To appear in ATLANTIS 46(1), Tan Arda Gedik

Constructional Emergence in A1–C1: A Bird's Eye Perspective and Alignment

This study aims to take a usage–based constructionist approach to observing the emergence of constructions in L2 speakers of English from a bird's eye perspective and aligns some of them with CEFR levels. To do this, five equally balanced subparts from the EFCAMDAT corpus were compiled and analyzed using TAASSC and SPSS. The findings present strong evidence for and confirm previous studies that speakers at lower proficiency levels use fixed or prototypical expressions and do not deviate as much from conventional ways of combining constructions, i.e., collexemes, possibly stemming from not having reached an abstract form yet. The top 10 high and low frequency constructions do not show a developmental path. In the end, the study presents a set of constructions aligned for each CEFR level to serve as a rudimentary table of alignment.

Keywords: construction grammar, constructional learning, L2 writing, CEFR levels

Emergencia constructiva en A1–C1: perspectiva y alineación a vista de pájaro

Este estudio tiene como objetivo adoptar un enfoque construccionista basado en el uso para observar la aparición de construcciones en hablantes de inglés L2 desde una perspectiva de ojo de pájaro y alinea algunas de ellas con los niveles del MCER. Para hacer esto, se compilaron y analizaron cinco subpartes igualmente equilibradas del corpus EFCAMDAT utilizando TAASSC y SPSS. Los hallazgos presentan una fuerte evidencia y confirman estudios previos de que los hablantes con niveles de competencia más bajos usan expresiones fijas o prototípicas y no se desvían tanto de las formas convencionales de combinar construcciones, es decir, colexemas, posiblemente debido a que aún no han alcanzado una forma abstracta. Las 10 principales construcciones de alta y baja frecuencia no muestran un camino de desarrollo. Al final, el estudio presenta un conjunto de construcciones alineadas para cada nivel del MCER para servir como una tabla rudimentaria de alineación.

Palabras clave: gramática de la construcción, aprendizaje constructivo, escritura L2, niveles del MCER

1. Introduction

One of the central objectives of second language acquisition (SLA) is understanding how language and more proficiency develops over time in a target language. As such, there have been many studies that took up second language development from a variety of different perspectives and approaches. Recently, with more linguists subscribing to approaches that do not separate lexis from grammar, SLA has seen many studies on the lexicogrammatical development of L2 speakers of many different languages.

This study seeks to observe the emergence of some constructions in L2 speakers of English across different proficiency levels and aligns them with CEFR levels. As such, the current research study deepens our understanding of constructional learning and presents further evidence. Subscribing to a usage–based constructionist approach, this study is similar in nature to Ellis (2008), Ellis and Ferreira-Junior (2009), Römer, Skalicky, and Ellis (2018), Römer et al. (2014) and Römer and Yilmaz (2019), to name a few. However, one difference there exists is that this study takes a more global perspective while above–mentioned studies focus on a specific set of verb–argument constructions in L2 speakers and their development across various CEFR levels. This paper is meant to be a complementary study to our understanding of how constructions emerge across proficiency levels and what the implications of it are for SLA. The research hypotheses and research questions that are investigated in this study are:

H1: there will be correlations between syntactic indices and CEFR levels
H2: highly frequent constructions will have their roots in earlier proficiency levels
H3: low frequency constructions will show a developmental path as proficiency increases
H4: as proficiency increases, attested constructions and lemma–construction combinations in the corpus will decrease
RQ1: Are there clear cut CEFR levels when constructions start to emerge? When and which constructions are they?

2. Usage–Based Theories

Within usage–based approaches, language is regarded as an emergent structure that arises in and through usage–events, i.e., repetition of sequences of words. By subscribing to an emergent understanding of language, a priori rules for language are no longer needed and thus language becomes an adaptive system, that is, a system which changes in regard to ambient input (e.g., Hopper 1987; Larsen-Freeman 1997; Ellis and Larsen-Freeman 2006). Usage–based approaches have almost become synonymous with cognitive linguistics, which is a result of overlapping findings in each field. Another reason is that counting frequency and investigating frequency effects have been done by cognitive linguists, whose results have confirmed usage–based assumptions (see for instance Divjak 2019, 40-95). These assumptions are namely (i) language is learned by general domain cognitive abilities, for instance hearing, perception, pattern recognition and joint attention to name a few (Tomasello 2003, 282-320) and that (ii) language arises from usage–events (e.g., Diessel 2016).

Construction grammar has been a successful theory with ample evidence to support its assumptions and claims. While it is important to acknowledge that construction grammar is a family of theories, most of them agree on several central tenets (see Hoffmann and Trousdale 2013, 1-14 for a discussion). These are (i) language consists of form–meaning pairings, i.e., constructions, (ii) constructions are learned, (iii) there are no deep and surface structures, in other words derivation is not applicable in construction grammar, (iv) constructions vary in shape, size and abstraction and finally (v) constructions are usage–driven. The line of thinking we subscribe to in this study is (usage–based) construction grammar which embodies the assumptions outlined above.

2.1 Construction Grammar

In a nutshell, construction grammar is a symbolic view of language that unifies form and meaning and constructions are learned without a separation of lexis from grammar and there is not an innate grammar regulatory system, i.e., Universal Grammar.

Figure 1: Lexicogrammatical Continuum (Adapted from Gedik 2022, 30)

Prefixes, suffixes *un–, re–, de–…*	Words *cat, dog, bird…*	Fixed expressions *Here you are*	Partially fixed expressions *The Xer, The Yer*	Fully abstract schemas *Subj Verb Obj Oblique*

Constructionists believe that there is no dividing line between lexis and grammar, but rather, they merge to create what is called the lexicogrammatical continuum. This continuum (figure 1) signifies the gradience of linguistic items, because some items behave more like lexis and some display properties that would be regarded as ‘grammar’. What this continuum also shows is that any construction taken up on the continuum is a pairing of form and meaning. For instance, *the Xer the Yer* construction is a partially filled construction, that is, it has fixed elements and slots and this form sequence is paired with the meaning of correlationality. In other words, the intensity or probability of one event described in one clause is dependent on the other.

