Towards an n-grammar of English

Bert Cappelle, Natalia Grabar

IV Constructing a constructicon for L2 learners

Bert Cappelle and Natalia Grabar

Towards an n-grammar of English

Abstract: In this chapter, it is shown how we can develop a new type of learner's or student's grammar based on n-grams (sequences of 2 or 3, 4, etc. items) automatically extracted from a large corpus, such as the Corpus of Contemporary American English (COCA). The notion of n-gram and its primary role in statistical language modelling is first discussed. The part-of-speech (POS) tagging provided for lexical n-grams in COCA is then demonstrated to be useful for the identification of frequent structural strings in the corpus. We propose using the hundred most frequent POS-based 5-grams as the content around which an 'n-grammar' of English can be constructed. We counter some obvious objections to this approach (e.g. that these patterns only scratch the surface, or that they display much overlap among them) and describe extra features for this grammar, relating to the patterns' productivity, corpus dispersion, functional description and practice potential.

Keywords: ESL/EFL, POS n-grams, frequency, construct-i-con, grammar teaching

1 Introduction: Words, words, words, but where's the grammar?

Linguists these days are being spoiled with increasingly large corpora. There is for instance Oxford University's popular British National Corpus (BNC), which contains 100 million words and which is freely available from Mark Davies's website, among other online services.[1] Davies's bigger and more up-to-date Corpus of Contemporary American English (COCA) contains 450 million words (Davies 2008-) and his more recently added Global Web-Based English (GloWbE) allows us to search through 1.9 billion words (Davies 2013). This web corpus is now dwarfed by others, such as ENCOW14, which contains almost 17 billion tokens.[2] And then there is the biggest 'corpus' of all, the indexable part of the World Wide Web itself, which as long ago as June 2006 was estimated to contain 14.3 billion web pages and to increase in size by 280 million web pages a day (De

1 http://corpus.byu.edu/, last accessed on 2 February 2015.

2 http://corporafromtheweb.org/encow14/#more-72, last accessed on 28 February 2015.

Kunder 2006). Whether we use a search engine such as Google or query a comparatively much smaller but still very large corpus designed for linguistic research, what we have at our finger tips in each case is a venerable treasure trove of data about real language use.

The availability of frequency-based word lists compiled from such large corpora of varied texts (e.g. Davies and Gardner 2010) may be of great benefit to practitioners in the field of teaching English as a second or foreign language (ESL/EFL). And indeed, for several decades, corpora have already served as a valuable aid in developing vocabulary teaching materials (see, e.g., McCarthy and O'Dell 2001 for a well-known product). Corpus-based vocabulary teaching prevents certain 'pet' expressions in ESL/EFL, such as *raining cats and dogs*, from being taught too vigorously, and common but less favorite ones, such as *right up your* (or *his*, *her*, etc.) *alley*, from being ignored altogether.

In sharp contrast to the teaching of lexis, grammar teaching does not involve much attention to frequency and focuses instead on, for example, how to construct interrogative or passive structures from canonical (declarative, active) ones. Very often, though, grammar is not even taught that explicitly, since this is felt to go against the prevailing functionally-oriented approach to language learning. It is our impression that when grammar is taught at all, explicitly or in task-based learning settings, the sequence and selection of grammar patterns is mostly a matter of convention and convenience.

Lexis and grammar, as we shall have the opportunity to see, are two sides of the same coin, in that concrete lexical items (words and collocations) belong to more abstract categories (word classes and phrasal structures). One might therefore assume that teaching specific words and expressions automatically results in teaching rules of grammar. Moreover, as there are patterns which combine concrete and more abstract pieces, a distinction between lexis and grammar is often claimed to be illusionary (cf. Ellis and Cadierno 2009). Nevertheless, abstract structures also have an existence which is not wholly reducible to the collocations and idioms that they represent. This is because grammar patterns are generalizations not just over idioms but over lexically rather mundane combinations as well. For instance, the passive construction is not 'just' used in expressions such as *to be cast in stone* or *to be caught between a rock and a hard place*. It is a structure which can be applied productively and it should therefore be taught as such. So, since abstract phrasal constructions do not only underlie frequently used lexical sequences but also provide blue-prints for creative combinations, they need to be focused on in their own right. A purely lexical approach cannot suffice in language teaching.

Most importantly, we need to know which abstract structures are most frequent in the language, because as it is, ESL/EFL is still in dire need of a reliable,

ordered inventory of the most frequently used grammatical patterns in English. Material developers would much appreciate linguists to provide them with a list of common grammar structures for active mastery, to be distinguished from less common patterns that learners can acquire more incidentally. This is, in any case, what the first author of this paper has heard first-hand from an educational advisor for Flemish secondary school teachers (Johan Delbaere, personal communication). The result of this lack of an objective standard of frequent patterns is that constructions that are typically taught may not actually be that frequent and, conversely, that frequent constructions may go unnoticed by material developers. The aim of this paper, therefore, is to show that we can exploit corpus data not just to identify frequent lexical items and lexical patterns of co-occurrence but also to find frequent grammar patterns. That is, just as lexicographers have been successful in detecting common words and collocations, grammarians should really start using corpora to find the most common structural patterns in a language.

To be fair, some existing grammars do take corpus frequencies into account. A prime example is the *Longman Grammar of Spoken and Written English* (Biber et al. 1999), which is entirely corpus-based and provides detailed frequency information (across registers), as well as the *Longman Student Grammar of Spoken and Written English* (Biber, Conrad, and Leech 2002), which is based on the latter. Another example of a corpus-based grammar is Cobuild's two-volume *Grammar Patterns* (Francis, Hunston, and Manning 1996, 1998), whose lexicogrammatical approach, outlined in Hunston and Francis (2000), is heavily influenced by work by Halliday and Sinclair (e.g. Halliday 1978; Sinclair 1991). Other early studies that comment on frequency of use are referenced in Celce-Murcia and Larsen-Freeman (1999). However, despite these valuable works, linguists so far have not yet produced any *ranking* of frequent grammar patterns for the benefit of EFL/ESL teachers, students and material developers.

We will show that this can be achieved by using *n-grams* – continuous sequences of *n* (i.e. any specified number of) items. Our demonstration will be restricted to n-grams extracted from the COCA corpus. This is entirely for practical reasons, as will become clear.[3] We believe that common lexical and grammatical n-grams are constructions, in a Construction Grammar sense: they are form-

3 Apart from COCA, there are other corpora which allow n-gram-based grammar studies. For instance, as shown in Cappelle (2014), using Google's *Ngram Viewer* (Michel et al. 2010), we can exploit the n-grams extractable from Google Books for (diachronic) research into grammar patterns, since this corpus has been tagged and allows part-of-speech searches (Lin et al. 2012). The COW corpora also provide n-gram data sets (http://hpsg.fu-berlin.de/cow/ngrams/, last accessed on 28 February 2015).

(Unicode 9 17/12/15 10:39) WDG-New (155mm×230mm) DGMetaSerifScience (OpenType) pp. 269–302 1695 Gilquin_10_Cappelle (p. 273)

function pairings which native speakers have memorized (and which learners of a language should acquire) as a result of their high frequency. For Construction Grammarians, frequency is only one of the criteria to identify constructions, another possible criterion being the unpredictable nature of the link between a unit's form and its function (e.g. Goldberg 2006). Yet, a great number, perhaps even a majority, of Construction Grammarians these days seem to take a usage-based approach to the study of patterns, which means that they consider a unit as a construction as soon as it has sufficient frequency (as evidenced by corpus data), regardless of whether or not that unit displays any sort of arbitrariness in the way its form links up with its function. This is also the approach taken here. We are less concerned with the potential unpredictability of a pattern's form or function than with its high frequency.

The structure of our paper is as follows. In Section 2, we will introduce the concept of n-grams. In Section 3 we will propose an application of n-grams to English language learning. Section 4 is devoted to some possible criticisms that could be levelled at this approach and to our rebuttal of them. Section 5 presents some further features of an envisaged n-gram-based grammar, or 'n-grammar', of English, which is a project-in-progress. Our conclusions can be found in Section 6.

2 What are n-grams, and what are they typically used for?

N-grams are sequences of n items, where *n* stands for any natural number (1, 2, 3, 4, etc.) of linguistic units. For example, the word string *the fool on the hill* contains five 1-grams (usually called 'unigrams'), namely *the*, *fool*, *on*, *the* and *hill*, four 2-grams (or 'bigrams'), namely *the fool*, *fool on*, *on the* and *the hill*, three 3-grams (or 'trigrams'), namely *the fool on*, *fool on the* and *on the hill*, two 4-grams, namely *the fool on the* and *fool on the hill*, and also one 5-gram, namely the string *the fool on the hill* itself. There are not just word-based n-grams but also character-based n-grams. Thus, the letter sequence *chat* consists of four unigrams (*c*, *h*, *a* and *t*), three bigrams (*ch*, *ha* and *at*), two trigrams (*cha* and *hat*) and one 4-gram (*chat*). The items in question that an n-gram has *n* adjacent instances of could be of any category. For instance, in Section 3, we will make use of n-grams whose items are word classes (determiner, noun, verb, etc.).

