Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective

Please cite as

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. (2024). Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

Abstract

This research explores the POS-tag sequences that shape the transition from upper intermediate (B2 CEFR) to near-native proficiency (C2 CEFR) in a corpus of essays (n=32,410) from the Cambridge Learner Corpus. Gilquin (2018) and others have shown that POS tag sequences offer a holistic approach to extracting the most commonly used patterns without a starting point of an *a priori* set of words and word sequences. Using corpus linguistics informed by usage-based theories of language learning, this paper examines the frequency and distribution of 4-slot POS-tag sequences in L2 English writing, drawing on the taxonomy of pattern grammar (Francis et al. 1996, 1998; Hunston & Francis, 2000). Findings point to the presence of both core and emergent POS-tag sequences in learner language in the two proficiency levels analysed. These sequences point to the presence of dynamic language restructuring processes as learners become more proficient and re-evaluate their understanding of frequency and distribution in English. This paper shows evidence of how language competence increases with proficiency. The research offers new evidence to our understanding of the development of L2 writing in EFL contexts.

Key words: learner corpora, language proficiency development, CEFR, usage-based, pattern-grammar, POS tag sequences, POS n-grams

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. 2024.Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

1. INTRODUCTION

Usage-based theories of language acquisition (Ellis et al., 2015) provide new ways to interpret learner corpora, offering a model that explains L2 learning as a fully abstracted system stored on a continuum of formulaicity, from heavily entrenched chunks to syntactically connected strings. L2 learning can be thus seen as the result of the abstraction of the statistical properties of language, including how variation operates in registers and how these abstractions shape up across L2 proficiency levels. Understanding how the statistical properties are acquired by L2 speakers is a major challenge for usage-based models (Ortega & Tyler, 2018).

Learner language research has explored L2 writing (1) in relatively small corpora (Aarts & Granger, 1998; Paquot & Granger, 2012) and (2) in cohorts where schooling year or age are used as proxies for language competence (Meunier, 2015) rather than attested performance levels (Green, 2010). The prevailing methodology used for the analysis of learner language involves comparing the results native corpus data and identify errors and patterns of learner 'over- and underuse of formulaic sequences' (Paquot & Granger, 2012: 132). We use the term 'learner' as it is understood within the context of learner corpora 'defined as electronic collections of natural or near-natural data produced by foreign or second language (L2) learners' (Granger et al. 2015: 1). The term 'learner' within the context of this study refers to exam takers whose writing scripts make up the Cambridge Learner Corpus (CLC) and encompasses the entire range of proficiency, from early stage users of an L2 to highly proficient users of language. Our research seeks to examine usage in the largest written learner corpus to date, the CLC, which has metadata for performance level, calibrated to the Common European Framework of Reference (CEFR) (see Section 3.1). Using a corpus of B2 and C2 level CLC writing (around 11.5 million words), this study adopts a bottom-up approach, aiming to investigate how POS-tag sequences (Aarts and Granger 1998; Granger and Rayson, 1998; Capelle and Grabar, 2016; Gilquin, 2018) and pattern grammar (Hunston and Francis, 1999; Hunston, 2019; Hunston & Su, 2019) characterise English for Foreign Language (EFL) writing.

This paper explores the transition from upper-intermediate (B2 CEFR) to advanced (C2 CEFR) proficiency levels by comparing the frequency, distribution and usage of POS-tag sequences by drawing on the potential of pattern grammar to show 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) and other morpho-syntactic units. The chosen proficiency levels represent in CEFR terms both independent language users (B2) and proficient users (C2) displaying the highest levels of proficiency according to the CEFR scale.

2. CORPUS-DRIVEN ANALYSIS OF LEARNER LANGUAGE DEVELOPMENT

2.1 Examining learner language beyond the word unit: Part of Speech (POS) sequences

Using a POS-tagged corpus, it is possible to search for POS n-grams (Capelle & Grabar, 2016), from which the high frequency of the sequences can inform expressions of syntactic patterning (Kennedy, 1996), phraseological patterns (Granger & Bestgen, 2014) and even constructions (Capelle & Grabar, 2016). 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. For Hunston & Su (2019: 568), pattern grammar 'generalizes from the patterning of individual words as observed through concordance lines from a large corpus'. To differenciate 'pattern grammar' from 'grammar pattern', the former is an *approach* to the grammar of English (Hunston & Francis, 2000), whereas the latter refers to the ways in which words are used (e.g. complementation of verbs). Hunston has argued that grammar patterns and the identification of the meaning groups (Francis et al., 1996, 1998) can be used as the basis for the ultimate identification of constructions at a consistent level of specificity. Grammar pattern coding uses abbreviated symbols to stand for word classes (verbs, nouns, etc.) or clause types (to-infinitive, that-clauses, etc.). Hunston notes that the term '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).

Granger & Rayson (1998: 125) used POS tags to mine 'significant patterns of over- and underuse' and suggested their approach offered an accurate picture of national interlanguages. Aarts & Granger (1998) highlighted the usefulness of POS tags in SLA research echoing Kennedy's (1996) observation that the co-occurrence of tags as expressions of syntactic patterning can provide further insight into the quantitative analysis of syntactic structures and processes. They found distinctive trigrams in three corpora of L2 English learners (the most

frequent being Preposition + Article + Noun) and identified specific sequences that characterised the language of Dutch, Finnish and French learners. They concluded that tag sequences can be used to 'rank interlanguages according to their degree of divergence from the native speaker norm' (1996: 140). Cross-sectional designs examine language as used by language learners at a single point in their development. They offer researchers the opportunity to capture some of the linguistic features that characterise usage though they impose limitations when looking at L2 usage in relation to development.

A wide range of studies have used n-grams to extract lexical bundles (e.g. Allen, 2009; Juknevičienn̄ė, 2009; Ping, 2009), collocation (e.g. Groom, 2009; Granger & Bestgen, 2014), formulaic sequence (e.g. Götz & Schilk, 2011), and clausal sequences (e.g. De Cock, 2007). Gilquin (2018) advocates the use of POS tags to probe into the sequences of L2 speaking when compared with L1 speakers. Using the LINDSEI corpus and its L1 counterpart, she extracted POS tag sequences (or POS n-grams as Capelle & Grabar (2016) have termed them) revealing that basic constructions such as [NP] and [Subj V] were more commonly used than complex constructions across both datasets. Gilquin (2018) holds that this approach can enhance the representativeness and generalizability of data. In our research, we have adopted the methodology proposed by Hunston (2019) where the term 'construction*'* is co-terminous with each of the meaning groups identified in grammar patterns.

2.2 Corpora and L2 writing development in EFL contexts

In instructed L2 contexts, students are expected to demonstrate their language competence by writing argumentative pieces. The development of argument is considered to be one of the key indicators of a successful writer (Lea & Street, 1998). Two main approaches have been used by corpus researchers: contrastive interlanguage analysis (CIA) (Granger, 1994, 2015) and studies of language development.

CIA has produced a considerable number of studies of learner language usage in EFL contexts (Gilquin & Paquot, 2008), suggesting that learners, teachers and material developers need more awareness of the frequency of L1 (English) linguistic features. By the 1990s, salient differences between L2 writers and L1 writers had already been explored. Silva (1993) scrutinised 72

empirical studies which involved a direct L1: L2 comparison. Overall, generalised findings were that L2 writing lacked fluency, was more erroneous than L1 writing, was simpler and contained patterns distinct from L1 writing. The International Corpus of Learner English (ICLE) (Granger 1996) has been widely used in CIA studies to look at various lexical and lexico-grammatical features (Gilquin & Granger, 2011; Thewissen, 2013)), including involvement features (Ädel, 2008), comparison of rhetorical functions in academic discourse (Gilquin & Paquot, 2008) and linguistic features of L2 essays in different components of the ICLE (Park, 2017).

Other studies have examined argumentative writing by observing various linguistic developmental patterns in L2 (Chen & Baker 2016; Staples *et al.* 2013) and L1 (Staples *et al.* 2016) separately, as well as comparing the performance of L2 English writers with that of L1 speakers (Chen & Baker 2010). Staples *et al*. (2013: 224) found that formulaic patterns in learner writing evolve towards 'self-constructed language as their proficiency increases' and that variability is very limited in terms of the functional use of bundles or the degree of fixedness. Mazgutova and Kormos (2015) observed the syntactic and lexical development in L2 argument writing from an English for Academic Purposes (EAP) programme, revealing noticeable differences in the lexical and syntactic development between lower-intermediate and upper-intermediate level proficiency groups, particularly in the lower-intermediate group.