Many linguists believe that speakers start their language learning journey with idiomatic phrases or verb–islands (Tomasello 2003, 117-121) and keep detailed records of constructions: the items that occur with them and their lexicosemantic features as well as extra–linguistic conditions (Bybee 2010, 14-32). Frequency helps learners distinguish constructions’ conventionalized forms from unconventionalized forms and produce them in line with the conventionalized usage patterns. As Herbst (2020, 84) makes it explicit by saying “layers of usage events… become linked on the basis of recognized similarities between them”. This means that usage events help learners identify conventionalized forms of a construction. Focusing on entrenchment, Divjak (2019, 51) illustrates it as “repeated presentations of a verb in particular constructions (e.g., *The rabbit disappeared*) cause a child [learner] to infer probabilistically that the verb cannot be used in non–attested constructions (e.g., **The magician disappeared the rabbit*)”. In Goldberg’s (2019, 77) account, this type of entrenchment is called simple entrenchment where frequency is “simply a proxy for familiarity”. Another, perhaps more important type of entrenchment that needs attention is what Goldberg (2019, 77) calls conservatism via entrenchment, i.e., statistical preemption. This ability is activated when “the more frequently a verb has been witnessed in a language in any other construction, the more resistant it should be to being used in any new way” (Goldberg 2019, 77). In other words, speakers will calculate how many times an item and a construction should have occurred together based on the frequency information of the item and the construction and based on this information arrive at a conclusion of generalizability of an item. This, however, does not mean we retain all the item–specific information for a construction, since memory is lossy (Goldberg 2019), but whenever we experience a construction it “can form a lossy structured representation that prioritizes what the word designates and includes various contextual aspects of the encounter" (Goldberg 2019, 16). For Goldberg (2019, 94), entrenchment also explains how "better–covered constructions are easier to access, which results in more conventional language being used more often, which further strengthens the association between conventional forms and particular messages–in–context". This is the reason why a positive correlation between increasing proficiency, that is mastery of the target constructions, and a higher accuracy of idiomatic speech is expected of L2 speakers of any language. Furthermore, it would be plausible to assume that there exists an inverse correlation between construction frequency, as in high and low frequency constructions, and proficiency. This stems from frequency effects, because arguably a speaker will experience rarer constructions as proficiency increases. In other words, a learner with growing proficiency is more likely to encounter more low frequency constructions, especially if the learner has print exposure in the target language. There is experimental evidence that those L2 with more print exposure outperformed L2 speakers with less print exposure in tasks requiring vocabulary, collocation, and grammatical knowledge, which require varying levels of constructional knowledge (Dąbrowska 2019). The connection between print exposure and low frequency constructions stems from how written language harbors more complex and subsequently rarer constructions (see Roland 2007 for English).

2.2 Constructional Knowledge in L1 and L2 Speakers

Various studies have proven that L2 learners do not differ in terms of their constructional knowledge (e.g., Römer et al. 2014) of constructions. As such, constructions have an ontological status for both L1 and L2 speakers and this indicates that grammar is just as meaningful as lexical items are. However, although there are many studies on analyzing the constructional knowledge of L1 speakers (Ambridge and Lieven 2015; Behrens 2009; Goldberg 2014; Goldberg, Casenhiser, and Sethuraman 2004; Lieven, Pine, and Baldwin 1997 to name a few), studies that analyze L2 constructional knowledge are fewer in comparison (Eskildsen 2012, 2014; Roehr-Brackin 2014; Tode and Sakai 2016). This difference, however, can be justified because of a lack of reliable L2 corpora until recently (see Meunier 2015 on this).

Previous research demonstrates that L2 speakers of English have constructional knowledge, differ in their verb–VAC (verb-verb argument construction) associations with regard to proficiency and

L1 background and there are systematic differences in their usage of certain constructions (Gries and Wulff 2005; Römer et al. 2014; Römer, Skalicky, and Ellis 2018). However, research on a general outlook without subscribing to a particular L1 background has been relatively scarce. Römer (2019) is one exception to this. She investigates the constructional development of Mexican and German speakers of English from A1 through C1 levels, using the same corpus used in the present study. Her analyses provide a detailed observation of how constructions develop at certain levels with specific items in two different learning groups. The analyses conclude that learners differ in their item-specific usage of certain constructions but become more productive over time. A similar study was conducted by Römer and Berger (2019), comparing the same learner groups sampled from the same corpus. In this study, they specifically focus on a list of prepositional constructions (i.e., V about N, V across N and others, see Römer and Berger 2019: 1095). Their analyses also show that there is a correlation between growing proficiency and productivity as well as an inverse correlation between growing proficiency and using fewer fixed expressions. This productivity may partially lead to unconventional attestations of a given construction, even at advanced levels. For instance, both Goschler and Stefanowitsch (2019) and Gedik and Uslu (2022) provide evidence for advanced German and Turkish speakers respectively. Their results show that at times advanced L2 speakers' choice for the verbal slot in the ditransitive construction may be affected by strongly entrenched items in the corresponding construction in the other language. Although this may not always be the case, it is still possible and therefore it is plausible to hypothesize if with growing proficiency the number of corpus attested item-construction combinations decreases.

It is however important to note that so far these studies have compared specific learner groups and have not mixed the sample with learners from different backgrounds (see Römer et al. 2014 for a similar study). Furthermore, there have not been many studies that systematically attempt to align CEFR levels to various constructions. This is understandable as it is quite a tedious task to do. McCarthy (2016) is the only study that partially aligns the ditransitive construction, i.e., *she gave him a book*, with CEFR levels, by analyzing error rates and data from learner corpora. He notes that the ditransitive construction is mastered by C1, i.e., produced with no errors. However, he does not necessarily claim that the ditransitive construction belongs to a specific CEFR level, otherwise by that token it would be an A1 level construction.

Scholars have demonstrated that L2 speakers start their language learning journey with a set of fixed and highly repetitive constructions just like L1 speakers, which then grow in complexity, productivity and become less fixed (Eskildsen 2009; Eskildsen and Cadierno 2007; Li et al. 2014). Studies also suggest that with increasing proficiency, the accuracy of constructional knowledge also increases (Crossley and Salsbury 2011; Bestgen and Granger 2014). In addition to this, there is evidence that L2 speakers' knowledge of constructions is also influenced by their L1 (Li et al. 2014; Eskildsen, Cadierno, and Li 2015; Goschler and Stefanowitsch 2019; Gedik and Uslu 2022; Römer and Yilmaz 2019). There is strong evidence that advanced L2 speakers are also influenced by strongly entrenched verb–VAC combinations in their L1 (Goschler and Stefanowitsch 2019; Gedik and Uslu 2022).[1]

In an experiment, Lee and Kim (2011) tested Korean speakers' knowledge of the English intransitive construction, the ditransitive, and resultative constructions, developmentally. They explain that Korean speakers of English did not show a developmental understanding of the intransitives. Put simply, the speakers did not start from the bottom of a taxonomical constructional family and construct the superordinate intransitive construction. Their performance on the ditransitive and the resultatives also varied, with most participants finding them difficult. This arguably shows that both L1 and also other personal factors can contribute to these variations. For instance, as the authors also argue, constructions that are similar across the two languages may be learned faster in a specific learner group in comparison to others because the L1 in that case may act as a training crutch. Personal factors include but are not limited to working memory capacity, attention, and motivation, to name a few (see Sparks 2022 for a detailed review).