N-grams can be automatically extracted from spoken and/or written corpora and primarily play a role in computational linguistics, where they are used for statistical language modelling. By 'language model', computational linguists

understand a set of probabilities (P's) which reflect, as accurately as possible, real language use. As Jurafsky (2012) puts it, "[i]t might have been better to call this 'the grammar'. I mean, technically, what this is, is telling us something about how [well] [...] words fit together, and we normally use the word 'grammar' for that, but it turns out that the word 'language model' [...] is standard". Based on n-grams extracted from a large corpus, a language model may compute the likelihood of an entire string of *n* items ('joint probability') and/or the likelihood of a single upcoming item given *n*-1 previous items ('conditional probability'). Estimates of these probabilities generated by an n-gram-based language model are used in a variety of practical applications. Table 1 gives some examples, drawn from Jurafsky (2012).

Table 1: Some applications of an n-gram-based probabilistic language model (based on Jurafsky 2012)

Application	Task	Example
Machine translation	Distinguishing between 'good' and 'bad' translations by their probabilities	*High winds tonight* may be a better translation than *large winds tonight*, based on: P(*high* winds tonight) > P(*large* winds tonight)
Spell correction	Detecting likely mistakes based on the probabilities of word sequences	*The office is about fifteen minuets from my house* likely contains a misspelling from *minutes*, based on: P(about fifteen *minutes* from) > P(about fifteen minuets from)
Speech recognition	Deciding between two sequences that sound phonetically similar by comparing their probabilities	*I saw a van* is likely to be a more accurate transcription than *eyes awe of an*, based on: P(*I saw a van*) >> P(*eyes awe of an*)

There are many other everyday applications. Word-based and character-based n-grams underlie features such as word suggestion and word completion available on search engines and on our smartphones' text messaging function.

While extracting n-grams from corpora is a common method of identifying recurrent formulae in discourse, other types of sequences are sometimes used apart from n-grams, such as so-called lexical bundles (Biber, Conrad, and Cortes 2004), p-frames (Römer 2010) and skip-grams (Guthrie et al. 2006).

3 Using COCA n-grams for a new kind of grammar

3.1 The problem of ubiquitous constructions

Finding out what the most frequently used constructions are in a short text may sound like an easy enough task. In fact, it is not. To begin with, we would have to decide on an appropriate definition of 'construction'. Secondly, suppose that we adopt a quite open definition of 'construction', as is common in Construction Grammar (e.g. Goldberg 2006), and count as construction every learned form-function pairing, ranging from individual words and morphemes to larger syntactic structures, it would then be hard not to overlook any of them, as any single sentence typically may contain one or several dozen constructions. This will become clear if we consider an example taken from Goldberg (2003):

(1) *What did Liza buy the child?*

This short sentence contains all of the following constructions:

(2) a. the *buy*, *child*, *did*, *Lisa*, *the* and *what* constructions (i.e. words)

b. the Ditransitive construction (i.e. double-object construction)

c. the Question construction (which is a fairly abstract construction, involving a certain intonation contour)

d. the Subject-Auxiliary Inversion construction (which is not only used in questions)

e. the VP construction

f. three cases of the NP construction (namely, *What*, *Liza* and *the child*)

For the time being, it is technically very hard, if not impossible, to detect and tally all these kinds of constructions automatically, which is what would be required if we wanted to count constructions in a whole corpus.

We propose to bypass the problem of scripting such a construction-detecting program by relying on readily available part-of-speech (POS) n-grams, which we will treat as constructions (or major parts thereof). This decision, of course, needs proper justification, which we will attempt to give in Section 4.1. At present, we are focusing on describing the methodology used.

3.2 The general idea

Via the website www.ngrams.info, one can download free lists of the most frequent 2-, 3-, 4- and 5-grams from COCA. Each list contains about 1,000,000 lexical n-grams. The lists are ordered from the most frequent to the least frequent n-grams. Table 2 gives some examples from the top of each list.

Table 2: N-grams from COCA, with some of the highest-frequency examples

N-grams	Examples
2-grams	*of the*, *is a*, *going to*, *I think*, ...
3-grams	*one of the*, *a lot of*, *the United States*, *as well as*, ...
4-grams	*I do-n't know*, *for the first time*, *on the other hand*, ...
5-grams	*I do-n't think so*, *the rest of the world*, *by the end of the*, ...

Observe, by the way, that the contracted negator (*-n't*) is treated as a separate word by the tagger.

Via the website mentioned above, it is also possible to download lists of n-grams where part-of-speech tags are presented together with the actual words making up each n-gram. What we claim here is that one can exploit this information to find common grammar structures in the corpus (and hence, to the extent that COCA is a representative corpus, in a major variety of the English language). For the purposes of illustration, Figure 1 shows the top section of the list of 4-grams containing part-of-speech information.

The left-most column gives us the number of occurrences ('tokens') of the lexical n-gram ('type') in question in COCA. Thus, *I don't know* is the most frequent 4-gram in COCA, occurring 54,632 times in the corpus, followed by *I don't think*, with 43,760 occurrences.

The four columns to the right contain the part-of-speech information, based on the CLAWS 7 tagset.[4] Thus, the tag *ppis1* stands for 'singular personal pronoun, first person, subjective case' (i.e. the word *I*), *vd0* for '*do* as a finite form (in declarative and interrogative clauses)', *xx* for '*not*' or its contracted form, and *vvi* for 'the base form of a lexical verb used as an infinitive'. As can be noticed, the first two n-grams have the same part-of-speech tags. They share this part-of-speech tagging with *I don't want*, a little further down the list (see the boxes with dotted lines). Similarly, the 4-grams *the end of the* and *the rest of the* (in 4th and 6th position) share their part-of-speech labelling (see the boxes with full lines). The idea now is to order all these POS 4-grams by their frequency,

4 http://ucrel.lancs.ac.uk/claws7tags.html, last accessed on 28 February 2015.

54632	I	do	n't	know	ppis1	vd0	xx	vvi
43760	I	do	n't	think	ppis1	vd0	xx	vvi
33968	in	the	United	States	ii	at	np1	np1
29848	the	end	of	the	at	nn1	io	at
27119	do	n't	want	to	vd0	xx	vvi	to
21537	the	rest	of	the	at	nn1	io	at
19864	at	the	end	of	ii	at	nn1	io
19165	for	the	first	time	if	at	md	nnt1
18632	I	do	n't	want	ppis1	vd0	xx	vvi
18115	at	the	same	time	ii	at	da	nnt1
16809	in	the	middle	of	ii	at	nn1	io
16681	one	of	the	most	mc1	io	at	rgt
16626	of	the	United	States	io	at	np1	np1
15857	is	one	of	the	vbz	mc1	io	at
14392	to	be	able	to	to	vbi	jk	to

…

Figure 1: Most frequent lexical 4-grams ('types') from COCA, together with their number of corpus occurrences ('tokens') and their part-of-speech tags

that is, by the number of different lexical 4-grams (lexical types) that instantiate them. What this reordering results in is shown in Figure 2.

The most frequent POS 4-gram in COCA is the one instantiated by *at the end of*, *in the middle of* and 6,984 other sequences of a preposition (other than *of*), the definite article, a singular common noun and the preposition *of*. Other than the list of lexical 4-grams (Figure 1), this list gives us direct information about what the most common syntactic structures are for 4-word sequences in COCA. We could use such a frequency list as the basis for an n-grammar of English. As an added bonus, we could combine this syntactic information with lexical information about the most common actual 4-grams (see also Section 3.4). Indeed, Construction Grammar assumes that both lexical chunks and the more general patterns they instantiate have their role to play in (first and second) language

6986	ii at nn1 io
5382	nn1 io at nn1
4645	ii at jj nn1
4235	nn1 ii at nn1
4177	at jj nn1 io
3847	at nn1 io at
3609	ii at1 jj nn1
3569	at1 jj nn1 io
3313	to vvi at nn1
3249	at nn1 ii at
3028	ii at nn1 nn1
2848	at nn1 io nn1
2797	ii at nn1 cc
2684	at1 nn1 ii at
2573	at1 jj nn1 ii
…	

Figure 2: Most frequently instantiated POS 4-grams in COCA with number of lexical instantiations ('types') for each POS 4-gram

acquisition (cf. Ellis 1996, 2003, 2013; Tomasello 2003; see also Lewis 1993, the papers in Cowie 1998 and Wray 2002, inter alia, on the role of chunks in acquisition). Learners can be said to master a target language all the more accurately the more they manage to use lexical items in their preferred constructional environment (Wulff and Gries 2011).