3. RESEARCH METHODOLOGY

Our research uses two subsets of the CLC to examine the transition from B2 to C2 language through the analysis of the most frequent POS tag sequences in both datasets. We illustrate how these sequences can serve as the basis to identify patterns and track the development from intermediate to advanced L2 writing. Using corpus linguistics informed by usage-based theories of language learning, we take a bottom-up, big-data approach to examine learner language and its development across proficiency levels. The backdrop of our approach is Hunston's (2019) language patterns and her suggestion that 'constructions' refer to a sub-set of instances of a grammar pattern following Francis et al. (1996, 1998).

3.1 Data: the Cambridge Learner Corpus

The CLC is a 55-million-word corpus consisting of over 250,000 exam scripts collected from 1993 to 2012. The largest learner corpus to date, the CLC includes exam scripts, all benchmarked to the CEFR, from exam candidates from over 200 countries, with more than 140 different first languages. Exams consist of the Cambridge English tests (e.g. BEC, CAE, CPE, KET and PET) and each script is tagged with metadata including first language, nationality, CEFR exam and learner performance level, year, task type and question prompt. The corpus is owned by Cambridge University Press (CUP) and is used for in-house research, including materials development. CLC was queried on a bespoke version of Sketch Engine provided by CUP, but access to the raw corpus was not possible due to personal data protection protocols in place. Other uses include McCarthy's (2006) study of the CEFR and learner language, Hawkins & Filipović's (2012) analysis of criterial features and O'Keeffe & Mark's (2017) profile of learner competence in grammar leading to the English Grammar Profile resource.

3.2 Sub-corpora used in this study

The study uses two sub-corpora of all B2 and C2 argumentative-tagged texts. In the CLC, argumentative writing is defined broadly as exam responses where learners offer their opinion or point of view. Examples of argumentative texts in the CLC include various formats such as letters, reports, notes, and emails. Overall, the exam scripts consist of a wide range of writing tasks and topics encompassing giving opinions about issues around the world such as disadvantages of single-sex schools, to discussing and providing solutions for environmental problems, to elaborating on the advantages and disadvantages of topical trends.

3.2.1 B2 sub-corpus

The B2 sub-corpus consists of 5,416,524 words from 12,684 candidates. 58% were written by university students, 28% by secondary school students and 2% by primary school children. Approximately 47% of the candidates were male and 53% female. There are over 46 different L1s in the sub-corpus with Chinese, Greek, Portuguese, Farsi, and Polish ranking most frequently. Figure 1 below shows the number of words of the top 15 L1 sub-corpus.

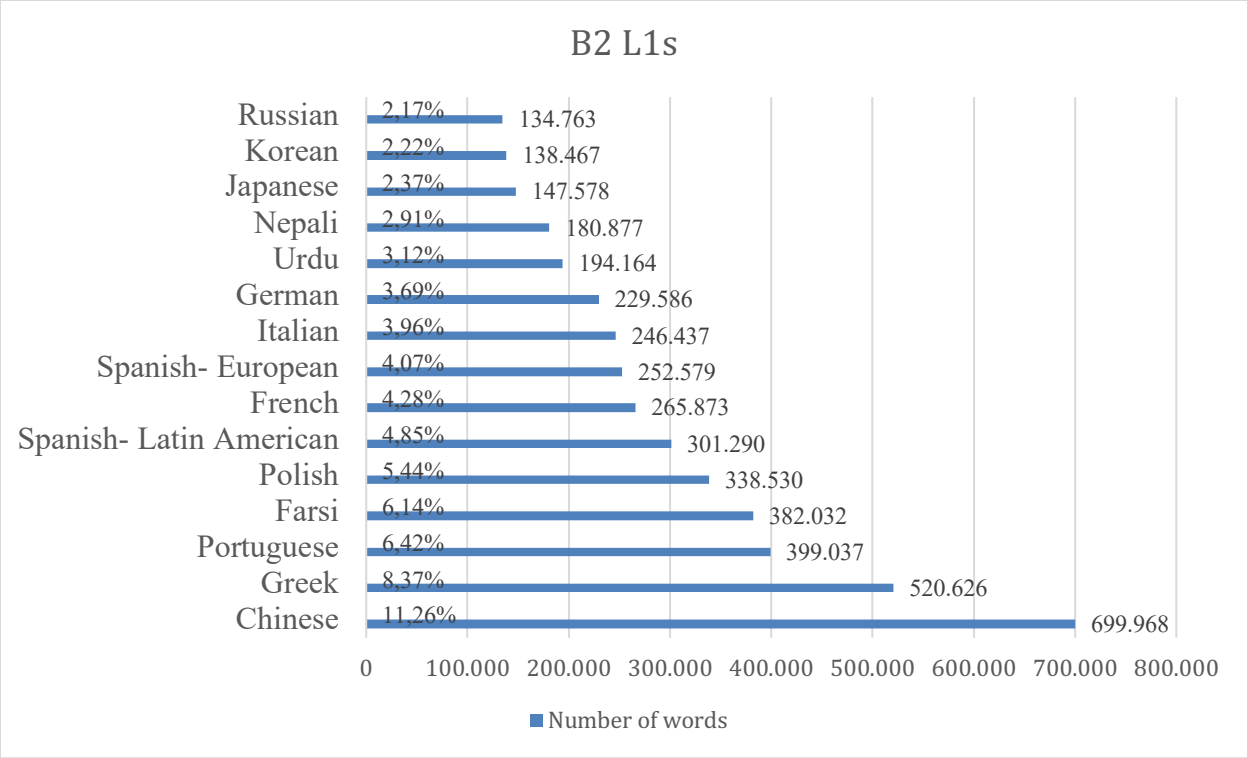

Figure 1. Top 15 B2 L1s in the CLC.

Although all of the written samples in the corpus include different types of argumentative tasks, there is a wide range of formats which include letter and composition (28%), letter/reference (26%), report (12%), informative/instructional text (11%), article (7%), emails (7%), story (3%), proposal (1%), and survey (0.9%). The data was tagged for task type by the CLC team.

3.2.2 C2 sub-corpus

The C2 sub-corpus consists of texts written by 9,096 participants, totalling 6,134,475 words. Among them, 56% were university students, 15% secondary school students and 0.4% primary school children. There were more female (53%) than male (47%). There are over 110 different L1s in the corpus including Greek, German, Portuguese French and Spanish. Figure 2 shows the token of the top 15 L1 learner groups in this sub-corpus.

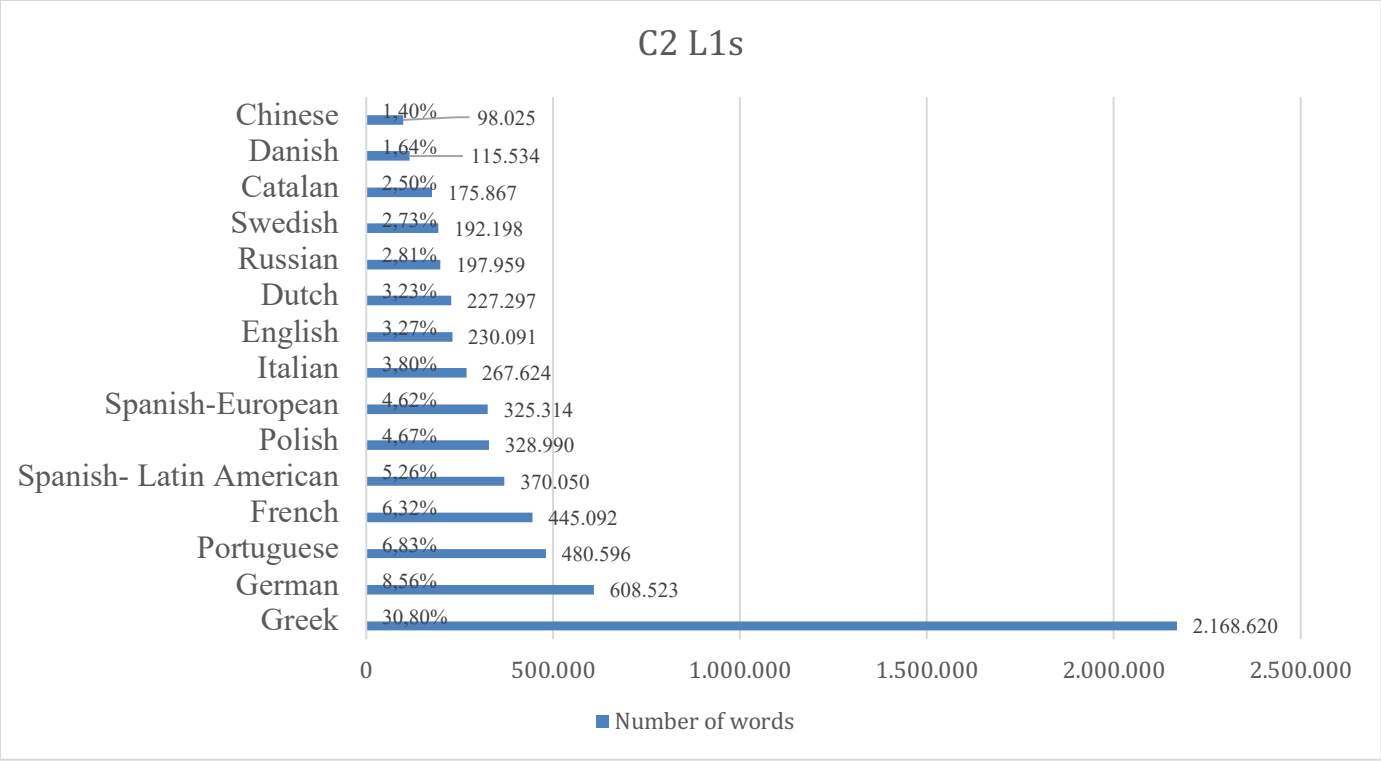

Figure 2. Top 15 C2 L1s in the corpus.