Figure 2: several VACs from TAASSC

VACs	Examples
Nsubj_verb_dobj (the transitive construction)	*I cook dinner*

[1] VAC is another name used to refer to constructions in natural language processing and L2 constructional acquisition studies, i.e., the caused-motion or the ditransitive constructions. To keep the study aligned with such fields and the tool used in the study, I also use VAC to refer to constructions in this study.

Modal_nsubj_verb_xcomp (the modal–question construction)	*Would you consider applying?*

3. Methodology & Association Measures

The study was carried out using the EFCAMDAT corpus in late 2021. In what follows, information about the corpus, the subcorpora created for this study, and the association measures used are introduced. The EFCAMDAT corpus is based on the texts that were submitted by the users. The website was previously known as Englishtown (now englishlive.com) where learners of English had to take a placement test to enroll in the courses offered by Education First online language school. The placement test would place students in one of the 16 proficiency levels available, all of which were aligned with CEFR levels (Council of Europe 2020). Students were regularly given writing tasks. Out of 128 tasks, they were regularly given tasks such as write an email, a movie review and introduce yourself, to name a few (Alexopoulou et al. 2015).

In the texts compiled here, because there was no data available for C2, the level had to be discarded. Each subcorpus was roughly a combination of 10,000 writing samples, all of which were balanced out across different writing tasks for reliable results and a variety of topics. However, keeping frequency counts equal was difficult as with increasing proficiency, there were more words per sample. Nevertheless, for each level, there was data from four different, randomly selected and equally balanced tasks.

In this study, I compiled 5 well–balanced subcorpora from the Education First–Cambridge Open Language Database (EFCAMDAT; Alexopoulou et al. 2015; Geertzen, Alexopoulou and Korhonen 2013), which is a large corpus of written texts by L2 speakers of English from different linguistic backgrounds. Ranging from A1 through C1, each subcorpus had roughly one million words (see table 1 for a detailed overview and had writings from different tasks that were given to the writers). The subcorpora were created as .txt files and were split up into 500 .txt files and were processed in batches of 50 files. This was necessary as TAASSC seems to have a limit on how many words it can process in one file (Kristopher Kyle, personal communication, February 8, 2022). The text files were analyzed using the syntactic sophistication setting of the tool, with minimum VAC frequency set to 5. When selecting indices to compare against the reference corpora, ALL_COCA was selected as in the case of this dataset, speakers did not have to use specialized language, i.e., academic, magazine or news. As such, a combination of all of the subsections of COCA is what makes language representative of all of its special uses, and it is assumed that speakers learn all of them in a piecemeal fashion. The indices were then imported into SPSS for a multiple regression analysis. Indices that did not meet the assumptions of a multiple regression analysis were discarded, e.g., due to collinearity and violations of normality (Tabachnik and Fidell 2014).[2] In the end, there were 15 indices to report. Indices are reported as corpus–based, which are calculated based on the subcorpus under analysis and against–reference–corpus, which is the COCA.

Table 1: Subcorpus Word Count

CEFR Level	Word count
A1	1142862
A2	1217936
B1	1202052
B2	1264006
C1	1305818

3.1 Association Measures and Automatized Tools

It is possible to measure constructional frequency in finer detail. So far, three main approaches have been employed in studies: (i) faith scores, (ii) delta p scores and (iii) collostructional strength. However, due to space related issues, only (iii) will be explained.

Collostructional analysis (Gries and Stefanowitsch 2003) predicts the likelihood of two items from the corpus appearing next to one another. Kyle and Crossley (2017, 525) employ the following formula in TAASSC, which is slightly different than the original formula, to calculate

[2] To detect collinearity, coefficients and VIF scores were used; Shapiro-Wilk test (the p values) was used to determine normality.

collostructions (Gries and Stefanowitsch 2003) as it is computationally lighter, and the authors claim that it is perfectly compatible with the original formula: $\left(\left(\frac{a}{a+b}\right) - \left(\frac{c}{c+d}\right)\right) * (a + b)$. This formula gives the output a+b+c+d for "approximate collexeme strength" (Kyle and Crossley 2017, 525).

The tool used in this study is TAASSC (available at https://www.linguisticanalysistools.org/taassc.html) developed by Kristopher Kyle (Kyle 2016). The tool automatically analyzes given texts in relation to a number of syntactic indices. However, the one used in this study is syntactic sophistication, which calculates usage–based indices and automatically detects VACs. Alongside the above–mentioned association measures, the tool also calculates the approximate percentage of constructional and lemma coverage in texts against a reference corpus, namely the COCA (Davies 2010), puts out type–token ratios for constructions, lemmas and lemma–construction combinations, among many other syntactic indices (See an exhaustive list of all the indices at https://docs.google.com/spreadsheets/d/1JtzWoR9CQzk4MWcv6amYZHuoZ1MzXsO2/edit?usp=share_link&ouid=109136599743509551186&rtpof=true&sd=true).[3] The tool has been used in quite a few studies (Gedik 2022; Kyle and Crossley 2017; Kyle and Crossley 2018 to name a few). Thus, the tool was selected as a means of automatically detecting and calculating constructional data.

4. Results & Analysis

Out of 35 indices, 20 of them had to be discarded as they violated the assumptions of the analysis, i.e., normality and collinearity. In table 2 and 3, the descriptive statistics for the rest of the indices are reported. See the supplementary material for a detailed descriptive statistics for each CEFR level and index. See the appendix for the descriptive statistics of corpus-based and against-corpus indices.

[3] In case the link breaks, the same file is hosted at TAASSC - NLP TOOLS FOR THE SOCIAL SCIENCES (linguisticanalysistools.org).

Table 2: Correlations for corpus–based indices

Correlations (corpus–based)

		LF	CF	LFC_per_mil	Collexeme
Pearson Correlation	CEFR	–.394	.108	–.115	–.200
Sig. (1–tailed)	CEFR	.000	.000	.000	.000

Lemma-frequency (LF) is calculated based on the items in the verbal slot of the constructions. These verbs that appear in the slot are then compared against their actual corpus frequencies from the reference corpus of TAASSC. Thus, LF shows the usage of high or low frequency verbs in the verbal slot of the constructions. Construction-frequency (CF) works in the exact same way as LF, but this time for identified VACs in the corpus and the reference corpus. LFC_permillion (lemma-construction frequency) is the calculation of verb-VAC combination frequency per million, i.e., normalized. In other words, it calculates how often *give* occurs in the ditransitive in the corpus, and then compares it against the reference corpus. The end result shows whether these verb-VAC combinations are highly frequent (common) or not (rare), i.e., *I begrudged him his affluence*. Finally, collexeme is also known as collostructional analysis (Gries and Stefanowitsch 2003). As it was pointed out, it calculates the joint probability that two items in a corpus will co-occur. Collexeme can give insight into the prototypicality or novelty of verb-VAC combinations in a corpus.

