prep det JJ NNS IN DT <i>many parts of the</i>	-15
48	adj prep det noun JJ IN DT NN <i>due to the fact</i>	-3
49	prep noun to-inf verb-base IN NN TO VV <i>in order to make</i>	-25
50	det noun prep adj DT NN IN JJ <i>a lot of different</i>	-22

Table 8.8 Top 50 4-gram POS sequences at C2, and their rank differences at C1

8.3.1 Core sequences

There are 20 core sequences that are highly convergent in ranking (within +/-5) in the C1 and C2 data (Table 8.9) indicating a high degree of stabilisation between sequence usage in these two proficiency levels:

C2 rank	POS tag sequences and examples	Rank difference C2-C1
1	noun prep det noun NN IN DT NN <i>aim of this proposal</i>	0

2	prep det adj noun IN DT JJ NN <i>at the same time</i>	0
3	prep det noun prep IN DT NN IN <i>in the middle of</i>	-1
4	prep det adj noun DT JJ NN IN <i>a great deal of</i>	-1
5	det noun prep det DT NN IN DT <i>the aim of this</i>	2
7	det noun prep noun DT NN IN NN <i>a lot of time</i>	-3
10	to-inf verb base det noun TO VV DT NN <i>to get a job</i>	1
11	adj noun prep det JJ NN IN DT <i>other side of the</i>	0
12	noun prep det adj NN IN DT JJ <i>cinema for a weekly</i>	-2
13	plural-noun prep det noun NNS IN DT NN <i>parts of the world</i>	0
14	noun prep poss-pronoun noun NN IN PPZ NN <i>response to your article</i>	-1
15	pronoun modal adverb verb-base PP MD RB VV <i>I will never forget</i>	3
17	verb-base det noun prep VV DT NN IN <i>improve the quality of</i>	1
19	prep det noun conj IN DT NN CC <i>of the world and</i>	-3
20	-ed-form prep det noun VVN IN DT NN <i>created by this situation</i>	-4
29	adverb prep det noun RB IN DT NN <i>all over the world</i>	-5
35	to-inf verb-base prep det TO VV IN DT <i>to go to the</i>	3
43	noun prep det plural-noun	1

	NN IN DT NNS <i>majority of the people</i>	
45	verb- <i>is</i> det adj noun VBZ DT JJ NN <i>is the only way</i>	-2
48	adj prep det noun JJ IN DT NN <i>due to the fact</i>	-3

Table 8.9 Core sequences: C2 sequences which are highly convergent in ranking at C1.

Only one of the sequences was not also in the top 50 C1 sequences. At rank #48 we see JJ IN DT NN (adjective+preposition+determiner+noun e.g. *due to the fact that*), barely missing the top 50, ranking at #51 at C1. This closeness in ranking of these core most frequently used sequences suggests a consistency in C1 and C2 learners' abstraction of structural regularities, which will be discussed further in Chapter 9.

8.3.2 Emerging sequences

There are 12 sequences which rank higher at C2 than at C1. These are sequences which have become increasingly more important for C2 learners (Table 8.10):

C2 rank	POS tag sequences and <i>examples</i>	Rank difference C2-C1
49	prep noun to-inf verb-base IN NN TO VV <i>in order to make</i>	-25
50	det noun prep adj DT NN IN JJ <i>a lot of different</i>	-22
38	noun prep adj plural-noun NN IN JJ NNS <i>number of old people</i>	-20
28	adj noun prep noun JJ NN IN NN <i>short period of time</i>	-17
46	adj plural-noun prep det JJ NNS IN DT <i>many parts of the</i>	-15
25	det noun prep poss-pronoun DT NN IN PPZ <i>the rest of their</i>	-14
42	prep noun prep det IN NN IN DT <i>in response to the</i>	-14
34	prep det adj plural-noun	-12

	IN DT JJ NNS <i>of the main reasons</i>	
40	prep poss-pronoun adj noun IN PPZ JJ NN <i>in our everyday life</i>	-12
22	det noun prep plural-noun DT NN IN NNS <i>a lot of people</i>	-9
23	det noun to-inf verb-base DT NN TO VV <i>the opportunity to learn</i>	-6
32	verb-base det adj noun VV DT JJ NN <i>play an important role</i>	-6

Table 8.10 Emerging sequences: C2 sequences which are higher ranked at C2 than C1 (with rank difference).

Six of these sequences are new to the top 50 at C2 #49, 50, 38, 46, 42, 40 (indicated in blue), the remainder were already emerging in the C1 data and are continuing to rise in ranking at C2 as they become more frequently used. Noun phrases with adjectives continue to dominate these emerging sequences, and noticeably there is an increase in plural noun use. These noun phrase sequences are characterised by both fragments of phrases, e.g. #50 DT NN IN JJ *a lot of different*, #46 JJ NNS IN DT *negative aspects of the*, #42 IN NN IN DT *in response to the* as well as sequences which constitute ‘complete’ phrases, #22 DT NN IN NNS *a lot of people* and #23 DT NN TO VV *the opportunity to learn*. Given the fact that both types are equally high ranking, they are both considered to be of interest in the C1 and C2 repertoire, both equally contributing evidence for the frequent use of preformulated routines and building blocks. See Chapter 9 for a discussion on ‘completeness’.

8.3.3 Decreasing sequences

There are only three sequences which decrease in ranking at C2 in comparison with C1, and the change in ranking is minimal (Table 8.11). This contributes to the picture of stabilisation that is emerging, at least from a frequency and distribution perspective, between the C2 and C1 levels.

C2 rank	POS tag sequences and examples	C2-C1
	Rank difference C2-C1	C1
26	prep det noun noun IN DT NN NN <i>with an internet café</i>	6
44	prep det noun pronoun IN DT NN PP <i>as a result they</i>	7
36	verb-base prep det noun VV IN DT NN <i>go for a walk</i>	6

Table 8.11 C2 sequences decreasing in ranking at C2 in comparison with C1 (with rank difference).

From this initial analysis, overall characteristics of the top 50 C2 sequences include:

- strong convergence and stabilisation of use between C1 and C2 sequences, evident through an increasing body of core sequences and a decrease in divergent sequences
- an increase in noun phrases in which nouns are premodified with adjectives and a wider range of determiners
- an increase in noun phrases with post-modifying non-finite verb sequences and prepositional phrases
- a scarcity of sequences containing tensed verbs

8.3.4 A developmental picture: summary from the C level perspective

The picture which emerges from the C1 and C2 level data is one of stabilisation and consolidation, building on the stabilisation of sequences which was beginning to become evident at B2. So far in this chapter I have looked at the distribution of POS tag sequences in both the C1 and C2, looking forward from C1 to C2 and back from C2 to C1. We have seen overall, in structural terms, that:

- There is a high degree of convergence between the sequences used at both levels, with an increasing body of core sequences. There is a decrease in divergent sequences and those which are seen to decrease in usage from C1 to C2 do with minimal divergence in ranking.
- Noun phrases continue to be on the increase, including those with adjectives, plural nouns, post-modifying prepositional phrases and post-modifying non-finite clauses.
- Following a trend seen at B2 level, sequences containing verbs are on the decrease, particularly those with tensed verb forms and modal verbs.

8.3.5 Background to case study selection

As was the case in Chapters 6 and 7 when we looked at A and B level sequences, in this first phase of analysis, comparing sequence ranking and distribution has revealed changes which warrant further investigation. In the following sections I explore the lexical and functional exponents of representative core and emerging sequences. Firstly two sequences with pronoun + tensed verb + to-inf+ verb-base PP VVD TO VV (*I started to cry, I decided to go*) PP VVP TO VV (*I want to apply, you need to go*).

The following sections (8.4 to 8.6) explore research questions RQ2 and RQ3 firstly to examine how representative sequences develop across proficiency levels and in doing so explore whether existing frameworks for classification of language patterning account for this development.

In these case studies the lexical and functional properties of the following three sequences are investigated:

- (1) IN NN IN DT prep + noun + prep + det, e.g. *in response to the*
- (2) VVN IN DT NN -ed-form + prep + det + noun, e.g. *created by this situation*
- (3) DT NN TO VV det + noun + to-inf + verb-base e.g. *the opportunity to meet*

These three sequences are of interest as they represent diverse sequences:

- (1) incorporates a prepositional phrase fragment, beginning with a preposition and ending with a determiner
- (2) is one of the few most frequent sequences at C1 and C2 that contains a verb form
- (3) is an example of a noun phrase post-modified by a non-finite verb

All three are examples of sequences which have increased considerably in frequency of usage as proficiency increases, and which as such, contribute to defining the C level data.

8.4 Case study 1: prep + noun + prep + det (IN NN IN DT)

IN NN IN DT prep + noun + prep + det, (e.g. *in response to the*) is an example of an emerging sequence in the top 50 sequences at C2 level. It is ranked at #42 at C2, increasing its ranking from #393 at A1, #208 at A2, #189 at B1, #89 at B2, and #56 at C1. It has been selected as it illustrates development of a repeated pattern occurring with a prepositional phrase fragment.

8.4.1 Occurrences by level: prep + noun + prep + det (IN NN IN DT)

Table 8.12 shows the breakdown of the raw and relative occurrences of this sequence over all six levels as well as the percentage of occurrences covered by the top 1000 types for each level.

	subcorpus size	raw occurrences	relative PMW occurrences	total occurrences	1000 types as % occurrences
				1000 types	
A1	2456971	872	355	872	100
A2	5703217	3052	535	2997	98
B1	3261473	1778	545	1778	100
B2	5263979	3817	725	2897	76
C1	6711568	5998	894	4117	69
C2	7698695	7607	988	4969	65

Table 8.12 Breakdown of occurrences by level of prep + noun + prep + det

As indicated in previous chapters, the change in the % figure in the right-hand ‘1000 types as % of occurrences’ column indicates that the number of types for this sequence increases with proficiency, i.e. the range of different lexical exponents for the sequences increases. 1000 types make up 65% of all types used at C2. The relative PMW occurrences show a steady increase in frequency of relative occurrences as proficiency increases. While initial indications in terms of frequency suggest increased usage between B1 and B2 and beyond, these frequency figures need to be complemented with a qualitative view and an analysis of the lexical patterning and functional profile.

8.4.2 Lexical and functional distribution by level: *prep + noun + prep + det (IN NN IN DT)*

The top 50 lexical exponents were extracted for all levels and reviewed, using the KWIC and concordance functions in Sketch Engine. This section provides a summary overview of the findings. The top 20 lexical exponents for all levels are shown in Table 8.13. A crude overview illustrates that the exponents in the A1, A2, and B1 data are dominated by topic-related results (shaded in grey), sequences referring to place (shaded in green, e.g. *in front of the*) as well a limited range of sequences which function as linking devices (unshaded, e.g. *in spite of this, with reference to*), while in B2, C1 and C2 there is increasing use of additional (unshaded) linking devices.

A1	A2	B1	B2	C1	C2
					in response to the
in front of the	in front of the	in front of the	in front of the	in front of the	
by bus to the	in front of a	in front of a	In addition to this	In addition to this	in front of the
by car to the	to school in the	In addition to this	in front of a	on behalf of the	In addition to this
in front of a	In front of the	on television on the	on behalf of the	in response to the	in front of a
of fun at the	by bus to the	on holiday with some	In spite of the	in front of a	on behalf of the
of fun in the	in front of that	in spite of the	in spite of the	with regard to the	in touch with the
at home at half	in front of this	In spite of the	with reference to the	In spite of the	In addition to that
by taxi to the	in centre of the	in love with a	in connection with the	In addition to that	with regard to the
by bus because the	by car to the	In front of the	In spite of this	in spite of the	in response to an

to meeting about the	in contact with the	on TV on the	with regard to the	with reference to the	On top of that
for go to the	in love with the	on holiday with the	in touch with the	in touch with the	in spite of the
by train because the	in front to the	In spite of this	in response to the	in charge of the	In addition to the
of information about the	of time in the	on top of the	In addition to that	in connection with the	with reference to the
to shopping after the	in fron of the	in love with the	On behalf of the	On top of that	In spite of the
in front off the	in love with a	in contact with the	per cent of the	In addition to the	in favour of the
on top of the	for example in the	on holiday with both	In addition to all	in contact with the	in response to a
to music at the	In spite of the	With reference to the	In addition to these	In spite of this	in contact with the
by train to the	like go to the	in touch with the	With reference to the	per cent of the	of life in the
for help with the	at night in the	with regard to the	in charge of the	In addition to these	in charge of the
on foot to the	In addition to this	of time in the	in love with a	With reference to the	in love with the
after school at the	in front of The	in spite of this	in contact with the	On behalf of the	on top of the
after school in the	on top of the	on holiday to the	in reference to the	in case of an	in connection with the
on foot because the	For example in the	in favour of the	of money for the	in favour of the	in relation to the

by bus at the	for go to the	on holiday for a	in addition to this	in relation to the	in order for the
on wednesday in the	of town because the	of time on the	In spite of that	In addition to all	in love with a

Table 8.13 Top 20 lexical exponents for IN NN IN DT for all six levels

The most frequently occurring sequence *in front of the* dominates the lower proficiency level results, with this one sequence constituting 46% of all occurrences at A1 and A2 and 21% of results at B1. It continues to drop at B2, to 10.3% of occurrences, dipping to 5.7% and 5.03% of all results at C1 and C2. Analysis of the concordance lines and collocational patterns at all levels shows that, at the A levels, the most common collocations are places in a town (*in front of the cinema/bank/supermarket*) while at C levels they are concrete items situated in a location (e.g. *in front of the/a television/computer/screen/TV*), reflecting topic and task at both ends of the proficiency scale. This provides evidence of usage of this sequence as a routinised pattern to refer to ‘person or thing at location’ as early as A1.

With the exception of *in front of the*, the lower levels are dominated by fragments spanning two phrases, for example, *of fun at the, to school in the*, shown in the extracts from the A level data below.

Extract 8.1

Hi Ally, We had a lot **of fun at the** party. (A1, KET 2004)

Extract 8.2

Next Tuesday we must change the time, because I have to go **to school in the** morning. (A2, PET 2008)

These are in stark contrast with the increasing range of routinised fragments that begin to appear at the B levels and dominate C levels (Table 8.13). In the next section I explore their functional behaviour.

8.4.3 Applying a functional categorisation

There are no equivalent patterns in the Pattern Grammar taxonomy for this type of prepositional phrase fragment. In lexical bundle (LB) terminology they fall within the category of referential bundles, and predominantly with a subcategorization of ‘intangible framing attributes’ (Biber *et al.* 2004, p.387, Biber 2006, p.159), e.g. *in addition to the*.