There is a wide range of task types, including composition (27%), letter (22%), informative/instructional texts (13%), article (12%), proposal (8%), story (8%), report (7%), review (2%), and note (0.05%).

In the B2 sub-corpus, the percentages of Chinese (11.26%) and Greek (8.3%) candidates are much higher than that of other L1s. In the C2 sub-corpus, the L1 Greek data constituted 30.80%. To test the effect of this on our results, we carried out our analyses by both including and excluding the top ranking L1 data, Chinese and Greek. The frequency and distribution ranks of the POS tag sequences remained the same for both B2 and C2 levels, regardless of whether these L1s were included or excluded. For this reason, all L1 data from the sub-corpora were included in our analyses.

3.3 Identification and discussion of POS sequences and patterns

Our research observes in micro-detail some of the sites of emergence of form and meaning combinations in L2 language development. After Gilquin (2018), we first identified the

frequencies and distribution of top 30 4-gram POS-tag sequences in the two sub-corpora and then analysed two tag sequences that are representative of developmental change in the two levels analysed. These instances detail the specific words that have been found to occur in each pattern, and which are divided into groups based on meaning. Following Hunston (2019), each lexical realisation was categorised using a pattern grammar approach, applying the taxonomy set out in Hunston and Francis (2000)[1]. This means first identifying form groupings or 'grammar patterns', e.g. N of n (noun of noun), N to n (noun to noun) and then meaning groupings (e.g. era/fraction/site, access/response).

4. RESULTS

In this section, we show how 4-gram POS tag sequences were distributed in the CLC sub-corpora across the B2 and the C2 levels. Appendix 1 shows the 30 most frequent sequences in the B2 and C2 sub-corpora.

Figures 3 and 4 show the most frequent sequences in the CLC B2 and C2 sub-corpora. On the horizontal axis of both figures we find the rank order of frequency in which the sequences appeared in the sub-corpora, while on the vertical axis we find the normalised frequency per 1 million words. While the 30 most frequent sequences account for 49.2% of all 4-gram POS-tag sequences in the B2 data, they account for 53.4% in the C2 sub-corpus.

[1] (See also https://grammar.collinsdictionary.com/grammar-pattern)

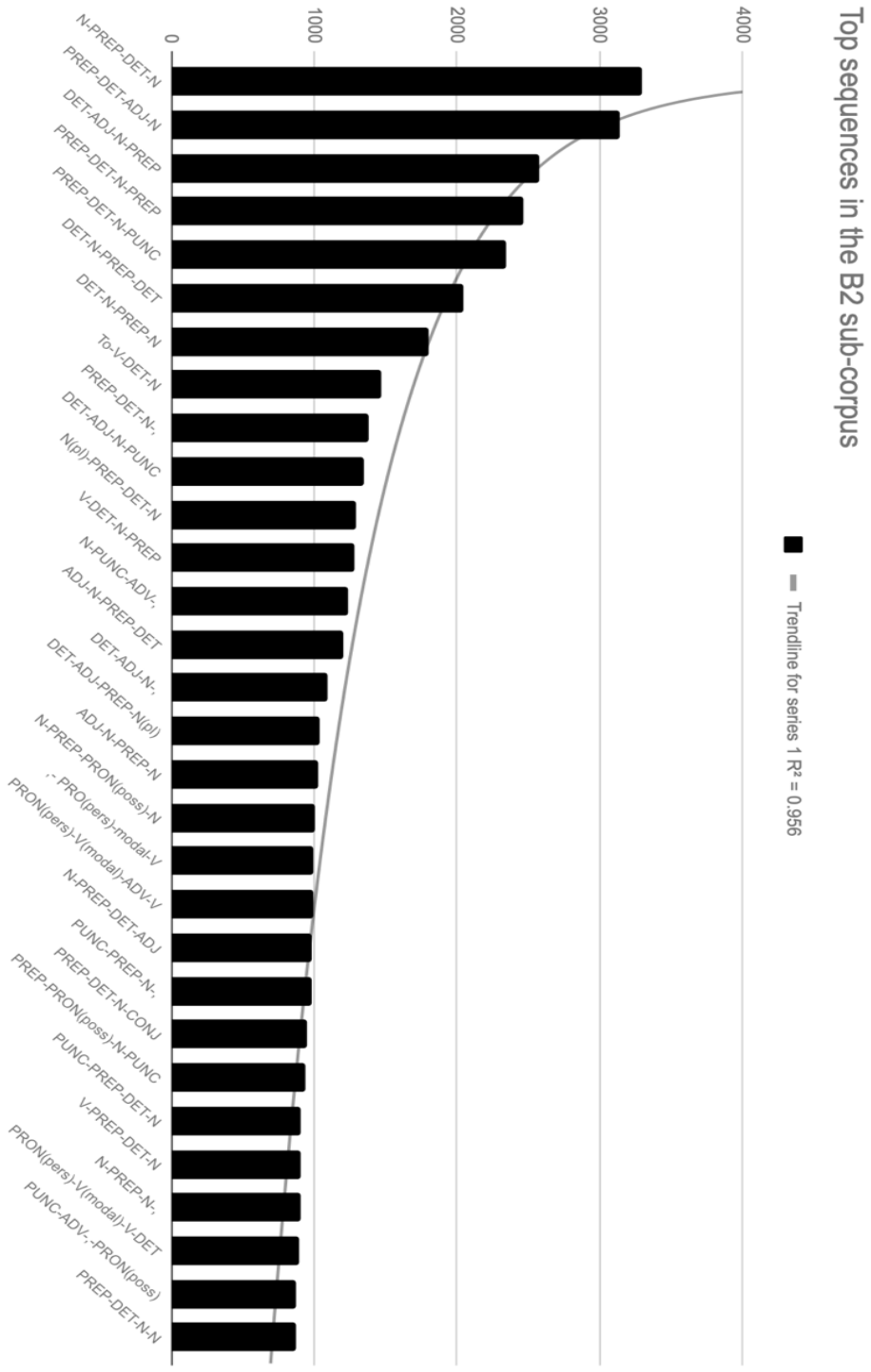

Figure 3. Most frequent POS- tag sequences in the B2 corpus

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. 2024.Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