Table 3: Correlations for against–reference–corpus indices

Correlations (against–reference–corpus)

	all_av_lemma_freq	all_av_constructi	all_av_lemma_construc	all_av_approx_colle	all_lemma_ttr	all_construction_ttr	all_lemma_construction_ttr	all_lemma_attested	all_construction_att	all_lemma_construction_attested

			on_freq	tion_freq	xeme					ested	
Pearson Correlation	CEFR	–.693	–.332	–.706	–.279	.853	.829	.850	.163	–.541	–.725
Sig. (1–tailed)	CEFR	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

All_av_lemma_freq compares lemma frequencies against the lemma frequencies in the reference corpus. all_av_construction_freq works in the same way as the previous index, but it calculates the VAC frequencies. All_av_lemma_construction_freq also works in the same as the previous indices, but it calculates frequencies for lemma-construction combinations. All_av_approx_collexeme calculates the average collostructional strength in the corpus against the reference corpus. All_lemma_ttr calculates the main verb lemma type-token ratio against the reference corpus for the corpus itself. All_construction_ttr does the exact same calculation as the previous index, but for constructions. Similarly, all_lemma_construction_ttr does the same calculation but for lemma-construction combinations. All_lemma_attested is the percentage of lemmas in the corpus that are in the reference corpus. All_construction_attested is the same ratio calculated for constructions. All_lemma_construction_attested is the same ratio but this time it is calculated for lemma-construction combinations.

4.1 Hypothesis 1: Correlations between indices and CEFR levels

TAASSC first identifies POS tags, and then establishes dependencies to identify constructions (see Kyle 2016, 35-43 for a detailed discussion). As reported in previous studies, e.g., Römer and Berger (2019), there were correlations between CEFR levels and lemma–frequency (LF) (r= –.394), construction–frequency (CF) (r= .108), lemma-construction frequency (LFC) per million (r= –.115) and collexemes (r= –.200). The statistically significant findings (*p= .00*) suggest that the correlation between proficiency and indices is real. The weak correlation shows that the contribution proficiency makes among other indices is relatively smaller. To interpret the practical effects of proficiency on these indices it is possible to observe the r^2 scores (LF r^2 = .155, CF r^2 = .012, LFC_permillion r^2 = .013, collexemes r^2 = .040). As such, an increasing proficiency would account for 15.5% of the variance in lemma-frequency, 1.2% in construction-frequency, 1.3% in lemma-construction frequency per million, and 4% in collexemes in learner corpora. Although the effect sizes are small, these correlations can be interpreted as follows. As proficiency increases, students' use of common or highly frequent verbs decreases. This has been reported as being a predictor of being a proficient writer, as lexical diversity essentially increases (e.g., McCarthy and Jarvis 2010). CF also shows a positive correlation and this indicates that learners use a higher variety of constructions as proficiency increases. This is not surprising as previous studies also report similar findings of low proficiency students using a set of fixed constructions (see Römer 2019). LFC per million supports the findings here regarding CF and LF, and suggests that on average, as proficiency increases, students gradually decrease their use of highly frequent lemma–construction combinations. Per million in this index is computed based on the corpus analyzed in the study and not against a reference corpus. Finally, there seems to be an inverse correlation between increasing proficiency and lemma-construction combinations, i.e., collexemes. It indicates that speakers use verbs that are less attracted to the constructions and as such they move away from formulaic and fixed expressions towards a more varied writing. To see if this increase or decrease for the above-mentioned indices holds across all CEFR levels, a multivariate analysis with the contrast option (K Matrix) is run. The contrast option helps with contrasting the findings for each variable across independent variables, i.e., CEFR levels. Table 4 shows CEFR level comparisons, all of which show statistical significance except for LFC_permillion from A2 onward. The p-values become clearer in light of descriptive statistics for each index (see appendix). LF drastically decreases until B2, where it increases again, but nosedives at C1 at an all time low. The increase at B2 may have been due to task requirements where learners may have used more high frequency words. CF increases until A2 and then decreases until B2, and reaches its all time high at C1.

Table 4: K Matrix Contrast Results

	A1 vs. A2	A2 vs. B1	B1 vs. B2	B2 vs. C1

LF (sig. 1-tailed)	.000–.693	.000–.332	.000–.706	.000–.279
CF (sig. 1-tailed)	.000	.000	.000	.000
LFC_permillion (sig. 1-tailed)	.000	.354	.877	.016
Collexeme (sig. 1-tailed)	.000	.000	.000	.000

Turning our attention to the indices in table 4, which compare the findings in the corpus against the COCA, first a description of each is necessary. Those that end in _freq are frequency–based indices, those with _ttr indicate the type–token ratio and those with _attested refer to how many items in the corpus analyzed here exist in the reference corpus. All indices are statistically significant (*p= .00*).

1) *All_av_lemma_freq* (r= –.693); *all_av_construction_freq* (r= –.332); *all_av_lemma_construction_freq* (r= –.706); *all_av_approx_collexeme* (r= –.279) show inverse correlations. The strongest correlation seems to be between lemma, lemma–construction and CEFR levels. This suggests that on average as proficiency increases, speakers move away from formulaic, fixed and highly repetitive lemmas and lemma–construction combinations. This supports the previous findings (Römer 2019; Goldberg 2006, 45-65). As for *approx_collexeme*, it supports the finding that with improving proficiency, students use less prototypical verbs in the constructions.

2) *All_lemma_ttr* (r= .853); *all_construction_ttr* (r= .829); *all_lemma_construction_ttr* (r= .850) demonstrate that with increasing proficiency, the type–token ratio per lemma, construction and lemma–construction combinations decrease. This indicates a more varied, or perhaps a lexicogrammatically richer output. As such, it is possible to argue that construction learning occurs via experience with highly repetitive and prototypical

examples at earlier stages (e.g., Goldberg 2006, 69-92), i.e., *I gave him a book* serves as a prototypical example for the ditransitive construction.

3) *All_lemma_attested* (r= .163); *all_construction_attested* (r= –.541); *all_lemma_construction_attested* (r= –.725) suggest that as CEFR levels increase, speakers' production of attested items decrease and attested lemmas suggest that as proficiency increases, there are fewer attested lemmas. When descriptive statistics are taken into account, while the first two do not demonstrate a considerably big leap between CEFR levels, *all_lemma_construction_attested* does. This might suggest that speakers may be acquiring productivity, resulting in highly schematic constructions, similar to the findings of Römer (2019) in relation to productivity. Thus, productivity may be an important factor to research in future L2 studies, especially in relation to producing unconventional constructions.

