

Exploring the spoken learner English constructicon: A corpus-driven approach

Gaëtanelle Gilquin
Université catholique de Louvain

Abstract

This study, which is set in the field of Applied Construction Grammar, seeks to identify the constructions that are typical of higher intermediate to advanced spoken learner English. It does so by relying on the recurrent sequences of part-of-speech (POS) tags extracted from the Louvain International Database of Spoken English Interlanguage (LINDSEI) and its native counterpart. This corpus-driven approach reveals that learner speech mainly consists of basic constructions like [NP] or [Subj V], although longer and more complex constructions can be found among the less frequent sequences. The chapter also discusses methodological issues (such as the link between POS tag sequences and constructions), as well as theoretical matters (including the place of speech in Construction Grammar).

1. Introduction: Construction Grammar and learner speech

Construction Grammar (CxG), as developed among others by Goldberg (1995, 2006), is a family of approaches that argue that constructions, defined as “conventionalized pairings of form and function” (Goldberg 2006: 3), are the fundamental units of language. What is traditionally referred to as the ‘mental lexicon’, that is, the repertoire of words and information about these words stored in the mind, therefore takes the form, in construction grammarians’ view, of a ‘constructicon’, a network of constructions that represent speakers’ knowledge of a language. In CxG, constructions cover a wide range of phenomena, not only syntactic structures (like the ditransitive construction), but also morphemes, words or idioms, with various degrees of specification (idioms, for instance, can be fully specified, as in *kick the bucket*, or partly specified, as in *kick <someone> when PRONOUN BE down*). Empirical studies, based on corpus data and/or experiments, have demonstrated the existence of speakers’ mental representations of constructions (e.g. Bencini & Goldberg 2000) and have provided very detailed information about the use of some constructions in naturally-occurring language (see Boas 2003 on resultative constructions or Hilpert 2008 on future constructions, among many other examples).

While constructions have been studied in English and, increasingly, in other languages as well (Hilpert’s 2008 study, for example, investigates future constructions in English, German, Dutch, Danish and Swedish, and some publications have focused on constructions in other languages, e.g. French in Bouveret & Legallois 2012), non-native language varieties have hardly been dealt with from a CxG perspective, although it was demonstrated over ten years ago that learners of a language do have constructions too (cf. Gries & Wulff 2005). In fact, the field of ‘Applied Construction Grammar’ (De Knop & Gilquin 2016), which adopts a CxG perspective to examine second/foreign language teaching and learning, is a very recent development within constructionist approaches. In a literature review on the subject, Gilquin & De Knop (2016) have identified ten studies or so that apply the

theory of CxG to the study of learner language, and the volume edited by De Knop & Gilquin (2016) includes another eleven studies representing Applied Construction Grammar. What is typical of these studies – and, one could add, of many studies in CxG – is that they rely on written data. De Knop & Mollica (2016), for example, apply among learners of German an experimental design (sorting task) that has often been used to test the existence of constructions in native speakers' mental representations: on the basis of a written questionnaire listing a number of sentences, the subjects are required to write down the sentences in different boxes according to their overall meaning. As for Valenzuela Manzanares & Rojo López (2008), they rely on the International Corpus of Learner English (ICLE; Granger et al. 2002), a corpus of argumentative essays written by learners of English, to investigate the use of the ditransitive construction by Spanish learners of English.¹

As a result of this bias towards written learner language, we have very little information about the use of constructions in learner speech, at least from a purely CxG perspective.² In an attempt to fill this gap in Applied Construction Grammar, the present chapter explores the spoken constructicon typical of English as a foreign language (EFL), by examining the spoken production of a large number of higher intermediate to advanced EFL learners.³ Its aim is to identify the constructions that are likely to be entrenched in the spoken EFL constructicon. The methodology adopted is corpus-driven and relies on the extraction of part-of-speech tag sequences from the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010) and its native counterpart, the Louvain Corpus of Native English Conversation (LOCNEC; De Cock 2004). In this respect, too, the study can be said to be exploratory, since it tests a recent proposal by Cappelle & Grabar (2016) to use part-of-speech tag sequences as an approximation to constructions. In the next section, it will be explained how a part-of-speech tagged corpus can provide information about the constructicon, while in Section 3 the corpora and methodology used in this study will be described. The results of the analysis can be found in Section 4, followed by some methodological afterthoughts in Section 5 and concluding remarks in Section 6.

¹ Studies in Applied Construction Grammar that have dealt with speech include Ellis & Ferreira-Junior (2009), Eskildsen (2014) and Roehr-Brackin (2014). What is characteristic of these studies, however, is that they investigate a small number of learners, from only one (in Eskildsen 2014 and Roehr-Brackin 2014) to seven (in Ellis & Ferreira-Junior 2009).

² We do have information about the use of linguistic phenomena in learner language that could be said to correspond to constructions in the CxG sense (e.g. clausal complementation in Tizón-Couto 2014, epistemic adverbial markers in Gablasova & Brezina 2015, or formulaic language in Wood 2010), but these studies are not theoretically embedded within CxG. Moreover, they are limited to the investigation of one or a small set of similar constructions and do not seek to adopt the kind of global approach that is aimed at here.

³ In that, it differs from Eskildsen (2014), a bottom-up constructionist study of the spoken production of one ESL (English as a second language) learner. While Eskildsen (2014) provides a qualitative analysis of an individual constructicon, showing for example how the emergence of a construction relates to previous utterances produced by the learner, the constructicon that will be described here is an abstraction based on the production of a large number of learners. This abstracted constructicon may not correspond to the actual constructicon of any of the individuals, but because it relies on many individual constructicons, it may be said to present a higher degree of representativeness than Eskildsen's (2014) description of a single constructicon.

2. Part-of-speech tagging to explore a construction

Part-of-speech (POS) tagged corpora are annotated in such a way that each token in the corpus is accompanied by a tag indicating the part of speech of the word. These tags are useful to disambiguate forms that can correspond to different word classes (e.g. *promise* as a noun or as a verb), but also to retrieve all items that belong to a specific word class (e.g. all adjectives). Interestingly, in the same way as one can extract clusters of words from a corpus (see, e.g., Conrad & Biber 2004 or Chen & Baker 2010), it is also possible, on the basis of a POS tagged corpus, to extract clusters of POS tags, that is, sequences of POS tags that are recurrent in a corpus (e.g. a sequence of an adjective followed by a noun). Such sequences are interesting because, as pointed out by Kennedy (1996: 225), they represent “expressions of syntactic patterning” and can form “the basis for quantitative studies of the use of syntactic structures and processes”.

The first corpus-based study that used POS tag sequences to investigate (written) learner language was Aarts & Granger (1998). They borrowed this technique from stylometry, in which POS tag sequences are used as a possible marker of authorship (cf. Spassova & Turell 2007 or Bel et al. 2012). Applying this technique to three components of ICLE (the Dutch, French and Finnish components) as well as a comparable native corpus, they sought to “uncover EFL learners’ fingerprints” (Aarts & Granger 1998: 132). They thus discovered that, in comparison with the native writers, the three learner populations overused patterns starting with a connective and underused patterns involving prepositions. They also showed that patterns specific to a certain mother tongue (L1) population were quite common, with French-speaking learners, for example, overusing sentence-initial *to*-infinitive clauses of purpose. This study was followed by a few others which sought either to find out more about the structure of interlanguage (cf. Tono 2000 on interlanguage development or Borin & Prütz 2004 on L1 transfer) or to automatically identify learners’ L1, in the spirit of the earlier studies in authorship attribution (cf. Golcher & Reznicek 2011).