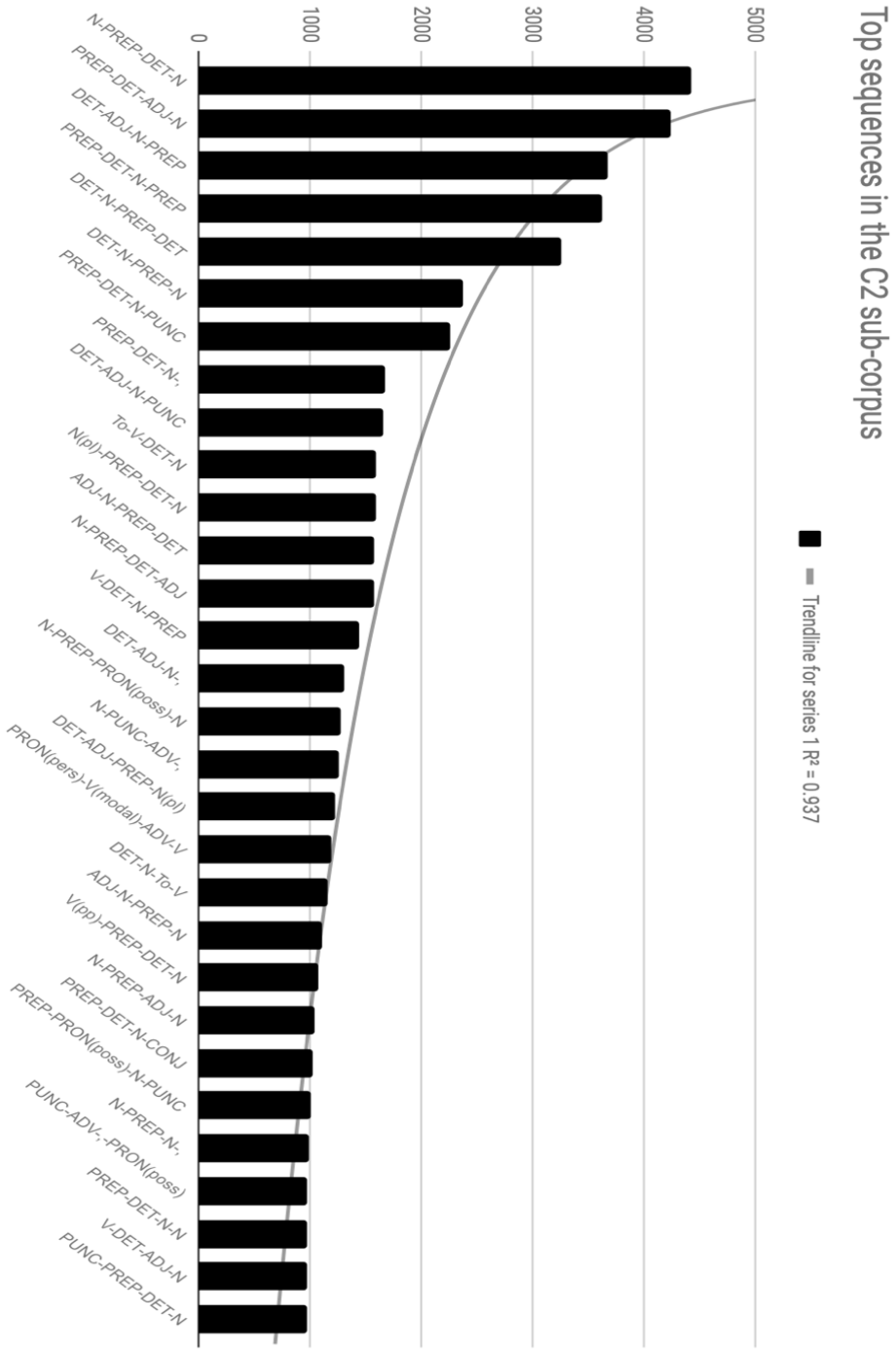

Figure 4. Most frequent POS-tag sequences in the C2 corpus.

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. 2024.Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

The most frequent sequence for both the B2 and C2 sub-corpora was the noun-preposition-determiner-noun (e.g. *member of the family*) sequence. In the B2 dataset, the sequence was used 17,850 times (3.83%), or 3,295 times per million words. The most frequently used phrase was *aim of this report* followed by *purpose of this report*, *end of the year* and *aim of this proposal*. In the C2 dataset, the noun-preposition-determiner-noun sequence appeared 27,937 times (4.50%), which is 4,424 per million words. At C2, the most frequently used phrase was *library with an internet*, followed by *aim of this proposal, purpose of this proposal, growth in the world* and *response to the article.*

The logarithmic trendline r^2 values were 0.96 and 0.94 for the B2 and C2 data, respectively, which suggests that both values are a good fit of the line to the data as the coefficient of determination explains over 90% of the variability.

4.1 Types of sequences: a developmental perspective

Three types of sequences characterise the transition from B2 to C2 writing. The first type is 'core sequences', those 4-gram POS tag sequences that appear both in the B2 and the C2 data in the top 30, of which there are 25 (Appendix 1). The second type are sequences that are used much less frequently by C2 writers than by B2 writers, usually descending tens of rank positions. There are 3 of these. The third type are those that emerge in the C2 top 30. In other words, these are sequences that are used more frequently in the C2 data. They are either not found in the B2 data in the same ranking or found further down the B2 top 30 ranking.

4.1.1 Type 1: Core sequences

A substantial part of the 30 sequences analysed can be regarded as core sequences in the writing. Four of these rank in the same order in the top 4 of the most frequent sequences in both datasets. The top 4 with examples from both B2 and C2 are:

#1 Noun-Preposition-Determiner-Noun

> *The main **aim of this proposal** is to suggest the best facility which will improve a quality of learning languages in the St. Paul's college.* (B2 performance, L1 Polish, CAE)

> *The **aim of this proposal** is to outline the three possible uses of the land as presented by your company.* (C2 performance, L1 Greek, CPE)

#2 Preposition-Determiner-Adjective-Noun

*But **on the other hand**, a lot of zoos in the world are in a very bad economic situation.* (B2 performance, L1 Polish, FCE)

***On the other hand**, there are some arguments against living longer.* (C2 performance, L1 Portuguese, CPE)

#3 Determiner-Adjective-Noun-Preposition

*The advantage of this company is that it offers **a wide range of** drinks including herbal teas and good coffee.* (B2 performance, L1 Russian, CAE)

*This very special part of the city combines anything a young person could wish for, offering **a wide range of** day-and-night-activities.* (C2 performance, L1 German, CPE)

#4 Preposition-Determiner-Noun-Preposition

*Finally, I think that the best idea would be to have a disco **at the end of the** course because most people want to do something different.* (B2 performance, L1 Greek, FCE)

***At the end of the** day you should be pleased at work and simultaneously ensure that you have enough free time*. (C2 performance, L1 German, CPE)

All in all, 25 of the 30 sequences (80%) are found in both B2 and C2 writing. In terms of frequency and rank (Appendix 1):

1. The four most frequent sequences are ranked identically in B2 and C2 sub-corpora;

2. The 10 most frequent sequences are identical in B2 and in C2.

3. Of the 20 most frequent sequences in C2, 18 are already in the top 20 B2 data.

The implication here is that B2 learners have already started to understand the probability that a set of sequences are appropriate when mapping the meaning demands (Ellis, 2012b) of the types of tasks found in the data analysed. As an aside, a search in the essay subset (728,000 words) of the Cambridge International Corpus (CIC) (1.597 million-word L1 corpus used by CUP to develop ELT materials) shows that nine of the ten most frequent sequences are found both in the CIC top and the C2 sub-corpus top 10 (with the exception of DT JJ NN SENT). This comparison has to be interpreted cautiously as writing tasks across the CLC and the CIC are not always comparable. However, it may indicate that C2 writers´ appreciation of the frequency and distribution of sequences is close to that of L1 English data.

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. 2024.Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

In Preposition-Determiner-Adjective-Noun (#2), we find that *on the other hand*, *at the same time* or *for a long time* top up the ranks in both B2 and C2:

> ***At the same time****, it cases pollution*. (B2 performance, L1 Italian, FCE)
>
> ***At the same time****, stubborn people are often very strong persons who have great difficulty to admit that they are wrong.* (C2 performance, L1 Swedish)

While in Preposition-Determiner-Noun-Preposition (#4) *at the end of* and *in the middle of* are the two exponents most frequently used by the writers in the two datasets:

> ***At the end of*** *my study I have to write a report and essay*. (B2 performance, L1 Polish, SfLL1)
>
> ***At the end of*** *the day you should be pleased at work and simultaneously ensure that you have enough free time.* (C2 performance, L1 German, CPE)

Lexical exponents on the top frequency rank orders vary little from B2 to C2 data. Equally, we can see that there is a task effect in play at B2 for the third ranking sequence (Determiner-Adjective-Noun-Preposition) where we find *the second quarter of* (e.g. *The passenger revenue increased quite steadily until* ***the second quarter of*** *2006, when it was at its peak moment*.) and *the third quarter of* (*There were some problem with punctuality in* ***the third quarter of*** *2005 year but even though 85% of the train arrived according to the plan*.) as part of the BECH exam across the world in 2007 in B2. All of the references with respect to the use of these particular phrases were followed by a year, indicating the presence of task effect. The most frequent lexical exponents found in C2 data suggest a transition towards more frequent use of formulaic language, featuring collocations more frequently in C2 (*a wide range of* is used 57.3 per million words in C2 vs 24.4 per million words in B2) and sequences in C2 that are rare in B2 (*the vast majority of, an integral part of)*. While most of these lexical sequences are concerned with the expression of quantity or with building argumentation, in B2 we find bundles such as *a new shop in*, *a new collection of* or *a special day in*, which are not formulaic.