On Hypothesis 4

These findings, especially _attested indices combined with _ttr, confirm that students start learning their L2 with a set of limited lemmas and constructions, which then expand to cover a higher variety. The change in *all_construction_attested* and *all_lemma_construction_attested* suggests that there is a probability that the high repetition of frequently–used construction and lemma-construction combinations help with suppressing creativity or productivity to those who are at lower proficiency levels, which guide them to use more idiomatic and fixed phrases, i.e., lexically-prefabricated chunks. This is arguably because at lower levels, learners may not have schematized constructions that are highly productive. With partial evidence from the indices in the present study and previous studies (Goschler and Stefanowitsch 2019; Gedik and Uslu 2022), H4 is confirmed. Increasing proficiency fosters productivity but this extension may be affected by some factors, strongly entrenched items in the corresponding construction in the L1 (Goschler and Stefanowitsch 2019), and may potentially result in unattested instantiations of constructions, i.e., *I explain you the book*.

4.2 Hypothesis 2: Highly frequent constructions in early CEFR levels

With approximately 1193 constructions identified in all levels, the frequency cut off for H2 and H3 was identified by using the *CF_per_million* data in the results files put out by TAASSC. As such, high frequency constructions were those which appeared >50000 per million and low frequency constructions were <1000 per million (the vertical axis in figure 3 represents per million). To reiterate, per million is calculated based on within text data, not against a reference corpus.

The top 10 highly frequent construction, then, were as follows:

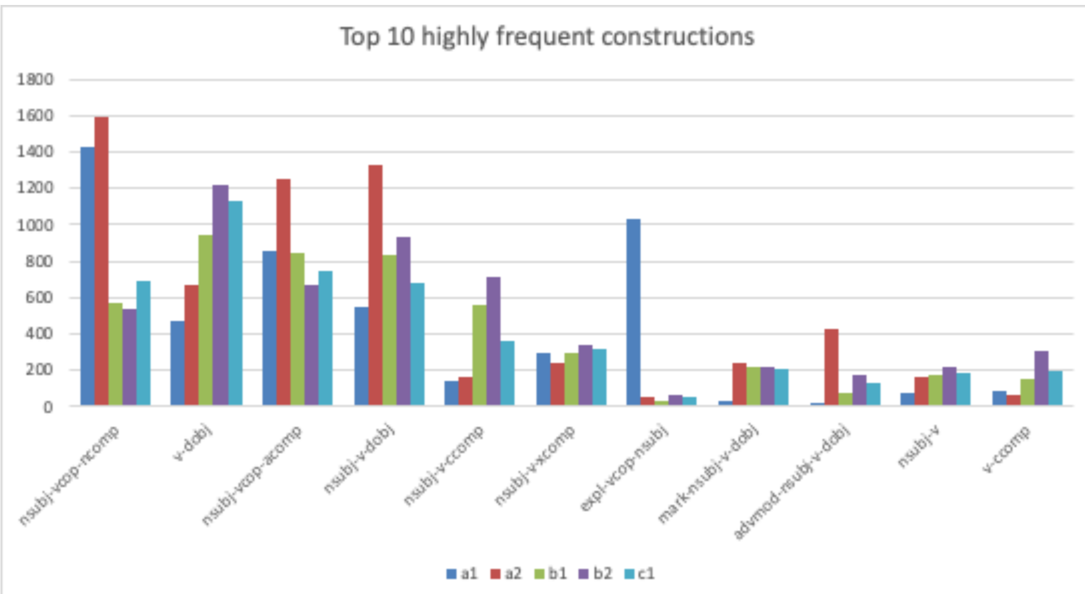

Figure 3: Top 10 highly frequent constructions

As seen in the figure, there does not seem to be a clear upward trend for any of the top 10 constructions. As such this hypothesis, that is, highly frequent constructions stem from earlier CEFR levels, is not confirmed. This may still be the case, however, this may not be possible to capture using just corpus evidence. It may need to be triangulated with experimental studies investigating language aptitude in L1 and L2, working memory, print exposure, and phonological abilities in L1. As for the idiosyncrasy of these constructions across CEFR levels, it is difficult to explain why that is the case. However, one possible explanation might be individual differences in grammatical knowledge (see Sparks 2022 for a detailed discussion), in which the assumption would be that each L2 learner would have constructions entrenched at different levels than other

L2 speakers of the same proficiency level. This different level of entrenchment stems not only from differences in exposure, but also individual differences in the cognitive machinery in speakers' L1 abilities. Sparks (2022) outlines fifty years' worth of research analyzing how good phonological abilities in the L1 significantly predict L2 success or L2 aptitude. Thus, while corpora, which we take as a pseudo-measure of quantifying exposure, can account for a lot, it is not the only factor that needs to be taken into account (see Dąbrowska 2016 on why individual differences should be taken into account in linguistic inquiry). Another possibility might be the communicative requirements of certain tasks, from which the subcorpora in the current study were sampled. These requirements may not have allowed for a need to use these constructions.

4.3 Hypothesis 3: Low frequency constructions in later CEFR levels

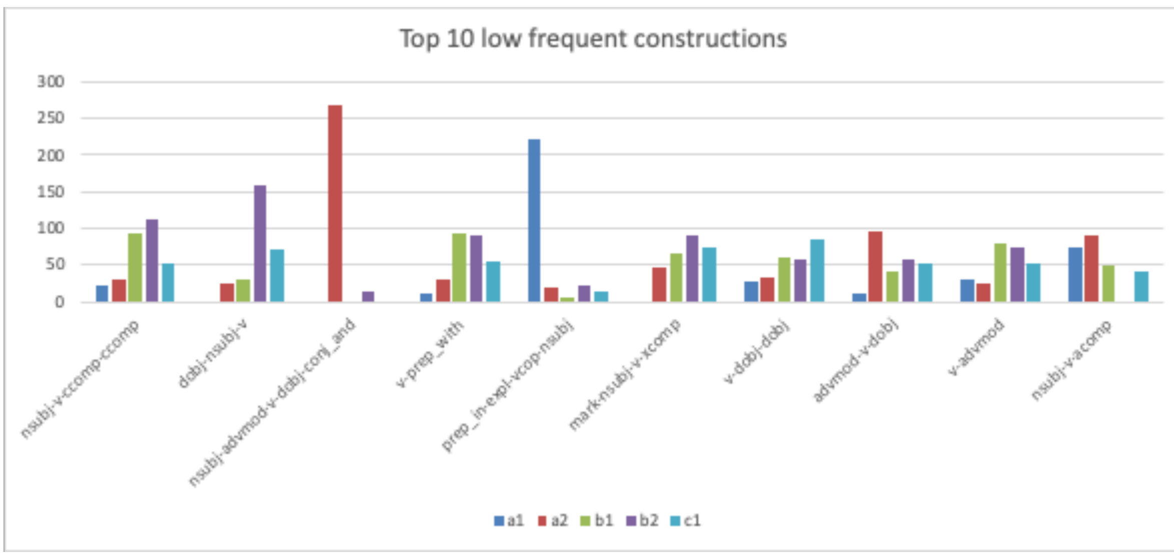

Figure 4: Top 10 low frequency constructions

Except for v-dobj-dobj, the other constructions do not follow an upward trend with increasing proficiency. Even then, the differences between levels are not statistically significant and there is a slight decrease between B1-B2. Therefore, the hypothesis that low frequency constructions emerge in later CEFR levels is denied. Once again, it is difficult to ascertain the idiosyncrasy of these constructions. It could be due to the communicative requirements of the tasks. Another explanation might be the fact that low frequency constructions possibly do emerge at earlier CEFR levels due to exposure in their lexically prefabricated instances. In other words, a low frequency construction may already be used by an A2 or B1 level student with a highly frequent item in the verbal slot of the construction. But the semanto-pragmatic features of the said construction may not be acquired fully or the construction is possibly not schematized until later CEFR levels.