Recently, it was suggested by Cappelle & Grabar (2016) that POS tag sequences, or ‘POS n-grams’, can be used to approximate constructions in a CxG sense.⁴ More precisely, the authors claim that “common (...) grammatical n-grams are constructions, in a Construction Grammar sense: they are form-function pairings which native speakers have memorized (and which learners of a language should acquire) as a result of their high frequency”. While in Goldberg (1995) non-compositionality was seen as the necessary condition for a construction to exist, later on, in Goldberg (2006), it is frequency that became the main criterion for a pattern to be recognized as a construction. It therefore makes sense, as Cappelle & Grabar (2016) do, to consider that frequent POS tag sequences can correspond to constructions – although not all POS tag sequences are necessarily constructions, as we will see below. Relying on this assumption and using the Corpus of Contemporary American English as a basis for the extraction of the 100 most frequent POS 5-grams,⁵ Cappelle and Grabar propose constructing an “n-grammar” of English, a

⁴ See also Wible & Tsao (2010) and Forsberg et al. (2014) for an automatic extraction of constructions partly based on POS n-grams.

⁵ It should be pointed out that Cappelle & Grabar (2016) define frequency in terms of types. For them, the most frequent POS tag sequences are those that correspond to the highest number of *different* lexical sequences. Here, frequency will be defined in terms of tokens, rather than types, since it is high token frequency that is said to promote entrenchment (see Bybee & Thompson 1997, Ellis 2013).

repertoire of POS n-grams that can serve as a useful resource for the teaching of the English language. More generally, and more importantly for our purposes, they establish a convincing link between frequent POS tag sequences and the construction of a language or language variety (although they do not use the term ‘construction’ as such). Building on this principle, the present study will consider POS tag sequences in a corpus of spoken learner English as a way of approaching the construction that is typical of EFL speech.

The reliance on POS n-grams has the obvious advantage that the extraction can be done fully automatically. The technique can thus be applied to large corpora including the production of many individuals, which ensures a high degree of representativeness and generalizability of the results (see also footnote 3). In addition, it allows for a global approach to constructions, with no a priori (and presumably subjective) selection of certain constructions for investigation, as would for example be the case for a technique like collostructional analysis, whose starting point has to be a (set of) specific construction(s). Here, all the POS tag sequences are considered to be of potential relevance. It should be underlined, however, that POS tag sequences only make it possible to approach, or approximate, a construction. For one thing, not all POS tag sequences are “units of language” (Goldberg 1995: 4). Cappelle & Grabar (2016: 281) recognize this too when they write that “n-grams are ‘blind’ to constituent structure. Sometimes, an n-gram does not contain enough (or one might say, it may contain too much) to make up what we would intuitively consider an ordinary linguistic sequence”. In order to overcome this problem, Cappelle and Grabar ‘complete’ the POS n-grams when necessary, turning for example the sequence “*to verb the X_{noun} of*” into “*to verb the X_{noun} of (Y_{NP})*” (as illustrated by *to improve the quality of / to improve the quality of life*). Here, the POS n-grams will not be completed, but the lexical sequences underlying them will be examined, as recommended by Aarts & Granger (1998: 135), so as to check the status of these POS n-grams. The second reason why POS tag sequences only paint an incomplete picture of what the construction looks like is that, as mentioned in Section 1, constructions in a CxG sense cover a variety of phenomena. By looking at POS n-grams, we mainly focus on the more syntactic types of constructions and neglect word-based constructions (like individual words or idioms). While this could be viewed as a limitation of the study, it can also simply be seen as a reflection of construction grammarians’ closer attention to syntactic constructions (cf. the so-called ‘argument structure constructions’) to the detriment of more word-based constructions.⁶ It can also be said that this predominantly syntactic approach is in fact an ideal complement to the corpus linguistic perspective on learner language, which is more often centred on word-based phenomena (cf., for example, Nesselhauf 2005 or Ädel 2006), and a good starting point for a first exploration of the construction of learner speech.

3. Corpora and methodology

The POS-based exploration of the spoken EFL construction is based on LINDSEI, the Louvain International Database of Spoken English Interlanguage, whose first

⁶ Morphemes, for instance, are considered as an extension of the category of constructions by Goldberg (1995: 4): “*expanding the pretheoretical notion of construction somewhat*, morphemes are clear instances of constructions in that they are pairings of meaning and form that are not predictable from anything else” (emphasis added).

version was released in 2010. LINDSEI is made up of the transcription of informal interviews of higher intermediate to advanced learners of English representing different L1 backgrounds. All the components of LINDSEI in its published version have been exploited in this study, namely eleven components corresponding to eleven L1 backgrounds (Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish) and a total of almost 800,000 words produced by 554 different learners (the interviewers' turns have been disregarded).⁷ To serve as a point of reference, LINDSEI has been used in combination with LOCNEC (Louvain Corpus of Native English Conversation), the (British) native counterpart of LINDSEI, corresponding to some 125,000 words for the interviewees' turns. All the components of LINDSEI and LOCNEC were compiled according to the same design criteria, which makes them perfectly comparable with each other. Thus, each of the interviews lasts about 15 minutes and includes three tasks: a warming-up activity, in which the interviewees had to talk about one of three set topics for a few minutes, a free informal discussion about topics of concern to young people, and a picture description, based on the same cartoon. The interviewees all have a similar profile, being students in their third or fourth year at university, with English as their main subject. The interviews were also transcribed with the same guidelines and were linked up with metadata about the interviewer, the interviewee and the context of the interview.

The released versions of LINDSEI and LOCNEC consist in raw text, with no annotation other than the tags that are part of the transcription conventions,⁸ e.g. the use of the tags and to open and close, respectively, the interviewees' turns, or the use of the equals sign to indicate word truncation (like *esp*=). Thanks to the help of the *Centre de Traitement Automatique du Langage* of the University of Louvain, however, all the data were POS tagged and it is these POS tagged versions of LINDSEI and LOCNEC that were used here.⁹ One of the problems with automatic POS taggers is that, with some rare exceptions, they have been designed to process standard written language (Gilquin & De Cock 2011: 149). Running them on spoken and/or learner corpus data, therefore, is not necessarily a straightforward matter. Yet, attempts to POS tag LINDSEI by means of CLAWS (the Constituent Likelihood Automatic Word-tagging System; Garside & Smith 1997) turned out to be relatively successful, with an accuracy rate of about 92% (see Gilquin 2016).¹⁰ In the process of POS tagging LINDSEI for this study, a simplified version of the CLAWS tagset was used that reduced the number of different tags, from 137 separate POS tags in the original tagset to 27 POS tags in the simplified version (the list of simplified tags and their meanings can be found in the Appendix).¹¹ In addition, the settings of CLAWS

⁷ A second version of LINDSEI is currently in preparation and should include twenty subcorpora.

⁸ The transcription conventions can be found at <<https://uclouvain.be/en/research-institutes/ilc/cecl/transcription-guidelines.html>>.

⁹ I am deeply indebted to Hubert Naets, from the *Centre de Traitement Automatique du Langage*, both for POS tagging the two corpora and for extracting the POS n-grams that served as a basis for the analysis presented in this chapter.

¹⁰ The latest version of the POS tagger, CLAWS4, was used in conjunction with the C7 tagset. I thank Paul Rayson for providing access to CLAWS locally.

¹¹ For example, "CC" (coordinating conjunction) and "CCB" (adversative coordinating conjunction) in the original tagset were combined into the tag "CCO" (coordinating conjunction). A tag for truncation ("TR") was added to account for incomplete words.

had to be adapted to take into account the specificities of the LINDSEI and LOCNEC transcription conventions.