Rather than using 2-, 3- or, as just demonstrated, 4-grams, we suggest using 5-grams as the basis of our n-grammar of English, which are the longest n-grams available from COCA's n-gram website. Traditional grammars tend to focus on shorter units, but if we want to target intermediate to advanced students, we believe that strings of 5 segments present an adequate size – neither too short, nor too long. While even longer n-grams could in principle have been used, if they had been available from COCA, we do not think such longer strings would

have provided many more relevant constructions, since longer strings are likely to be made up of shorter component structures, which we will show to be the case for 5-grams already. Moreover, in those cases in which a 5-gram does not coincide with a complete syntactic phrase, it can still be extended 'by hand' with a phrasal category, something which will also be illustrated below. In short, our choice of using complete or extended 5-grams is thus motivated by the aim to use units that are as long and complete as possible, but which at the same time still allow manipulation and combination to form even larger structures in the language.

3.3 The method in detail

We restrict our selection to the 100 most frequent POS 5-grams based on the COCA list of lexical 5-grams containing part-of-speech information. In an n-gram-based grammar of English, each such pattern could and should be presented together with some of its frequent lexical instantiations, so as to show how the skeletal structures can be fleshed out in actual language use. Why 100 patterns? This is a somewhat arbitrary choice, motivated less by linguistic factors than by reasons related to learner motivation: learners might consider 100 patterns an achievable target. Needless to say, one could also select 200 patterns, 500 patterns, etc., or alternatively 365 patterns, one for each day of the year.

The list of lexical 5-grams with part-of-speech tags that can be downloaded from Mark Davies's website mentioned above (www.ngrams.info) contains exactly 1,293,537 types of lexical strings. The list is cut off at 5-grams with a minimum frequency of 5 occurrences in the corpus (presumably because the list was meant to contain ca. one million types). Remember from Table 2 that this list contains such sequences as *I don't want to* or *the rest of the world*. We grouped these lexical strings according to the syntactic patterns they instantiate (i.e., their part-of-speech tag sequence). We thus obtained a total of 325,552 POS 5-grams. The number of lexical strings ('types') per POS 5-gram varies from 7,272 for *at nn1 io at nn1*, a structure shared by *the rest of the world*, *the side of the road* and thousands more (where *at* stands for *the*, *nn1* for a singular common noun and *io* for *of*), to just 1, for instance in the case of *appge cc appge jjt nn1* for *his or her best interest* (in which *appge* stands for a possessive determiner, *cc* for a coordinating conjunction and *jjt* for a superlative adjective).

We then took the 100 most frequent POS 5-grams as main content for the n-grammar. By 'most frequent' POS 5-grams, we mean those that represent the highest number of types, that is, the highest number of different lexical 5-grams

that have the structure specified by them. These patterns (for which we also made the part-of-speech labels more transparent) now range in frequency from still 7,272 types for the sequence [*the* X_{noun} *of the* Y_{noun}] to 499 types for the sequence made up of a complex 3-word preposition followed by *the* and an adjective (e.g. *in front of the whole*).

As is clear from this last example, a 5-gram does not necessarily form a complete constituent, since 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. We therefore added an element to the right in those cases where the right-most boundary of a 5-gram does not coincide with a closing bracket, so to speak. Thus, in the case of the pattern [*at/in/to/… the* X_{noun} *of the*], we just add an element to the right of the determiner. This element could be a noun, but it could also be an adjective which precedes a noun, among other possibilities. Technically, the grammatical category covering all of these is what is called a 'nominal' (nom) in Huddleston, Pullum et al.'s (2002) grammar, or an 'N-bar' ($N^{|}$) in X-bar theory (Chomsky 1970; Jackendoff 1977) – that is, a noun phrase minus the determiner. We always completed with a category label that stands for the widest range of possible continuations. Additions are between parentheses in what follows. Table 3 shows the result of the procedure for the ten most common POS 5-grams, along with an example of each.

Table 3: Ten most frequent syntactic (completed) 5-grams in COCA

Syntactic pattern	Example
the X_{noun} *of the* Y_{noun}	*the rest of the world*
at/in/to/… the X_{noun} *of the* (Y_{nom})	*at the end of the (day)*
X_{noun} *at/in/to/… the* Y_{noun} *of* (Z_{NP})	*increase in the number of (students)*
the adj X_{noun} *of the* (Y_{nom})	*the other side of the (room)*
to verb *the* X_{noun} *of* (Y_{NP})	*to improve the quality of (life)*
at/in/to/… the adj X_{noun} *of* (Y_{NP})	*on the other side of (the room)*
a(n) adj X_{noun} *at/in/to/… the* (Y_{nom})	*a far cry from the (original proposal)*
at/in/to/… the X_{noun} *of one's* (Y_{nom})	*at the top of his (lungs)*
the X_{noun} *of the* adj (Y_{nom})	*the end of the Cold (War)*
at/in/to/… the X_{noun} *at/in/to/… the* (Y_{nom})	*on the way to the (hospital)*

We never added any symbol to the left of the 5-gram, as the left-most symbol always constitutes the first element of a constituent. However, in some cases, a lexical instantiation of a POS n-gram might benefit from one or more added elements at the left. For instance *the benefit of the doubt* is a syntactically complete

unit – it is an NP – but it typically occurs in lexically larger environments, involving verbs such as *give*, *get* or *deserve*.

3.4 The medium-level and hybrid nature of the POS n-grams

As the reader will have noticed, the syntactic information associated with the n-grams from COCA consists of quite specific part-of-speech tags. While there are only eight or nine word classes traditionally recognized in English, the CLAWS 7 tagset, which was used for tagging COCA, distinguishes between 137 different categories. For instance, the preposition *of* is treated not as any preposition but as the word *of*, all the forms of the verb *be* are treated differently from each other, singular nouns are treated as different from plural nouns, common nouns as different from proper nouns, and so on. As a result, the POS n-grams are not maximally general, as would have been the case if they were of the type 'Det N Prep Det N'. Nor of course are these POS n-grams maximally specific, which is the case only for purely lexical n-grams such as *the name of the motel*, where each item is an actual word. While it would be possible for us to come up with more general patterns based on the specific part-of-speech tags and to calculate their frequencies, our POS n-grams as they are may in fact come close to having the ideal grain size of a construction: neither too schematic nor too concrete. We do not want to claim here that there are no such things as very general constructions or that some specific lexical strings cannot have the status of stored language units; indeed, a standard assumption of Construction Grammar is that generalizations over exemplars and the (sufficiently frequent) exemplars themselves are stored in the speaker's mind (cf. Section 4.1). Yet, while it is obvious that specific items *have to* be stored if they are formally or semantically unpredictable – storage is required for words and idioms – there is no equally compelling reason why we would need to store the most schematic rules of language. As Croft (1998: 168) formulates it, "[s]peakers do not necessarily make the relevant generalizations, even if clever linguists can".[5]

5 In actual fact, Croft's (1998) quoted sentence is lifted from an article that deals more with semantics (polysemy and homonymy) than with the level of generality at which speakers store constructions. However, these issues are not unrelated and lie at the heart of the difference between a Goldbergian (1995) approach to argument structure constructions (i.e. one in which they are treated as highly schematic form-meaning patterns) and a Boas-style (2003) or Iwata-style (2008) approach to them (i.e. one in which so-called 'mini-constructions' or specific lexical constructions are associated with individual verbs or even individual verb senses). See also Levshina and Heylen (2014) for related findings about the optimality of medium-level granularity in the context of semantic classes of predicates governing the choice between competing constructions.

Apart from, or as a corollary of, being somewhat below the maximum level of generality, our POS n-grams are also somewhat hybrid in nature, that is, they are *partially* rather 'syntactic' and *partially* rather 'lexical'. The mixing of levels results purely from the rich, fine-grained tagset that is used for the COCA corpus, but there are computationally more sophisticated methods for automatically generating linguistically 'interesting' n-grams which combine lexical items and formal categories: see Wible and Tsao (2010), Lyngfelt et al. (2012) and Forsberg et al. (2014). Again, by mixing more general and more specific items in a single template, we may approximate the ideal of constructions viewed as language units that actually operate in the mind of speakers. For instance, in our top hundred POS n-grams, we find the following hybrid structures:[6]

(3) a. *the* X_{noun} *of the* Y_{noun} (pattern No. 1), e.g. *the rest of the world*

b. *the* X_{noun} *of a(n)* Y_{noun} (pattern No. 20), e.g. *the son of a bitch*

c. *a(n)* X_{noun} *of the* Y_{noun} (pattern No. 30), e.g. *a thing of the past*

The trained linguist may need some convincing to see that these are distinct patterns and not *just* different realizations of a single more general pattern. But notice, first of all, the difference in frequency. The POS n-gram in (3a), as we noted above, covers 7,272 lexical types (with at least five tokens, i.e. corpus occurrences, each), while the ones in (3b) and (3c) only cover 1,448 and 968 lexical types, respectively. The pattern [*a(n)* X_{noun} *of a(n)* Y_{noun}], with two indefinite articles, does not even rank among the hundred most frequent POS 5-grams. Secondly, while each pattern provides open slots for nouns, they do not allow the same nouns in these slots. For instance, we would not find ?*the son of the bitch*, ??*the thing of the past* or ??*a rest of the world*. This suggests that each pattern has its own particular properties, causing it to attract certain nouns and to repel certain others. We will come back to this in Section 5.3.