See an example for v-dobj-dobj :

- v–dobj–dobj (*get*$_{v}$ *$ 50,000 salary*$_{dobj}$ *and 1 month holiday*$_{dobj}$)

4.4 Aligning Constructions to CEFR levels

In this section, constructions will be filtered to align with CEFR levels, using raw frequency, to identify when they first emerge. In other words, how many times a construction was produced at a certain level. It is important to know that this analysis here is not clear–cut and may differ based on many cognitive or frequency related factors in different L2 speakers of English. It is meant to serve as a rudimentary analysis of general tendencies in this study. For that reason, the top 50 highly frequent constructions will be aligned with CEFR levels. A construction was considered to have emerged in a specific level if it had not occurred more than 50% (based on raw frequency) in the previous level than that of the current level. Similarly, constructions that emerge in one level and continue occurring in other levels were eligible. Those that did not meet the criteria were not added to the list. The examples were taken from the respective levels.

Table 5: Constructional Alignment

CEFR Levels	Constructions (raw/normalized frequency)
A1	Expl–vcop–nsubj (A1: 1,182/1035; A2: 55/66.98): *there is a goat* Expl–vcop–nsubj–nsubj (A1: 316/361.14; A2: 10/12.17): *There are three windows and a chair* Prep_in–expl–vcop–nsubj (A1: 220/251.43; A2: 18/21.92): *In my office, there are many people*
A2	Nsubj–v–dobj (A1: 549/627; A2: 1333/1623.51): *I have a daughter* Mark–nsubj–vcop–acomp (A1: 16/18.28; A2: 215/261.85): *so I am busy* V–prep_at (A1: 0/0; A2: 71/86.47): *study at*

	Expl–vcop–ncomp (A1: 0/0; A2: 67/81.60): *There is a kitchen* Nsubj–v–prep_to (A1: 0/0; A2: 80.38): *I go to Canada*
B1	V–dobj–prep_on (A2: 0/0; B1: 119/143.04): *take me on a holiday* Nsubj–v–ccomp–ccomp (A2: 29/35.32; B1: 215/76): *I hope I will understand and speak this language* Nsubjpass–v (A2: 0/0; B1: 48/57.69): *the girl was hit* Mark–nsubj–v–dobj–xcomp (A2: 8/9.7; B1: 16/19.23): *Because I try my best to learn*
B2	Advmod–nsubj–v–prep_for (B1: 0/0; B2: 70/88.48): *Also I waited for the exam* Nsubj–advmod–v (B1: 0/0; B2: 52/65.72): *I can also watch* Dobj–nsubj–v–xcomp (B1: 15/18.03; B2: 37/46.76): *that he tried to learn* Nsubj–v–iobj–dobj (B1: 13/15.62; B2: 36/45.50): *He gave her a box* Mark–nsubj–vcop–xcomp (B1: 5/6.01 ; B2: 32/40.44): *Since she is to go*
C1	V–dobj–prepc_by (B2: 0/0; C1: 39.17): *had a dream by creating* V–prep_into (B2: 0/0; C1: 29/37.86): *got me into* Mark–dep–v–dobj (B2: 13/16.43; C1: 28/36.56): *in order to know more friends* Mark–nsubj–v–prep_for–ccomp (B2: 0/0; C1: 27/35.25): *If you vote for me, I will appeal* V–prt–prep_on (B2: 0/0; C1: 27/35.25): *followed up on this* Mark–nsubj–v–dobj–prep_as (B2: 0/0; C1: 26/33.95): *Whether they learned it as a second language* Nsubj–v–prep_on–prep_for–mwe–prep_with (B2: 0/0; C1: 28.72): *I hope I can count on you for support because with your help we can make a difference*

Constructions presented here are of varying specificity and abstractness. This table demonstrates that with increasing proficiency, the use of more arguably complex constructions increases. That is, each level has a new construction in comparison to the previous one which has a subordinating word, except for A1 and A2. For B1 that is because, B2 *that* and *since*, and for C1 is *in order to* and *whether*. This classification does not necessarily mean that learners do not use such subordinating constructions at earlier levels, but in the current study the current classification arises with the data at hand. Therefore, a more robust and valid attempt at classifying constructions per CEFR levels should triangulate data from different corpora and also include other indices apart from frequency only.

5. Conclusions

With growing proficiency, speakers expand their constructional knowledge and combine constructions with different items. Furthermore, as we have seen in frequency–based indices, speakers mostly use fixed expressions and stay relatively loyal to them, from which they slightly depart in upcoming proficiency levels.

Starting with the hypotheses, in H1, we reconfirmed some of the findings of previous studies, namely that there is a statistically significant correlation, especially considering the size of the subcorpora here, between selected syntactic indices and CEFR levels. More specifically, there is evidence that speakers with growing proficiency move away from fixed and highly repetitive expressions to cover more lower–frequency lemma–construction combinations, which is backed up by LFC per million and collexeme indices. Furthermore, when the writing samples are compared against a reference corpus, there is once again strong evidence that speakers expand their mental constructicon and the exemplar representations of lemma–construction combinations, which is supported by point 3 under section 4.1. Furthermore, there is evidence that as speakers gain more proficiency, the type token ratios show a positive trend, meaning that output becomes lexicogrammatically more diverse. This can provide partial evidence for how constructions are learned with highly repetitive and prototypical examples in them at earlier stages. We denied H2 and H3 on the basis of lack of evidence, that is, high frequency constructions do not necessarily stem from earlier CEFR levels and low frequency constructions do not emerge in later CEFR levels. Finally, we confirmed H4 as a natural consequence and byproduct of H1, namely that as proficiency increases there will be a lower score of attested constructions in the texts against a reference corpus.

As such, the findings presented here confirm the central tenets of usage–based approaches. Namely, that language is an experience–based phenomenon and that constructions are first learned with high–frequency items in them, which are then expanded onto other lower–frequency items. The indices confirm the findings of previous studies and present evidence that constructions are learned in a piecemeal fashion and speakers use the highly repetitive and fixed expressions as training wheels (or item–islands) to cover and learn more constructions in upcoming proficiency levels and acquire more productivity, i.e., the constructions become more highly schematic. These findings can inform the teaching of foreign languages in the following ways:

a) repetition and recycling of constructions is important at earlier levels
b) speakers will not always show a developmental behavior with all constructions
c) presenting learners highly repetitive and prototypical examples of constructions should ideally be of help in learning constructions at especially earlier stages.