However many of the sequences found in the learner data are not accounted for in the LB taxonomy. A closer look at the sequences found in the C1 and C2 data shows that they play a predominantly cohesive role. In an attempt to find generalisations, I devised and applied the following simple taxonomy:

category	function	example
ref_link	a sequence which signals or links to something or someone within or out of the text	<i>in response to the</i>
ref_place	a sequence which points to or specifies a place	<i>in front of the</i>
add_link	a sequence which signals and links to additional information	<i>in addition to the</i>
contrast_link	a sequence which signals and links to contrast	<i>in spite of the</i>
purpose_link	a sequence which signals and links to a purpose or explanation	<i>in order for the</i>
topic	a sequence which is recurrent because it is topic-related	<i>by bus to the</i>

Table 8.14 Functional taxonomy for IN NN IN DT sequences

The taxonomy was applied to the top 25 at each of the B2, C1 and C2 levels and their usage at each level is shown in Table 8.15:

		B2	C1	C2
add_link	<i>In addition to this</i>	*	*	*
	<i>In addition to that</i>	*	*	*
	<i>In addition to all</i>	*	*	
	<i>In addition to these</i>	*	*	
	<i>in addition to this</i>	*		
	<i>In addition to the</i>		*	*
	<i>On top of that</i>		*	*
contrast_link	<i>In spite of the</i>	*	*	*
	<i>in spite of the</i>	*	*	*
	<i>In spite of this</i>	*	*	
	<i>In spite of that</i>	*		
	<i>in favour of the</i>		*	*
ref_link	<i>in response to the</i>	*	*	*
	<i>in response to an</i>			*
	<i>in response to a</i>			*
	<i>on behalf of the</i>	*	*	*
	<i>with regard to the</i>	*	*	*
	<i>with reference to the</i>	*	*	*
	<i>in reference to the</i>	*	*	
	<i>in connection with the</i>	*	*	*
	<i>in relation to the</i>		*	*
ref_place	<i>in front of the</i>	*	*	*
	<i>in front of a</i>	*	*	*
	<i>on top of the</i>			*
purpose_link	<i>in order for the</i>			*

Table 8.15 Breakdown of functions for the top 25 B2, C1, C2 IN NN IN DT sequences

Observations from this analysis are that:

- *In addition to all* features in the top 25 at B2 (4.0 times PMW) and C1 (3.4 times PMW) but decreases in usage at C2 (2.1 times PMW), while *in addition to the/that/this* persists in use, particularly in sentence initial position to link back and introduce additional information:

Extract 8.3

It is a city of great historical heritage and natural beauty too. **In addition to this**, local people are more than friendly and hospitable with foreigners, something that will be a guarantee for the conference's success (C2, CAE, 2002)

- There is an increase in the repertoire of the *add_link* category: *On top of that* appears at C1 (6.0 times PMW) and increasingly C2 (7.4 times PMW), with the same additional linking function, again in sentence initial position:

Extract 8.4

It might be true that what is considered to be healthy or unhealthy changes over time, however, or science progresses, the margin of error becomes smaller, and experts are then able to have a more accurate and precise say on all matters, including health. It is safe to say that at the present moment, science is at a level that is advanced enough for it to be taken seriously. **On top of that**, the consensus on what is healthy may also be modified according to the ever changing habits of humanity and evolutionary factors. (C2, CPE, 2006)

- *In favour of the* appears in the top 25 at C1 (3.7 times PMW) and increases in use at C2 (6.2 times PMW)
- There is an awareness of the subtle collocational patterning and structural generalisations within the *ref_link* sequences where the selection of the preposition *with/in* varies and *with/in response/reference regard/relation to* and a settling of this pattern as proficiency increases.
- *In response to the/a/an* becomes the dominant sequence as a cohesive device for an external reference at C2. At C2 the *in response to* the sequence appears consistently with a specialised function after *I am writing* and collocates strongly with *campaign, article, comments, announcement, letter(s), discussion, invitation* whereas the less frequent *with reference to* has two collocates: *article* and *radio programme*:

Extract 8.5

Dear Sir I am writing **in response to the** recent article published in your newspaper concerning education (C2, CPE, 2003)

- There is an additional function at C2, with the inclusion of a sequence expressing a link to purpose, *in order for the*. The patterning after this particular sequence shows an understanding of how it affects the subsequent syntactic pattern resulting in a noun + to+infinitive structure, as shown in the concordance lines in Figure 8.3:

the advantages and disadvantages of such a move ,	in order for the	final decision to prove beneficial both to our firm and its employ
DT NNS CC NNS IN PDT DT NN ,	IN NN IN DT	JJ NN TO VV JJ CC IN PPZ NN CC PPZ NNS
great achievers .	in order for the	exhibition to best reflect his achievements , a sample of a long
JJ NNS SENT	IN NN IN DT	NN IN JJS VVP PFZ NNS , DT NN IN DT JJ
more , alternative sources of energy should be used	in order for the	conventional ones not to disappear .
JJ NNS IN NN MD VB VVN	IN NN IN DT	JJ NNS RB TO VV SENT
my firm conviction that you should take some action	in order for the	museum to be improved .
PPZ NN NN IN/that PP MD VV DT NN	IN NN IN DT	NN TO VB VVN SENT
We should change the time we start our meetings	in order for the	students to have time for dinner and come on time .
PP MD VV DT NN PP VVP PPZ NNS	IN NN IN DT	NNS TO VH NN IN NN CC VV IN NN SENT
One has to show the sense in being tidy	in order for the	child to accept it as a useful behaviour .
PP VHZ TO VV DT NN IN VBG JJ	IN NN IN DT	NN TO VV PP IN DT JJ NN SENT
I am willing to embrace .	in order for the	development to take place , some changes should happen , not r
PP VBP JJ TO VV SENT	IN NN IN DT	NN TO VV NN , DT NNS MD VV , RB
ould be in harmony with the way people live nowadays ,	in order for the	city , to achieve a result of a modern image .
VB IN NN IN DT NN NNS VVP RB	IN NN IN DT	NN , TO VV DT NN IN DT JJ NN SENT
om 10 am to 8 p.m. It should stay open more hours	in order for the	students to come and train after their lessons .
N CD VBP IN CD NN PP MD VV RP JJR NNS	IN NN IN DT	NNS TO VV CC VV IN PPZ NNS SENT
is the main ingredients used while preparing dishes ,	in order for the	visitors to understand the importance of this aspect .
VBZ DT JJ NNS VVN IN VVG NNS ,	IN NN IN DT	NNS TO VV DT NN IN DT NN SENT
some recommendations that are felt of significance	in order for the	film studio to be improved .
DT NNS WDT VBP VVN IN NN	IN NN IN DT	NN NN TO VB VVN SENT
I believe that there is nothing that should be done	in order for the	programme to become better and of a better quality .
P VVP IN/that EX VBZ NN WDT MD VB VVN	IN NN IN DT	NN TO VV RBR CC IN DT JJR NN SENT

Figure 8.3 Concordance lines of the *in order for the* sequence

In summary, a brief analysis of this sequence is revealing. It illustrates a refinement in the use of sequences at C2 in the *add_link*, *contrast_link* and *ref_link* categories, alongside an increase in the functions (*purpose_link*).

It shows that users are increasingly sensitive to cohesive demands of the co-text and context. They demonstrate this internally within a sequence, externally beyond the sequence to the selection of patterning beyond the sequence and to the wider textual cohesion.

It demonstrates sensitivity to the demands of register and specialisation of the functions of fixed and semi-fixed sequences.

It demonstrates an increasing awareness of word co-selection, moving from independent lexical choice at the lower levels, to a fixedness of patterning at the higher levels.

8.5 Case study 2: -ed-form + prep + det + noun (VVN IN DT NN)

The second case study looks at a sequence which is core to both C1 and C2: VVN IN DT NN -ed-form + prep + det + noun (e.g. *created by this situation*). It is ranked at #20 at C2,

increasing its ranking from #407 at A2, #211 at B1, #50 at B2, and #24 at C1, with no occurrences at A1. It has been selected as it illustrates development of a pattern occurring with a past participle verb phrase. As we have seen in previous chapters verb phrases occur less frequently as proficiency increases.

8.5.1 Occurrences by level: *-ed form + prep + det + noun (VVN IN DT NN)*

Table 8.16 shows a breakdown of the raw and relative occurrences of this sequence over all six levels as well as the percentage of occurrences covered by the top 1000 types for each level.

	subcorpus size	raw occurrences	relative PMW occurrences	total occurrences	1000 types as % occurrences
				1000 types	
A1	2456971	299	122	299	100
A2	5703217	1789	314	1449	81
B1	3261473	1673	513	1239	74
B2	5263979	5167	982	1972	38
C1	6711568	9208	1372	2689	29
C2	7698695	11206	1456	2768	25

Table 8.16 Breakdown of occurrences by level of *-ed form + prep + det + noun*

The relative PMW occurrences show a leap in frequency of occurrences between B1 and B2, with almost twice as many occurrences at B2 than B1. Another leap between B2 and C2 shows this sequence increases by almost 50%. This is also reflected in the % of occurrences found in the top 1000 types. The % figures in the right-hand column suggest that this sequence becomes more and more productive as proficiency increases. It is composed of 4 slots, two of which are open (VVD and NN) and which in theory can allow any *-ed* verb form and any noun form and two which are closed (IN and DT).

8.5.2 Lexical and functional distribution by level: *ed-form + prep + det + noun (VVN IN DT NN)*

When the individual lexical exponents are extracted and analysed in more detail, the frequencies are small, as shown in Table 8.17 which gives the top 20 for C1 and C2. However given that the sequences are composed of four elements, two of which are open

word classes, a wide range of lexical exponents is to be expected. Despite the low occurrences, some generalisations emerge.

The data shows a mix of lexical exponents displaying degrees of fixedness between the elements. Some of the prepositions (in the IN slot) following the -ed forms (VVN slot) are fixed collocates (*included/mentioned in* (+publication), *satisfied with* (+amenity), *located/situated in* (+place)). These are highlighted in bold. The focus here is on the differences between C1 and C2, rather than across all levels. More of the C2 sequences are whole or part of fixed formulaic sequences (e.g. *come to the conclusion*, *come to an end*, *based on the fact*, *stuck in a traffic*). These are shaded in grey.

C1	Freq	PMW	C2	Freq	PMW
<i>given to the hospital</i>	67	9.98	<i>created by this situation</i>	65	8.44
<i>included in the course</i>	49	7.30	<i>given to the hospital</i>	33	4.29
<i>satisfied with the transport</i>	32	4.77	<i>stuck in a traffic</i>	31	4.03
<i>located in the centre</i>	28	4.17	<i>come to the conclusion</i>	28	3.64
<i>mentioned in the advertisement</i>	27	4.02	<i>situated in the centre</i>	25	3.25
<i>arranged at the hotel</i>	23	3.43	<i>mentioned in the article</i>	20	2.60
<i>situated in the centre</i>	23	3.43	<i>located in the centre</i>	20	2.60
<i>included in the price</i>	22	3.28	<i>included in the price</i>	19	2.47
<i>come to the conclusion</i>	21	3.13	<i>based on the fact</i>	16	2.08
<i>written in the advertisement</i>	19	2.83	<i>killed in a car</i>	15	1.95
<i>included in the offer</i>	17	2.53	<i>raised in the article</i>	15	1.95
<i>satisfied with the accommodation</i>	15	2.23	<i>caused by the noise</i>	14	1.82
<i>offered in the college</i>	15	2.23	<i>situated in the middle</i>	14	1.82
<i>entered for an exam</i>	14	2.09	<i>come to an end</i>	13	1.69
<i>satisfied with the staff</i>	14	2.09	<i>located in the city</i>	13	1.69
<i>included in the programme</i>	13	1.94	<i>included in the exhibition</i>	13	1.69
<i>created by this situation</i>	12	1.79	<i>combined with the fact</i>	12	1.56

<i>played for an hour</i>	12	1.79	<i>stuck in the traffic</i>	12	1.56
<i>satisfied with the location</i>	11	1.64	<i>caused by the fact</i>	12	1.56

Table 8.17 Top 20 VVN IN DT NN sequences at C1 and C2

Three of the top 20 at C2 sequences end with *the fact*. Further exploration of this reveals that 61.5% of all occurrences of -ed form + prep + *the fact* are found at C2, 30.1% at C1 and the rest at B2. 96% of these are followed by a that-clause. The most frequent exponents are *based on the fact that / attributed to the fact that / combined with the fact that / caused by the fact that*.

Extract 8.6

Nevertheless, there is also another point of view, which supports the opposite idea. This argument is **based on the fact that** what we know about famous people is not always everything they are. (C2, CAE 2013)

Extract 8.7

What this reader believes is that cars offer people security and privacy, **combined with the fact that** they are extremely useful. (C2, CPE 2010)

The sequence *based on the fact that* would appear to have a specialised function of justifying an opinion or giving an explanation, and provide frames for what is to follow. They serve to create cohesion within the text and to provide signals to the reader. Other semi-fixed sequences with ed-form + preposition + *the fact that* sequences also demonstrate this function. While C1 level data also contains this usage, it appears less frequently.

What appears to be happening between the C1 and C2 level is an increasing use of prefabricated, routinised patterns, with specialised functions of these sequences such as *based on the fact that, come to the conclusion, come to an end, stuck in a traffic (jam)*.

On a separate note, this sequence highlights a limitation of the POS tagging. The VVN tag also results in words which might otherwise have been tagged as adjective, e.g. *satisfied*. The issue of tagging is discussed again in Chapter 9.

8.6 Case study 3: det + noun + to-inf + verb-base (DT NN TO VV)

The final case study is an example of an emerging sequence. It has been selected for further investigation as an example of a sequence containing a noun phrase post-modified with a non-finite verb. DT NN TO VV det + noun + to-inf + verb-base (e.g. *the opportunity to meet*)

is ranked at #518 at A1, jumping to #175 at A2, #110 at B1, into the top 50, #43, at B2, #29 at C1 and #23 at C2. It becomes increasingly more important to the C level repertoire.

8.6.1 Occurrences by level: *det + noun + to-ing + verb base (DT NN TO VV)*

Table 8.18 shows a breakdown of the raw and relative occurrences of the sequence *det + noun + to-ing + verb base (DT NN TO VV, e.g. the opportunity to meet)* over all six levels as well as the percentage of occurrences covered by the top 1000 types for each level.

	subcorpus size	raw occurrences	relative PMW occurrences	total occurrences	1000 types as %
				1000 types	occurrences
A1	2456971	703	286.12	703	100.00
A2	5703217	3368	590.54	2465	73.19
B1	3261473	2387	731.88	1658	69.46
B2	5263979	5737	1089.86	3207	55.90
C1	6711568	8504	1267.07	4350	51.15
C2	7698695	10420	1353.48	5048	48.45

Table 8.18 Breakdown of occurrences by level of *det + noun + to-inf + verb base*

The relative PMW occurrences show a leap in frequency of occurrences between A1 and A2, and between B1 and B2, with almost a third as many occurrences at B2 than B1 (5737 vs 2387). Another leap between B2 and C2 shows that this sequence increases by a third again (5737 vs 10420). The % figures in the right-hand column suggest that this sequence becomes more productive as proficiency increases.

However in comparison with other structures, it is less productive. For example the top 1000 types of the sequence in case study 2 *VVN IN DT NN* (e.g. *situated in the centre, came to the conclusion*) at C2 level constitute only 25% of all types, whereas, for this case study 3 sequence, the top 1000 types at C2 constitute almost half of all types (48.5%). This warrants further investigation.

The sequence *DT NN TO VV* (*the opportunity to meet*) contains two closed word class slots DT (determiner) and TO (to-inf), and two open slots, NN (noun) and VV (verb base). With

two open noun and verb slots there is scope for a wide range of lexical items. However closer analysis of the lexical patterning suggests that the noun (NN) slot in this sequence is semi-fixed, with a narrow range of candidates. This is explored in 8.6.2.