Once LINDSEI and LOCNEC were POS tagged, recurrent sequences of POS tags were extracted from the interviewees' turns. Sequences of two to ten POS tags were retrieved, together with the raw frequency of these sequences in the corpus and the lexical sequences corresponding to these POS tag sequences. It should be noted that filled pauses were excluded from the analysis. Theoretically, they could easily be accommodated by a constructionist approach. Like any other construction, they consist in a pairing of form and function, the form being something like *er* or *erm*, and the function being, for example, stalling for time or segmenting discourse (see Clark & Fox Tree 2002 for an overview of some of the functions of filled pauses). However, from a practical point of view, the occurrence of filled pauses within sequences of POS tags prevents the automatic detection of the types of constructions that have traditionally been recognized in CxG. In (1), for instance, the NP construction cannot be identified due to the filled pause found between the determiner and the noun. While I believe that, ultimately, such phenomena should be taken into account in a CxG of speech, for a first CxG-based exploration of spoken interlanguage it might be safer to rely on constructions that are close enough to the classic repertoire of constructions in CxG. This does not mean that features typical of spontaneous speech are totally excluded. Thus, truncations, which represent parts of words in the traditional sense, have been retained, as have repetitions of POS tags, which can correspond to disfluent sequences but also to standard constructions (compare *exam exams* and *bus station* for the combination of two nouns). Unfilled pauses, corresponding to blanks in the recording, are not assigned any POS tags and are therefore not included in the analysis.

- (1) she shows off before her er friends (LINDSEI-PL017)¹²

The approach adopted here can be described as 'corpus-driven' or bottom-up, in that it starts from the free exploration of the corpus data to make generalizations about language, and more particularly about the spoken EFL constructicon. This approach can be contrasted with a 'corpus-based' or top-down approach, which looks at the corpus data through the prism of a specific idea or hypothesis. The difference between corpus-based and corpus-driven should be seen as a continuum, though, with studies being more or less corpus-based or corpus-driven. Purely corpus-driven studies, in particular, are difficult to set up as researchers often have some sort of (even vague) idea before embarking on an analysis. The corpus-driven approach, in fact, has been characterized as an "idealized extreme" along the continuum by McEnery et al. (2006: 8). In the present case, POS tagging could be said to limit the corpus-driven scope of the analysis. Relying on a pre-existing POS tagset means that one starts from a definition of word classes that, to a certain extent, will guide the

Foreign words were tagged as "unclassified words" ("FU"). See <http://ucrel.lancs.ac.uk/claws7tags.html> for the original C7 tagset.

¹² In the examples, the relevant part corresponding to the POS tag (sequence) being discussed is underlined. Dots represent unfilled pauses (of various lengths), and the equals sign marks truncated words. The code between brackets after each example provides information about the corpus from which the sentence is taken (LINDSEI or LOCNEC) as well as the number of the interview. In the case of LINDSEI, the code also indicates the interviewee's mother tongue background (BG=Bulgarian, CH=Chinese, DU=Dutch, FR=French, GE=German, GR=Greek, IT=Italian, JP=Japanese, PL=Polish, SP=Spanish, SW=Swedish).

analysis and the interpretation of the results, while, as pointed out by Biber (2010: 201), “[i]n its most extreme form, the corpus-driven approach assumes only the existence of word forms; grammatical classes and syntactic structures have no a priori status in the analysis”. However, Biber (2010) himself classifies as “corpus-driven research” (though of a hybrid type) the studies in ‘pattern grammar’ which “assume the existence of some grammatical classes (e.g., verb, noun) and basic syntactic structures” (Biber 2010: 202). In the present case, the fact that, on the basis of pre-existing POS tags, patterns are made to emerge automatically from the data, with no human control over what should be kept and what should be left out, suggests that the initial stages of the analysis are sufficiently atheoretical to qualify as a corpus-driven study. In addition, we will focus on the most frequent sequences of POS tags, which also contributes to the more ‘corpus-driven’ orientation of the study, since “recurrent patterns” and “frequency distributions” are the two elements cited by Tognini-Bonelli (2001: 87) as constituting the foundation of a corpus-driven approach. Such a corpus-driven approach is rare in CxG, which has tended to start from hypotheses about specific constructions that are tested by examining instances of these constructions in corpora. Yet, it is ideally suited for the purposes of the present study, which seeks to explore a language variety that has hardly been dealt with in CxG.

4. A corpus-driven analysis of LINDSEI’s constructicon

4.1. Single POS tags

As a first overview of the constructicon emerging from LINDSEI, we can consider Table 1, which provides a list of the POS tags found in the corpus, in decreasing order of (raw) frequency. A similar list is provided for LOCNEC by way of comparison. What we see is that the three most frequent POS tags in LINDSEI are word classes that compose noun phrases: personal pronouns (“PRONpers”), common nouns (“N”) and determiners (“DET”). They are followed by lexical verbs (“Vlex”), which come in fourth position in LINDSEI. This top four seems to point to a rather basic structure in learner language, made up of noun phrases and verbs.

If we compare the list of POS tags for LINDSEI with that for LOCNEC, we notice that, aside from the obvious differences in frequency, which are essentially due to the differing sizes of the two corpora, there is quite some overlap in the ranks occupied by the POS tags. Most of them are ranked similarly in the two corpora, either having exactly the same rank or being just one rank apart. The POS tags that present a difference of two ranks or more are in bold in the table. Among these, we can mention adverbs and determiners. While, as mentioned above, determiners round out the top three in LINDSEI, in LOCNEC it is adverbs (“ADV”) that occupy this position. This suggests that adverbs are more important in the native spoken constructicon than they are in the learner spoken constructicon. The verb *have* (“Vhave”) and existential *there* (“EX”) are ranked higher in native than in non-native speech, whereas for the infinitive marker *to* (“TO”), proper nouns (“Nprop”) and truncated words (“TR”), it is the opposite. That “TR” is ranked higher in LINDSEI than in LOCNEC seems to indicate that disfluency as expressed through truncation is a comparatively more typical phenomenon in learner English than in native English.

LINDSEI				LOCNEC		
Rank	POS	Meaning	Freq.	POS	Meaning	Freq.
1	PRONpers	Pers. pron.	102454	PRONpers	Pers. pron.	16869
2	N	Common noun	90727	N	Common noun	14132
3	DET	Determiner	84590	ADV	Adverb	13921
4	Vlex	Lexical verb	77605	Vlex	Lexical verb	13265
5	ADV	Adverb	76653	DET	Determiner	12066
6	PREP	Preposition	55376	PREP	Preposition	8975
7	Vbe	Verb <i>be</i>	45582	Vbe	Verb <i>be</i>	7816
8	CCO	Coord. conj.	42826	CCO	Coord. conj.	6442
9	ADJ	Adjective	36038	ADJ	Adjective	5267
10	CSU	Subord. conj.	24022	UH	Interjection	3769
11	UH	Interjection	18076	CSU	Subord. conj.	3287
12	NEG	Negation	15417	Vhave	Verb <i>have</i>	2219
13	TO	Inf. marker <i>to</i>	13675	NEG	Negation	2202
14	Vhave	Verb <i>have</i>	11165	Vdo	Verb <i>do</i>	2059
15	Vdo	Verb <i>do</i>	11104	TO	Inf. marker <i>to</i>	1926
16	Nprop	Proper noun	10684	NUM	Numeral	1674
17	NUM	Numeral	10473	Vmod	Modal verb	1640
18	Vmod	Modal verb	9926	Nprop	Proper noun	1554
19	TR	Truncation	9525	PRONindef	Indef. pron.	928
20	PRONindef	Indef. pron.	5353	EX	Exist. <i>there</i>	484
21	ZZ	Letter	3961	TR	Truncation	381
22	EX	Exist. <i>there</i>	2513	ZZ	Letter	346
23	PRONwh	<i>wh</i> -pronoun	1549	PRONwh	<i>wh</i> -pronoun	199
24	GE	Gen. marker	530	GE	Gen. marker	109
25	FU	Unclassified	340	FU	Unclassified	22