Because of a pattern's close association with *some* lexical items and not with others, we feel that it is worthwhile to provide this information to learners. This is fully in line with a constructionist and usage-based approach to language learning, which stresses the importance of exemplars in acquisition (Abott-Smith and Tomasello 2006; Ellis 2006, 2013). So, ideally, an n-grammar should present not just semi-schematic and hybrid patterns but also some of

6 The first segment is not always the determiner *the* but could also be the quantifier *no*, as in *no mention of the fact (that…)*. Because this determiner is used far less frequently than *the*, we use the latter as a transparent substitute for the tag 'at'. A rare example in which the quantifier *no* is used before the second noun is *the point of no return*.

(Unicode 9 17/12/15 10:39) WDG-New (155mm×230mm) DGMetaSerifScience (OpenType) pp. 269–302 1695 Gilquin_10_Cappelle (p. 283)

the frequent lexical instantiations that they generalize over. In the case of the pattern [*the* X_{noun} *of the* Y_{noun}], this would mean that the learner also gets to see some fully lexical sequences, possibly even with corpus frequencies (number of tokens) added to them, as shown in Table 4.

Table 4: Most frequent lexical realizations of the pattern [*the* X_{noun} *of the* Y_{noun}] in COCA

the rest of the world (3,618)	*the benefit of the doubt* (547)
the side of the road (1,217)	*the edge of the bed* (530)
the rest of the country (1,174)	*the center of the room* (526)
the fact of the matter (825)	*the State of the Union* (495)
the end of the world (764)	*the back of the room* (463)
The fact of the matter (717)	*the back of the head* (450)
the end of the war (670)	*the middle of the room* (448)
the rest of the way (597)	...

Note, incidentally, that *the fact of the matter* appears twice. This is due to the fact that the downloadable lexical n-grams with POS information are case-sensitive, which means that a word with a capital letter and the same word without a capital letter are treated as belonging to different n-grams. This may seem like a nuisance, but it actually provides useful information about where that n-gram is found in the sentence (sentence-initially or not).

4 Possible points of criticism and their rebuttal

We are aware of some immediate objections that one might raise against the approach we take to selecting patterns to be included in a new, radically usage-based grammar. We can think of at least the following four points of criticism:

(i) Not all of these POS 5-grams are constructions.
(ii) There is a lot of (and perhaps too much) overlap between them.
(iii) The top hundred POS 5-grams are but the tip of the iceberg.
(iv) By restricting ourselves to 5-grams, we may miss out on interesting 2-, 3- and 4-grams.

In the following subsections, we will defend our approach against this possible criticism.

4.1 Not all of them constructions?

One might wonder what is so special about, for example, the pattern [*the* X_{noun} *of the* Y_{noun}], which appears to be formed on the basis of some general phrase-structure rules, namely the ones listed in (4), combined with the knowledge that *the* is a determiner and *of* a preposition:

(4) a. NP → det nom

b. nom → noun (PP)

c. PP → prep NP

It is true that on Goldberg's (1995: 4) original definition of the term, this first pattern and many (if not most) of the other patterns in our top hundred POS 5-grams would not qualify as constructions. This definition stated that a form-function pairing is a construction only if there is something about its form or function that is not strictly predictable from what is already available in the grammar. However, in Goldberg's (2006: 5) later work, this requirement is loosened: "Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. *In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency*" [emphasis ours]. The view that constructions are psychologically entrenched form-function pairs is also expressed by Croft and Cruse (2004: 288), Langacker (2005: 140) and Bybee (2006: 715).

Given their high frequency, which is the basis of their selection in the first place, there is no doubt that all of the hundred POS 5-grams meet this definition of 'construction': even if some of them are compositional, their sheer frequency makes it unlikely that they are formed anew each time they are used. It is much more plausible that these patterns are directly retrieved from what construction grammarians, using a term coined by Jurafsky (1991), call the 'construct-i-con'.

4.2 Too much overlap between them?

It will have been noted that many of the patterns shown in Table 3 look like variations on a theme. That is, they do not all represent fully distinct constructions. For example, the two most frequent POS 5-grams ([*the* X_{noun} *of the* Y_{noun}] and [*at/in/to*/… *the* X_{noun} *of the* (Y_{nom})]) share four fifths of their component elements (namely '*the* X_{noun} *of the*').

As we see it, however, this overlap is not a problem. On the contrary, it allows us to integrate in this new type of learner's and student's grammar an important feature of language: the possibility of reusing parts of structures over and over again in slightly different environments or with slight modifications. In other words, one of the main properties of language is that it involves structures that are partially reusable and adaptable. This feature can be visualized in chart form, which for each pattern shows its relatedness to some of the other patterns in the grammar. Figure 3 is an example of what such a chart could look like for the first pattern in the n-grammar.

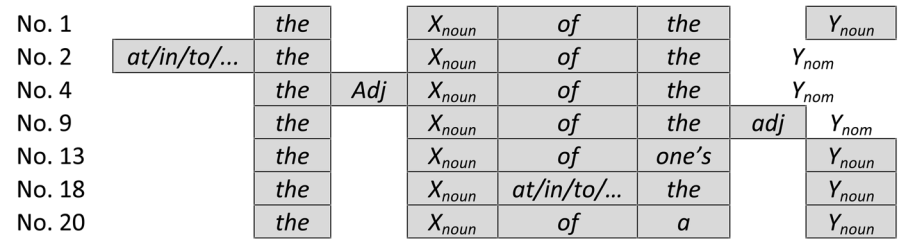

Figure 3: 'Chop and change' chart for the first pattern in the n-grammar (top row)

The chart shows how a pattern can be chopped up to allow for the insertion of elements in the right position (e.g. adjectives before nouns) and be changed by replacing elements in a structural position by alternatives in that position (e.g. an indefinite article or a possessive determiner instead of a definite article). As such, this 'chop and change chart' (a term whose rights of use in grammar instruction we hereby reserve) directly represents the syntagmatic and paradigmatic relations between structural elements in the grammar. The overlap between patterns reflects the fact that grammar is a combinatorial system, which operates on classes of discrete string segments such as adjectives or nouns. This is not just how linguists see it, but how the human brain treats grammar (Pulvermüller and Knoblauch 2009; Pulvermüller, Cappelle, and Shtyrov 2013).

The n-grammar proposed here is thus not a maximally parsimonious system to generate word sequences. Instead, it represents structural information in a way that is full of redundancy. This redundancy may be helpful for learners to master the structures: though they could have been captured more economically, this would have been at the expense of a range of learning opportunities spread out through time, which is needed for consolidated acquisition. This is as true for humans as it is for the simplest of organisms. To cite the Nobel prize winning neuroscientist Eric R. Kandel: "Conversion of short-term to long-term

memory storage requires spaced repetition – practice makes perfect, even in snails" (Kandel 2001: 1031).

4.3 Just the tip of the iceberg?

Remember that for the n-grammar proposed here, we retained a mere 100 POS 5-grams out of 325,552 such patterns in COCA (themselves generalizing over more than a million lexical 5-grams with at least 5 occurrences each in the corpus). Put differently, our top hundred most frequently occurring syntactic templates only represent 0.03% of all possible POS 5-grams based on lexical 5-grams with at least 5 tokens in COCA. Put differently still, our selection does not seem to count for much. Using familiar imagery, if the part of an iceberg that appears above water is only one tenth of its total volume, then our selection does not represent the proverbial tip of the iceberg, and not even the tip of the tip of the tip of the iceberg.