8.6.2 Lexical distribution by level: *det + noun + to-inf + verb-base (DT NN TO VV)*

Analysis of the top 50 at levels C1 and C2 showed a preference for repeated lexical sequences, in particular *the opportunity to + verb* (shaded in green) *the chance to + verb* (in blue). Other repeated sequences also include *a lot/place/way to + verb*. The top 20 are illustrate in Table 8.19:

C1	Freq	PMW	C2	Freq	PMW
<i>the opportunity to meet</i>	61	9.17	<i>the opportunity to meet</i>	71	9.29
<i>the chance to meet</i>	56	8.42	<i>the opportunity to learn</i>	60	7.85
<i>the opportunity to learn</i>	53	7.97	<i>a lot to offer</i>	56	7.33
<i>the opportunity to see</i>	49	7.37	<i>a place to stay</i>	51	6.67
<i>no time to see</i>	41	6.17	<i>the chance to meet</i>	50	6.54
<i>a place to stay</i>	40	6.02	<i>a place to live</i>	48	6.28
<i>the chance to learn</i>	40	6.02	<i>the chance to learn</i>	44	5.76
<i>the opportunity to take</i>	40	6.02	<i>the opportunity to get</i>	43	5.63
<i>the opportunity to visit</i>	40	6.02	<i>the opportunity to go</i>	43	5.63
<i>the opportunity to go</i>	40	6.02	<i>the chance to see</i>	41	5.37
<i>the chance to visit</i>	35	5.26	<i>the opportunity to see</i>	38	4.97
<i>the opportunity to get</i>	34	5.11	<i>the opportunity to do</i>	38	4.97
<i>the chance to do</i>	32	4.81	<i>the opportunity to express</i>	38	4.97
<i>the chance to see</i>	31	4.66	<i>this letter to express</i>	38	4.97
<i>the opportunity to improve</i>	31	4.66	<i>the opportunity to visit</i>	37	4.84
<i>a lot to do</i>	30	4.51	<i>the opportunity to enjoy</i>	37	4.84
<i>the chance to get</i>	29	4.36	<i>the proposal to build</i>	35	4.58
<i>a lot to offer</i>	29	4.36	<i>a lot to do</i>	34	4.45
<i>the opportunity to travel</i>	28	4.21	<i>a way to relax</i>	33	4.32
<i>the opportunity to do</i>	26	3.91	<i>the opportunity to travel</i>	33	4.32

Table 8.19 Top 20 C1 and C2 lexical exponents of *det + noun + to + verb*

The initial view revealed some obvious generalisations and fixed patterning in the first three slots in the 3-gram POS tag sequence DT NN TO. To get a narrower picture of functional development the distribution of this 3-gram was analysed across A2 to C2 levels. Figure 8.4 shows a snapshot of this, illustrating how the top 10 C2 lexical exponents are distributed across other levels in percentage terms. The figures in the graph indicate the % of their distribution across all DT NN TO occurrences at each level, for example *the opportunity to* constitutes 12.92% of all occurrences at C2 and 1.74% of all A1 occurrences of this 3-gram sequence. The top 10 at C2 make up 40.15% of all occurrences of this 3-gram, decreasing to 13.3% at A2.

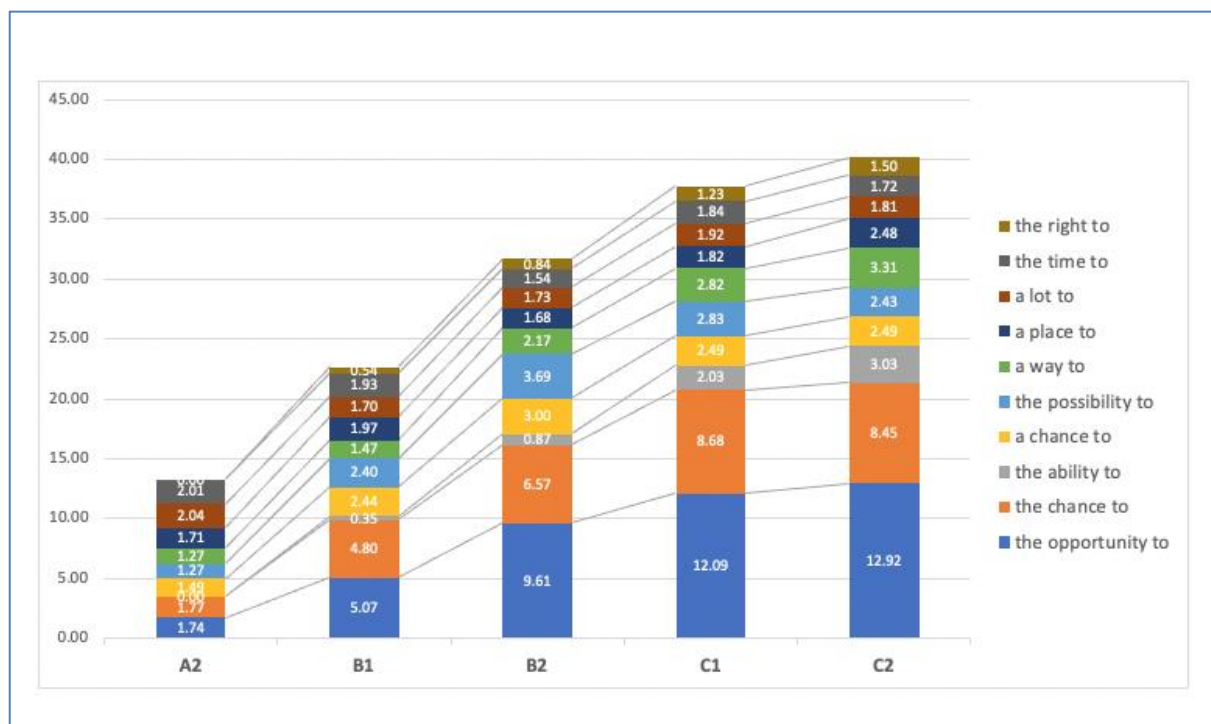


Figure 8.4 Top 10 C2 DT NN TO sequences distributed across A2-C2 levels

Some initial observations show that there is a steady increase in the use of:

- *the opportunity to* across all levels, with it becoming the dominant form at C2.
- *the chance to* until C1, with a slight decrease at C2.
- *the possibility to* until C1, with a decrease at C2.
- *a chance to* until B2, with a decrease at C1 and C2.
- *a way to* and *the right to* across all levels.

The remaining sequences are fairly consistent in their distribution.

The distribution of frequencies of use become more and more Zipfian as proficiency increases, with frequency dropping as the ranking of each form decreases, illustrated in Figure 8.5. In this graph the top 10 C2 DT NN TO (det + noun + to-inf) sequences are plotted across all levels. The dotted power line shows a good fit to the C2 data. It appears that, in line with UB theory, as proficiency increases one of two forms become pioneering / pathbreaking forms for this sequence (see Chapter 3).

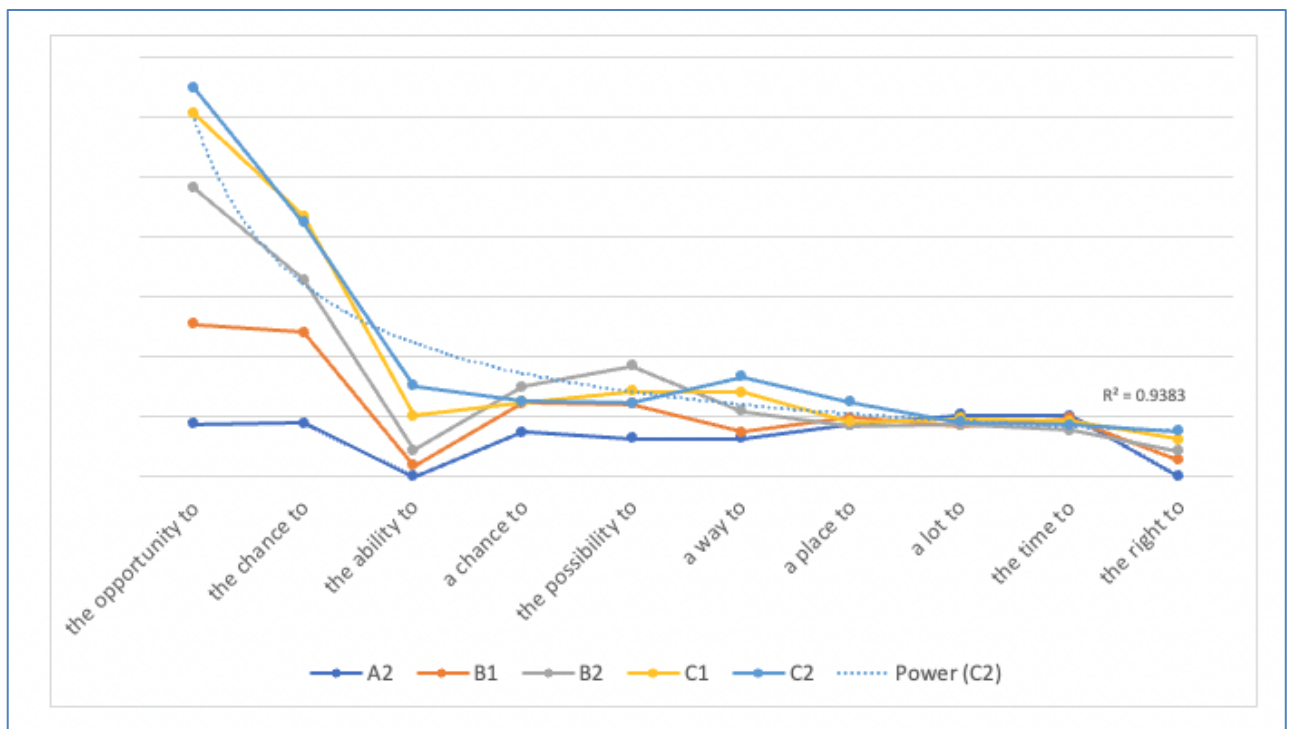


Figure 8.5 Distribution of top 10 C2 DT NN TO sequences across all levels

8.6.3 Functional distribution by level: *det + noun + to-inf + verb-base (DT NN TO VV)*

As with previous sequences a pattern grammar approach (Hunston and Francis 2000) was used to look at their functional profile. This pattern falls into the N + to-inf pattern, and there are 15 meaning groups within this pattern (desire, arrangement, promise, proposal, attempt, ability, permission, request, responsibility, reason, tendency, claim, nouns with other meanings, productive uses, other related patterns)

<https://grammar.collinsdictionary.com/grammar-pattern/n-to-inf>

The B2, C1 and C2 lexical sequences were categorised according to the pattern grammar meaning groups, as sample of this is illustrated in Table 8.20:

pattern grammar meaning group	B2	C1	C2
ability	<i>the opportunity to</i>	<i>the opportunity to</i>	<i>the opportunity to</i>
	<i>the chance to</i>	<i>the chance to</i>	<i>the chance to</i>
	<i>the possibility to</i>	<i>the possibility to</i>	<i>the ability to</i>
	<i>a chance to</i>	<i>a chance to</i>	<i>a chance to</i>
	<i>the time to</i>	<i>the ability to</i>	<i>the possibility to</i>
	<i>an opportunity to</i>	<i>the time to</i>	<i>the time to</i>
	<i>no time to</i>	<i>an opportunity to</i>	<i>an opportunity to</i>
	<i>the ability to</i>	<i>no time to</i>	<i>the power to</i>
	<i>a possibility to</i>	<i>some time to</i>	<i>some time to</i>
		<i>a possibility to</i>	
permission	<i>the right to</i>	<i>the right to</i>	<i>the right to</i>
desire	<i>no need to</i>	<i>no need to</i>	<i>the need to</i>
		<i>the need to</i>	<i>no need to</i>
pleasure	<i>a pleasure to</i>		
productive use	<i>a lot to</i>	<i>a lot to</i>	<i>a place to</i>
	<i>a place to</i>	<i>a place to</i>	<i>a lot to</i>
nouns with other meanings	<i>a way to</i>	<i>a way to</i>	<i>a way to</i>
	<i>the way to</i>	<i>the way to</i>	
attempt			<i>an effort to</i>
			<i>an attempt to</i>
			<i>the courage to</i>
uncat_TOPIC	<i>this letter to</i>	<i>this letter to</i>	<i>this letter to</i>
	<i>a company to</i>	<i>the world to</i>	<i>a person to</i>
	<i>the idea to</i>	<i>a pleasure to</i>	
	<i>no problem to</i>		

Table 8.20 Functional categorisation of Top 20 DT NN TO sequences

Looking more closely at the lexical realisations of the meaning groups, we see that the dominant *ability* group, B2 learners used *a chance*, *the opportunity/chance/possibility to* + verb, whereas at C2 there are decreasing instances of *the possibility to* + verb and a reliance

on *the opportunity to verb*. The *possibility* sequences do not appear in the pattern grammar categorisation, suggesting that they are not frequent sequences in the COBUILD L1 data that pattern grammar was based on. The decrease at C2 suggests that C2 users are sensitive to this usage or lack of. There is also an increase in the C2 ability group examples with the addition of *the power to* and an increase in the meaning groups with addition of the attempt group. Overall for C2 writers, there seemed to be a narrowing in on a more fixed formula in *the opportunity to*, alongside a broadening of verbs when looking beyond the 3-grams. C2 learners appear to do more with the same pattern (a type of grammatical polysemy). Several of the lexical exponents did not fit into any categorisation (labelled uncat_TOPIC) and appear to be sequences that span two phrases and are generated by the topic.

For all meaning groups, beyond the top 20, there was an overall movement away from task or topic-oriented, often concrete, head nouns, towards increased use of abstract or figurative ‘shell’ nouns (Hunston and Francis, 2000) in semi-fixed frames (e.g. *a lot to do/learn, the time to do, a proposal to build, the right to live*).

In terms of the development of form-meaning pairings within this emerging pattern at C2, results from the top 20 suggest that C2 learners:

- can do more with the same patterns.
- deploy more form-meaning mappings.
- show a tendency for one or two ‘pioneering’ forms and shed less frequent forms.
- rely on more semi-fixed structures and less topic-oriented language.

8.6.4 Applying pattern grammar: *det + noun + to-inf + verb-base (DT NN TO VV)*

The application of pattern grammar as a framework for functional analysis seems to be largely successful at this level. It has limitations in that the patterns are not categorised in terms of their frequency and no indication of the frequency of the group members is given.

For example the group containing *opportunity* gives no indication of the frequency of its members.

N to-inf

The 'ability' group

These nouns refer to the ability or opportunity to do something, or the inability to do something.
*She is confident in her **ability** to cope with whatever weather conditions may arise.*
*Children should have the **opportunity** to explore ideas, follow their interests, expand their horizons.*
*He knows exactly what he is saying and will soon have the **power** to see his wishes through.*
*Always make concessions while you still have **room** to manoeuvre.*

• ability	• facility	• means	• power
• capability	• freedom	• opportunity	• room
• capacity	• inability	• option	• scope
• chance	• incapacity	• potential	• space
	• instinct		• time

Figure 8.7 N to-inf pattern and ability meaning group

https://grammar.collinsdictionary.com/grammar-pattern/n-to-inf_7

It tends to overgeneralise, grouping some of the verbs into generic categories such as the productive group (*a lot to*) and nouns with other meanings (*a way to*). Additionally it does not account for a higher level conceptual meaning of this structure, which sits above the individual groups: C1 and C2 learners appear to be sensitive to the modal nature of noun + to-inf (e.g. *the need to, an attempt to, an ability to, an opportunity to, the chance to, the power to*). Many of the meaning groups for the N to-inf pattern identified in pattern grammar carry a modal function, e.g. ability, attempt, permission, This requires further exploration.

8.7 C1 to C2: Summary

As evidenced, at the C levels there is continued convergence of usage of POS tag sequences, a continued growth in lexical and functional usage and a developing awareness of fixedness of patterning in relation to specificity of meaning.

This concludes the detailed analysis of levels A to C, reviewing patterns of convergence and divergence across all levels. In the final chapter I take stock of the findings and discuss some of the insights and considerations for future research.

Chapter 9 Discussion and conclusions: Mapping the routes

This chapter returns to the aims of the study and the research questions posed in Chapter 1 and summarises how they have been answered. It then considers the limitations of the study and the avenues for future investigation that this research has uncovered.