Future studies are encouraged to take up this line of research and explore some of the shortcomings of this study, for instance the number of words in each subset and low number of constructions analyzed in constructional alignment and hypotheses.

The current study presented a bird's eye perspective and insight into the implications of constructional emergence at different proficiency levels and aligned them with CEFR levels. As such, findings suggest and confirm that there is a correlation between usage–based syntactic indices and CEFR levels, that language is learned in a piecemeal fashion and speakers use highly repetitive fixed expressions, which can range anywhere from small constructions such as collocations to argument structure rules such as the transitive construction [*nsubj–v–dobj*] and with growing proficiency move towards an expanded mental constructicon. Finally, some constructions were aligned following CEFR levels and compared against expected skills of respective levels. This alignment suggests that speakers' knowledge of constructions partially follows a developmental sequence, with previous constructions combined with newer ones. The findings can prove useful for future SLA studies and applied construction grammar, whereby the teaching of constructions is concerned.

References

Alexopoulou, Theodora, Jeroen Geertzen, Anna Korhonen, and Detmar Meurers. 2015. "Exploring Big Educational Learner Corpora for SLA Research: Perspectives on Relative Clauses." *International Journal of Learner Corpus Research* 1 (1): 96–129.

Ambridge, Ben, and Elena Lieven. 2015. "A Constructivist Account of Child Language Acquisition." In *The Handbook of Language Emergence*, edited by B. MacWhinney and W. O'Grady, 478–510. Malden, MA: Wiley-Blackwell.

Behrens, Heike. 2009. "Usage-Based and Emergentist Approaches to Language Acquisition." *Linguistics* 47: 383–411.

Bencini, Giulia M.L, and Adele E Goldberg. 2000. "The Contribution of Argument Structure Constructions to Sentence Meaning." *Journal of Memory and Language* 43 (4): 640–51. https://doi.org/10.1006/jmla.2000.2757.

Bestgen, Yves, and Sylviane Granger. 2014. "Quantifying the Development of Phraseological Competence in L2 English Writing: An Automated Approach." *Journal of Second Language Writing* 26: 28–41.

Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511750526.

Council of Europe, ed. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment ; Companion Volume*. Strasbourg: Council of Europe Publishing.

———. n.d. "English Profile." Accessed January 19, 2022. http://www.englishprofile.org/english-grammar-profile/egp-online.

Crossley, Scott, and Thomas Lee Salsbury. 2011. "The Development of Lexical Bundle Accuracy and Production in English Second Language Speakers." *International Review of Applied Linguistics in Language Teaching*, 1–26.

Davies, M. 2010. "The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English." *Literary and Linguistic Computing* 25 (4): 447–64. https://doi.org/10.1093/llc/fqq018.

Dąbrowska, Ewa. 2016. "Cognitive Linguistics' Seven Deadly Sins." *Cognitive Linguistics* 27 (4): 479–91. https://doi.org/10.1515/cog-2016-0059.

———. 2019. "Experience, Aptitude, and Individual Differences in Linguistic Attainment: A Comparison of Native and Nonnative Speakers: Experience, Aptitude, and Individual Differences." *Language Learning* 69 (March): 72–100. https://doi.org/10.1111/lang.12323.

Diessel, Holger. 2016. "Frequency and Lexical Specificity in Grammar: A Critical Review." In *Experience Counts: Frequency Effects in Language*, edited by Heike Behrens and Stefan Pfänder, 209–38. Berlin: De Gruyter Mouton.

Divjak, Dagmar. 2019. *Frequency in Language: Memory, Attention and Learning*. Cambridge University Press.

Ellis, Nick C. 2008. "The Dynamics of Second Language Emergence: Cycles of Language Use, Language Change, and Language Acquisition." *The Modern Language Journal* 92 (2): 232–49. https://doi.org/10.1111/j.1540-4781.2008.00716.x.

Ellis, Nick C., and Fernando Ferreira-Junior. 2009. "Construction Learning as a Function of Frequency, Frequency Distribution, and Function." *The Modern Language Journal* 93 (3): 370–85. https://doi.org/10.1111/j.1540-4781.2009.00896.x.

Ellis, Nick C, and Diane Larsen-Freeman. 2006. "Language Emergence: Implications for Applied Linguistics—Introduction to the Special Issue." *Applied Linguistics* 27 (4): 558–89.

Eskildsen, Søren W. 2009. "Constructing Another Language—Usage-Based Linguistics in Second Language Acquisition." *Applied Linguistics* 30 (3): 335–57.

———. 2012. "L2 Negation Constructions at Work." *Language Learning* 62 (2): 335–72.

Eskildsen, Søren W, and Teresa Cadierno. 2007. "Are Recurring Multi-Word Expressions Really Syntactic Freezes? Second Language Acquisition from the Perspective of Usage-Based Linguistics." In *Collocations and Idioms*, edited by M Nenonen and S Niemi, 1:86–99.

Eskildsen, Søren W, Teresa Cadierno, and Peiwen Li. 2015. "On the Development of Motion Constructions in Four Learners of L2 English." *Usage-Based Perspectives on Second Language Learning*, 207–32.

Eskildsen, Søren Wind. 2014. "What's New?: A Usage-Based Classroom Study of Linguistic Routines and Creativity in L2 Learning." *International Review of Applied Linguistics in Language Teaching* 52 (1): 1–30.

Geertzen, Jeroen, Theodora Alexopoulou, Anna Korhonen, and others. 2013. "Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCAMDAT)." In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, 240–54. Citeseer.

Gedik, Tan Arda. 2022. "An Analysis of Lexicogrammatical Development in English Textbooks in Turkey: A Usage-Based Construction Grammar Approach." *Explorations in English Language and Linguistics* 9 (1): 26–55. https://doi.org/10.2478/exell-2022-0002.

Gedik, Tan Arda, and Fatıma Uslu. 2022. "L1-L2 Transfer in the Ditransitive Construction: A Usage-Based Replication Study with Turkish Speakers of English." In *Advances in Second/Foreign Language Acquisition*. Nicosia, Cyprus. https://asefola.weebly.com/uploads/1/4/0/2/140216775/book_of_abstracts_-_asefola_conference.pdf.

Goldberg, Adele E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Goldberg, Adele E. 2014. "Fitting a Slim Dime between the Verb Template and Argument Structure Construction Approaches." *Theoretical Linguistics* 40 (1–2): 113–35.

Goldberg, Adele E. 2019. *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions.* Princeton, New Jersey: Princeton University Press.