If we disregard POS 5-grams that are instantiated by fewer than 5 different lexical 5-grams ('types'), there are no longer 325,552 POS 5-grams in COCA, but 36,617 of them. Our selection then represents 0.27% of these. This is admittedly still a very small portion, which is barely visible in the left-most stacked bar in Figure 4. However, if we now look at the types represented by these 36,617 POS 5-grams, we find that there is a total number of 823,683 different lexical 5-grams in COCA. Our top hundred POS 5-grams represent a non-negligible portion of these: 105,184 types, which is 12.77% (cf. middle stacked bar in Figure 4). In terms of tokens (individual occurrences), there are 11,248,178 sequences of 5 words corresponding to the 36,617 POS 5-grams. Our top hundred POS 5-grams cover 1,615,199 tokens. This high number represents 14.36% of the total number of tokens for all 36,617 POS 5-grams (cf. the stacked bar on the right in Figure 4). What we find here is something akin to what Zipf (1935) noted for lexical items, namely that the most frequent items (types) cover a large part of the occurrences (tokens) in usage. In the Brown corpus, for instance, half of the word volume is accounted for by only 135 vocabulary items (Fagan and Gençay 2010).

In sum, even our very small set of POS 5-grams (just one hundred out of more than thirty thousand in the corpus) appears to have quite large coverage in terms of number of lexical strings (both the types and their tokens) that correspond to these syntactic templates. It exceeds the percentage of an iceberg that extends above the water surface.

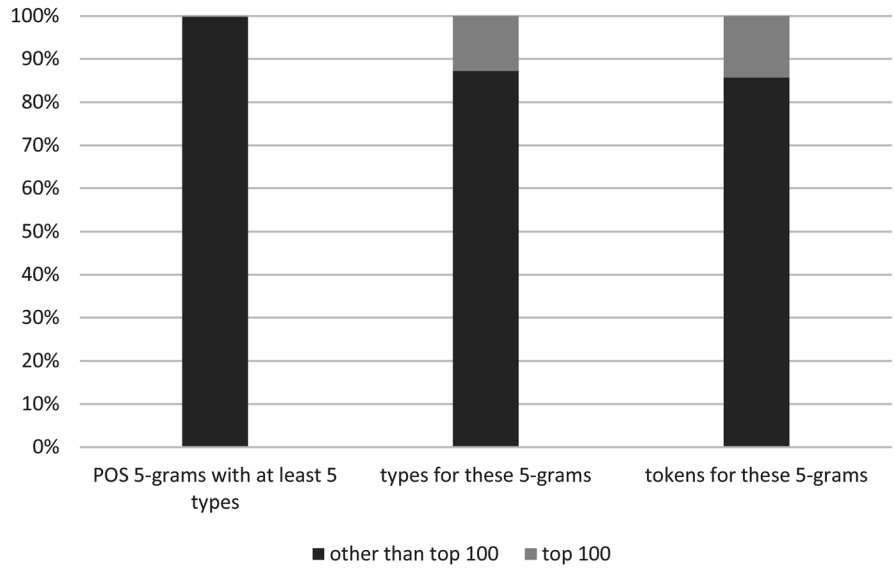

Figure 4: Share of the selected highest-frequency POS 5-grams in COCA and its coverage in terms of types and tokens

4.4 Neglecting important 2-, 3- and 4-grams?

It may seem an odd choice to take 5-grams as our basis for a grammar of English. Even quite apart from the fact that some constructions are longer than five words or may involve discontinuities and hence cannot be captured as 5-grams, there is the risk of overlooking interesting grammar patterns that are shorter than 5 words. For instance, among the top ten most frequently used 2-grams, we found the pattern [X_{noun} $Y_{noun\ plur}$], that is, a noun-noun compound in the plural, such as *family members*, *interest rates*, *phone calls*, *college students*, *side effects*, and several thousands more. Unfortunately, the two-word POS sequence [X_{noun} $Y_{noun\ plur}$] is not a part of any of our top hundred POS 5-grams. Does this example not suggest that we may fail to integrate some vital grammar patterns by focusing on 5-grams only?

While any common lower-n-gram pattern that is not included in any of our 5-gram patterns is surely a missed opportunity to capture all that is essential in grammar, the actual situation does not give reason for too much concern. Table 5 shows that the five highest-frequency POS 2-grams taken together are included 102 times in our selection of POS 5-grams, and that the five highest-frequency POS 3-grams and POS 4-grams are still also included 64 times and 35 times, respectively. (Obviously, shorter sequences have a higher likelihood of being included than longer ones.)

Table 5: Inclusion of high-frequency POS 2-, 3- and 4-grams from COCA in the hundred most frequent POS 5-grams from COCA

N-grams with n < 5	Top 5 most frequent POS n-grams	Number of inclusions in top 100 POS 5-grams
2-grams	1. adj X_{noun}	19
	2. X_{noun} Y_{noun}	3
	3. adj $X_{noun\ plur}$	2
	4. X_{noun} *{at/in/to/. . .}*	20
	5. *the* X_{noun}	58
3-grams	1. *{at/in/to/. . .} the* X_{noun}	27
	2. *a(n)* adj X_{noun}	9
	3. *the* adj X_{noun}	7
	4. X_{noun} *{at/in/to/. . .} the* (X_{nom})	14
	5. *{at/in/to/. . .} an* X_{noun}	7
4-grams	1. *{at/in/to/. . .} the* X_{noun} *of* (Y_{NP})	17
	2. X_{noun} *of the* Y_{noun}	6
	3. *{at/in/to/. . .} the* adj X_{noun}	3
	4. X_{noun} *{at/in/to/. . .} the* Y_{noun}	6
	5. *the* adj X_{noun} *of* (Y_{noun})	3

Clearly, it is not the case that by looking at 5-grams only, we ignore <5-grams. As explained in Section 2, a 5-gram by its very nature simultaneously harbors two 4-grams, three 3-grams and four 2-grams. As a consequence, if we study a hundred 5-grams, we actually get to see a thousand n-grams, not even counting the individual words. This is not to say that we get nine hundred *different* 2-, 3- and 4-grams for free with our 5-grams, as many of these lower-n-grams will be included several times. This is not a problem, in light of our discussion of redundancy and repetition in Section 4.2.[7]

7 To look at this from a somewhat different perspective, we might say that the high frequency of certain 5-grams accounts to some extent for the frequency of some of its included <5-grams. This is what O'Donnell (2011) points out to be the case for lexical n-grams: for instance, at the top of the list of lexical 5-grams in COCA we find *I don't want to*. This occurs 12,659 times in the corpus and thereby contributes in no small way to the high frequency of its multiple component parts (*I*, *do*, *n't*, *want*, *to*, *don't*, *want to*, *I don't want*, etc.). Therefore, by studying high-frequency 5-grams, the learner is given a glimpse into some of the reasons why smaller combinations are so frequent. This is true, we feel, for both n-gram templates of the sort discussed in our text and lexically 'filled-in' n-grams, of the sort O'Donnell (2011) focuses on.

By selecting 5-grams, we automatically retrieve more complex structures. This is why a '5-grammar' of English may be more ideally suited for intermediate to advanced learners of English than for absolute beginners. For lower-level learners, n-grams other than 5-grams (namely, 2-grams, 3-grams and 4-grams) might be a better way to start. In other words, we do not want to claim that using 5-grams is the only valid way of constructing an n-grammar of English.

5 Further features of the n-grammar

5.1 Adding a visual measure of productivity

Pedagogical grammars do not generally contain any statistical information about frequency, unlike modern dictionaries, many of which provide an indication of how common a word is, or in which genre or register it is typically used. Our proposed n-grammar can easily include such information. In Sections 3.2 and 3.4 we already suggested that individual lexical n-grams associated with the more schematic POS n-grams may be shown with their actual corpus frequencies, thus giving the learner some idea of their usefulness as chunks in the target language. We believe that if learners see, for example, that *The fact of the matter* at the beginning of a sentence is used more than 700 times in a corpus of native-speaker English, this kind of knowledge may cause them to take note of this expression more consciously and stimulate them to use it themselves (cf. e.g. Schmidt (1990) and Robinson (2006) on the 'noticing' hypothesis and the role of conscious attention in second language acquisition). But the patterns themselves could also be provided with frequency information. Thus, the most frequent POS 5-gram, [*the* X_{noun} *of the* Y_{noun}], might be stated explicitly to have 7,272 types (with at least 5 tokens each). In addition, we might mention that these types together represent 126,077 tokens in the corpus.