9.1 Recapping: aims of the study

This research seeks to bring together elements of second language acquisition studies and corpus linguistics methodology, proposing a bottom-up data-first approach to shed light on second language development in largescale data. One of the driving forces of this study is methodological, to explore a way to capture, on a global level, how L2 English learners put together sequences of words from early stages of proficiency to advanced levels and to try to observe and map out how structure emerges through development. It investigates the usage-based notion of “structural regularities which emerge” from a lifetime of making sense of the distributional characteristics of language experience (Ellis 2013, p.89). While acknowledging the role of frequency in usage-based theories, the approach in this research is intended to be led by the data, being corpus-driven, with no preselection of items for exploration.

The study is deliberately exploratory. It uses POS tag sequences as a starting point and trawls the 52-million-word CEFR-benchmarked Cambridge Learner Corpus from the outermost syntactic layer available in corpus tools, as a means to observe learner language change and development across the proficiency levels. In using POS tags, one could argue, as Biber does, that the approach is not entirely driven by the data, since POS tags “assume the existence of some grammatical classes (e.g. verb, nouns) and basic syntactic structures” (Biber 2010, p.202) which may in themselves be perceived as having an element of preselection. However, while not a perfect solution, it does allow for the extraction of recurrent structural generalisations. It takes a mixed methods approach, first examining the frequency and distribution of POS sequences by level, identifying convergence and divergence in ranking of the sequences, and secondly looking qualitatively at form-meaning mappings of these sequences. It seeks to observe if there are sequences which characterise levels and the transition between levels, and explores whether an analysis of the accumulation of their use at a lexical and functional level can contribute to our understanding of how a generic repertoire of learner language develops. It aims to contribute to the theoretical debate by looking critically at current theories and descriptions of language development. It responds to the call to look at largescale learner data, and benefits from privileged access to such longitudinal

data, acknowledging the limitations of any corpus data and the need to triangulate across different datasets.

It set out to explore the following research questions:

RQ1 Is development in L2 writing observable through the frequency and distribution of POS sequences across proficiency levels?

RQ2 How does POS sequence usage develop across proficiency levels?

RQ3 Can existing frameworks for classification of language patterning account for a description of development in L2 writing?

In the next three sections (9.2, 9.3, 9.4) I summarise how the study has responded to the three research questions.

9.2 RQ1 Is development in L2 writing observable through the frequency and distribution of POS sequences across proficiency levels?

This study has adopted an understanding of ‘development’ along an axis of quality, as discussed in chapter 2 (Durrant *et al.* 2021). The implication here is that changes between increasing levels of proficiency in L2 are developmental in nature. This understanding is set in the context of a usage-based premise that language learning and usage is frequency based. Taking both a prospective *front view* and retrospective *rear view* of usage, Chapter 5 gave an overall perspective on how the rank distribution of the frequency of POS tag sequences changes across A1 to C2. It set out to explore a novel methodology. It showed how taking a bottom-up approach, capturing all 4-tag POS sequences across six CEFR levels of proficiency, facilitates an open view on development. From this analysis three types of sequences were observable: core, emerging and decreasing. These three sequence types continued to be clearly observable in the analysis offered in chapters 6 to 8 in which levels A, B and C were explored in more detail. These observations showed evidence of how movement through proficiency levels involves the restructuring of the frequency and distribution of these sequences. Change across levels, and therefore development, was characterised by a body of convergent sequences which grew as proficiency levels increased. The change in sequence usage showed that, at each level, learners acquired a sense of what is core at the next level, what was more useful to them and what became less useful. Figures 9.1, 9.2 and 9.3 give an overall summary of these core, emerging and decreasing sequence changes.

9.2.1 Core sequences

Figure 9.1 shows evidence of a growing body of core sequences from A1 to C2. These core sequences are those that are found in the top 50 at each level and are highly convergent in ranking with the next adjacent higher level and therefore frequently used. (Sequences in blue are those that are new to the top 50 in each level, not appearing in the top 50 of the previous level.)

A1 rank	A2	A2 rank	A1	B1	B1 rank	A2	B2	B2 rank	B1	C1	C1 rank	B2	C2	C2 rank	C1		
2	PP MD VV IN	-4	1	NN IN DT NN	-9	0	1	NN IN DT NN	0	1	1	NN IN DT NN	0	0	1	NN IN DT NN	0
11	NN IN PPZ NN	0	2	IN DT JJ NN	-60	-1	3	IN DT JJ NN	1	0	2	IN DT JJ NN	0	0	2	IN DT JJ NN	0
19	VV IN DT NN	5	4	IN DT NN IN	-11	-1	5	IN DT NN IN	1	-1	4	IN DT NN IN	-2	-2	3	IN DT NN IN	-1
26	PP MD VV DT	2	9	DT JJ NN IN	-100	0	7	DT NN IN DT	-9	2	5	DT NN IN DT	-2	2	4	DT JJ NN IN	-1
29	PP VBP VVG TO	-5	11	NN IN PPZ NN	0	-2	9	DT JJ NN IN	0	3	6	DT JJ NN IN	-3	1	5	DT NN IN DT	2
37	VBP VVG TO VV	-4	18	PP MD VV PP	-28	0	10	TO VV DT NN	-18	1	9	TO VV DT NN	0	-1	9	TO VV DT NN	-3
			24	PP MD VV DT	-2	-4	13	NN IN PPZ NN	2	-2	13	PP MD RB VV	-6	1	10	DT NN IN NN	1
			31	DT NN IN NN	-53	-1	25	VV DT NN IN	-11	5	14	NN IN PPZ NN	1	-1	11	JJ NN IN DT	0
			33	PP VVP PP MD	-62	2	28	PP MD VV DT	4	1	20	VV DT NN IN	-5	4	12	PP MD RB VV	-2
			42	IN DT NN CC	-137	2	39	IN DT NN PP	-23	1	31	TO VV IN DT	-14	-1	13	NNS IN DT NN	0
											41	IN DT NN PP	2	4	14	NN IN DT JJ	-1
											42	DT NN IN PPZ	7	3	15	NN IN PPZ NN	1
											48	VBZ DT JJ NN	-25	1	16	VV DT NN IN	1
											49	MD VV DT NN	-20	0	22	IN DT NN CC	-3
														24	VVN IN DT NN	-4	
														32	TO VV IN DT	-5	
														34	RB IN DT NN	3	
														42	NN IN DT NNS	1	
														47	VBZ DT JJ NN	-2	
														48	JJ IN DT NN	-3	

Figure 9.1 Core sequences across all levels

One key observation relevant to development shows that there is a core of consistently used sequences which is seen to grow as proficiency increases, implying a growing understanding of which sequences are most used and therefore most useful, a statistical structuring and restructuring of usage. They lend evidence to how abstraction, frequency and statistical learning takes place in the process of language learning, aligning with usage-based theory.

This core of top ranking sequences is identical across all the levels from A2 onwards.

Another key point is that they are dominated by noun sequences which remain highly ranking and become increasingly relevant to the developing repertoire as proficiency increases. Core sequences containing verbs (particularly modal verbs) are characteristic of the A1 repertoire with 5 of the 6 core sequences containing verbs. Sequences containing tensed verbs peak at B1 and become less and less important as proficiency increases, with only one core sequence containing a tensed verb remaining in the core sequences at C2 (e.g. VBZ DT JJ NN *is a great opportunity*).

There is stabilisation of a limited range of core forms between A2 and B1, and a leap in development between B1 and B2. There is accumulation and increase in the core sequences between B2 and C1 and a slight increase between C1 and C2. There is greater convergence between the highly ranked sequences at C2 and other levels, than between the highly ranked sequences at A1 and other levels, i.e. other levels are observed to be gradually aligning with the C2 top rankings.

Key observations relevant to development show that there is a steady increase in emerging sequences in adjacent levels from A1 to B2 until C1 where there appears to be a stabilisation, where there are already more sequences that are core to C1 and C2. At C2 more new sequences emerge in the top 50. At B2 there is an increase in the number and range of sequences that have emerged at this level which were not high ranking in the adjacent lower level B1, once more confirming a leap in development between B1 and B2. The emerging sequences from B2 onwards are predominantly sequences containing nouns.

9.2.3 Decreasing sequences

Figure 9.3 shows the change in decreasing sequences. These are sequences that are in the top 50 at each level but decrease in ranking in comparison with the next adjacent higher level decrease in rank and therefore become less used. The sequences in red are those that do not appear in the top 50 of the next adjacent level.

Key observations relevant to development show that as proficiency increases divergence decreases. Each level becomes more and more like the next level. Decreasing sequences are dominated by those containing verbs, indicating that sequences with noun phrases become more and more dominant in the repertoire of higher proficiency levels.

A1 rank	A2	A2 rank	A1	B1	B1 rank	A2	B2	B2 rank	B1	C1	C1 rank	B2	C2	C2 rank	C1
9	PP VVP TO VV	6	4	-6	4	-9	-8	12	8	-6	18	6	-33	26	6
12	VV IN PPZ NN	14	-5	-15	12	6	-24	15	-27	-13	20	-10	-6	36	6
16	IN PPZ NN IN	15	6	-7	18	0	-7	16	-25	-17	28	13	-27	44	7
18	PP VHP TO VV	25	-11	-18	22	7	-12	22	-7	-8	30	8	-6		
22	CD NN "" NN	30	-31	-15	27	-19	-33	25	7	-25	33	17	-38		
24	IN CD : CD	34	5	-16	31	-2	-35	27	-1	-8	35	8	-26		
30	PP VVD DT NN	35	-184	-25	34	-6	-38	34	12	-29	37	-4	-7		
31	PP VBP VVG IN	37	-337	-10	35	-9	-7	36	24	-24	49	0	-13		
35	IN CD NN ""	38	-43	-17	36	-47	-10	46	10	-36	50	25	-86		
41	NN IN CD NN	39	-27	-15	38	-346	-197								
48	MD VV IN PPZ	41	4	-15	48	-38	-40								
49	PP VVP DT NN	45	-8	-40	50	16	-73								
50	DT NN VBZ IN	49	-61	-76											

Figure 9.3 Decreasing sequences across all levels

9.2.4 Summary of RQ1

Summary: sequence change from A1 to C2

- The nature of the observed changes in sequence usage across levels provides evidence for the emergence of regularities, and sensitivity to structural conventions, through a structuring and restructuring of developing language systems.
- The sequences that are core to adjacent levels increase as proficiency increases.
- Convergence begins to emerge between A2 and B1. There is greater stabilisation of usage between A2 and B1 than between A1 and A2.
- Sequences with nouns and noun phrases in high ranking positions start to dominate the highest ranks, at A2. They are less prevalent at A1, pointing to an increase in noun phrase development from A1 to A2.
- Some sequences with modal verbs and present progressive forms are core to A1 and A2 and start to become less central to B1 repertoire, though sequences with modal verbs are the most frequent and consistently highly ranked sequences with verbs at all levels.
- Sequences containing verb phrases other than modals increase at B1. Tensed verbs following pronouns peak at B1 but begin to decrease at B2.
- There is greater similarity in distribution of the POS tag sequences between A2 and B1 and between B2 and C1 than there is between B1 and B2, pointing to a leap in development at the B level.
- Noun phrases containing adjectives increase at B1 and continue to increase to C levels.
- There is an increase in core sequences with greater syntactic complexity at B2, including post-modified noun phrases.
- Noun phrases continue to be on the increase in the C levels, including those with adjectives, plural nouns, post-modifying prepositional phrases and post-modifying non-finite clauses.

The overwhelming conclusion is that development across levels is observable through the analysis of frequency and distribution of POS-tag sequences and a growing core of convergent usage. Adjacent higher proficiency levels show overall greater convergence than non-adjacent levels. As outlined, these findings suggest that learners are sensitive to structural regularities in the language input, proposed by a usage-based theory.

They also point to a need to examine the status of the noun sequence in relation to development. Current valuable UB research on development has tended to centre on verb-based sequences as the object of focus (Ellis *et al.* 2016). However initial conclusions from RQ1 indicate that verb-based sequences decrease as proficiency increases. Noun-based sequences increase as proficiency increases. The B1 level is a turning point where verb-based and noun-based frequencies come together

In the next section I address how sequence usage develops in terms of formal and functional terms.

9.3 RQ2 How does POS sequence usage develop across proficiency levels?

In many ways the observations relating to RQ1 partly provide answers to RQ2. However there are two levels of development observable in POS sequence usage: changes in form and changes in function. Added to this, the first level of change involves not only distributional changes in POS sequences as seen in 9.2 above but also changes in the lexical exponents of those sequences. This second research question considers both quantitative and qualitative findings relating to lexical exponents and their functions. These lexical exponents are also subject to changes in frequency and distribution. RQ2 has been addressed through analysis of a representative sample of case studies at each level. It needs pointing out that the observations in this section represent part of a much greater picture.

9.3.1 From filling slots to using frames to abstracting formulae

Lexical development: types and tokens

The representative case studies have shown that as learners move through levels of proficiency, development is observable through both the lexical growth and functional growth of POS tag sequences. First and foremost lexical growth and diversity is statistically observable in each study when comparing the proportion that the top 1000 types constitutes of the total number of occurrences. The pattern that is consistently observable is a steady decrease in percentage proportion as proficiency increases, pointing to an increasing repertoire of types.

Pioneering sequences

Alongside this there is evidence of an understanding of the restrictions on the co-selection of lexis as well as the emergence of pathbreaking or pioneering sequences. Depending on the composition of the POS tag sequence some of the elements of a sequence have a greater

number of potential candidates than others, as seen, for example, in the noun sequence NN IN DT NN (*centre of the town, aim of this report*). At one level of abstraction, there are several observable structural regularities, such as noun *in the/a* noun, noun *on the/a* noun, noun *of the/a* noun, etc. As we have seen at the A1 level the sequence occurrences are shared across several of these forms (e.g. noun *in the* noun *clock in the morning*, noun *about the* noun *meeting about the concert*, noun *of the* noun *price of the ticket*) (Chapter 5). However from A2, one exponent, noun *of the/a* noun (e.g. *centre of the town*) becomes ‘pioneering’, and continues to increase until at C2, it constitutes 66% of all forms. In lexical terms, A level learners rely on a limited range of exponents, often relating to topic where individual ‘slots’ in the sequence are open, and the compositionality of the sequence is transparent (*table in the kitchen, middle of the town*), and they employ a limited range of functions, using words as building blocks, mostly driven by the prepositional meaning (e.g. *in* defining location). By B2/C1 level, lexical choice has expanded for each functional category and additional functions are employed. However within this expansion, there also appears to be a filtering process, as we observe one lexical form for each function taking the lion’s share (in this case by C2 *aim of the proposal* dominates *purpose of the proposal*).

At lower levels the prepositional phrase tends to have an adverbial function as clausal elements whereas by C levels they appear to be fixed constituents of the noun phrase (e.g. *table in the kitchen* (A2) compared with *aim of the proposal* (C1/C2))

Fixedness of patterning

Additionally at C2 a movement towards fixedness of patterning and co-selection is observable, with for example, the increased use of shell nouns with post-modifier (*majority of the population*), which seems to suggest that, at C2 level, learners are sensitive to the collocational restrictions. Alongside this as sequences of words become more fixed and more formulaic so do their functions become more specific and this is observable in the emergence of fixed and semi-fixed phraseological exponents with specialised meanings and functions (e.g. C2 examples (*without a*) *shadow of a doubt*, (*in*) *spite/view of the fact*, (*have a*) *whale of a time*, (*the other*) *side of the coin*).

Slots and frames to formulaic expressions

These characteristics of development are seen again and again, as represented in the case studies and illustrated by the following sequence containing a past tense verb VVD IN DT NNN (*went to the cinema, arrived at the airport*). The forms used at A2 are predominantly

limited to a formula 'went to the/a + noun', with 19 of the top 20 most lexical exponents of VVD IN DT NN containing this sequence of words. At B2 there is an increase in the range of verbs (*included, arrived, knock*), prepositions and nouns in each of the syntactic 'slots'. At C2, there is an even greater lexical range in all slots, along with the introduction of some formulaic sequences, seen in the delexical use of 'came' (*came to the conclusion; came as a surprise, came to no surprise, came as a shock*), a feature not observed in the top 20 at A2 and B2.