Table 1. Single POS tags and their frequency in LINDSEI and LOCNEC

4.2. Top POS *n*-grams

In this section, we examine the top thirty POS *n*-grams, that is, the most frequent sequences of POS tags, whatever their length. The list for LINDSEI and LOCNEC can be found in Table 2, which also includes the absolute frequency of the POS tag sequences as well as a concrete example of a lexical instantiation for each sequence.¹³

As can be expected, bigrams represent the bulk of the sequences, with longer *n*-grams being much less frequent. In a similar way to lexical bundles (cf. Altenberg 1991: 127), the longer the POS tag sequence, the less likely it is to occur frequently in a corpus. In fact, only bigrams and trigrams (in bold in the table) can be found among the top thirty POS tag sequences. The first 4-gram is ranked 54th in LINDSEI and 66th in LOCNEC. A comparison of the top thirty POS *n*-grams in LINDSEI and LOCNEC shows that, while the first trigram is ranked higher in LINDSEI (5th rank) than in LOCNEC (8th rank), there is a smaller number of distinct POS *n*-grams (types) in the

¹³ For the meaning of the POS tags, see Appendix.

former (three types) than in the latter (five types). Native speakers thus appear to have assimilated a higher number of longer, and hence presumably more complex, constructions than non-native speakers, but among the latter such constructions might be more entrenched, since they are ranked higher than in native speech.

Rank	LINDSEI			LOCNEC		
	POS n-gram	Freq.	Example	POS n-gram	Freq.	Example
1	DET N	46177	a bird	DET N	6380	my sister
2	PRONpers Vlex	30008	I agree	PRONpers Vlex	4754	he plays
3	PRONpers Vbe	25320	he is	PRONpers Vbe	4605	they are
4	PREP DET	24276	on the	PREP DET	3690	from a
5	PREP DET N	15220	after our exam	N PREP	2677	group of
6	N PREP	15180	essay about	ADV ADV	2601	quite early
7	N CCO	14560	sheet and	Vlex PRONpers	2580	call them
8	ADJ N	14414	clean city	PREP DET N	2361	for a day
9	Vlex PRONpers	14091	dislike it	Vlex ADV	2246	came up
10	Vlex DET	13856	chose my	ADJ N	2233	bad guy
11	CCO PRONpers	13626	and they	CCO PRONpers	2201	but she
12	CSU PRONpers	13578	that we	Vbe ADV	2132	are still
13	ADV PRONpers	12517	maybe it	CSU PRONpers	2115	because we
14	ADV ADV	12397	only then	Vlex PREP	2111	known as
15	Vlex PREP	12158	come to	N CCO	2086	day or
16	Vbe ADV	11164	was very	ADV PRONpers	2053	sometimes we
17	ADV ADJ	10383	just nice	Vlex DET	1973	heard the
18	Vlex ADV	10216	eat well	ADV ADJ	1592	quite big
19	TO Vlex	9885	to speak	PREP N	1585	on holiday
20	DET ADJ	9294	a fine	ADV Vlex	1566	never heard
21	N PRONpers	9253	books we	PRONpers Vhave	1460	they had
22	DET N PREP	8585	the list of	DET ADJ	1437	a real
23	Vlex DET N	8306	buy a house	PRONpers Vmod	1392	he can
24	PREP N	8005	for lunch	PRONpers Vbe ADV	1362	it is really
25	NEG Vlex	7610	not say	TO Vlex	1334	to see
26	PRONpers Vmod	7563	I should	DET N PREP	1320	the head of
27	Vbe DET	7527	is the	N PRONpers	1318	country I
28	PRONpers Vdo	7488	she does	PRONpers Vlex PRONpers	1212	we saw him
29	PRONpers Vhave	7468	we have	DET ADJ N	1195	a good sign
30	DET DET	7426	its own	PRONpers ADV	1179	I hardly

Table 2. Top thirty POS n-grams and their frequency in LINDSEI and LOCNEC

When considering the actual POS tag sequences, one important element to underline is that they do not necessarily correspond to constructions in the sense of complete structural units (see Section 2). The POS bigram “PREP DET” (preposition and determiner, ranked 4th in LINDSEI and LOCNEC), for example, is incomplete, in that it normally requires a noun to form a prepositional phrase. The POS trigram

“PREP DET N”, ranked 5th in LINDSEI and 8th in LOCNEC, is one way in which the sequence can be completed (with a common noun),¹⁴ but there are others, which can be found among the longer POS tag sequences, cf. “PREP DET ADJ N”, a sequence illustrated by *for the whole year*, but which, because of its length and structural complexity, only appears much later in the list of POS tag sequences (rank 151 in LINDSEI and 140 in LOCNEC). Some of these POS n-grams are also incomplete due to the spontaneous and unrehearsed nature of speech. Thus, in the lexical sequence *in my in my hometown*, taken from LINDSEI-BG012, the first “PREP DET” sequence is left incomplete, and it is only when the sequence is repeated that it is completed by a noun. In other cases, the sequence is simply interrupted and never taken up again, cf. the “PREP DET” sequence in *the sands on the . and the sands is very soft* (LINDSEI-CH001).

Table 2 also reveals that, quite interestingly, the top four POS n-grams are identical across native and non-native speech, with the combination of a determiner and a common noun (“DET N”) being the most frequent sequence, followed by the combination of a personal pronoun and a lexical verb (“PRONpers Vlex”), and that of a personal pronoun and the verb *BE* (“PRONpers Vbe”), and finally the (incomplete) “PREP DET” sequence mentioned above. This suggests that both native and non-native speakers’ constructions rely, in the first place, on short and simple constructions of the type [NP], [Subj V] and [PP],¹⁵ whose internal structure seems relatively basic: among the complete sequences, the NP is made up of a determiner and a noun, the Subject consists in a personal pronoun, and the Verb is either a lexical verb or the verb *BE*. Examples from LINDSEI illustrating each of these constructions are provided in (2) to (4).

- (2) and this was just it was just a guesthouse (LINDSEI-SW042)
- (3) she was . in a coma for .. a year .. and then she awoke (LINDSEI-GR039)
- (4) I read the story before the representation but er it was eh very touching to see it (LINDSEI-IT035)

Variants of each of these three constructions can be found further down the list, both in LINDSEI and LOCNEC. In the case of [NP], there are two other n-grams in Table 2 that can correspond to complete constructions, namely “ADJ N”, which occurs in both corpora and is illustrated in (5), and “DET ADJ N”, found only in the top thirty of LOCNEC and exemplified in (6). However, full [NP] constructions can also take the form of a single word, either a noun or a pronoun, and as such their presence in Table 2 can be detected whenever a personal pronoun (“PRONpers”) or a common noun (“N”) is part of an n-gram, e.g. “N PREP” or “CCO PRONpers”. It should also be underlined that nouns and personal pronouns need not be part of a recurrent sequence to function as an [NP] construction, which means that Table 2, which lists n-grams only, underestimates the predominance of the [NP] construction

¹⁴ It must be emphasized that a POS tag sequence that is structurally complete could still have other elements added to it, cf. “PREP DET N PREP N”, an extension of the “PREP DET N” sequence which, like the former, can correspond to a [PP] construction (see below on the [PP] construction), e.g. *from my point of view*.