Of the core sequences which are more frequently used by C2 writers and which occur higher up in the rank, one in particular involves the use of verbs. Personal Pronoun-Modal-Adverb-Verb (Rank 20 in B2, rank 19 in C2). Here we find that C2 writers use *I would also like*, or *I will*

never forget:

> ***I will never forget*** *the gratitude in her eyes and relief in her facial expression* (C2 level performance, L1 Bulgarian, CPE)
>
> ***I would also like*** *to remind you that Tall Trees Campsite is a popular British holiday resort and we get free accommodations.* (C2 level performance, L1 Chinese, CAE)
>
> *I would also like* in the C2 data is mainly followed by *comment, point out, mention*, while B2 writers prefer to use *know* or *thank you*.
>
> ***I'd also like*** *to know what is the exact price if it is half of the normal price*. (B2 level performance, L1 Chinese, FCE)
>
> ***I'd also like*** *to thank you for your advice* (B2 level performance, L1 Portuguese, FCE)

4.1.2 Type 2: Sequences increasingly less relevant in C2 writing

C2 candidates drop some sequences from their core repertoire in favour of more potent formulaic sequences. A sequence that has potency at B2 but which is less used at C2 is Preposition-Modal Auxiliary-Verb-Determiner (rank 28 in B2, rank 68 in C2). This is used by writers to interact with referential meaning and argumentation by means of exponents such as *we can see the*, *you can find a* or *we can see that*:

> *From the big mountain of the island* ***you can see the*** *sunset which is very beautiful and romantic* (B2 performance, L1 Greek, FCE)

This sequence (rank 28 at B2) is widely used when describing graphs and pictures. For C2 writers, this sequence is found in rank 68, which suggests that other means of expression have become more frequent in their writing. Sequences such as To-Verb-Determiner-Noun (*to find a job*) or Adjective-Noun-Preposition-Noun (*large amount of money*) also decrease a few rank positions. Further qualitative investigation of these sequences will be necessary to draw conclusions from the shift in these ranks.

4.1.3 Type 3: Emerging sequences in C2 writing

Five sequences emerge in C2 writing. Noun-Preposition-Determiner-Adjective (Rank 13 in C2, rank 21 in B2) is used at C2 to offer discussion of visual elements, typically figures and charts in some of the tasks. Some of the language underscores change, differences and similarities

(*percentage of the total*, *percentage in the third*, *millions in the third*), but also in the context of complex noun phrases where the headword is post modified by a prepositional phrase:

> *Not every profession has the* ***knowledge of a foreign*** *language as a prerequisite.* (C2 performance L1 German, CPE),

The lexical items used at C2 in Determiner-Noun-'TO'-Verb (Rank 19 in C2, rank 39 in B2) (see 4.3 for case study) include *the opportunity to learn*, *a lot to offer* or *the chance to meet*. Here C2 writers use non-finite to-clauses (Biber et al, 1999: 604) as postmodifiers:

> *we rarely have* ***the opportunity to learn*** *new things beyond the standard curriculum.* (C2 performance, L1 Greek, CPE)

> *Taking a holiday, also gives us* ***the opportunity to spend*** *time on ourselves and acquiring new experiences.* (C2 performance, L1 Spanish, CPE)

We take a more in-depth look at this sequence in section 4.3.

V(past participle)-Preposition-Determiner-Noun (Rank 22 in C2, rank 61 in B2) includes exponents such as *created by this situation*, *stuck in a traffic* or *come to the conclusion*. Some of these sequences are found after auxiliary verb *be* as in the following example:

> *The lunch shouldn't be* ***included in the price****! Some of us are vegetarians!* (C2 performance, L1 Romanian, CAE)

Noun-Preposition-Adjective-Noun(plural) (Rank 23 in C2, rank 36 in B2) is used with nouns such as *number*, *problem*, *majority*, *discussion* or *use* followed by a prepositional phrase where the noun complement headword is premodified by an adjective:

> *Countries all over the world are banning the* ***use of private cars*** *in town centres to decrease the levels of pollution.* (C2 performance, L1 Spanish, CPE)

The final emerging sequence is Verb-Determiner-Adjective-Noun (Rank 29 in C2, rank 33 in B2). C2 writers used formulaic sequences such as *play an important role*, *start a new life* or

make the right choice when the noun is premodified by an adjective and the NP follows a transitive verb:

> *Videos and tape-recordings complement the book collection. Both **play a major role** in modern language education.* (C2 performance, L1 Dutch, CPE)

In the following two sections we discuss in more detail the top core sequence (type 1) (Noun-Preposition-Determiner-Noun) in 4.2, and the top emerging sequence (type 3) (Determiner-Noun-'TO'-Verb) in 4.3. These two examples illustrate the type of analysis that could be carried out with each of the POS tag sequences in the sub-corpora.

4.2 Core sequence case study: Noun-Preposition-Determiner-Noun (NN IN DT NN)

Noun-Preposition-Determiner-Noun is the most frequent POS tag sequence (NN IN DT NN) in both B2 and C2 sub-corpora (Appendix 1). By way of case study, we looked at the top 20 most frequent lexical realizations of this sequence, illustrated in Table 1:

	B2	**C2**
1	part of the world	library with an internet
2	aim of this report	aim of this proposal
3	work for a company	purpose of this proposal
4	end of the day	growth in the world
5	end of the course	response to the article
6	part of the film	aim of this report
7	side of the island	understanding of the world
8	sailing on the lake	use of the land
9	part of any group	end of the day
10	publicity for the club	response to the campaign
11	rest of the world	part of the world
12	% of the energy	centre of the town
13	visit to a night club	solution to this problem
14	part of the city	access to the internet
15	time of the day	side of the coin
16	result from this energy	rest of the world

17	front of the TV	centre of the city
18	spite of the fact	society as a whole
19	society as a whole	solution to the problem
20	season of the year	response to an article

Table 1. Top 20 most frequent lexical realisations of Noun-Preposition-Determiner-Noun at B2 and C2

We note here some obvious task-related phrases which were part of question rubrics, for example ‘sailing on the lake’, ‘library with an internet (connection)’. In these two cases, representing, the task is heavily reliant on repetition of these phrases from the rubric and could not be undertaken without frequent reference to them. For example in the case of *library with an internet*, the task revolves around selection and discussion of three possible uses of an area of land for community use, one of which, and the most popular answer is *a library with an internet café*. With access to the question bank of the CLC we are able to take this into consideration and view them as outliers. Neither removal nor inclusion of these task-related phrases has any effect on the findings described below.

As mentioned in 3.3, for each lexical realization, we first identified form groupings, also known as grammar patterns, and then the meaning groups. In the cases where we found no corresponding meaning group in Hunston and Francis’ (2000) taxonomy, items were tagged as ‘uncategorised’. (Table 2) We found that the patterns spanned six form groupings at B2 and five at C2. At B2 N of n (e.g. *aim of this report*) accounts for 70% of the examples, across six different meaning groups; N for n (e.g. *publicity for the club*) accounts for 10% and the remaining categories of N to n, N as n, N from n and N on n account for 5% of groupings. In summary, the N of n pattern was dominant at B2. In the top 20 C2 patterns and meanings, there was a shift away from the prevalence of the N of n pattern. C2 candidates appear to have broadened how they deploy their form-meaning pairings, with the patterns N of n and N to n accounting for 55% and 30% respectively, with N as n, N in n, N with n accounting for 5% each (see also Table 2). Although the N of n group accounted for fewer examples at C2, the range of meanings it expressed increased.

Grammar pattern	Meaning group	Top 20 B2 examples	Grammar pattern	Meaning group	Top 20 C2 examples
N of n (70%)	aim	aim of this report	N of n (55%)	aim	aim of this proposal aim of this report purpose of this proposal
	era	end of the course end of the day season of the year time of the day		era	end of the day
	fraction	part of any group part of the city part of the film part of the world rest of the world		fraction	part of the world rest of the world
	percentage	% of the energy		issue	use of the land
	site	front of the TV side of the island		site	centre of the city centre of the town side of the coin
	uncategorised	spite of the fact		support	understanding of the world
N to n (5%)	journey	visit to a night club	N to n (30%)	access	access to the internet
N as n (10%)	uncategorised	society as a whole		response	response to an article response to the article response to the campaign
N for n (10%)	spokesman	publicity for the club		solution	solution to the problem solution to this problem
	uncategorised	work for a company	N as n (5%)	uncategorised	society as a whole
N from n (5%)	emissions	result from this energy	N in n (5%)	increase and decrease	growth in the world
N on n (5%)	uncategorised	sailing on the lake	N with n (5%)	uncategorised	library with an internet

Table 2. Top 20 B2 and C2 occurrences of Noun-Preposition-Determiner-Noun with corresponding pattern grammar groupings

In terms of the development of form-meaning pairings within core patterns common to both B2 and C2, results from the top 20 in both datasets suggest that:

- Though patterns remain core across the learning pathway from B2 to C2, there is inherent development in terms of the range, type and nature of the form-meaning pairings.
- B2 learners rely more on one pattern across a range of meaning categories. 70% of the top 20 core patterns used at B2 were N of n across 6 meaning groupings. By C2, N of n accounts for fewer pattern realizations (55%) but it is across 6 meaning groupings. This suggests some kind of honing within the same pattern.
- At C2, there appears to be some movement towards fixedness of patterning, which suggests a type of a sensitivity to item co-selection and formulaic abstractions. Whereas at B2 we see quite a few literal references, possibly driven by tasks in the exam (*aim of this report, end of the course, time of the day, part of the film, front of the tv, side of the island, visit to a nightclub*), by C2, we see the emergence of more formulaic use. For instance, we see more shell nouns, nouns 'which can be identified on the basis of their behavior' (Hunston & Francis, 1999: 185) followed by a post modifier (*understanding of the world, access to the internet, solution to the problem, society as a whole*). This seems to suggest that C2 writers 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.