On a semantic or functional level in the top 20, the 'went to the/a' sequence performs a 'movement to place' function, with a very specific fixed formula at A2. At B2 a greater range of verbs of movement is seen in the 'movement to place' function, alongside an increase in other functions; at C2 the 'went to the' sequence continues to be seen as the most dominant sequence, though with a lower number of examples, while the range of functions increases.

To generalise, as part of development, at A2 and B1 we see independent paradigmatic choices at a POS item level. Looking forward beyond B1, a pioneering form and function for a sequence starts to stabilise at B2 and dominates the most highly ranking lexical sequences by C2. Although we see a variety of candidates continuing to 'fill' the POS tag slots at B2 and C1, there is increasing distillation of 'slot candidates' so that by C2 level there is evidence, on the one hand, of increasingly specialised functions alongside increasing fixedness and constraint on the selection and combination of lexical items, with concrete formulas giving way to more figurative formulaic sequences. This lends evidence to the usage-based notion of the development of a syntactic slot and frame system to a fully abstracted system of 'constructions' (Ellis *et al.* 2018). As proficiency increases learners appear to show greater ability to abstract linguistic patterns from the input they are exposed to. This builds from low levels where learners first identify holophrastic lexical sequences (such as *I'd like* or *I went to the*). They then identify which structural units can go together and try out lexical items to fill these structural units (slots and frames) while developing a growing understanding of the frequency of use of items in slots, and degrees of fixedness between sequences of lexical exponents. A growing understanding of fixity (e.g. *a huge amount of* selected over *a big amount of*) might lend evidence to greater abstraction of the fine detail of the characteristics of patterning in the input.

9.3.3 *From description to evaluation: topics and framing*

General functional development is also observable, as illustrated by the case studies. A representative example DT JJ NN IN (det + adj + noun + prep) demonstrates how sequences are often dominated by topic-based lexical exponents at the lower levels, prompted by task (e.g. *the yellow door in / a black shirt for* vs *a good idea for / a little bit of*). It could be argued that it is the task which results in the high level of topic-related exemplars and the recency effect of encounter with rubric and topic. However it is possible that lower level learners hold on to topic and topic vocabulary because that is what they have in their repertoire. The building blocks of adjective+noun as in these examples, suggest that, at A levels, the adjective modifies the noun and performs a descriptive function. At higher levels the adjective+noun combination is part of a larger frame, which performs a series of functions: for example, setting the scene for the ‘main event’, guiding the reader through the discourse as in (*the main aim of the proposal is*); signalling an evaluation (*a good idea for, a great opportunity to*); signalling quantity (*a great deal of, a huge amount of*). This functional development seen in movement through levels in the data is characterised by a steady increase in frequency from topic-based building blocks with a descriptive function to referential and discourse-organising bundles with framing and quantifying functions. It displays a growing awareness of the sensitivity of register and requirements of task, an avenue for further investigation. It also raises questions about the nature of development and its dependence on input. Do L2 learners need to ‘pass through’ a literal descriptive, topic-based experience of language as building blocks before being able to pull out a repertoire of prefabricated and sometimes figurative language routines with a range of specialised functions? Are these topic-dependent stages reflective of L2 instruction or L2 development?

9.3.4 *Beyond the sequence: future work*

There is a further aspect of usage development evident from the data. Alongside the lexical and functional, there is growing evidence of the sensitivity to genre, as well as awareness of the needs of the wider text, beyond the sequences under observation. In a subsequent study (Mark, forthcoming) we see evidence that as proficiency increases learners gain more and more understanding of the discourse-management, orientation and signposting needs of writing.

9.4 RQ3 Can existing frameworks for classification of language patterning account for a description of development in L2 writing?

Quite simply, yes and no. On a case by case basis or level by level basis, we have observed, from the selected case studies, that some of the existing frameworks offer classification for some of the language patterning, and some offer a classification which can be applied to development. None offer adequate classification of development of language patterning across all levels and all sequences. Next I look at the frameworks applied in turn.

9.4.1 Applying existing frameworks for analysis: lexical bundle and p-frames

We have observed a tendency for verb-based sequences at lower levels and noun based sequences at higher levels. Evidence for this was successfully demonstrated through using a phrasal categorisation system adapted from Gray and Biber (2013), described in 5.4. This is also coherent with word class distributions found in Biber *et al.* (1999) across registers and Chen and Baker (2010).

At the phase in the methodology where lexical exponents and their functions are examined, a lexical bundle driven categorisation is applied with limited success. The limitations of applying this approach are partly due to the fact that lexical sequences are filtered to only include those that are considered to be ‘the building blocks’ of language, and to remove any context-dependent combinations, or bundles that do not carry an obvious meaning. Given the fact that many of the sequences in this study are (1) either incomplete in a semantic sense, or (2) at the lower proficiency are topic or context dependent, a categorisation can only be applied to sequences with clear referential, stance and discourse-organising functions (Chapter 5 and 6). Similar limitations apply when applying a p-frame approach. P-frames are recurrent word sequences that differ by only one word. The approach was adequate where only one open word class was present in a sequence, but even with a limited lexical repertoire, lower level data demonstrate greater fluidity in the co-selection of items, and less fixedness of form and meaning, which meant that at lower levels p-frames were difficult to assign.

9.4.2 Applying existing frameworks for analysis: Pattern grammar

The groupings described in Pattern Grammar provide a partial categorisation for some of the sequences under investigation. Many of the most frequent combinations of words are not accounted for. The groupings for categorising form-meaning relationships do not account for all forms and associated meanings, nor for the changes in form-meaning relationships across

levels, nor for emerging generalisations. We have observed that POS tag sequences are often not structurally complete. Pattern grammar does not accommodate fragments of (noun) phrases that 4-gram sequences often produce. Additionally where a fragment or part of a fragment is categorised, both structurally and semantically (e.g. adj N), many of the lexical items found in the learner data are not specified under any of the meaning categories in pattern grammar, or are categorised under a general heading. It does not account for subtle differences in compositionality.

As acknowledged by Hunston and Francis (2000), it is the occurrence of repeated forms that drive the pattern grammar categorisation. Meanings are arrived at intuitively and subjectively and are of secondary importance to form in this framework. Added to this, the relative frequency of one pattern over another is not central, which means that there is no indication in this framework whether one pattern or meaning group occurs more frequently than another. One result of this is that some of the most frequently occurring lexical realisations of the sequences in this study are not accounted for in the pattern grammar meaning groups (cf. 6.5.3 and 7.5.3). In summary pattern grammar provides a descriptive framework for some of the structural and functional elements in some 4-gram sequences but does not accommodate all, nor does it account for emerging generalisations.

9.4.2 *Applying existing frameworks for analysis: verb argument constructions*

While making great strides towards our enhanced understanding of emerging knowledge, the focus on the verb in VACs studies does not account for any development beyond the verb clause. Given the importance of the noun phrase in relation to proficiency and development identified above, it would seem the net for capturing structural development beyond the verb clause needs to be cast wider. VACs do not feature in the highest rank sequences in any of the repertoires analysed in this study. VACs are often low frequency and suffer from issues of findability even in large scale data.

Some POS-tag sequence development can reveal aspects of form-meaning mapping that constructions do not reveal, e.g. in the sequence VVD IN DT NN (e.g. *went to the cinema*), *went* dominates the past simple slot at lower levels, in sequences with a literal function (movement to place), and *came* is the verb which overwhelmingly features in the formulaic sequences at C2 level (e.g. *came to the conclusion, came as no surprise*).

POS-tag sequence analysis on both a quantitative and qualitative level may help to contribute to our understanding of VACs. For example, in Romer *et al.* 2015, *come* is identified as the

lead verb in the construction *V across n*, motion construction, with *walk, move* etc. ranking in frequency below *come*. While *come across* clearly has a literal motion meaning (*He came across the room to talk to me*) in the data observed in this study *come across* is overwhelmingly used with a figurative ‘happen upon’ meaning. The suggestion here might be that it still retains a prototypical motion meaning while taking on a figurative usage not seen in for example *walk across* or *move across*.

In summary, some of the existing categorisations are more successfully applied to sequences seen in the higher level data from B2 upwards (p-frames, lexical bundles, pattern grammar). The VAC approach offers considerable insight into the process of abstraction from slot and frame to formulaic and the emergence of the pioneering sequence, but falls short when looking beyond predefined verb-driven constructions, which as evidenced are the tip of the iceberg when exploring development.

Overall the limited success in applying these frameworks may be partly due to the fact that these taxonomies have been developed using L1 frequencies rather than looking at L2 data. This is explored in Monteiro *et al.* 2020 who point to the fact that most studies examining L2 production (and, in their case, lexical sophistication) do so using L1 norms (Ortega 2016) and that L2 production data might prove a more effective source of benchmarks for L2 analysis. A more in-depth L1: L2 comparative analysis is a promising area for future research.

9.5 Methodological and theoretical considerations

9.5.1 POS tag sequence as a way into analysis of development

This study has hoped to offer support to Granger and Rayson’s assertion that analysis through POS tagging can help to ‘form a quick picture of the interlanguage of a given learner population and that it opens up interesting avenues for future research.’ (1998, p. 138).

In some ways it aligns methodologically with Gilquin’s POS tag sequence approach to analysing spoken learner language. As with Gilquin (2018), results from this study show that the top ranking sequences are shared across the data sets, and that any one POS tag sequence hides a ‘great variety of linguistics instantiations’ (2018, p.14). This is not surprisingly since many of the tags represent an open word class. What this approach allows is a view on structural generalisation, which for example a lexical bundle approach does not. Like Gilquin, it also illustrates that the most frequently used sequences that emerge as proficiency

increases are phrasal constructions, e.g. PP MD VV TO *I would like to*, PP VVP TO VV *I want to +verb*, PP VVD TO VV *I decided to + verb*, DT JJ NN IN *the yellow door in, the wide range of*, etc. and not the sequences which are often at the centre of studies on constructions, as we have seen (e.g. caused motion construction, V *across n*).

While this study provides no more than a snapshot and there is clearly much work to be done, it builds on Gilquin (2018) in looking across the whole range of proficiency, not just the higher levels, but also in the qualitative way that she urged, through combination of the analysis of POS-tag sequences with an exploration of their lexical exponents, albeit in written rather than spoken data.

9.5.2 POS tag sequence vs constructions vs patterns vs p-frames vs bundles

One of the points of deviation in this study in comparison with Gilquin (2018) is that she set out to identify constructions. In this study I set out to cast as wide a net as possible to see what emerged without foreclosing on the type of pattern I wanted to find. It emerged that a POS tag sequence approach can be used as the starting point to capture various levels of abstraction. Several generalisations from this approach have become evident, relating to form and function:

All POS tag sequences are one of three kinds:

(i) complete and meaningful units (e.g. PP VVD TO VV *I decided to go*; IN DT JJ NNS *for a few months*; RB IN DT NN *early in the morning*)

(ii) part of a meaningful unit: not all POS tag sequences yield form-meaning mappings. (e.g. DT JJ NN IN *a wide range of, the other side of, an important role in*; TO VV IN DT *to escape from the*)

(iii) contain a meaningful unit with them (e.g. VBZ DT JJ NN *is a major problem*)

The choice of the number of tags in the POS tag sequence is arbitrary and is simply a starting point. (The rationale for using four is described in 4.4). POS tag sequences can be overlapping or extended, (e.g. DT JJ NN IN, JJ NN IN DT, NN IN DT NN) (((det + (adjective + ((noun + preposition))) + det) + noun))

Some POS tag sequences yield more forms and functions than others. For example the NN IN DT NN (*centre of the town* (referential_place), *aim of the proposal*).

Some POS-tag sequences can be categorised as p-frames (*the * of the*)

Some POS-tag sequences can be classified using Pattern Grammar, some not.

Some POS tag sequences are equivalent to constructions in the form-meaning mapping sense, for example verb + to-inf + verb base *decide/want to do something* might be classified as a construction expressing volition. However we have observed that there are two POS tag sequences which might be part of the same construction PP VVD TO VV (pronoun + past simple + to-inf + verb base) and PP VVP TO VV (pronoun + present simple + to-inf + verb-base, but that tense has a subtle effect on meaning. Are these two constructions or part of the same?

9.5.3 Summarising development through a usage-based lens

The overall findings and individual case studies explore both core and emerging sequences at each level. Initial analysis has shown that learners at A1 and A2 levels rely heavily on topic to put together sequences, for example concrete adjectives and nouns relating to the topic or task are the building blocks for sequences. This may lend evidence for the early slot and frame stage of the developmental sequence proposed by a usage-based theory of language learning and proof of consistent form-meaning mappings. However existing frameworks for structural and functional classification do not adequately account for early output at the A1/A2 levels and the growth in the lexical and functional diversity of the sequence as it increases with proficiency.

Looking forward beyond B1, we start to observe pioneering forms and functions sequences which emerge at B2 and continue to be the most highly ranking lexical sequences by C2. Although we see a variety of candidates continuing to 'fill' the POS tag slots at B2 and C1, there is increasing distillation of 'slot candidates' so that by C2 level there is evidence, on the one hand, of an increasingly specialised function (quantity) alongside increasing fixedness and constraint on the selection and combination of lexical items. At A2 and B1 we see independent paradigmatic choices at a POS item level. Beyond this we see the appearance of more and more linguistic routines. The syntactic sequences used by B2 learners, in the main, have not become more complex structurally by C2, but the 'patterns' increase in terms of their functional and lexical instantiations. Although this claim would require further examination, syntactic patterning appreciation (Ellis, 2017; Wulff and Ellis, 2018) seems to be activated earlier than collocational knowledge. Learners, as they encounter more and more opportunities to increase their performance through practice, seem to acquire first the most frequent sequences. Syntactic pattern appreciation seems to have stabilised at B2, it is a

collocational awareness and knowledge mapped to subtle functional awareness that is developing at B2. This resonates with Thewissen (2013), who found syntactic stabilisation at B2, and with Gilquin (2018) as described above, who in a comparison between L2 LINDSEI and L1 LOCNEC also found that among the top 30 most frequent POS tag sequences 25 were the shared by both groups. Change in sequence usage across levels suggests that as learners are exposed to more and more evidence they reach points where they have had sufficient input so as to allow for a significant understanding of the frequency and distribution of the most important sequences.

9.6 Current limitations and future avenues for investigation

Every study comes with limitations and this one clearly is no exception. Some limitations are the result of the scale of the project and can be addressed through future enquiry. As an attempt to consider development across six proficiency levels, from five exams, multiple tasks, over a 20 year period and 148 L1 backgrounds, across 52 million words of data, this study cannot consider every variable. Instead because of the rigour involved in benchmarking proficiency in the data, it has chosen proficiency levels as the variable under scrutiny. Two of the elephants in the room, namely the effect of task and L1 background on the data, are topics identified for future work.

Task effect

Task effect has been addressed to a degree by the use of performance level data (which allows for a wider range of tasks and avoids a direct correspondence between specific task and level) and by the analysis in Chapter 6. However, as noted, the lexical exponents of the higher frequency sequences at lower level data often reflect the topic of the exam task, whereas at the higher level we see a greater use of routine sequences which act as frames for content, vehicles for evaluation, and discourse organisation. Further investigation is already underway (Mark, forthcoming) and while initial findings already point to task effect on sequence use this is bound up in developing awareness of and sensitivity to the communicative demands of register.

We have also seen in previous chapters the shift from verb-based sequences to noun-based sequences. Lower proficiency level tasks are often centered around the themes of recounting and narrating actions and events, centering around topics of ‘where I live’, ‘what I do’, ‘what I did,’ ‘what I’m going to do’ and this may require more use of the verb form, whereas higher

level tasks often require analysis and evaluation. Reasons for this are not clear and require further investigation.