Such figures may not mean much by themselves to the learner. Though such high numbers might of course be impressive and therefore encourage the learner to devote due attention to the pattern, they will vary from corpus to corpus. A more general indication of frequency, similar to what can be found in certain dictionaries (e.g. high-frequency, medium-frequency, low-frequency) could be sufficient, if it were not for the fact that the top hundred most frequent POS 5-grams are naturally all at the high end of the frequency scale anyhow. It would probably be more beneficial to the learner to have a direct visual indication of a pattern's *usefulness*. If by 'useful' we mean how many different lexical realizations the pattern allows the learner to form, we should include a measure of the pattern's productivity. Productivity can be defined in terms of the ratio of

types per tokens: the more types per number of tokens, the more productive a pattern is. This is clear if we consider the other extreme case: if all the corpus occurrences of a pattern were instances of only one lexical string, that 'pattern' would have no productivity at all. Alternatively, productivity can be expressed in terms of the ratio of unique corpus occurrences ('hapax legomena', i.e. types with only one token) per tokens (cf. Baayen 1989): the more such single-occurrence types, the higher the probability that also 'outside' the corpus the pattern will be used to form novel creations and so the more productive the pattern. We propose here to combine the two measures (type-to-token ratio and hapax-to-token ratio) in a single graph.

There is one slight problem to overcome. Remember that all the lexical n-grams used for our n-grammar have at least five corpus occurrences, so that, strictly speaking, there are no hapax legomena among them. Therefore, we need to rely on a related statistic, which we could call 'pentakis legomena', that is, sequences that occur only five times in the corpus. The ratio of these, too, just like the ratio of hapax legomena, can give us an idea of how readily novel combinations are formed based on a given pattern. The cut-off of five occurrences per type (cf. Section 3.3) also results in a somewhat skewed type/token ratio: above this cut-off point, there is a smaller type/token ratio (as here we find types with comparatively many tokens) than below that cut-off point (where we find types with relatively few tokens). Our solution to compensate for this skewing is to multiply the type/token ratio by 5. This makes mathematical sense: suppose all the lexical types had just five occurrences, then the unadjusted type/token ratio would be 0.2, and by multiplying this by 5, we would obtain the maximal productivity score of 1, which would be just what we would like to find in that situation. Likewise, the pentakis/token ratio is also skewed compared to the more commonly used hapax/token ratio, since for any pattern, if there are many 'pentakises', one could expect there to be even more hapaxes. So, there is naturally a smaller 'pentakis'/token ratio given a cut-off restriction of 5 occurrences than there would be a 'hapax/token' ratio if the cut-off restriction was removed (and this is so even if the total number of tokens would of course also increase if we removed the cut-off restriction). The solution, here too, exists in multiplying the pentakis/token ratio by 5. The formula for the combined and adjusted type/token and pentakis/token measure of productivity of a pattern is given below, whereby n stands for the total number of lexical types instantiating the pattern with at least five occurrences, N the total number of tokens for all these types and p the number of lexical types with just five occurrences:

$$Productivity = \frac{\left(\frac{n}{N}\right).5 + \left(\frac{p}{N}\right).5}{2}$$

What this says is that the productivity of a pattern can be calculated by taking the average of its type/token ratio multiplied by five and its pentakis/token ratio multiplied by five. Thus, for the pattern [*the* X_{noun} *of the* Y_{noun}], the number of lexical types *n* is 7,272 types, the total number of corpus occurrences *N* is 122,685 and the number of pentakis legomena *p* is 1,670. If we feed these numbers in the formula above, we get the following result:

$$Productivity = \frac{\left(\frac{7,272}{122,685}\right).5 + \left(\frac{1,670}{122,685}\right).5}{2} = 0.186$$

This result can be represented in graph form on a scale from 0 to 1. A theoretical zero value of productivity would be obtained for a pattern where all types are prefabricated chunks. The value 1, for full productivity, would be the score for a pattern where all types are novel creations (or at least, where they are all pentakises). For ease of visual interpretation and comparison with other patterns, we use a logarithmic scale of 10, with the minimum value approximating zero and the maximum value 1.

Figure 5 charts the productivity of the first and second most frequent POS 5-gram in COCA. The lower productivity of [*at/in/to/*... *the* X_{noun} *of the* (Y_{nom})] is explained by the fact that the part (Y_{nom}) is not actually included in the 5-gram and so plays no role in the type and token data used for the calculation. This 'extended' 5-gram thus contains only two open slots, for a preposition and a noun, only the latter of which is an open word class. In the first pattern, there are two slots for an open word class, so the productivity of this pattern is obviously much higher. One might wonder whether it makes sense to add items that are not taken into account when counting the number of occurrences. The reason why we did this is that we want to show learners how a POS 5-gram can be used grammatically. If a POS 5-gram ends in a determiner or an adjective, we find it useful to state what the next element will be (a nominal). It should be obvious, though, that this element cannot be taken into consideration when we want to compare POS-grams for frequency, as there is no easy way to list up all the possible instantiations of this element, which could be a bare noun, a noun preceded by one or more adjectives, a noun followed by a prepositional phrase of any length, and so on.

It is important to make the learner see that a low-productivity pattern does not necessarily equal an uninteresting one. The lower the productivity, the greater the role of strongly entrenched sequences, which are responsible for the high token frequency. Thus, while this second pattern is clearly less productive, its most common type ([*at the end of the* Y_{nom}]) has 10,663 occurrences in COCA, against only 3,618 occurrences for the most frequently used lexical sequence

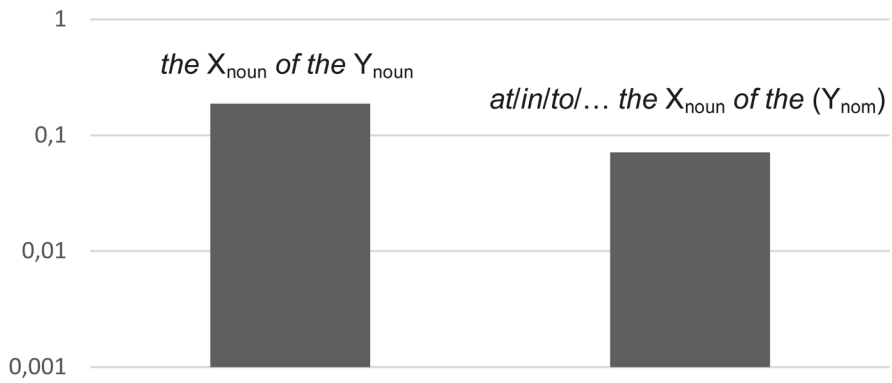

Figure 5: Productivity of two grammar patterns, visualised on a logarithmic scale

instantiating the first pattern (*the rest of the world*). To avoid the automatic association of higher with 'better', a suitable alternative visual representation might be one that plots the productivity score on a horizontally-oriented scale, where patterns towards the left margin are more 'chunkified' or 'lexical' and patterns towards the right margin are more 'gridlike' or 'syntactic'.

5.2 Adding a visual measure of dispersion

5-grams may not appear as frequently in some genres or registers as they do in others. To indicate how evenly or how skewed a language item appears in different sections of a corpus, we can (or even should) use measures of dispersion (cf. Gries 2008a). Figure 6 illustrates a visually attractive way of showing which of the large components of COCA make use of the pattern [*the* X_{noun} *of the* Y_{noun}] the most and the least. The data were obtained by entering the query 'the [*nn1*] of the [*nn1*]' in the COCA search interface and looking up how many hits we retrieve in each of the main components of the corpus (spoken, fiction, magazine, newspaper and academic). This is information which the COCA search interface provides at a click of the mouse. Note that this search retrieves results for types whose token frequency is also lower than five. As could be expected of a rather complex NP, we find this grammatical structure used least frequently in spoken English and most frequently in academic writing. The spread through the corpus is clearly uneven.

Besides making immediately clear the uneven frequency of the pattern across broad corpus components, the graph can be adapted to indicate in which of these components, if any, the pattern in question occurs *much* more/less fre-

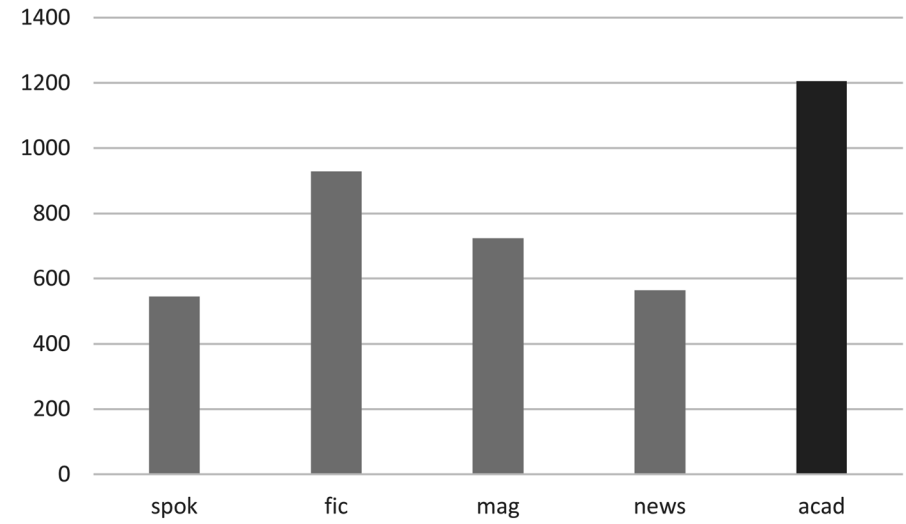

Figure 6: Dispersion of *the* X_{noun} *of the* Y_{noun} across COCA components. Bars indicate number of tokens per million words.

quently than expected (given an average). Specifically, we can use darker grey and lighter grey for markedly higher or lower frequency, respectively. This allows the learner to quickly identify the genre(s) in which the pattern is more conspicuously present or absent than what could have been expected under the assumption that it had an equal chance of occurrence in each component. We here define a markedly higher or lower frequency as a difference of at least 50% (for a positive difference, i.e. a surplus) or at least 33.33% (for a negative difference, i.e. a shortage) compared to the expected frequency. In this case, the expected frequency is 790 hits per million words (a figure obtained by dividing the total number of occurrences by the total corpus size multiplied by one million). Only in academic written discourse does the pattern display a marked difference (namely, an overuse of 53%) between what is observed and what is expected. In spoken discourse, the pattern is not sufficiently underrepresented – there is an underuse of (only) 31% – for its relative infrequency in this component to be considered of significance to the learner. This is why only the bar corresponding to academic writing has been given a different grey shade in Figure 6.