¹⁵ Throughout this chapter, constructions will be enclosed in square brackets. The abbreviations used in these constructions are as follows: N = noun; NP = noun phrase; Obj = object; Obl = oblique; PP = prepositional phrase; Prt = particle; S = sentence/clause; Subj = subject; V = verb; VP = verb phrase; Xcomp = predicative complement.

in native and non-native speech. It will be reminded from Table 1, which lists single POS tags, that “PRONpers” and “N” are the two most frequent POS tags in LINDSEI and LOCNEC, and all of these personal pronouns and common nouns are potential [NP] constructions, whether they are part of highly recurrent POS n-grams like those in Table 2 or not. As for the incomplete sequences that belong (and are common) to the top thirty of LINDSEI and LOCNEC, some seem to point to the presence of longer and more complex [NP] constructions, involving post-modification introduced by, e.g., prepositions or (zero) relative pronouns. The former is illustrated by the bigram “N PREP” and the trigram “DET N PREP”, whose (possible) status as an [NP] construction is confirmed by examples (7) and (8). As for the latter, it is suggested by the presence of a POS tag sequence that, at first sight, might be difficult to parse, namely “N PRONpers” (common noun and personal pronoun, ranked 21st in LINDSEI and 27th in LOCNEC). When examining in context the lexical sequences that underlie this POS bigram, it appears that in a number of cases they correspond to nouns followed by a bare relative clause, as shown in example (9).

- (5) yes of course we have beautiful landscapes (LINDSEI-IT027)
- (6) but erm that was that was definitely a new experience (LOCNEC-E030)
- (7) I think if . one director is really good em . he can talk about women and make films about them (LINDSEI-BG025)
- (8) it cannot work because the rest of the people doesn’t do it (LINDSEI-SP024)
- (9) that’s . one thing I have to do of course (LINDSEI-DU021)

The [Subj V] construction is also very prominent in the list through its several variants. Next to the “PRONpers Vlex” and “PRONpers Vbe” sequences that belong to the top three, Table 2 includes the combinations of a personal pronoun with the verb *HAVE* (“PRONpers Vhave”) and with a modal auxiliary (“PRONpers Vmod”) – two sequences that are ranked higher in native speech than in non-native speech – and the combination of a personal pronoun with the verb *DO* (“PRONpers Vdo”), a sequence that is only part of the top thirty of LINDSEI. In addition, the list for LOCNEC includes two trigrams which can both involve a [Subj V] construction: “PRONpers Vbe ADV”, with the addition of an adverb, and “PRONpers Vlex PRONpers”, which can correspond to a [Subj V Obj] construction, as illustrated in (10).

- (10) I don’t really want to go back to do that .. I enjoyed it but I don’t want I want to do something different now (LOCNEC-E013)

Among the four most frequent POS tag sequences shared by native and non-native speakers, we also noticed an incomplete sequence that could correspond to a [PP] construction, since it is made up of a preposition followed by a determiner (“PREP DET”, see example (4)). Next to this incomplete sequence, Table 2 includes two POS n-grams that have the structure of full [PP] constructions, combining, respectively, a preposition and a noun (“PREP N”, example (11)), and a preposition followed by a determiner and a noun (“PREP DET N”, example (12)). The latter, in fact, is the first POS trigram of the list and it happens to be ranked higher in LINDSEI than in LOCNEC. This suggests that the [PP] construction is relatively well entrenched in the spoken construction of EFL learners, allowing them to produce

with a certain degree of automaticity sequences of three words that, in certain cases, will function as post-modifiers of another phrase, thus making the sequence even longer and more complex. Compare, in this respect, (12), where the “PREP DET N” sequence functions independently as an adverbial, with (13), where the underlined “PREP DET N” sequence post-modifies another similar sequence (*at the end*) and is thus part of a longer [PP] construction.

- (11) there w= there was more . er friendship that could be felt among students (LINDSEI-PL018)
- (12) first everyone had to become quiet and at that moment . you saw how he got more nervous actually (LINDSEI-DU022)
- (13) eventually at the end of the book the two women . erm come together again (LINDSEI-FR033)

It was suggested earlier that the [Subj V] construction occupies a prominent position in the construction as it emerges from LINDSEI and LOCNEC. Among the top thirty POS tag sequences, we also find some that seem to point to the prominence (though to a lesser extent) of the [V Obj] construction. The “Vlex PRONpers” sequence, illustrated in (14), is a case in point. The sequence is shared by native and non-native speakers, but is ranked slightly higher among the former (7th rank in LOCNEC and 9th rank in LINDSEI). In addition, LINDSEI and LOCNEC each have a POS trigram which potentially involves a [V Obj] construction and which is not found in the top thirty list of the other corpus: a sequence made up of a lexical verb followed by a determiner and a noun (“Vlex DET N”) in LINDSEI (rank 23), and the “PRONpers Vlex PRONpers” sequence referred to above in LOCNEC, which combines with the [Subj V] construction to form a [Subj V Obj] construction. The trigrams are exemplified in (15) and (16), respectively.

- (14) since it’s not compulsory attending I . try to avoid it (LINDSEI-GR020)
- (15) that’s what I realised cos a friend of mine he bought a car there (LINDSEI-GE015)
- (16) I went travelling in Israel for a week and it impressed me cos it was so different to England (LOCNEC-E032)

In Section 4.1 it was pointed out that the POS tag for adverbs is ranked higher in LOCNEC than in LINDSEI. This predominance of adverbs in native speech is also visible in the list of POS tag sequences. As appears from Table 2, quite a few sequences include an adverb (“ADV”), but there are more such sequences in LOCNEC (8 sequences) than in LINDSEI (5 sequences), and the first one is ranked higher in LOCNEC (rank 6) than in LINDSEI (rank 13). Adverbs being often optional elements in the sentence, they are rarely included in the type of constructions that are typically described in CxG. However, some interesting findings emerge from the top thirty POS tag sequences including an adverb, especially when we compare the native and non-native sequences. I would like to focus on two sequences in particular, namely “ADV PRONpers” and “Vlex ADV”, where the adverb precedes a personal pronoun and follows a lexical verb, respectively. While the former is ranked higher in LINDSEI than in LOCNEC, the latter presents the reverse profile. The prominence of the “ADV PRONpers” sequence in LINDSEI could be related to the tendency of learners, documented for written English (e.g. Granger & Tyson 1996, see also Aarts

& Granger 1998: 137), to overuse connectors like *however* or *therefore* (which are tagged as adverbs by CLAWS) in initial position, that is, before the subject. This feature is exemplified in (17), which illustrates the “ADV PRONpers” sequence. Interestingly, LOCNEC, unlike LINDSEI, includes in its top thirty list of POS n-grams the combinations of an adverb followed by a lexical verb (“ADV Vlex”) and a personal pronoun followed by an adverb (“PRONpers ADV”), which both correspond to an alternative positioning of the adverb within the sentence, cf. (18) and (19). As for the “Vlex ADV” sequence, which is more characteristic of LOCNEC than of LINDSEI, it could partly be explained by the underuse of the phrasal verb construction [V Prt (Obj)] by learners of English, which characterizes both speech and writing but is particularly striking in speech (see Gilquin 2015). (20) provides an example of this construction in LOCNEC.