Triangulation with other corpora and creation of POS tag sequence database

Bound up with the issue of task, it is important to bear in mind Hunston's observation that "a statement about evidence in a corpus is a statement about that corpus" (Hunston 2002, p.23). The Cambridge Learner Corpus is a large collection of written exams and can shed light on learner use of language only within the confines of the exam. Triangulation with other corpora is another obvious avenue for further work. For this purpose one of the outputs of this study has been the development of a largescale database of (1) POS tag sequence usage at each proficiency level, and (2) the lexical exponents of sequences at each level. This database will be openly available for comparison with other corpora. It is also hoped that it will provide practical insights for language teaching (see below).

L1 background

In relation to the effect of L1 background, there are two obvious areas to explore. Firstly the effect of the L1 background and transfer on the development of sequence usage and secondly the effect of the distribution of data from different L1 backgrounds. In anticipation of this some initial work was undertaken at the B2 and C2 levels to test whether larger amounts of data from some L1s might skew results, by excluding two of the top ranking L1s (Chinese and Greek) from the data and comparing both with and without (Lim *et al.* forthcoming). Initial findings have indicated there are no major changes in the frequency ranking of the POS tag sequences and the top 10 continue to remain the same for both B2 and C2 levels. As the exclusion of L1 Greek or Chinese data did not make significant changes in the frequency and distribution ranks of the POS sequences analysed, all data were included in our analyses.

Access to data

As outlined, this research was made possible through access to the Cambridge Learner Corpus. Because of the commercially sensitive nature of the data, access to the data was made possible through a bespoke version of the Sketch Engine platform. In an ideal world a study of this nature would have made use of a range of statistical measures, and particularly association measures when looking at degrees of fixedness between sequence elements. Direct access to the raw data was not possible. Some work in collaboration with the ALTA institute on the raw data is in the planning.

POS tagging

As already noted, Gilquin and Granger (2015) cite lack of POS tagging and parsed data as one of reasons why studies of ‘grammatical features’ have been underrepresented in LCR in comparison with studies of lexis and phraseology. We have come a long way since manual POS tagging days and now take it for granted that data comes ready tagged. We have seen in this study that there are limitations to this. Tagging is not always accurate or consistent (*because* is sometimes tagged as a preposition, *many* is sometimes tagged as an adjective) but with the speed of technological development and big data training models this limitation is likely to be a transient one.

Implications for teaching

As O’Keeffe notes (2020, 2022) there is a resonance between instructional approaches - such as a data-driven learning approach (DDL) centered around exposing learners to patterns in data – and the pattern-finding and meaning mapping findings seen in a second language development study such as this. Through identification of the most frequently used sequences and patterns at different levels of proficiency, at a structural, lexical and functional level, it might be possible to create a clearer developmental pathway for learners. In ongoing work, O’Keeffe and Mark (forthcoming) point out that DDL has the potential, to drive a type of intensification of the cognitive process through “grappling” with patterns (O’Keeffe 2021). O’Keeffe (2020) makes the case for an urgent need to aggregate findings on second language development so as to guide the curation of patterns in the process.

9.7 Concluding remarks

I recognise that this study is merely a starting point. I have set out to try to fulfil something of the task set out by Hopper “to study a whole range of repetition in discourse, and in doing so to seek out those regularities which promise interest as incipient sub-systems” (1987). I hope to have demonstrated a non-linearity in language learning, evidence for a growing repertoire of structural, functional and lexical development, and sensitivity to the fixedness of patterning of usage, and continuous shuffling and reshuffling of the frequencies in the language encounter.

References

- Aarts, J., and S. Granger. (1998). 'Tag sequences in learner corpora: A key to interlanguage grammar and discourse,' in S. Granger (ed): *Learner English on computer*. Addison Wesley Longman, 132-141.
- Aijmer, K. (2002). *English discourse particles: Evidence from a corpus*. Amsterdam: John Benjamins Publishing.
- Alexopoulou, T., Geertzen, J., Korhonen, A., and Meurers, D. (2015). 'Exploring big educational learner corpora for SLA research: Perspectives on relative clauses', *International Journal of Learner Corpus Research*, 1(1), 96-129.
- Allen, D. (2009). 'Lexical bundles in learner writing: An analysis of formulaic language in the ALESS Learner Corpus', *Komaba Journal of English Education*, 1, 105-127.
- Arnon, I., and Christiansen, M. H. (2017). 'The Role of Multiword Building Blocks in Explaining L1–L2 Differences', *Topics in cognitive science*, 9(3), 621-636.
- Bestgen, Y., and Granger, S. (2014). 'Quantifying the development of phraseological competence in L2 English writing: An automated approach', *Journal of Second Language Writing*, 26, 28-41.
- Biber, D. (2009a). 'A corpus-driven approach to formulaic language in English: Extending the construct of lexical bundle', in L. Eckstein and C. Reinfandt (eds.), *Anglistentag 2008 Proceedings*, 367-377. Berlin: Wissenschaftlicher Verlag Trier.
- Biber, D. (2009b). 'A corpus-driven approach to formulaic language: Multi-word patterns in speech and writing', *International Journal of Corpus Linguistics*, 14, 275-311.
- Biber, D., Conrad, S., and Leech, G. (2002). *Student grammar of spoken and written English*. London: Longman.
- Biber, D., Conrad, S. and Cortes, V. (2004). 'If you look at...: Lexical bundles in university teaching and textbooks', *Applied Linguistics*, 25, 371-405.
- Biber, D., and Gray, B. (2011). 'Grammatical change in the noun phrase: The influence of written language use', *English Language and Linguistics*, 15(2), 223-250.
- Biber, D., and Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

- Biber, D., Gray, B, and Poonpon, K. (2011). 'Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?,' *TESOL Quarterly* 45(1), 5-35.
- Biber, D., Gray, B. and Staples, S. (2014). 'Predicting patterns of grammatical complexity across language exam task types and proficiency levels', *Applied Linguistics*, 7(5), 639-66
- Biber, D., Gray, B. Staples, S. and Egbert, J. (2020). 'Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement', *Journal of English for Academic Purpose*, 46.
- Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Biber, D., Reppen, R., Staples, S., Egbert, J. (2020). 'Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students', *International Journal of Learner Corpus Research*, 6, 38-71.
- Bley-Vroman, R. (1983). 'The comparative fallacy in interlanguage studies: The case of systematicity 1', *Language learning*, 33(1), 1-17.
- Brown, R. (1973). *A first language: The early stages*. London: George Allen and Unwin.
- Buerki, A. (2018). 'Formulaic sequences: a drop in the ocean of constructions or something more significant?', in MacKenzie, I, and Kayman, M. (eds.) *Formulaicity and Creativity in Language and Literature*. Taylor and Francis.
- Bybee, J. (1998). 'The emergent lexicon', *Chicago Linguistic Society*, 34, 421-435.
- Bybee, J. (2008). 'Usage-based grammar and second language acquisition', in P. Robinson and N. Ellis (eds.) *Handbook of cognitive linguistics and second language acquisition*. Routledge, 226-246.
- Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.
- Callies, M. (2008). 'Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety', in G. Gilquin, S. Papp and M.B. Díez-Bedmar, (2008). (eds.) *Linking up contrastive and learner corpus research*, Amsterdam: Rodopi, 199-226.

- Callies, M. (2015) 'Learner corpus methodology', in Granger, S., Gilquin, G., and Meunier, F. (eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press, 35-55.
- Callies, M., Diez-Bedmar, M.B. and Zaytseva, E. (2014). 'Using learner corpora for testing and assessing L2 proficiency', in Leclercq, P., H. Hilton and A. Edmonds (eds.), *Measuring L2 proficiency: Perspectives from SLA* (Second Language Acquisition series). Clevedon: Multilingual Matters, 71-90.
- Cappelle B. and N. Grabar. (2016). 'Towards an n-grammar of English,' in S. De Knop and G. Gilquin (eds.): *Applied Construction Grammar*. De Gruyter Mouton, 271-302.
- Carter, R., and McCarthy, M. (2006). *Cambridge grammar of English: a comprehensive guide*, Cambridge: Cambridge University Press.
- Carter, R., McCarthy, M., Mark, G., and O'Keeffe, A. (2011). *English Grammar Today*, Cambridge: Cambridge University Press.
- Chen, Y. and P. Baker. (2010). 'Lexical bundle in L1 and L2 academic writing,' *Language Learning and Technology*, 14(2), 30-49.
- Chen, Y. H., and Baker, P. (2016). 'Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1', *Applied Linguistics*, 37(6), 849-880.
- Clark, H. H., and Clark, E. V. (1977). *Psychology and language*, Cambridge: Cambridge University Press.
- Corder, S. P. (1967). 'The significance of learners' errors', *International Review of Applied Linguistics*, 5, 161-170.
- Corder, S. P. (1973). *Introducing applied linguistics*, Penguin Group.
- Corder, S. P. (1981). *Error analysis and interlanguage*, Oxford University Press.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge: Cambridge University Press.
- Council of Europe. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. URL: <https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989>

- De Cock, S., Granger, S., Leech, G., and McEnery, T. (1998). 'An automated approach to the phrasicon of EFL learners', in S. Granger (ed.), *Learner English on computer*, London and New York: Addison Wesley Longman, 67-79.
- DeCock, S. 2007. 'Routinized building blocks in native speaker and learner speech: Clausal sequences in the spotlight', in M. C. Campoy and M. J. Luzón (eds.) *Spoken corpora in applied linguistics*, Peter Lang, 217-233.
- Díez-Bedmar, M. B. (2012). 'The use of the Common European Framework of Reference for Languages to Evaluate Compositions in the English Exam Section of the University Admission Examination', *Revista de Educación*, 357, 55-79.
- Díez-Bedmar, M. B. (2018). Fine-tuning descriptors for CEFR B1 level: insights from learner corpora. *ELT Journal*, 72, 2, 199-209.
- Díez-Bedmar, M. B., and Papp, S. (2008). 'The use of the English article system by Chinese and Spanish learners', in G. Gilquin, S. Papp and M.B. Díez-Bedmar, (2008). (eds.) *Linking up contrastive and learner corpus research*, Amsterdam: Rodopi, 147-175.
- Díez-Bedmar, M. B., and Pérez-Paredes, P. (2010). 'La investigación del discurso escrito en el aprendizaje de idiomas en entornos colaborativos y wikis', *Revista de Educación a Distancia (RED)*.
- Díez-Bedmar, M. B., and Pérez-Paredes, P. (2020). 'Noun phrase complexity in young Spanish EFL learners' writing: Complementing syntactic complexity indices with corpus-driven analyses', *International Journal of Corpus Linguistics*, 25, 1, 4-35.
- Douglas Fir Group (2016). 'A transdisciplinary framework for SLA in a multilingual world', *Modern Language Journal*, 100 (Supplement 1, Centenary Anniversary), 19-47.
- Dulay, H. C., and Burt, M. K. (1973). 'Should we teach children syntax?', *Language learning*, 23, 2, 245-258.
- Durrant, P. and Schmitt, N. (2009). 'To what extent do native and non-native writers make use of collocations?' *International Review of Applied Linguistics*, 47, 157-177.
- Durrant, P., Brenchley, M., and McCallum, L. (2021). *Understanding development and proficiency in writing: quantitative corpus linguistic approaches*, Cambridge: Cambridge University Press.

- Ellis, N. C. (1996). 'Sequencing in SLA: Phonological Memory, Chunking and Points of Order', *Studies in Second Language Acquisition*, 18, 91-126.
- Ellis, N. C. (1998). 'Emergentism, connectionism and language learning', *Language Learning*, 48, 631-664.
- Ellis, N. C. (2002). 'Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition', *Studies in Second Language Acquisition*, 24, 143-188.
- Ellis, N. C. (2003). 'Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty and M. H. Long (eds.), *Handbook of second language acquisition*' Oxford: Blackwell, 33-68.
- Ellis, N. C. (2008). 'Usage-based and form-focused language acquisition', in P. Robinson and N. C. Ellis (eds.), *Handbook of cognitive linguistics and second language acquisition*, London: Routledge, 372-405.
- Ellis, N. C. (2009). 'Optimizing the input: Frequency and sampling in usage-based and form-focused learning', in M. H. Long and C. Doughty (eds.), *Handbook of language teaching*, Oxford: Blackwell, 139-158.
- Ellis, N. C. (2011). 'The emergence of language as a complex adaptive system', in J. Simpson (ed.), *Handbook of applied linguistics*, London: Routledge, 666-679.
- Ellis, N. C. (2012). 'Formulaic language and second language acquisition: Zipf and the phrasal teddy bear', *Annual Review of Applied Linguistics*, 32, 17-44.
- Ellis, N. C. (2013). 'Frequency-based grammar and the acquisition of tense-aspect in L2 learning', in Rafael Salaberry and Llorenç Comajoan (eds.) *Research Design and Methodology in Studies on Second Language Tense and Aspect*, Berlin: Mouton de Gruyter, 89-118.
- Ellis, N. C. (2015) 'Implicit and explicit language learning: Their dynamic interface and complexity', in P. Rebuschat (ed.) *Implicit and explicit learning of languages*, John Benjamins, 1-24.
- Ellis, N. C. (2017). 'Cognition, corpora, and computing: Triangulating research in usage-based language learning', *Language Learning*, 67, S1, 40-65.

- Ellis, N. C. (2019a). 'Essentials of a theory of language cognition', *Modern Language Journal*, 103 (Supplement 2019), 39-60.
- Ellis, N. C. (2019b). 'Usage-based theories of Construction Grammar: Triangulating Corpus Linguistics and Psycholinguistics', in Jesse Egbert and Paul Baker (eds.), *Using corpus methods to triangulate linguistic analysis*, New York and London: Routledge, 239-267.
- Ellis, N. C. and Ferreira-Junior, F. (2009a). 'Constructions and their acquisition: Islands and the distinctiveness of their occupancy', *Annual Review of Cognitive Linguistics*, 7, 188-221.
- Ellis, N. C. and Ferreira-Junior, F. (2009b). 'Construction learning as a function of frequency, frequency distribution, and function', *Modern Language Journal*, 93, 370-385.
- Ellis, N. C. and Ogden, D. C. (2017). 'Thinking about multiword constructions: Usage-based approaches to acquisition and processing', *Topics in Cognitive Science*, 9(3), 604-620.
- Ellis, N. C., O'Donnell, M. B., and Römer, U. (2013). 'Usage-Based Language: Investigating the Latent Structures that Underpin Acquisition', *Currents in Language Learning*, 1, *Language Learning*, 63, Suppl 1., 25-51.
- Ellis, N., Römer, U. and O'Donnell, M. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar*, Wiley.
- Ellis, N. C. and Simpson-Vlach, R. (2009). 'Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education', *Corpus Linguistics and Linguistic Theory*, 5, 61-78.
- Ellis, N., Simpson-Vlach, R., Römer, U., O'Donnell, M., and Wulff, S. (2015). 'Learner corpora and formulaic language in second language acquisition research', in S. Granger, G. Gilquin, and F. Meunier (eds.), *The Cambridge handbook of learner corpus research* (Cambridge Handbooks in Language and Linguistics, Cambridge: Cambridge University Press, 357-378.
- Ellis, R. (2021). 'A short history of SLA: Where have we come from and where are we going?', *Language Teaching*, 54(2), 190-205.
- Elman, J. L. (2009). 'On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon', *Cognitive Science*, 33, 547-582.