4.3 Emerging sequence case study: Determiner-Noun-TO-Verb (DT NN TO VV)

The emerging sequence Determiner-Noun-TO-Verb (DT NN TO VV) becomes more frequently used in the C2 data, moving in rank order of frequency from 39 in B2 to 19 in C2 data (Appendix 1). As with the core sequence case study (4.2), each of the top 20 lexical realisations was categorised using a pattern grammar approach (Hunston and Francis, 2000). Table 3 lists the top 20 lexical realisations of this pattern at B2 and C2: Overall, we identified more meaning groups at C2 (5) than at B2 (3). At B2 we found a dominance of both the ability group (35% of the top 20) and the productive group (45%). We found that use of the ability group increased to 60% at C2, and that other meaning groups (resources, proposal, rights) had emerged:

Meaning group	Top 20 B2 examples

Meaning group	Top 20 C2 examples

ability (35%)	a chance to see a chance to survive the chance to go the chance to see the opportunity to see the possibility to go the possibility to see
productive uses (45%)	a lot to do a bicycle to go a car to get the bicycle to go the car to get the car to go a meeting to discuss a place to live a place to park
Nouns with other meanings (10%)	a way to get a way to learn
Uncategorised (10%)	a zoo to see the zoo to see

ability (60%)	the opportunity to do the opportunity to enjoy the opportunity to get the opportunity to listen the opportunity to meet the opportunity to see the opportunity to travel a chance to see the chance to do the chance to know the chance to meet the chance to see
productive uses (20%)	a lot to do a lot to learn an advantage to live a place to live
resources (5%)	the time to do
proposal (10%)	a proposal to build the proposal to build
rights (5%)	the right to live

Table 3. Top 20 B2 and C2 occurrences of Determiner-Noun-TO-Verb with meaning groupings

Turning to the lexical realisations of the meaning groups: for the dominant *ability* group we found that B2 writers used *a chance, the opportunity/chance/possibility to* + verb, whereas at C2 we saw no instances of *the possibility to* + verb and a reliance on *the opportunity to verb*. However, this decrease in the noun choice is offset by expansion in the verb slot. In terms of the verb choice, we saw an increase from 3 verb types at B2 (*see, survive, go*) to 9 different verb types at C2. Overall at C2, there seemed to be a narrowing in on a more fixed formula in *the opportunity to* with a broadening of verbs. In other words, C2 writers appear to do more with the same pattern.

For all meaning groups, we observed an overall movement away from task or topic, often concrete, head nouns, (*bicycle, car, zoo*) towards increased use of abstract or figurative 'shell'

nouns (Hunston & Francis, 1999) 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, results from the top 20 suggest that C2 writers:

- can do more with the same patterns.
- deploy more form-meaning mappings.
- show a tendency for one or two 'pioneer' forms and shed less frequent forms.
- rely on more semi-fixed structures and less topic-oriented language.

5. DISCUSSION

Usage-based research has shown that language is a dynamic, developing system which restructures and grows as our experience of it broadens (Pérez-Paredes et al., 2020). Speakers develop an accumulating repertoire of form-meaning mappings which they draw on for productive and receptive use. We have examined a large corpus of L2 writing (11.5 million words) that spans two decades of examinations conducted by Cambridge Assessment across 21,780 learners of English, with tens of different L1s, age groups and exams. Rather than examining a set of *a priori* specific features of language, either lexical (e.g. stance adverbs) or grammatical (modal verbs, transitive verbs, etc.), we have adopted a bottom-up, data-driven exploratory approach in which we examine the language used by L2 writers at different stages of proficiency through the use of 4-gram POS tag sequences, irrespective of their L1 background.

Taking a bottom-up approach drives an open view on development, capturing all of the sequences used at both B2 and C2 performance levels and identifying those that are key to development. Our findings show evidence of how the acquisition of an L2 implies the restructuring of the frequency and distribution of language items (Ellis, 2006). Language development can be characterised through the presence of groups of sequences, as exemplified in this study. On the one hand, we find a stabilisation in the distribution of core sequences across both levels and, on the other, the emergence of and increase in use of sequences at C2 that play a lesser role at B2. Following Hunston (2019), through close analysis of the instantiations of the POS-grams, we have been able to observe where the sequences found in the learner data

correspond to groupings and meanings catalogued in pattern grammar. Applying pattern grammar taxonomy to these core and emerging sequences makes use of an existing robust framework on which to categorise form and meanings.

In the following paragraphs, we discuss how a dynamic 'restructuring' of sequences is manifested in the transition from upper-intermediate (B2) to proficiency (C2) level in the CLC.

We have seen that B2 and C2 writing is characterised by the presence of a core group of sequences that remain stable during the transition from one level to the next (see Section 4.1.1). In all, 83% of the top 30 sequences at B2 are also in the top 30 at C2. At the B2 level, learners have acquired a sense of what is also core to C2 writing. Gilquin (2018) also found that, in spoken language, among the top 30 most frequent POS tag sequences, 25 were shared by both groups. The presence of core sequences in our two sub-corpora suggests that B2 writers have been exposed to sufficient input so as to allow for a significant understanding of the frequency and distribution of the most important sequences used at C2. The ways in which the writers deploy core sequences become on the one hand more diverse and on the other more selective and formulaic. The same appears to be true for the emerging sequences. In the two exemplifications in this study (see 4.2 and 4.3), the range of meaning groups for the patterns increases while lexis previously used at B2 goes through a selection process where one or two nouns are preferred at C2. This has a resonance with a tendency, seen in both first and second language development studies focusing on verb argument constructions (VACs), for the highly frequent occurrence of one pioneering verb in each VAC (Goldberg et al. 2004; Ellis & Ferreria-Junior 2009; Ellis & Larsen-Freeman, 2009, Romer, 2019). Although this claim would require further examination, syntactic patterning appreciation (Wulff & Ellis, 2018) seems to be activated earlier than collocational knowledge. Language learners, as they encounter more and more opportunities to increase their performance through practice, seem to acquire first the most frequent sequences and then the most frequent lexical instantiations of these sequences. This is corroborated by Ninio (2011: 130) who, based on L1 corpus evidence, maintains that children first acquire the 'formal building blocks of sentence structure'. Syntactic pattern appreciation seems to be evident at B2. It is a collocational awareness and knowledge mapped to functional awareness

that is developing at B2, with movement towards more abstraction at C2. Interestingly, Thewissen (2013) also found syntactic stabilisation at B2.

Our findings also align with Gilquin (2018) in two further important respects. Firstly, the core sequences identified display head nouns in noun phrases. The types of constructions that are the focus of corpus-based construction studies (e.g. VACS) do not feature in the most used sequences in our data. In fact, our B2 learners seem to have already acquired stages 3 and 4 of the developmental framework of NP complexity advocated in Biber et al. (2011). Stages 1 and 2 involves the use of finite and non-finite complement clauses with an increasing range of verbs. The overall use of nouns in both the B2 and C2 sub-corpora is close to the percentage of use of nouns in academic writing reported in Biber et al. (1999). The presence of postmodification and the frequency of nouns in our two sub-corpora call for a re-assessment of Parkinson & Musgrave's (2014: 58) claim that postmodification is relatively infrequent 'in the writing of less proficient compared to more proficient L2 learners'. Secondly the most frequent sequences carry a vast array of instantiations. This is illustrated in the analysis of the core structure Noun-Preposition-Determiner-Noun (NN IN DT NN) (see 4.2), in which two of the four tags represent open word classes. It is perhaps not surprising that, when the tag represents an open word class, each tag sequence can be exemplified by thousands of different lexical instantiations, and particularly noteworthy are nouns (NN), which are the highest ranked word class in terms of individual tokens by a long way.

In the transition from B2 to C2 performance, there is considerable re-structuring that affects sequences in C2. On one level there is a reshuffling of the rankings in which emerging sequences become more prevalent. For the core sequences, we have seen through the analysis of Noun-Preposition-Determiner-Noun (NN IN DT NN) that when a sequence has potential for many instantiations, there appear to be several levels of restructuring, first of trial and then selection. First at a form level where B2 writers try out a wider variety of lexical items for a reduced range of forms, and secondly where at C2, writers select preferred lexical items for an increased range of forms. This often results in the use of sequences with 'shell' nouns carrying a greater degree of formulaicity and fixedness between items. This affords a view of development from slots and

frames to a fully abstracted system and a growing awareness of statistical frequencies at all levels of specificity.