- (17) actually he played a trick on everybody (LINDSEI-BG039)
- (18) we actually ended up in a s= very tiny village on the coast not far from Dubrovnik (LOCNEC-E039)
- (19) and I suddenly thought well I’m enjoying teaching (LOCNEC-E022)
- (20) through rehearsing it you find out more about .. what . the the different layers of meaning are (LOCNEC-E004)

Finally, Table 2 provides some insight into the use of coordination and subordination in native and non-native speech. Given the fact that subordination is usually considered to be syntactically more complex than coordination (cf. Beaman 1984: 45, Pallotti 2015: 124), we might expect coordinate constructions to be more typical of learner speech and subordinate constructions to be more typical of native speech. This expectation seems to be confirmed by the much higher rank of the “N CCO” sequence (noun followed by coordinating conjunction) in LINDSEI (7th rank) as compared to LOCNEC (15th rank). This sequence can be used to coordinate two nouns ([N *and* N], cf. (21)) or two clauses ([S *and* S], cf. (22)), but in both cases the structure is relatively simple. However, the “CCO PRONpers” sequence (coordinating conjunction followed by a personal pronoun), exemplified in (23), does not show any difference in ranking between LINDSEI and LOCNEC (both are ranked in 11th position). As for the only top thirty POS n-gram including a subordinating conjunction, “CSU PRONpers”, contrary to expectations it is ranked slightly higher in LINDSEI (rank 12) than in LOCNEC (rank 13). This result and the example in (24) demonstrate that learners are capable of syntactic complexity in speech, just like (and sometimes even more than) native speakers, and that these complex constructions can be well entrenched in the learner constructicon.

- (21) I could see the advantages and disadvantages of both systems really (LINDSEI-GE011)
- (22) we just met at the university centre and we . had lunch all together (LINDSEI-FR021)
- (23) I’m not so professional in in these these fields but I like it very much (LINDSEI-CH007)
- (24) she looks at it and she’s not very happy with the result although it does look like her (LINDSEI-SW047)

What this analysis reveals about the spoken EFL constructicon is first of all that it is not necessarily so different from its native equivalent. Among the thirty most

frequent POS tag sequences, twenty-five are shared by native and non-native speakers – despite having different ranks, for the most part. And while the more complex nature of native speech transpires for example from the proportion of POS trigrams in the top thirty, it appears that learners can deal with the complexity of certain structures at least as well as native speakers do. The differences that emerge mainly concern individual sequences, like the phrasal verb construction or the positioning of adverbs within a construction.

A second finding is that the constructions that are most highly entrenched in the native and non-native spoken constructions are relatively basic constructions of the phrase type (NP, PP, etc. or a combination thereof). The top thirty list of POS tag sequences includes very few constructions that could form a complete clause (which, of course, is related to the length of the n-grams). If we exclude cases of ellipsis (which could result in a clause status for, e.g., “PRONpers Vbe” as in *I am* or “PRONpers Vmod” as in *He should*) and imperatives (with which POS n-grams like “Vlex PRONpers” or “Vlex DET N” could be complete clauses, cf. *Check it!* or *Answer this question!*), we find possible instances of the intransitive construction [Subj V] in LINDSEI and LOCNEC through the POS n-gram “PRONpers Vlex” (e.g. *I sympathise*), as well as the [Subj V Obj] construction through the POS n-gram “PRONpers Vlex PRONpers” (e.g. *I like it*), which however is only found in the top thirty of LOCNEC. The sorts of constructions that are typically discussed in CxG, like the ditransitive construction, the caused motion construction or the resultative construction, do not seem to rank among the most commonly produced constructions in (native or non-native) speech. If we examine longer and less frequent POS n-grams, however, we discover sequences that can correspond to some of these constructions, e.g.

- (25) “PRONpers Vlex PRONpers DET N” ~ ditransitive double object construction [Subj V Obj₁ Obj₂]
I just er I gave her some yogurt . and eh she was so happy she was like smiling all the time (LINDSEI-PL011)
- (26) “PRONpers Vlex DET N PREP DET N” ~ caused motion construction [Subj V Obj Obl]
she takes the picture . to her home . and . all her . friends or or family look at the picture and . admire it (LINDSEI-GE045)
- (27) “PRONpers Vlex PRONpers ADJ” ~ resultative construction [Subj V Obj Xcomp]
because it’s horrible you know they drive you crazy (LINDSEI-SP017)

A third element worth underlining, which is perhaps not so apparent from the results outlined above but which has been constantly observed during the analysis, is that one and the same POS tag sequence hides a great variety of linguistic instantiations (especially, as can be expected, for open word classes). Some of these are lexically unique, in that the exact words of the sequence are not repeated elsewhere in the corpus, but they share a syntactic pattern which is brought to light thanks to the POS tagging. Without this syntactic level of abstraction, it would have been impossible to group these sequences together and take them as evidence that some construction might be entrenched in a construction of spoken (learner) English. This does not mean, of course, that the lexical sequences are irrelevant for a CxG-based analysis. For one thing, it is only through the careful examination of these sequences that we can interpret the constructions (see also Section 5). For another, it

would be useful to combine the analysis of POS n-grams with an analysis of lexical n-grams in order to try and distinguish cases where it is the fully abstract construction that seems to be entrenched from cases where, arguably, it is lexically specified constructions with the same syntactic structure that are entrenched.

5. Methodological afterthoughts

Since one of the aims of this chapter was to test the use of POS n-grams as a way of identifying entrenched constructions, a few methodological afterthoughts are in order. The first element to emphasize is that, as demonstrated in the preceding section, POS n-grams do provide valuable insights into the construction by highlighting recurrent patterns that, in some cases, correspond to self-contained constructions as defined in CxG. That this is only true *in some cases*, however, already points to one of the limitations of the methodology, namely that some of the POS tag sequences extracted from the corpora do not qualify as constructions as they have traditionally been recognized in CxG, since they are not structurally complete. The remaining sequences are *potential* constructions in the CxG sense. Sometimes, however, this potential status is not confirmed when we look at (some of) the lexical instantiations of the POS n-gram. For example, it was noted above that the “PRONpers Vlex PRONpers” can correspond to a [Subj V Obj] construction. But next to actual [Subj V Obj] constructions such as *I called her* or *he invited us*, the list of lexical sequences underlying this POS n-gram includes examples like *I believe she*, which in fact introduces a subordinate clause with an ellipted *that*. We also have to take account of the possible embedding of constructions. A “DET N” sequence can be a complete [NP] construction, but it can also be an NP that is embedded in a [PP] construction (cf. *a boy* in *about a boy*). And finally, a POS n-gram could correspond to several distinct constructions depending on the lexical items that are used in the concrete realizations. The “PRONpers Vlex DET ADJ N” sequence, for instance, points to the presence of a transitive construction of the type [Subj V Obj], as illustrated by *she cooked a nice meal*. However, provided a certain kind of verb is used, the sequence can also correspond to a “copular construction” (Goldberg 2006: 8), as in *it became a major success*. What this suggests is that POS n-grams are only an approximation to constructions. Since constructions in CxG are usually expressed in the form of phrases and/or functions, parsing would probably be a more reliable basis than POS tagging for the identification of constructions. The problem is that parsing of learner language is still in its infancy. Schneider & Gilquin (2016) propose an analysis of a parsed learner corpus, but this is a corpus of written English, and we may assume that parsing a corpus of learner speech would be even more challenging.