- Erman, B., and Warren, B. (2000). 'The idiom principle and the open choice principle' *Text-Interdisciplinary Journal for the Study of Discourse*, 20(1), 29-62.
- Eskildsen, S. W. (2009). 'Constructing another language. Usage-based linguistics in second language acquisition', *Applied Linguistics*, 30(3), 335-357.
- Eskildsen, S. W. (2012). 'L2 negation constructions at work', *Language Learning*, 62(2), 335-372.
- Eskildsen, S. W. (2015). 'What counts as a developmental sequence? Exemplar-based L2 learning of English questions', *Language Learning*, 65(1), 33-62.
- Eskildsen, S.W., and Cadierno, T. (2007). 'Are recurring multi-word expressions really syntactic freezes? Second language acquisition from the perspective of usage-based linguistics', in M. Nenonen and Niemi, S. (eds.). *Collocations and Idioms 1. Papers from the First Nordic Conference on Syntactic Freezes*, Joensuu: Joensuu University Press, 86-99.
- Fillmore, C. J. (1988, October). 'The mechanisms of construction grammar', in *Annual Meeting of the Berkeley Linguistics Society* (Vol. 14) 35-55.
- Francis, G. (1993). 'A corpus-driven approach to grammar: Principles, methods and examples', *Text and technology: In honour of John Sinclair*, 1, 137-156.
- Francis, G., S. Hunston, and E. Manning. (1998). *Collins Cobuild Grammar Pattern 2: Nouns and Adjectives*. HarperCollins.
- Francis, G., S. Hunston, and E. Manning. (1996). *Collins Cobuild Grammar Patterns 1: Verbs*. HarperCollins Publisher.
- Gablasova, D., Brezina, V., and McEnery, T. (2017). 'Exploring learner language through corpora: comparing and interpreting corpus frequency information', *Language Learning*, 67(S1), 130-154.
- Gablasova, D., Brezina, V., and McEnery, T. (2019a). 'The Trinity Lancaster Corpus: Applications in language teaching and materials development', in S. Götz and J. Mukherjee (eds.), *Learner corpora and language teaching*, Amsterdam: John Benjamins, 7-28.

- Gablasova, D., Brezina, V., and McEnery, T. (2019b). 'The Trinity Lancaster Corpus. Development, description, and application', *International Journal of Learner Corpus Research*, 5(2), 126-158.
- Garner, J. (2016). 'A phrase-frame approach to investigating phraseology in learner writing across proficiency levels', *International Journal of Learner Corpus Research*, 2(1), 31-67.
- Gilquin, G. (2018). 'Exploring the spoken learner English construction: A corpus-driven approach,' in R. Alonso (ed.): *Speaking in a second language*, John Benjamins, 127-152.
- Gilquin, G., Papp, S. and Díez-Bedmar, M. B. (2008). (eds.) *Linking up contrastive and learner corpus research*, Amsterdam: Rodopi.
- Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.
- Goldberg, A. E. (2006). *Constructions at Work. The Nature of Generalization in Language*, Oxford: Oxford University Press.
- Götz, S. and Schilk, M. (2011). 'Formulaic sequences in spoken ENL, ESL, and EFL: Focus on British English, Indian English and learner English of advanced German learners', in J. Mukherjee and M. Hundt (eds.): *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*. John Benjamins, 79-100.
- Granger, S. (1994). 'The learner corpus: A revolution in applied linguistics', *English Today*, 10, 25-33.
- Granger, S. (1996). 'Learner English around the world,' in S. Greenbaum (ed.): *Comparing English worldwide*. Clarendon Press, 13-24.
- Granger, S. (1998). 'The computer learner corpus: A versatile new source of data for SLA research,' in S. Granger (ed.): *Learner English on computer*. Addison Wesley Longman, 3-18.
- Granger, S. (2002). 'A bird's-eye view of learner corpus research', *Computer learner corpora, second language acquisition and foreign language teaching*, 6, 3-33.
- Granger, S. (2009). 'The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation', in K. Aijmer (ed.), *Corpora and language teaching*, Amsterdam: John Benjamins, 13-32.

- Granger, S. (2015). 'Contrastive interlanguage analysis: A reappraisal', *International Journal of Learner Corpus Research*, 1(1), 7-24.
- Granger, S. (2021). 'Commentary: Have Learner Corpus Research and Second Language Acquisition Finally Met?', in B. Le Bruyn and M. Paquot (eds.), *Learner corpus research meets second language acquisition*, Cambridge Applied Linguistics, Cambridge: Cambridge University Press, 243-257.
- Granger, S., and Bestgen, Y. (2014). 'The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study,' *International Review of Applied Linguistics in Language Teaching* 52, 3, 229-252.
- Granger, S., Dupont, M., Meunier, F., Naets, H. and Paquot, M. (2020). *The International Corpus of Learner English*. Version 3. Louvain-la-Neuve: Presses universitaires de Louvain.
- Granger, S., Gilquin, G., and Meunier, F. (eds.) (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.
- Granger, S. and Rayson, P. (1998). 'Automatic profiling of learner texts,' in S. Granger (ed.): *Learner English on computer*. Addison Wesley Longman, 119-131.
- Gray, B., and Biber, D. (2013). 'Lexical frames in academic prose and conversation', *International Journal of Corpus Linguistics*, 18(4).
- Gray, B., and Biber, D. (2015). 'Phraseology', in D. Biber and R. Reppen (eds.), *Cambridge handbook of corpus linguistics*, Cambridge: Cambridge University Press, 125-145.
- Gries, S. T., and Wulff, S. (2005). 'Do foreign language learners also have constructions?', *Annual Review of Cognitive Linguistics*, 3(1), 182-200.
- Groom, N. (2009). 'Effects of second language immersion on second language collocational development,' in A. Barfield and H. Gyllstad (eds.), *Researching collocations in another language: Multiple interpretations*. Palgrave Macmillan, 21-33.
- Harrison, J., and Barker, F. (eds.) (2015). *English profile in practice. English Profile Studies, Vol. 5*, Cambridge: Cambridge University Press.
- Hasko, V. and Meunier (eds.) (2013). 'Capturing L2 development through learner corpus analysis', *The Modern Language Journal*, 97(S1), 1-10.

- Hawkins, J. A., and Buttery, P. (2009). 'Using learner language from corpora to profile levels of proficiency: Insights from the English Profile programme', in L. Taylor and C. J. Weir (eds.) *Language testing matters: Investigating the wider social and educational impact of assessment*, Cambridge: Cambridge University Press, 158-175.
- Hawkins, J. A., and P. Buttery. (2010). 'Criterial features in learner corpora: Theory and illustrations', *English Profile Journal*, 1(1), 1-23.
- Hawkins, J. and Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge: Cambridge University Press.
- Heuboeck, A., Holmes, J., and Nesi, H. (2008). *The BAWE corpus manual*.
http://www.reading.ac.uk/internal/appling/bawe/BAWE_documentation.pdf.
- Hopper, P. (1987). 'Emergent grammar', in *Annual Meeting of the Berkeley Linguistics Society* (Vol. 13) 139-157.
- Hunston, S. and G. Francis. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins.
- Hunston, S. (2019). 'Patterns, constructions, and applied linguistics,' *International Journal of Corpus Linguistics* 24(3), 324-353.
- Jarvis, S. (2000). 'Methodological rigor in the study of transfer: Identifying L1 influence in them interlanguage lexicon', *Language learning*, 50(2), 245-309.
- Johnson, K., and Johnson, H. (eds.). (1999). *Encyclopedic dictionary of applied linguistics*, Oxford: Blackwell.
- Juknevičienė, R. (2009). 'Lexical bundles in learner language: Lithuanian learners vs. native speakers,' *KaLBOTYRa* 61(3), 61-72.
- Kellerman, E. (1995). 'Crosslinguistic influence: Transfer to nowhere?', *Annual review of applied linguistics*, 15, 125-150.
- Kennedy, G. (1996). 'The corpus as a research domain,' in S. Greenbaum (ed.), *Comparing English worldwide*, Clarendon Press, 217-226.
- Kyle, K., and Crossley, S. (2017). 'Assessing syntactic sophistication in L2 writing: A usage-based approach', *Language Testing*, 34(4), 513-535.

- Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites* (Vol. 1). Stanford University Press.
- Larsen–Freeman, D. (2006). ‘The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English’, *Applied Linguistics*, 27, 590-619.
- Larsen-Freeman, D. (2015). ‘Saying what we mean: Making a case for “language acquisition” to become “language development”’. *Language Teaching*, 48(4), 491-505.
- Lenko-Szymanska, A. (2014). ‘The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective’, *International Journal of Corpus Linguistics*, 19(2), 225-251.
- Li, P., Eskildsen, S.W., and Cadierno, T. (2014). ‘Tracing an L2 learner’s motion constructions over time: A usage-based classroom investigation’, *The Modern Language Journal*, 98(2), 612-628.
- Lieven, E., V.M., and Tomasello, M. (2008). ‘Children's first language acquisition from a usage-based perspective’, in P. Robinson and N. C. Ellis (eds.), *Handbook of cognitive linguistics and second language acquisition*, London: Routledge, 168-196.
- Lim, J., Mark, G., Pérez-Paredes, P., O’Keeffe, A. (journal submission, awaiting response). ‘Exploring key Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective.’
- Little, D. (2007). ‘The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy’, *The Modern Language Journal*, 91(4), 645-655.
- Logan, G. D. (1990). ‘Repetition priming and automaticity: Common underlying mechanisms?’, *Cognitive Psychology*, 22(1), 1-35.
- Long, M. H., and Robinson, P. (1998). ‘Focus on form: Theory, research, and practice’, in C. Doughty, and J. Williams (eds.), *Focus on form in classroom second language acquisition*, Cambridge: Cambridge University Press. 15-41.
- Lundell, F. F. (2020). ‘Formulaicity’, in N. Tracy-Ventura, and M. Paquot, (eds.). (2020). *The Routledge handbook of second language acquisition and corpora*, Routledge, 370-381.

- MacWhinney, B. (1992). 'Transfer and competition in second language learning', in *Advances in psychology* (Vol. 83), North-Holland, 371-390.
- McEnery, T., Brezina, V., Gablasova, D., and Banerjee, J. (2019). 'Corpus linguistics, learner corpora, and SLA: Employing technology to analyze language use', *Annual Review of Applied Linguistics*, 39, 74-92.
- Meunier, F. (2015). 'Developmental patterns in learner corpora', in S. Granger, G. Gilquin and F. Meunier (eds.), *The Cambridge handbook of learner corpus research*, Cambridge: Cambridge University Press, 379-400.
- Meunier, F. and Littré, D. (2013). 'Tracking learners' progress: Adopting a dual "corpus cum experimental data" approach', *The Modern Language Journal*, 97(S1), 61-76.
- Myles, F. (2005). 'Interlanguage corpora and second language acquisition research', *Second Language Research*, 21(4), 373-391.
- Myles, F. (2015). 'Second language acquisition theory and learner corpus research', in S. Granger, G. Gilquin and F. Meunier, (eds.), *The Cambridge handbook of learner corpus research*, Cambridge University Press, 309-331.
- Myles, F., Mitchell, R. and Hopper, J. (1999). Interrogative chunks in French L2: a basis for creative construction?. *Studies in Second Language Acquisition*, 21(1), 49-80.
- Negishi, M., Takada, T., and Tono, Y. (2013, January). 'A progress report on the development of the CEFR-J', in *Exploring language frameworks: Proceedings of the ALTE Kraków Conference*, 135-163.
- Nemser, W. (1971). 'Approximative systems of foreign language learners', *International Review of Applied Linguistics*, 9, 115-123.
- Ninio, A. (1999). 'Pathbreaking verbs in syntactic development and the question of prototypical transitivity', *Journal of child language*, 26(3), 619-653.
- Ninio A. (2005). 'Testing the role of semantic similarity in syntactic development', *Journal of Child Language* 32(1), 35-61.
- Norris, J. M., and Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning*, 50(3), 417-528.
- Odlin T. (1998). 'On the affective and cognitive bases for language transfer', in R. Cooper (ed.), *Compare or contrast?* Tampere, Finland: University of Tampere, 81– 106.

- O'Keeffe, A. (2020). Data-driven learning—a call for a broader research gaze. *Language Teaching*, 54(2), 259-272.
- O'Keeffe, A. (2021). Data-driven learning, theories of learning and second language acquisition, in P. Pérez-Paredes and G. Mark (eds.) *Beyond concordance lines: Corpora in language education*, 35-55.
- O'Keeffe, A., and Mark, G. (2017). 'The English Grammar Profile of learner competence: Methodology and key findings', *International Journal of Corpus Linguistics*, 22(4), 457-489.
- O'Keeffe, A. and Mark, G. (forthcoming) 'Principled grammar curation: using a corpus to identify grammar competencies and enhance data-driven learning design'.
- Ortega, L. (2013). 'SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn', *Language learning*, 63, 1-24.
- Ortega, L., Tyler, A. E., Park, H. I., and Uno, M. (eds.). (2016). *The usage-based study of language learning and multilingualism*, Georgetown University Press.
- Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(2), 324.
- Paquot, M. and S. Granger. (2012). 'Formulaic Language in Learner Corpora,' *Annual Review of Applied Linguistics* 32, 130-149.
- Pawley, A., and Syder, F. H. (1983). 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency', in J. C. Richards and R. W. Schmidt (eds.), *Language and communication*, London: Longman, 191-226.
- Pendar, N. and Chapelle, C. A. (2008). 'Investigating the promise of learner corpora: Methodological issues', *CALICO journal*, 25(2), 189-206.
- Perek, F. (2015). *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives* (Vol. 17), John Benjamins Publishing Company.
- Pérez-Paredes, P., and Díez-Bedmar, M. B. (2019a). 'Researching learner language through POS keyword and syntactic complexity analyses', in S. Götz and J. Mukherjee (eds.), *Learner corpora and language teaching*, Amsterdam: John Benjamins, 101-127.

- Pérez-Paredes, P., and Díez-Bedmar, M. B. (2019b). 'Certainty adverbs in spoken learner language. The role of tasks and proficiency', *International Journal of Learner Corpus Research*, 5(2), 252-278.
- Pérez-Paredes, P., G. Mark, and A. O'Keeffe. (2020). The impact of usage-based approaches on second language learning and teaching. Cambridge Education Research Reports. Cambridge University Press.
- Piantadosi, S. T. (2014). 'Zipf's word frequency law in natural language: A critical review and future directions', *Psychonomic bulletin and review*, 21(5), 1112-1130.
- Pine, J. M., and Lieven, E. V. M. (1997). 'Slot and frame patterns in the development of the determiner category', *Applied Psycholinguistics*, 18, 123-138.
- Ping, P. (2009). 'A study on the use of four-word lexical bundles in argumentative essays by Chinese English: A comparative study based on WECCL and LOCNESS', *CELEA journal*, 32(3), 25-45.
- Renouf, A., and Sinclair, J. (1991). 'Collocational frameworks in English', in K. Aijmer and B. Altenberg, *English corpus linguistics*, 128-143.
- Robinson, P. (1997). 'Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions', *Studies in second language acquisition*, 19(2), 223-247.
- Robinson, P. (ed.). (2002). *Individual differences and instructed language learning* (Vol. 2). John Benjamins Publishing.
- Römer, U. (2010). 'Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews', *English text construction*, 3(1), 95-119.
- Römer, U. (2019). 'A corpus perspective on the development of verb constructions in second language learners', *International Journal of Corpus Linguistics*, 24(3), 268-290.
- Römer, U., and Berger, C. M. (2019). 'Observing the emergence of constructional knowledge: Verb patterns in German and Spanish learners of English at different proficiency levels', *Studies in Second Language Acquisition*, 41(5), 1089-1110.
- Römer, U., and Garner, J. (2019). 'The development of verb constructions in spoken learner English: Tracing effects of usage and proficiency', *International Journal of Learner Corpus Research*, 5(2), 206-229.