C2 writing shows more sensitivity to noun phrasal structures. The emerging sequences we have seen in the C2 data involve post-modification slots in the NP and encompass structures such as prepositional phrases and non-finite clauses. This is consistent with Biber & Gray's (2016) finding that prepositional phrases in postmodifying slots create 'dense information structures' that may contribute to the use of fewer verbs than in other registers (2016: 192). They also argue that academic writing is characterised by the presence of phrasal complexity features much more characteristic, and thus more important than clausal complexity features (2016). Recent research has found that in beginner and pre-intermediate EFL levels, the use of countable nouns and prepositional phrases shape up the transition from lower to higher secondary school language learning (Pérez-Paredes & Díez-Bedmar, 2019). This is an interesting finding as the type of writing most widely used in EFL contexts differs from the writing found in academic settings and academic writing more generally (Hardy & Römer, 2013; Biber & Gray, 2016) where, typically, there is no time limitation and more planning and editing time is available. Written registers reflect careful planning, revision, and editing (Biber, 2006) in contrast with real-time online editing found in spoken communication (Biber et al., 1999.) In its current form, those learners taking the CPE exam, C2 level target, have 90 minutes to complete the two parts of the writing paper and write around 560 words. Learners writing under exam conditions may be accessing linguistic sequences that are easier to retrieve from subconscious statistical knowledge, drawing on their cognitive understanding of lexical behaviour and the chunks acquired.

Using a bottom-up data-driven approach, we have been able to uncover the most frequent sequences in the transition from B2 to C2 writing. The use of precise ways to address level performance can only benefit our understanding and descriptions of learner language. The C2 level is perceived as near-native, with a high degree of fluency, demonstrating precise language, displaying sensitivity to most contexts (Council of Europe, 2018). However, there is a general lack of understanding of how frequency and distribution of features generally affect how languages are learnt.

6. CONCLUSION

In this paper, through a data-driven approach, we have identified some of the most accessible options that language learners draw on (Tachihara & Goldberg, 2020) when using the core and emerging sequences analysed in 4.2 and 4.3. We see the identification of sequences as a 'way in' to observing how language is constructed and this in turn sheds light on how L2 learners restructure their mental representation of English. In usage-based terms, this is conceptualised in terms of 'constructions' which, as illustrated by Ellis et. al 2016, are what language learners learn. Hunston posits that if these are indeed what learners learn, 'then it makes sense that they are also what learners should be taught and teacher know about' (2019:1). However, this may be an overgeneralisation as we as yet do not know the degree to which these 'grammar patterns' (Hunston & Francis, 2000) are implicitly or explicitly acquired. Further experimental work is needed in this regard. Indeed, though there has been a considerable amount of empirical work on constructions to date, as Hunston notes, it offers detailed descriptions of a relatively small number of constructions (2019).

A POS-driven analysis allows for a level of generalisation and abstraction at both a syntactic and functional level that would not have been seen through a lexical lens. Applying a pattern grammar taxonomy first to formal groupings (categories) and then to meanings allows us to see that the C2 candidates make more frequent use of collocations and formulaic language and that a wider range of semantic meanings emerge in C2 writing. The relationship between the L1 and L2 linguistic store merits further investigation. For example, it remains to be seen whether the core sequences indentified in this study would be displayed in similar ways by L1 English speakers writing similar tasks. Additionally, as suggested by Staples et al. (2013), lexical bundles may be a reflection of the limited input that learners receive during learning. Triangulation from different corpora and of different language extraction techniques and units of analysis will be necessary so as to gain further understanding of whether these are global developmental features observed in other learner corpora. Similarly, further work is needed to test these findings in learner data using inferential statistics. We suggest that future work could focus on some of the core and emerging sequences identified in this study by adopting a longitudinal research design and a variety of tasks that complement the exam tasks in the CLC.

The frequency and distribution of syntactic sequences and their realisations need pedagogical consideration. In this work we offer an approach to observing development in a way that (a) shows evidence that language learning 'involves the distributional analysis of the language stream' (Ellis & Larsen-Freeman, 2009:95) and (b) offers practical ways to advance usage-based pedagogies that understand that 'syntax itself is meaningful and that syntactic patterns are templates abstracted' from usage (Tyler, 2010: 285). There is potential to use this process to identify which patterns and meanings occur most frequently so as to curate a pathway of developing patterns and meanings for the learner based on frequency and distribution.

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. 2024.Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

Acknowledgements

We would like to thank the anonymous reviewers for their useful comments on early versions of this paper and Cambridge University Press for providing access to the Cambridge Learner Corpus.

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. 2024.Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

REFERENCES

Ädel, A. 2008. 'Involvement features in writing: Do time and interaction trump register awareness?,' in G. Gilquin, S. Papp and M. Díez-Bedmar (eds): *Linking up contrastive and learner corpus research.* Brill, pp. 35-53.

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, pp. 132-141.

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.

Biber, D. 2006. *University language. A corpus-based study of spoken and written registers.* John Benjamins.

Biber, D., and B. Gray. 2016. *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Biber, D., B. Gray, and K. Poonpon. 2011. 'Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?,' *TESOL Quarterly* 45/1: 5–35.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman grammar of spoken and written language*. Longman.

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, pp. 271-302.

Chen, Y. and P. Baker. 2010. 'Lexical bundle in L1 and L2 academic writing,' *Language Learning & Technology* 14/2: 30-49.

Chen Y. and P. Baker. 2016. 'Investigating critical discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics* 37/6: 849-880.

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

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, pp. 217-233.

Ellis, N. 2006. 'Cognitive perspectives on SLA: The associative cognitive CREED,' *AILA Review* 19: 100-121.

Ellis, N. and F. Ferreira-Junior. 2009. 'Construction learning as a function of frequency, frequency distribution, and function,' *Modern Language Journal* 93/3: 370-385.

Ellis, N. and D. Larsen-Freeman. 2009. 'Constructing a Second Language: Analyses and computational simulations of the emergence of linguistic constructions from usage,' *Language Learning* 59/S1: 90-125.

Ellis, N., M. O'Donnell, and U. Römer. 2015. 'Usage-based language learning,' in B. MacWhinney and W. O'Grady (eds.): *The Handbook of Language Emergence*. Wiley, pp. 163-180.

Ellis, N., U. Römer, and M. O'Donnell. 2016. *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations of construction grammar.* Wiley.

Francis, G., S. Hunston, and E. Manning. 1996. *Collins Cobuild Grammar Patterns 1: Verbs.* HarperCollins Publisher.

Francis, G., S. Hunston, and E. Manning. 1998. *Collins Cobuild Grammar Pattern 2: Nouns and Adjectives.* HarperCollins Publisher.

Gilquin, G. 2018. 'Exploring the spoken learner English construction: A corpus-driven approach,' in R. Alonso (ed.): *Speaking in a second language*. John Benjamins, pp.127-152.

Gilquin, G. and S. Granger. 2011. 'From EFL to ESL: Evidence from the International Corpus of Learner English,' in J. Mukherjee and M. Hundt (eds.): *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*. John Benjamins, pp. 55-78.

Gilquin, G. and M. Paquot. 2008. 'Too chatty: Learner academic writing and register variation,' *English Text Construction* 1/1: 41-61.

Götz, S. and M. Schilk. 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, pp. 79-100.

Granger, S. 1994. 'The learner corpus: A revolution in applied linguistics,' *English Today* 10/3: 25-33.

Granger, S. 1996. 'Learner English around the world,' in S. Greenbaum (ed.): *Comparing English worldwide.* Clarendon Press, pp. 13-24.

Granger, S. 2015. 'Contrastive interlanguage analysis: A reappraisal,' *International Journal of Learner Corpus Research* 1/1: 7-24.

Granger, S., and Y. Bestgen. 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. and P. Rayson. 1998. 'Automatic profiling of learner texts,' in S. Granger (ed.): *Learner English on computer.* Addison Wesley Longman, pp. 119-131.

Goldberg, A. E., D. M. Casenhiser, and N. Sethuraman. 2004. 'Learning argument structure generalizations,' *Cognitive Linguistics* 15: 289-316.

Granger, S., Gilquin, G., & Meunier, F. (eds.). 2015. *The Cambridge handbook of learner corpus research*. Cambridge University Press.

Green, A. 2010. 'Requirements for Reference Level Descriptions for English,' *English Profile Journal* 1: E6.

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, pp. 21-33.

Hardy, J. A. and U. Römer. 2013. 'Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP),' *Corpora* 8/2: 183-207.

Hawkins, J. A. and L. Filipović. 2012. *Critical Features in L2 English.* Cambridge University Press.

Hunston, S. 2019. 'Patterns, constructions, and applied linguistics,' *International Journal of Corpus Linguistics* 24/3: 324-353.