This uneven dispersion suggests that if we had used another corpus (containing for instance no academic writing at all), the frequency ranking of our POS n-grams might have been very different. The same may be true if we had used a corpus representing another variety of English. Obviously, corpus results

depend on (and vary with) the corpus used. This is also valid for the tagset (see Section 3.4), whose choice will have an important influence on the patterns that are extracted, as well as for the settings (e.g. case sensitivity, mentioned also in Section 3.4). There is no such thing as *the* n-grammar of English.

5.3 Providing a functional description

The quantitative measures discussed in Sections 5.1 and 5.2 may be interesting to the numerically-minded learner. However, especially if we want to adopt a constructionist approach to language pedagogy, we should also attempt to show how each pattern has its own functional properties, unless one is prepared to argue for "the legitimacy of semantically null constructions" (Fillmore et al. 2012: 326; see Hilpert 2014: Chapter 3 for discussion). Providing a semantic or functional characterization of a POS n-gram may need some inventiveness on the part of the grammarian. Yet, by looking at the most frequent lexical n-gram instantiations, we often get a clue as to what the pattern is predominantly used for. In the case of the by now familiar pattern [*the* X_{noun} *of the* Y_{noun}], we could formulate its function along the lines shown in (5):

(5) [*the* X_{noun} *of the* Y_{noun}]
This pattern allows speakers to link two entities: the noun phrase following *of* (e.g. *the road*) and the noun preceding it (*side*). Among the most frequent instantiations of this pattern, there are quite a few sequences where the first noun denotes a portion (e.g. *rest*) or a position or dimension in space or time (e.g. *end*, *side*, *edge*, *center*, *middle*, *back*, *top*) which 'zooms in' on a part of the larger whole expressed by the noun phrase after *of*. Not surprisingly, this pattern overlaps with the next most productive one, namely [*at/in/to/…* *the* X_{noun} *of the* Y_{noun}] (pattern No. 2), which adds a preposition to indicate a relation to this spatial or temporal portion or location, e.g. *at the end of the Y*, *in the middle of the Y*, *at the top of the Y*, *by the end of the Y*.

Note that this pattern's functional description is not only informed by its frequent lexical instantiations; it also brings out the formal and functional relatedness of this pattern with another one.

5.4 Providing opportunities for practice

Finally, let us offer some thoughts on how the selected grammar patterns can be integrated into language learning activities aimed at consolidating the syntactic

structures and their common instantiations. We hope that developers of language learning materials might come up with a full range of concrete ideas, but one obvious possibility of a practice activity for a given pattern is to encourage learners to use suitable lexical instantiations in particular usage contexts. This could be implemented as a simple fill-in exercise, which may or may not take the form of a multiple-choice task whereby, given a particular sentence, learners have to use the most suitable lexical n-gram from a set. These lexical n-grams themselves, too, could be presented with gaps, which learners have to fill in with contextually suitable items. Ideally, the sentences to be completed should be taken from carefully selected authentic spoken or written discourse (although one may have to clean up and/or simplify attested examples if learners are to benefit from them optimally; cf. Gries 2008b); the fillers should be chosen from the set of high-frequency sequences provided with the pattern (cf. Table 4).

Another sort of exercise could take the form of a role play between pairs of students, who in a particular usage situation have to use a number of preselected n-grams. For instance, two students could be asked to act as people of influence in international politics, such as the Secretary of State of the US and the British Prime Minister, discussing one or other rogue state's presumed possession of weapons of mass destruction. Student A has to use *(give X / get / deserve) the benefit of the doubt*, *the rest of the world*, *the State of the Union* and *(on/off/from) the face of the earth*. Student B has to use *the end of the world*, *the fact of the matter*, *(reach) the end of the line* and *(just) the tip of the iceberg*. The teacher should not stop the role play until he or she is satisfied these sequences have been used accurately and naturally (i.e. in a correct syntactic environment and, whenever relevant, taking account of the idiomatic or encyclopedic meaning of an expression, which of course should first have been illustrated by means of authentic examples).

Once such common sequences have been mastered, a further exercise could consist in using n-grams flexibly. For instance, teams could compete against each other to produce the highest number of phrases that instantiate a POS n-gram. This would allow them to learn the language by thinking in more general categories, to exploit the combinatorial flexibility of grammar and use and reuse at best available linguistic chunks. Other creatively-oriented exercises could be to use patterns to form rhymes (for more advanced students), but of course, such an exercise should not replace tasks that appeal more directly to functional needs. As an alternative focusing again more on realistic language use, students could be asked to detect instantiations of a set of POS n-grams, say 5 different ones provided to them, in an authentic text. Some of these POS n-grams, and accordingly their instantiations, may overlap (e.g. *with the tip of the* and *the tip of the tongue*), demonstrating how n-grams can incrementally

combine to form full sentences. An easy related exercise could be to ask learners which lexical 5-grams in a text form complete constituents and which lexical 5-grams do not. Such an exercise would raise students' awareness of language structure and might enhance linguistic insight.

For students of linguistics, the 'linear' approach to grammar proposed here could be offset by digressions on how the seemingly purely sequential structure of grammar patterns is actually hierarchically organized. An n-gram-based approach to grammar need not be incompatible with a more traditional linguistic reflection about constituent structure. For instance, we may explain to students that the sequence *the rest of the world* has the structure shown in Figure 7.

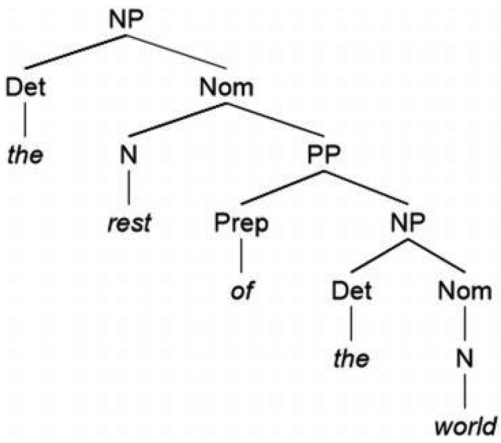

Figure 7: Tree diagram showing the hierarchical structure of a linear sequence

Such a tree diagram could lead to a discussion of recursion – the fact that a noun phrase can contain a noun phrase, for instance – or of why a preposition should not be simplistically defined as a word which comes before a noun – since in that case, a determiner would also be a preposition. An exercise could be to form complex trees by means of tree fragments constituting a toy grammar, such as one in which NP branches into Det and Nom, another in which Nom branches into N and PP (or just N), and yet another in which PP branches into P and NP.

6 Conclusion

These are exciting times. We have access to online corpora, specifically designed for linguistics or otherwise, containing staggering amounts of words. Corpora

have opened our eyes to lexical frequencies, collocations, and so on, but they need not close our eyes to syntactic patterns. With the help of automatic taggers that reach high precision rates, we can now hold our descriptions of grammar to the same empirical standard as our descriptions of the lexicon.