Related to this last point is the fact that trying to describe the construction of spoken production brings its own share of difficulties. Filled pauses have been mentioned earlier as one type of element that, by interrupting structural units, makes it impossible to extract certain constructions automatically – at least if we take the position, implicit in CxG, that constructions should not include disfluency features. Truncated words, though less common, can have the same effect. While such phenomena can easily be disregarded in the identification of POS n-grams, as has been done here for filled pauses, other disfluent phenomena are more difficult to detect, and thus remove from the data, as they present the same pattern as standard, fluent phenomena. Compare, for example, *a card every day* and *a country a country*, which are both linguistic instantiations of the “DET N DET N” sequence, but only the

second of which presumably corresponds to a disfluent repetition. False starts are another example of a typically spoken phenomenon that could not easily be distinguished from fluent sequences with the same succession of POS tags (unless such phenomena have been annotated beforehand in a special way, probably manually). Solving this problem would involve developing more sophisticated techniques for the automatic treatment of disfluency in corpora, or perhaps simply recognizing the specificity of speech in CxG and admitting that filled pauses or false starts, for example, should have their place, not only as constructions, but also within constructions.¹⁶

Another feature of this study is that it has adopted an essentially corpus-driven approach. While such an approach comes with a commitment to “the integrity of the data as a whole” (Tognini-Bonelli 2001: 84), since it does not start with (possibly biased) assumptions, hypotheses or theories which could cast light on certain data only, in practice the analyst may be overwhelmed by the “wealth of data” (Aarts & Granger 1998: 135). The extraction of the POS n-grams from LINDSEI and LOCNEC provided a huge quantity of data, which could all be examined at different levels of analysis: the level of the POS n-grams (e.g. “DET ADJ N”), their realizations in the form of different lexical sequences (e.g. *the best way*), and the use of these lexical sequences by a certain speaker in a specific context (e.g. *perhaps it’s not the best . erm it’s not the best way to put it* (LINDSEI-PL033)). In this chapter, only a tiny proportion of these data could be examined, and for the top thirty POS n-grams that were analyzed more thoroughly it was not possible to look at all the lexical instantiations of these POS n-grams in context. This also means that manual disambiguation of all the data is not feasible and that the frequencies that have been provided in this study are raw, not only in the sense of being absolute rather than relative frequencies, but also in the sense of being unedited. This is the reason why, in this chapter, it was decided not to place too much emphasis on frequencies, and on quantitative results in general. While we were not able to consider the top 100 POS n-grams that Cappelle & Grabar (2016) suggest should be included in an “n-grammar” of English and that represent, in their own words, less than “the tip of the tip of the tip of the iceberg” (Cappelle & Grabar 2016: 287), this exploratory study has made it possible to draw a first sketch of the constructicon as it emerges from the spoken production of EFL learners.

6. Concluding remarks

In this chapter, a methodology recommended some 20 years ago for the analysis of learner grammar, and recently applied to the identification of constructions in a CxG sense, has been tested on spoken learner English with a view to exploring the constructicon of this language variety, whose study has so far largely been neglected in CxG. Despite its limitations, the methodology has offered some new insights into the structure of learner speech. To those readers who are familiar with constructionist approaches, the results may seem disappointing because they do not reveal the presence, among the top-ranking POS n-grams, of the type of constructions that are

¹⁶ The recognition of typically spoken phenomena is of course not totally absent from CxG or CxG-inspired works (see, e.g., Fried & Östman 2005, Fischer 2010, Fischer & Alm 2013). However, the “bias away from spoken language” that Fried & Östman (2005: 1753) referred to over ten years ago is still very much a feature of most constructionist approaches.

typically dealt with in CxG. However, the results are a reflection of the fact that learner speech – and, to a large extent, native speech too – mainly relies on relatively basic constructions of the type [NP], [PP] or [Subj V]. Note that this does not necessarily say anything about the quality of the speech produced. A basic construction may be instantiated by more or less sophisticated sequences, lexically speaking, depending on the choice of words. Compare, for example, *a really nice place* and *this wonderfully accurate picture*, which are both examples of the “DET ADV ADJ N” POS n-gram but which differ in their degree of lexical sophistication. In addition, POS n-grams do not normally provide information about the correct or idiomatic nature of a sequence. Thus, the “DET ADJ N” POS n-gram includes, next to perfectly appropriate sequences such as *a bright future* or *a famous singer*, less acceptable or idiomatic lexical sequences like *a academical world*, *a beautiful hair* or *a big amount* (5 occurrences of *big amount(s)* in the British National Corpus, as opposed to 714 occurrences of *large amount(s)*). Claiming that learner speech mostly relies on basic constructions, therefore, does not imply any judgment about this language variety. Besides, the list of the top twenty POS trigrams provided in Aarts & Granger (1998: 141) for native and non-native (argumentative) writing suggests that, even in native writing, which arguably represents the kind of default language variety that is (implicitly or explicitly) relied on to make theoretical claims in CxG, syntactically elaborate constructions might not make up the largest proportion of the constructicon.

As can be expected from an exploratory study such as this one, there are many avenues of research that open up in order to refine the preliminary results that have been obtained. Next to the analysis of more and longer POS n-grams in LINDSEI and LOCNEC, we could compare the findings for speech with similar ones for writing, in order to determine what the specificities of each constructicon are. A more targeted approach to the data could also be adopted. While LINDSEI has been treated as an aggregate here, in an attempt to access ‘the’ spoken learner English constructicon, an obvious next step would be to consider certain L1 components of LINDSEI individually in order to pinpoint constructions that are specific to these L1 populations (cf. Aarts & Granger 1998) and discover possible traces of L1 transfer in the use of these constructions (cf. Borin & Prütz 2004). The constructicons thus identified would still be ‘collective’ constructicons, however, emerging from the combined use of language by several learners. Only by analyzing each LINDSEI interview separately would it be possible to identify individual constructicons. Finally, it has been assumed in this chapter that learners’ constructicon is reflected in their language production, and that the corpus frequency of constructions provides information about their degree of entrenchment. However, a constructicon is a mental repertoire of constructions and it might be, for example, that a construction is found in a person’s constructicon but is not instantiated in their language production, especially within the specific context of a 15-minute interview, or that the frequency of a construction in a corpus does not serve as the most accurate indication of how strongly entrenched the construction is in mental representations. It has also been assumed that there is a distinct spoken constructicon, but again, even if the constructions typical of speech and of writing differ from each other, it might not be the case that such a distinction actually exists in people’s minds. In order to answer questions of this type, a more experimental kind of approach should be adopted. Because so little research has been conducted on spoken interlanguage within constructionist approaches, the range of issues that can be investigated is wide. It is to

be hoped that the exploration started here can be taken further and provide fresh insights into both learner language and the nature of speech.

References

- Aarts, J. & Granger, S. 1998. Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In *Learner English on Computer*, S. Granger (ed.), 132-141. London & New York: Addison Wesley Longman.
- Ädel, A. 2006. *Metadiscourse in L1 and L2 English*. Amsterdam & Philadelphia: John Benjamins.
- Altenberg, B. 1991. Amplifier collocations in spoken English. In *English Computer Corpora: Selected Papers and Research Guide*, S. Johansson & A.-B. Stenström (eds), 127-147. Berlin & New York: Mouton de Gruyter.
- Beaman, K. 1984. Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In *Coherence in Spoken and Written Discourse*, D. Tannen (ed.), 45-80. Norwood: Ablex.
- Bel, N., Queralt, S., Spassova, M. & Turell, M. T. 2012. The use of sequences of linguistic categories in forensic written text comparison revisited. In *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*, 192-209. Centre for Forensic Linguistics, Aston University, Birmingham.
- Bencini, G. M. L. & Goldberg, A. E. 2000. The contribution of argument structure constructions to sentence meaning. *Journal of Memory and Language* 43(4): 640-651.
- Biber, D. 2010. Corpus-based and corpus-driven analyses of language variation and use. In *The Oxford Handbook of Linguistic Analysis*, B. Heine & H. Narrog (eds), 193-223. Oxford: Oxford University Press.
- Boas, H. C. 2003. *A Constructional Approach to Resultatives*. Stanford, CA: CSLI Publications.
- Borin, L. & Prütz, K. 2004. New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In *Corpora and Language Learners*, G. Aston, S. Bernardini & D. Stewart (eds), 67-87. Amsterdam & Philadelphia: John Benjamins.
- Bouveret, M. & Legallois, D. (eds). 2012. *Constructions in French*. Amsterdam & Philadelphia: John Benjamins.
- Bybee, J. & Thompson, S. 1997. Three frequency effects in syntax. In *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure*, 378-388. Berkeley: Berkeley Linguistics Society.
- Cappelle, B. & Grabar, N. 2016. Towards an n-grammar of English. In *Applied Construction Grammar*, S. De Knop & G. Gilquin (eds), 271-302. Berlin: De Gruyter.
- Chen, Y.-H. & Baker, P. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14(2): 30-49. <<http://llt.msu.edu/vol14num2/chenbaker.pdf>>
- Clark, H. H. & Fox Tree, J. E. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84: 73-111.
- Conrad, S. M. & D. Biber. 2004. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica* 20: 56-71.
- De Cock, S. 2004. Preferred sequences of words in NS and NNS speech. *Belgian*