Hunston, S. and G. Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins.

Juknevičienė, R. 2009. 'Lexical bundles in learner language: Lithuanian learners vs. native speakers,' *KaLBOTYRa* 61/3: 61-72.

Kennedy, G. 1996. 'The corpus as a research domain,' in S. Greenbaum (ed.): *Comparing English Worldwide*. Clarendon Press, pp. 217-226.

Lea, M., and B. Street. 1998. 'Student writing in higher education: An academic literacies approach,' *Studies in Higher Education* 23/2: 157-172.

McCarthy, M. 2016. 'Putting the CEFR to good use: Designing grammars based on learner-corpus evidence,' *Language Teaching* 49/1: 99-115.

Mazgutova, D. and J. Kormos. 2015. 'Syntactic and lexical development in an intensive English for academic purposes programme,' *Journal of Second Language Writing* 29: 3-15.

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 University Press, pp. 379-444.

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. 2024.Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

Ninio, A. 2011. *Syntactic development, its input and output*. Oxford University Press.

O'Keeffe, A. and G. Mark. 2017. 'The English Grammar Profile of learner competence: Methodology and key findings,' *International Journal of Corpus Linguistics* 22/4: 457-489.

Paquot, M. and S. Granger. 2012. 'Formulaic Language in Learner Corpora,' *Annual Review of Applied Linguistics* 32: 130-149.

Park, M. 2017. 'Native language identification of learner essays based upon ICLE-KR,' *Proceedings of the Korea Association of Teachers of English (KATE) 2017*: 269-276.

Parkinson, J. and J. Musgrave. 2014. 'Development of noun phrase complexity in the writing of English for Academic Purposes students,' *Journal of English for Academic Purposes* 14/C: 48-59.

Pérez-Paredes, P. and B. Díez-Bedmar. 2019. 'Researching learner language through POS Keyword and syntactic complexity analyses,' in S. Götz and J. Mukherjee (eds.): *Learner Corpora and Language Teaching.* John Benjamins, pp. 101-128.

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.

Ping, P. 2009. 'A study on the use of four-word lexical bundles in argumentative essays by Chinese English majors- A comparative study based on WECCL and LOCNESS,' *CELTA* 32/3: 25-45.

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.

Silva, T. 1993. 'Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications,' *TESOL Quarterly* 27/4: 657-677.

Staples, S., J. Egbert, D. Biber, and A. McClair. 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., J. Egbert, D. Biber, and B. Gray. 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.

Thewissen, J. 2013. 'Capturing L2 development through learner corpus analysis: Insights from an error-tagged learner corpus,' *The Modern Language Journal* 97/1: 77-101.

Tyler, A. 2010. 'Usage-based approaches to language and their applications to second language learning,' *Annual Review of Applied Linguistics* 30: 270-291.

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. 2024.Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

Tylor, 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, pp. 316-321.

Wulff, S. 2016. 'A friendly conspiracy of input, L1, and processing demands: *that*–variation in German and Spanish learner language,' in L. Ortega, A. E. Tyler, H. I. Park and M. Uno (eds.): *The usage-based study of language learning and multilingualism*. Georgetown University Press, pp. 115-136.

Appendix 1. Top 30 most frequent 4-gram POS tag sequences in CLC B2 and C2 sub-corpora (per million words)

Rank	B2	Norm'd freq	%	C2	C2	Norm'd freq	%	B2
1	NN IN DT NN Noun-Preposition - Determiner-Noun	3295	3.83	1	NN IN DT NN Noun-Preposition-Determiner-Noun	4424	4.50	1
2	IN DT JJ NN	3142	3.66	2	IN DT JJ NN Preposition-Determiner-Adjective-Noun	4241	4.31	2

Lim, J., Mark, G., Pérez-Paredes, P. & O'Keeffe, A. 2024.Exploring Part of Speech (POS)-tag sequences in a large-scale learner corpus of L2 English: A developmental perspective. *Corpora*, 19(1).

	Preposition-Determiner-Adjective-Noun							
3	DT JJ NN IN Determiner-Adjective-Noun-Preposition	2580	3.00	3	DT JJ NN IN Determiner-Adjective-Noun-Preposition	3681	3.74	3
4	IN DT NN IN Preposition-Determiner-Noun-Preposition	2464	2.87	4	IN DT NN IN Preposition-Determiner-Noun-Preposition	3629	3.69	4
5	IN DT NN SENT Preposition-Determiner-Noun-Punctuation	2339	2.72	7	DT NN IN DT Determiner-Noun-Preposition-Determiner	3252	3.31	6
6	DT NN IN DT Determiner-Noun-Preposition-Determiner	2047	2.38	5	DT NN IN NN Determiner-Noun- Preposition-Noun	2380	2.42	7
7	DT NN IN NN Determiner-Noun-Preposition-Noun	1802	2.10	6	IN DT NN SENT Preposition-Determiner-Noun- Punctuation	2255	2.29	5
8	TO VV DT NN To- Verb - Determiner-Noun	1473	1.71	10	IN DT NN , Preposition-Determiner-Noun- ,	1674	1.70	9
9	IN DT NN , Preposition-Determiner-Noun- ,	1382	1.61	8	DT JJ NN SENT Determiner-Adjective-Noun- Punctuation	1663	1.69	10
10	DT JJ NN SENT Determiner- Adjective-Noun- Punctuation	1344	1.56	9	TO VV DT NN To- Verb-Determiner-Noun	1598	1.62	8
11	NNS IN DT NN Noun (Plural)- Preposition-Determiner-Noun	1291	1.50	11	NNS IN DT NN Noun (Plural)- Preposition-Determiner-Noun	1592	1.62	11
12	VV DT NN IN Verb-Determiner-Noun-Preposition	1281	1.49	14	JJ NN IN DT Adjective-Noun-Preposition- Determiner	1579	1.61	14
13	NN SENT RB , Noun-Punctuation- Adverb- ,	1239	1.44	17	NN IN DT JJ Noun-Preposition- Determiner-Adjective	1569	1.60	21
14	JJ NN IN DT Adjective-Noun- Preposition-Determiner	1203	1.40	12	VV DT NN IN Verb- Determiner-Noun-Preposition	1437	1.46	12
15	DT JJ NN , Determiner-Adjective-Noun- ,	1095	1.27	15	DT JJ NN , Determiner-Adjective-Noun- ,	1315	1.34	15

16	DT NN IN NNS Determiner-Noun-Preposition-Noun (Plural)	1041	1.11	18	NN IN PP$ NN Noun-Preposition- Possessive pronoun-Noun	1272	1.29	18
17	JJ NN IN NN Adjective-Noun- Preposition-Noun	1024	1.19	21	NN SENT RB , Noun-Punctuation- Adverb- ,	1264	1.29	13
18	NN IN PP$ NN Noun-Preposition- Possessive pronoun-Noun	1003	1.17	21	DT NN IN NNS Determiner-Noun- Preposition-Noun (plural)	1227	1.25	16
19	, PP MD VV ,- Personal pronoun- Modal- Verb	997	1.16	16	PP MD RB VV Personal pronoun- Modal-Adverb- Verb	1195	1.22	20
20	PP MD RB VV Personal pronoun- Modal- Adverb-Verb	995	1.16	19		1167	1.19	39
21	NN IN DT JJ Noun- Preposition-Determiner- Adjective	987	1.15	13	JJ NN IN NN Adjective-Noun-Preposition-Noun	1113	1.13	17
22	SENT IN NN , Punctuation- Preposition-Noun- ,	979	1.14	59		1068	1.09	61
23	IN DT NN CC Preposition - Determiner-Noun-Conjunction	946	1.10	24		1035	1.05	36
24	IN PP$ NN SENT Preposition- Possessive Pronoun-Noun-punctuation	936	1.09	25	IN DT NN CC Preposition-Determiner-Noun-Conjunction	1020	1.04	23
25	SENT IN DT NN Punctuation- Preposition-Determiner-Noun	907	1.06	30	IN PP$ NN SENT	1017	1.03	24
26	VV IN DT NN Verb- Preposition-Determiner-Noun	905	1.05	37	NN IN NN SENT	989	1.01	27
27	NN IN NN SENT Noun- Preposition-Noun-,	902	1.05	26	SENT RB , PP	979	1.00	29

28	PP MD VV DT Personal pronoun- Modal-Verb- Determiner	897	1.04	68	IN DT NN NN Preposition- Determiner-Noun-Noun	974	0.99	30
29	SENT RB , PP Punctuation- Adverb- ,-Personal pronoun	875	1.02	27		968	0.98	33
30	IN DT NN NN Preposition- Determiner-Noun-Noun	870	1.01	28	SENT IN DT NN	968	0.98	25

Core constructions
B2 Constructions less used in C2 (Below 1-30 rank)