- Römer, U., and Garner, J. (2022). ‘What can corpus linguistics tell us about second language acquisition?’ in A. O’Keeffe and M. McCarthy (eds.) (2nd edition), *The Routledge handbook of corpus linguistics*, Routledge, 328-340.
- Römer, U., O'Donnell, M. B., and Ellis, N. C. (2014). ‘Second language learner knowledge of verb–argument constructions: Effects of language transfer and typology’, *The Modern Language Journal*, 98(4), 952-975.
- Römer, U., Skalicky, S. C., and Ellis, N. C. (2018). ‘Verb-argument constructions in advanced L2 English learner production: Insights from corpora and verbal fluency tasks’, *Corpus Linguistics and Linguistic Theory*, 16(2), 303-331.
- Ryland Williams, J., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M., and Sheridan Dodds, P. (2015). Zipf's law holds for phrases, not words. *Scientific reports*, 5,1, 1-7.
- Selinker, L. (1972). ‘Interlanguage’, *International Review of Applied Linguistics in Language Teaching*, vol. 10, no. 1-4, 209-232.
- Seretan, V., Nerima, L., and Wehrli, E. (2004). ‘Multi-word collocation extraction by syntactic composition of collocation bigrams’, *Recent Advances in Natural Language Processing III*, 91-100.
- Simpson-Vlach, R., and Ellis, N. C. (2010). ‘An academic formulas list: New methods in phraseology research’, *Applied Linguistics*, 31(4), 487-512.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.
- Slobin, D. I. (1978). ‘Suggested universals in the ontogenesis of grammar’, *Vladimir Honsa (Hg.): Papers on linguistics and child language, Den Haag*, 249-364.
- Staples, S., Egbert, J., Biber, D. and McClair, A. (2013). ‘Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section,’ *Journal of English for academic purposes* 12(3), 214-225.
- Staples, S., Egbert, J., Biber, D. and Gray, B. (2016). ‘Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre,’ *Written Communication* 33(2), 149-183.

- Swain, M. (2006). 'Languaging, agency and collaboration in advanced second language proficiency', *Advanced language learning: The contribution of Halliday and Vygotsky*, 95-108.
- Tarone, E. (2018). 'Interlanguage', in C.A. Chapelle (ed.) *The encyclopedia of applied linguistics*, 1-7.
- Taylor, L., and Jones, N. (2006). 'Cambridge ESOL exams and the Common European Framework of Reference (CEFR)', *Research Notes*, 24(1), 2-5.
- Thewissen, J. (2013). 'Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus', *The Modern Language Journal*, 97, 77-101.
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*, Studies in Corpus Linguistics. Amsterdam: John Benjamins.
- Tomasello, M. (1992). *First Verbs: A Case Study of Early Grammatical Development of Cognition and Action*. Cambridge, MA: MIT Press.
- Tomasello, M. (2003). *Constructing a Language. A Usage-based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.
- Tomasello, M. (2006). 'Acquiring Linguistic Constructions', in D. Kuhn, R. S. Siegler, W. Damon, and R. M. Lerner (eds.), *Handbook of child psychology: Cognition, perception, and language*, John Wiley and Sons Inc. 255-298.
- Tomasello, M., and Brooks, P. J. (1999). 'Early syntactic development: A construction grammar approach', in M. Barrett (ed.) *The development of language*, 161-190.
- Tono, Y. (2000). 'A corpus-based analysis of interlanguage development: analysing POS tag sequences of EFL learner corpora', *Practical Applications in Language Corpora*, 123-132.
- Tono, Y. (2013). 'Automatic extraction of L2 criterial lexico-grammatical features across pseudo-longitudinal learner corpora: using edit distance and variability-based neighbour clustering', *L2 vocabulary acquisition, knowledge and use*, 149-176.
- Tono, Y. and Díez-Bedmar, M. B. (2014). 'Focus on learner writing at the beginning and intermediate stages: The ICCI corpus', *International Journal of Corpus Linguistics*, 19(2), 163-177.

- Tyler, A., and Ortega, L. (2016). 'Usage-based approaches to language and language learning: An introduction to the special issue', *Language and Cognition*, 8(3), 335-345.
- Tyler, A., and L. Ortega. (2018). 'Usage-inspired L2 instruction: Some reflections and a heuristic,' in A. Tyler, L. Ortega, M. Uno, and H. Park (eds.) *Usage-inspired L2 instruction: Researched Pedagogy*. John Benjamins, 316-321.
- Vidakovic, I., and Barker, F. (2010). 'Use of words and multi-word units in Skills for Life Writing examinations'. University of Cambridge ESOL Examinations *Research Notes*, (41), 7-14.
- Vyatkina, N. (2013). 'Specific syntactic complexity: Developmental profiling of individuals based on an annotated learner corpus', *The Modern Language Journal*, 97(S1), 11-30.
- Wray, A. (2000). 'Formulaic sequences in second language teaching: Principle and practice', *Applied linguistics*, 21(4), 463-489.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
- Wulff, S., and Ellis, N. C. (2018). 'Usage-based approaches to second language acquisition' in D. Miller, F. Bayram, J. Rothman and L. Serratrice (eds.) *Bilingual Cognition and Language: The state of the science across its subfields*. *Studies in Bilingualism*, Vol. 54, 37-56. Amsterdam: John Benjamins.
- Zipf, G. K. (1935). *The psycho-biology of language*. Oxford: Houghton, Mifflin.

Appendices

Appendix 1 Tasks at each exam level of the Cambridge mainsuite exams

	LEVEL	A2	B1	B2	C1	C2
	EXAM	KET	PET	FCE	CAE	CPE
STYLE						
Informative/news		✓	✓	✓	✓	✓
Complaint/apology/response		✓	✓	✓	✓	✓
Business			✓	✓	✓	
Descriptive/creative autobiographical			✓	✓	✓	✓
Advice			✓	✓	✓	✓
Argumentative/opinion			✓	✓	✓	✓
Critical				✓	✓	✓
Application/response				✓	✓	
FORMAT						
Note/email/memo		✓	✓	✓	✓	
Informative/instructional text		✓	✓	✓	✓	✓
Letter/reference		✓	✓	✓	✓	✓
Story			✓	✓		✓
Survey/questionnaire/form			✓	✓	✓	
Composition/essay				✓	✓	✓
Article				✓	✓	✓
Report				✓	✓	✓
Proposal				✓	✓	✓
Review				✓	✓	✓

Appendix 2 English Penn TreeBank tagset

The table shows English Penn TreeBank tagset with Sketch Engine modifications (earlier version).

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there is
FW	foreign word	les
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
PDT	predeterminer	both the boys
POS	possessive ending	friend's
PP	personal pronoun	I, he, it
PPZ	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best

RP	particle	give up
SENT	Sentence-break punctuation	. ! ?
SYM	Symbol	/ [= *
TO	infinitive 'to'	togo
UH	interjection	uhhuhhuhh
VB	verb be, base form	be
VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, sing. present, non-3d	am, are
VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have
VHZ	verb have, 3rd person sing. present	has
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, sing. present, non-3d	take
VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP\$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

Main differences to the default Penn tagset

[In TreeTagger](#)

Distinguishes be (VB) and have (VH) from other (non-modal) verbs (VV)

For proper nouns, NNP and NNPS have become NP and NPS

SENT for end-of-sentence punctuation (other punctuation tags may also differ)

[In TreeTagger tool + Sketch Engine modifications](#)

the word 'to' is tagged IN when used as a preposition and TO when used as an infinitive marker

Bibliography

M. Marcus, B. Santorini and M.A. Marcinkiewicz (1993). [Building a large annotated corpus of English: The Penn Treebank](#). In *Computational Linguistics*, volume 19, number 2, 313–330.

Appendix 3 Sample of the master cohort of the top 1000 sequences at all levels and their rankings at other levels

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	All 4-grams	A1 rank	A2 rank	B1 rank	B2 rank	C1 rank	C2 rank	BNC		A1-A2	A1-B1	A1-B2	A1-C1	A1-C2
2	SENT PP MD VV	1	7	17	21	36	70	286		-6	-16	-20	-35	-69
3	PP MD VV IN	2	6	12	36	60	95	269		-4	-10	-34	-58	-93
4	IN DT NN SENT	3	3	2	3	6	6	9		0	1	0	-3	-3
5	IN PPZ NN SENT	4	5	6	8	17	18			-1	-2	-4	-13	-14
6	NN SENT PP VVP	5	8	15	24	43	90	479		-3	-10	-19	-38	-85
7	NN SENT PP MD	6	26	37	38	53	99	321		-20	-31	-32	-47	-93
8	DT NN SENT PP	7	12	8	11	21	30	126		-5	-1	-4	-14	-23
9	NP NP , PP	8	17	26	68	182	519			-9	-18	-60	-174	-511
10	PP VVP TO VV	9	15	22	34	63	94	490		-6	-13	-25	-54	-85
11	NN IN DT NN	10	1	1	1	1	1	2		9	9	9	9	9

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	All 4-grams	A1 rank	A2 rank	B1 rank	B2 rank	C1 rank	C2 rank	BNC		A2 -A1	A2-B1	A2-B2	A2-C1	A2-C2
2	NN IN DT NN	10	1	1	1	1	1	2		-9	0	0	0	0
3	IN DT JJ NN	62	2	3	2	2	2	3		-60	-1	0	0	0
4	IN DT NN SENT	3	3	2	3	6	6	9		0	1	0	-3	-3
5	IN DT NN IN	15	4	5	4	4	3	1		-11	-1	0	0	1
6	IN PPZ NN SENT	4	5	6	8	17	18			1	-1	-3	-12	-13
7	PP MD VV IN	2	6	12	36	60	95	269		4	-6	-30	-54	-89
8	SENT PP MD VV	1	7	17	21	36	70	286		6	-10	-14	-29	-63
9	NN SENT PP VVP	5	8	15	24	43	90	479		3	-7	-16	-35	-82
10	DT JJ NN IN	109	9	9	6	5	4	5		-100	0	3	4	5
11	DT JJ NN SENT	34	10	11	7	7	9	19		-24	-1	3	3	1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	All 4-grams	A1 rank	A2 rank	B1 rank	B2 rank	C1 rank	C2 rank	BNC		B1-A1	B1-A2	B1-B2	B1-C1	B1-C2
2	NN IN DT NN	10	1	1	1	1	1	2		-9	0	0	0	0
3	IN DT JJ NN	62	2	3	2	2	2	3		-1	-1	-1	-4	-4
4	IN DT NN SENT	3	3	2	3	6	6	9		-59	1	1	1	1
5	NP NP NP NP	40	13	4	12	18	51	16		-36	-9	-8	-14	-47
6	IN DT NN IN	15	4	5	4	4	3	1		-10	1	1	1	2
7	IN PPZ NN SENT	4	5	6	8	17	18			2	1	-2	-11	-12
8	DT NN IN DT	32	16	7	5	3	5	4		-25	-9	2	4	2
9	DT NN SENT PP	7	12	8	11	21	30	126		1	-4	-3	-13	-22
10	DT JJ NN IN	109	9	9	6	5	4	5		-100	0	3	4	5
11	TO VV DT NN	43	28	10	9	9	10	18		-33	-18	1	1	0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	All 4-grams	A1 rank	A2 rank	B1 rank	B2 rank	C1 rank	C2 rank	BNC		B2-A1	B2-A2	B2-B1	B2-C1	B2-C2
2	NN IN DT NN	10	1	1	1	1	1	2		-9	0	0	0	0
3	IN DT JJ NN	62	2	3	2	2	2	3		-60	0	-1	0	0
4	IN DT NN SENT	3	3	2	3	6	6	9		0	0	1	-3	-3
5	IN DT NN IN	15	4	5	4	4	3	1		-11	0	-1	0	1
6	DT NN IN DT	32	16	7	5	3	5	4		-27	-11	-2	2	0
7	DT JJ NN IN	109	9	9	6	5	4	5		-103	-3	-3	1	2
8	DT JJ NN SENT	34	10	11	7	7	9	19		-27	-3	-4	0	-2
9	IN PPZ NN SENT	4	5	6	8	17	18			4	3	2	-9	-10
10	TO VV DT NN	43	28	10	9	9	10	18		-34	-19	-1	0	-1
11	IN DT NN ,	89	23	21	10	8	8	15		-79	-13	-11	2	2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	All 4-grams	A1 rank	A2 rank	B1 rank	B2 rank	C1 rank	C2 rank	BNC		C1-A1	C1-A2	C1-B1	C1-B2	C1-C2
2	NN IN DT NN	10	1	1	1	1	1	2		-9	0	0	0	0
3	IN DT JJ NN	62	2	3	2	2	2	3		-60	0	-1	0	0
4	IN DT NN IN	15	4	5	4	4	3	1		-11	0	-1	0	1
5	DT JJ NN IN	109	9	9	6	5	4	5		-104	-4	-4	-1	1
6	IN DT NN SENT	3	3	2	3	6	6	9		3	3	4	3	0
7	DT JJ NN SENT	34	10	11	7	7	9	19		-27	-3	-4	0	-2
8	IN DT NN ,	89	23	21	10	8	8	15		-81	-15	-13	-2	0
9	TO VV DT NN	43	28	10	9	9	10	18		-34	-19	-1	0	-1
11	DT NN IN NN	84	31	32	18	10	7	8		-74	-21	-22	-8	3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	All 4-grams	A1 rank	A2 rank	B1 rank	B2 rank	C1 rank	C2 rank	BNC		All 4-grams	C2-A1	C2-A2	C2-B1	C2-B2	C2-C1
2	NN IN DT NN	10	1	1	1	1	1	2		NN IN DT NN	-9	0	0	0	0
3	IN DT JJ NN	62	2	3	2	2	2	3		IN DT JJ NN	-60	0	-1	0	0
4	IN DT NN IN	15	4	5	4	4	3	1		IN DT NN IN	-12	-1	-2	-1	-1
5	DT JJ NN IN	109	9	9	6	5	4	5		DT JJ NN IN	-105	-5	-5	-2	-1
6	DT NN IN DT	32	16	7	5	3	5	4		DT NN IN DT	-27	-11	-2	0	2
7	IN DT NN SENT	3	3	2	3	6	6	9		IN DT NN SENT	3	3	4	3	0
8	DT NN IN NN	84	31	32	18	10	7	8		DT NN IN NN	-77	-24	-25	-11	-3
9	IN DT NN ,	89	23	21	10	8	8	15		IN DT NN ,	-81	-15	-13	-2	0
10	DT JJ NN SENT	34	10	11	7	7	9	19		DT JJ NN SENT	-25	-1	-2	2	2
11	TO VV DT NN	43	28	10	9	9	10	18		TO VV DT NN	-33	-18	0	1	1

Appendix 4 Lexical bundle classification (Biber *et al.* 2004)

1. Stance bundles	A. Epistemic stance <i>the fact that the</i>	
	B. Attitudinal/modality stance	B1) Desire <i>I want you to</i> B2) Obligation/directive <i>it is important to</i> B3) Intention/ Prediction <i>we are going to</i> B4) Ability <i>to be able to</i>
2. Discourse organizers	A. Topic introduction <i>in this chapter we</i>	
	B. Topic elaboration/clarification <i>on the other hand</i>	
3 Referential expressions	A. Identification/ focus <i>one of the things</i>	
	B. Imprecision <i>or something like that</i>	
	C. Specification of attributes	C1) Quantity specification <i>a little bit of</i> C2) Tangible framing <i>in the form of</i> C3) Intangible framing <i>on the basis of</i>
	D. Time/Place/Text reference	D1) Place reference <i>in the United States</i> D2) Time reference <i>at the same time</i> D3) Text-deixis <i>as shown in table</i> D4) Multi-functional reference <i>in the middle of</i>