Goldberg, Adele E., Devin M. Casenhiser, and Nitya Sethuraman. 2004. "Learning Argument Structure Generalizations." *Cognitive Linguistics* 15 (3). https://doi.org/10.1515/cogl.2004.011.

Goschler, Julia, and Anatol Stefanowitsch. 2019. "Generalization and Transfer in L2 Acquisition: The Role of Entrenchment in L1 and L2." In . https://iclc2019.site/wp-content/uploads/abstracts/applied/ICLC-15_paper_202.pdf.

Gries, Stefan Th., and Anatol Stefanowitsch. 2003. "Collostructions: Investigating the Interaction of Words and Constructions." *International Journal of Corpus Linguistics* 8 (2): 209–43. https://doi.org/10.1075/ijcl.8.2.03ste.

Gries, Stefan Th., and Stefanie Wulff. 2005. "Do Foreign Language Learners Also Have Constructions?" *Annual Review of Cognitive Linguistics* 3 (1): 182–200.

Guasti, MT. 2004. *Language Acquisition. The Growth of Grammar*. Cambridge, MA: The MIT Press.

Herbst, Thomas. 2020. "Constructions, Generalizations, and the Unpredictability of Language: Moving towards Colloconstruction." *Constructions and Frames* 12 (1): 56–95. https://doi.org/10.1075/cf.00035.her.

Hoffmann, Thomas, and Graeme Trousdale, eds. 2013. *The Oxford Handbook of Construction Grammar*. Oxford ; New York: Oxford University Press.

Hopper, Paul. 1987. "Emergent Grammar." In *Annual Meeting of the Berkeley Linguistics Society*, 13:139–57.

Kim, Jong-Bok, and Laura A Michaelis. 2020. *Syntactic Constructions in English*. Cambridge University Press.

Kyle, Kristopher. 2016. "Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication." Georgia State University. http://scholarworks.gsu.edu/alesl_diss/35/.

Kyle, Kristopher, and Scott Crossley. 2017. "Assessing Syntactic Sophistication in L2 Writing: A Usage-Based Approach." *Language Testing* 34 (4): 513–35. https://doi.org/10.1177/0265532217712554.

Kyle, Kristopher, and Scott A Crossley. 2018. "Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices." *The Modern Language Journal* 102 (2): 333–49.

Larsen-Freeman, Diane. 1997. "Chaos/Complexity Science and Second Language Acquisition." *Applied Linguistics* 18 (2): 141–65.

Lee, Jin-Hwa, and Hye Min Kim. 2011. "The L2 Developmental Sequence of English Constructions and Underlying Factors." 영어학 11 (3): 577–600.

Li, Peiwen, Søren W Eskildsen, and Teresa Cadierno. 2014. "Tracing an L2 Learner's Motion Constructions over Time: A Usage-Based Classroom Investigation." *The Modern Language Journal* 98 (2): 612–28.

Lieven, E. V., J. M. Pine, and G. Baldwin. 1997. "Lexically-Based Learning and Early Grammatical Development." *Journal of Child Language* 24 (1): 187–219. https://doi.org/10.1017/s0305000996002930.

McCarthy, Michael. 2016. "Putting the CEFR to Good Use: Designing Grammars Based on Learner-Corpus Evidence." *Language Teaching* 49 (1): 99–115. https://doi.org/10.1017/S0261444813000189.

McCarthy, Philip M, and Scott Jarvis. 2010. "MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment." *Behavior Research Methods* 42 (2): 381–92.

Meunier, Fanny. 2015. "Developmental Patterns in Learner Corpora." In *The Cambridge Handbook of Learner Corpus Research*, edited by Sylviane Granger and Gaëtanelle Gilquin, 379–400. Cambridge University Press.

Roehr-Brackin, Karen. 2014. "Explicit Knowledge and Processes from a Usage-Based Perspective: The Developmental Trajectory of an Instructed L2 Learner." *Language Learning* 64 (4): 771–808.

Römer, Ute. 2019. "A Corpus Perspective on the Development of Verb Constructions in Second Language Learners." *International Journal of Corpus Linguistics* 24 (3): 268–90. https://doi.org/10.1075/ijcl.00013.roe.

Römer, Ute, and Cynthia M. Berger. 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. https://doi.org/10.1017/S0272263119000202.

Römer, Ute, Audrey Roberson, Matthew B. O'Donnell, and Nick C. Ellis. 2014. "Linking Learner Corpus and Experimental Data in Studying Second Language Learners' Knowledge of Verb-Argument Constructions." *ICAME Journal* 38 (1): 115–35. https://doi.org/10.2478/icame-2014-0006.

Römer, Ute, Stephen C. Skalicky, and Nick C. Ellis. 2018. "Verb-Argument Constructions in Advanced L2 English Learner Production: Insights from Corpora and Verbal Fluency Tasks." *Corpus Linguistics and Linguistic Theory* 0 (0). https://doi.org/10.1515/cllt-2016-0055.

Römer, Ute, and Selahattin Yilmaz. 2019. "Effects of L2 Usage and L1 Transfer on Turkish Learners' Production of English Verb-Argument Constructions." *Vigo International Journal of Applied Linguistics*, no. 16: 107–34.

Sparks, Richard L. 2022. *Exploring L1-L2 Relationships: The Impact of Individual Differences*. Multilingual Matters. https://doi.org/10.21832/9781800411807.

Tabachnik, Barbara G., and Linda Fidell S. 2014. *Using Multivariate Statistics* . Harlow, Essex: Pearson Education.

Tode, Tomoko, and Hideki Sakai. 2016. "Exemplar-Based Instructed Second Language Development and Classroom Experience." *ITL-International Journal of Applied Linguistics* 167 (2): 210–34.

Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Appendix

Table 6: Descriptive Statistics for corpus–based indices

Descriptive Statistics (corpus–based)

	Mean	Std. Deviation	N
CEFR	3.15	1.418	14120
LF	1517.63	1921.708	14120
CF	217.27	372.579	14120
LFC_per_mil	2447.806654771	11321.0685942004	14120
collexeme_approx	717.03977783535	1094.503550113793	14120

Table 7: Descriptive statistics for against–reference–corpus indices

Descriptive Statistics (against–reference–corpus)

	Mean	Std. Deviation	N
all_av_lemma_freq	2130726.451654	1089642.7990616	2347
all_av_construction_freq	578675.096749	117719.4343716	2347
all_av_lemma_construction_freq	236276.3095257	119155.86437477	2347

all_av_approx_collexeme	33032.2840899431	41865.47914957857	2347
all_lemma_ttr	.23181847222	.094126285048	2347
all_construction_ttr	.45105434755	.095545569046	2347
all_lemma_construction_ttr	.62566841864	.148603619021	2347
all_lemma_attested	.99317114786	.005132110376	2347
all_construction_attested	.94516094431	.022173839298	2347
all_lemma_construction_attested	.86364407124	.044220413735	2347