- Journal of English Language and Literatures (BELL)*, New Series 2: 225-246.
- De Knop, S. & Gilquin, G. (eds). 2016. *Applied Construction Grammar*. Berlin: De Gruyter.
- De Knop, S. & Mollica, F. 2016. A construction-based analysis of German ditransitive phraseologisms for language pedagogy. In *Applied Construction Grammar*, S. De Knop & G. Gilquin (eds), 53-87. Berlin: De Gruyter.
- Ellis, N. 2013. Construction grammar and second language acquisition. In *The Oxford Handbook of Construction Grammar*, Th. Hoffmann & G. Trousdale (eds), 365-378. Oxford: Oxford University Press.
- Ellis, N. C. & Ferreira-Junior, F. 2009. Constructions and their acquisition. Islands and the distinctiveness of their occupancy. *Annual review of Cognitive Linguistics* 7: 187-220.
- Eskildsen, S. W. 2014. What's new? A usage-based classroom study of linguistic routines and creativity in L2 learning. *International Review of Applied Linguistics* 52(1): 1-30.
- Fischer, K. 2010. Beyond the sentence. Constructions, frames and spoken interaction. *Constructions and Frames* 2(2): 185-207.
- Fischer, K. & Alm, M. 2013. A radical construction grammar perspective on the modal particle-discourse particle distinction. In *Discourse Markers and Modal Particles: Categorization and Description*, L. Degand, B. Cornillie & P. Pietrandrea (eds), 47-87. Amsterdam & Philadelphia: John Benjamins.
- Forsberg, M., Johansson, R., Bäckström, L., Borin, L., Lyngfelt, B., Olofsson, J. & Prentice, J. 2014. From construction candidates to constructicon entries: An experiment using semi-automatic methods for identifying constructions in corpora. *Constructions and Frames* 6(1): 114-135.
- Fried, M. & Östman, J.-O. 2005. Construction Grammar and spoken language: The case of pragmatic particles. *Journal of Pragmatics* 37(11): 1752-1778.
- Gablasova, D. & Brezina, V. 2015. Does speaker role affect the choice of epistemic adverbials in L2 speech? Evidence from the Trinity Lancaster Corpus. In *Yearbook of Corpus Linguistics and Pragmatics 2015: Current Approaches to Discourse and Translation Studies*, J. Romero-Trillo (ed.), 117-136. Dordrecht: Springer.
- Garside, R. & Smith, N. 1997. A hybrid grammatical tagger: CLAWS4. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, R. Garside, G. Leech & A. McEnery (eds), 102-121. London: Longman.
- Gilquin, G. 2015. The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory* 11(1): 51-88.
- Gilquin, G. 2016. POS-tagging LINDSEI: An experiment. Presentation at the *LINDSEI workshop on the POS-tagging of spoken interlanguage*, Louvain-la-Neuve, 15 October 2016.
- Gilquin, G. & De Cock, S. 2011. Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics* 16(2): 141-172.
- Gilquin, G., De Cock, S. & Granger, S. 2010. *Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.
- Gilquin, G. & De Knop, S. 2016. Exploring L2 constructionist approaches. In *Applied Construction Grammar*, S. De Knop & G. Gilquin (eds), 3-17. Berlin: De Gruyter.
- Golcher, F. & Reznicek, M. 2011. Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus. In *Proceedings of Quantitative Investigations in Theoretical Linguistics 4 (QITL-4)*, 29-34. Berlin.

- Goldberg, A. E. 1995. *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago & London: The University of Chicago Press.
- Goldberg, A. E. 2006. *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Granger, S., Dagneaux, E. & Meunier, F. 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain
- Granger, S. & Petch-Tyson, S. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes* 15(1): 17-27.
- Gries, S. Th. & Wulff, S. 2005. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3: 182-200.
- Hilpert, M. 2008. *Germanic Future Constructions. A Usage-based Approach to Language Change*. Amsterdam & Philadelphia: John Benjamins.
- Kennedy, G. 1996. The corpus as a research domain. In *Comparing English Worldwide: The International Corpus of English*, S. Greenbaum (ed.), 217-226. Oxford: Clarendon Press.
- McEnery, T., Xiao, R. & Tono, Y. 2006. *Corpus-Based Language Studies. An Advanced Resource Book*. Oxon & New York: Routledge.
- Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam & Philadelphia: John Benjamins.
- Pallotti, G. 2015. A simple view of linguistic complexity. *Second Language Research* 31(1): 117-134.
- Roehr-Brackin, K. 2014. Explicit knowledge and processes from a usage-based perspective: The developmental trajectory of an instructed L2 learner. *Language Learning* 64(4): 771-808.
- Schneider, G. & Gilquin, G. 2016. Detecting innovations in a parsed corpus of learner English. *International Journal of Learner Corpus Research* 2(2): 177-204.
- Spassova, M. & Turell, M. T. 2007. The use of morpho-syntactically annotated tag sequences as markers of authorship. In *Proceedings of the Second European IAFL Conference on Forensic Linguistics, Language and the Law*, 229-237. Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona.
- Tizón-Couto, B. 2014. *Clausal Complements in Native and Learner Spoken English. A Corpus-Based Study with LINDSEI and VICOLSE*. Bern: Peter Lang.
- Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam & Philadelphia: John Benjamins.
- Tono, Y. 2000. A corpus-based analysis of interlanguage development: Analysing part-of-speech sequences of EFL learner corpora. In *PALC'99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Łódź, 15-18 April 1999*, B. Lewandowska-Tomaszczyk & P. J. Melia (eds), 323-340. Frankfurt am Main: Peter Lang.
- Valenzuela Manzanares, J. & Rojo López, A. M. 2008. What can language learners tell us about constructions? In *Cognitive Approaches to Pedagogical Grammar: A Volume in Honour of René Dirven*, S. De Knop & T. De Rycker (eds), 197-230. Berlin: Mouton de Gruyter.
- Wible, D. & Tsao, N.-L. 2010. StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, 25-31. Los Angeles: Association for Computational Linguistics.

Wood, D. 2010. *Formulaic Language and Second Language Speech Fluency: Background, Evidence and Classroom Applications*. London: Continuum.

Appendix: Simplified version of the C7 tagset

POS tag	Meaning
ADJ	adjective
ADV	adverb
CCO	coordinating conjunction
CSU	subordinating conjunction
DET	determiner
EX	existential <i>there</i>
FO	formula
FU	unclassified word
GE	genitive marker
N	common noun
Nprop	proper noun
NEG	negation <i>not</i> or <i>n't</i>
NUM	numeral
PREP	preposition
PRONindef	indefinite pronoun
PRONpers	personal pronoun
PRONwh	<i>wh</i> -pronoun
PUNC	punctuation
TO	infinitive marker <i>to</i>
TR	truncated word
UH	interjection
Vbe	verb <i>be</i>
Vdo	verb <i>do</i>
Vhave	verb <i>have</i>
Vlex	lexical verb
Vmod	modal verb
ZZ	alphabetical symbol (letter)