Espousing a radically usage-based approach to grammar, we have shown here how we can make use of relatively schematic templates derived from the Corpus of Contemporary American English as the basis of an 'n-grammar' of English. The selection of patterns to be included in such a grammar is based on the corpus frequency of part-of-speech 5-grams, which we consider to be constructions. We have demonstrated that a small number of high-frequency 5-grams – just one hundred out of several tens of thousands – can cover quite a large portion of actually used 5-word strings (as well as strings of 2, 3 and 4 words). This leads to a rather revolutionary approach to developing a pedagogical grammar: it breaks with traditional sequencing in grammars, which often deal first with everything related to the verb, then the noun, etc. – in this or another convenient (for largely conventional) order. Our motivation for following the order suggested by corpus frequencies is that it seems to us only common sense that those patterns that underlie the highest number of concrete word sequences should be presented before any others. While Leech (2011) may be right in taking a somewhat more considered stance regarding the principle "more frequent = more important to learn", this is mainly because some researchers (cf. De Cock and Granger 2004) have noted that learners may overuse vocabulary items, such as *big* or *nice*. It is true that there are very common words which learners will soon come across, use themselves successfully and as a result start feeling rather too comfortable with. Leech's reservation, however, applies less to more complex grammar patterns of the sort we have considered here, consisting of as many as five segments.

Our proposal to rank frequent and productive n-gram templates may help EFL/ESL material developers to select form-function patterns for active mastery. We hope to have illustrated or at least suggested how the construction of a book-length construct-i-con, discussing the hundred most frequent 5-gram templates, can lead to a fresh, empirically based and ultimately perhaps more relevant approach to teaching grammar.

Acknowledgements

Earlier versions of this paper were presented not just at the international conference Constructionist Approaches to Language Pedagogy (CALP, 8-9 November 2013, Brussels, Belgium) but also at the University of Erlangen-Nurnberg,

Germany (1 July 2014) and the University of Tsukuba, Japan (8 July 2014). We thank the organizers of these events for giving us a stage, as well as members of the audience whose questions helped us formulate our ideas more clearly. We are also especially grateful to the editors of this volume and the anonymous reviewers for their constructive comments. All remaining inadequacies are ours alone.

References

Abott-Smith, Kirsten & Michael Tomasello. 2006. Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review* 23. 275–290.

Baayen, R. Harald. 1989. *A corpus-based approach to morphological productivity: Statistical analysis and psycholinguistic interpretation*. Ph.D. dissertation, Free University of Amsterdam.

Biber, Douglas, Susan Conrad & Viviana Cortes. 2004. *If you look at...*: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3). 371–405.

Biber, Douglas, Susan Conrad & Geoffrey Leech. 2002. *Longman student grammar of spoken and written English*. London: Longman.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Boas, Hans C. 2003. *A constructional approach to resultatives*. Stanford: CSLI Publications.

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82 (4). 711–733.

Cappelle, Bert. 2014. Review of Stefan Thim, *Phrasal verbs: The English verb-particle construction and its history*. Berlin & New York: Mouton de Gruyter. *English Language and Linguistics* 18(3). 572–586.

Celce-Murcia, Marianne & Diane Larsen-Freeman. 1999. *The grammar book: An ESL/EFL teacher's course*. Boston, MA: Heinle & Heinle.

Chomsky, Noam. 1970. Remarks on nominalization. In Roderick A. Jacobs & Peter S. Rosenbaum (eds.), *Reading in English transformational grammar*, 184–221. Waltham: Ginn.

Croft, William. 1998. Linguistic evidence and mental representations. *Cognitive Linguistics* 9(2). 151–173.

Croft, William & D. Allen Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Cowie, A.P. (ed.). 1998. *Phraseology: Theory, analysis, and applications*. Oxford: Oxford University Press.

Davies, Mark. 2008–. *The Corpus of Contemporary American English: 450 million words, 1990–present*. Available online at http://corpus.byu.edu/coca/.

Davies, Mark. 2013. *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. Available online at http://corpus2.byu.edu/glowbe/.

Davies, Mark & Dee Gardner. 2010. *A frequency dictionary of contemporary American English: Word sketches, collocates, and thematic lists*. London/New York: Routledge.

De Cock, Sylvie & Sylviane Granger. 2004. Computer learner corpora and monolingual learners' dictionaries: The perfect match. *Lexicographica* 20. 72–86.

De Kunder, Maurice. 2006. *Geschatte grootte van het geïndexeerde World Wide Web* [Estimated size of the World Wide Web]. MA dissertation. Tilburg University.

Ellis, Nick C. 1996. Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition* 18. 91–126.

Ellis, Nick C. 2003. Constructions, chunking, and connectionism: The emergence of second language structure. In Catherine J. Doughty & Michael H. Long (eds.), *Handbook of second language acquisition*, 63–103. Malden, MA: Blackwell.

Ellis, Nick C. 2006. Cognitive perspectives on SLA: The Associative-Cognitive CREED. *AILA Review* 19: 100–121.

Ellis, Nick C. 2013. Second language acquisition. In Thomas Hoffmann & Graeme Trousdale (eds.), *Oxford handbook of Construction Grammar*, 365–378, Oxford: Oxford University Press.

Ellis, Nick C. & Teresa Cadierno. 2009. Constructing a second language: Introduction to the special section. *Annual Review of Cognitive Linguistics* 7. 111–139.

Fagan, Stephen & Ramazan Gençay. 2010. An introduction to textual econometrics. In Aman Ullah & David E. A. Giles, *Handbook of empirical economics and finance*, 133–153. Boca Raton: Chapman & Hall/CRC.

Fillmore, Charles J., Russell R. Lee-Goldman & Russell Rhodes. 2012. The FrameNet Constructicon. In Hans C. Boas & Ivan A. Sag (eds.), *Sign-Based Construction Grammar*, 283–299. Stanford: CSLI.

Forsberg, Markus, Richard Johansson, Linnéa Bäckström, Lars Borin, Benjamin Lyngfelt, Joel Olofsson & Julia Prentice. 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.

Francis, Gill, Susan Hunston & Elizabeth Manning. 1996. *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins. Available online at http://arts-ccr-002.bham.ac.uk/ccr/patgram/.

Francis, Gill, Susan Hunston & Elizabeth Manning. 1998. *Collins COBUILD Grammar Patterns 2: Nouns and adjectives*. London: HarperCollins.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5). 219–224.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Gries, Stefan Th. 2008a. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437.

Gries, Stefan Th. 2008b. Corpus-based methods in analyses of second language acquisition data. In Peter Robinson & Nick Ellis (eds.), *Handbook of Cognitive Linguistics and second language acquisition*, 406–431. New York: Taylor & Francis.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie & Yorick Wilks. 2006. A closer look at skip-gram modelling. *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)*, 1222–1225. Genoa, Italy.

Halliday, M.A.K. 1978. *Language as social semiotic: The social interpretation of language and meaning*. London: Arnold.

Hilpert, Martin. 2014. *Construction Grammar and its application to English*. Edinburgh: Edinburgh University Press.

Huddleston, Rodney, Geoffrey K. Pullum et al. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Hunston, Susan & Gill Francis. 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Iwata, Seizi. 2008. *Locative alternation: A lexical-constructional approach*. Amsterdam: John Benjamins.

Jackendoff, Ray S. 1977. *X-bar syntax: A study of phrase structure*. Cambridge, MA: MIT Press.

Jurafsky, Dan. 1991. *An on-line computational model of human sentence interpretation: A theory of the representation and use of linguistic knowledge*. Ph.D. Dissertation, University of California, Berkeley.

Jurafsky, Dan. 2012. Language modeling: Introduction to n-grams. Lecture slides of Stanford University's online course Natural Language Processing. https://class.coursera.org/nlp/lecture/14 (last accessed on 7 August 2014).

Kandel, Eric R. 2001. The molecular biology of memory storage: A dialogue between genes and synapses. *Science* 294. 1030–1038.

Langacker, Ronald W. 2005. Construction grammars: Cognitive, radical, and less so. In Francisco J. Ruiz de Mendoza Ibáñez & M. Sandra Peña Cervel (eds.), *Cognitive Linguistics: Internal dynamics and interdisciplinary interaction*, 101–159. Berlin: Mouton de Gruyter.

Leech, Geoffrey. 2011. Frequency, corpora and language learning. In Fanny Meunier, Sylvie De Cock, Gaëtanelle Gilquin & Magali Paquot (eds.), *A taste for corpora: In honour of Sylviane Granger*, 7–32. Amsterdam: John Benjamins.

Levshina, Natalia & Kris Heylen. 2014. A radically data-driven Construction Grammar: Experiments with Dutch causative constructions. In Ronny Boogaart, Timothy Colleman & Gijsbert Rutten (eds.), *Extending the Scope of Construction Grammar*, 17–46. Berlin: Mouton de Gruyter.

Lewis, Michael. 1993. *The Lexical Approach: The State of ELT and the Way Forward*. Hove, UK: Language Teaching Publications.

Lin, Yuri, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, William Brockman & Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram Corpus. *Proceedings of the 50th annual meeting of the Association for Computational Linguistics, Volume 2: Demo papers (ACL '12)*, 169–174. Stroudsburg, PA: Association for Computational Linguistics.

Lyngfelt, Benjamin, Lars Borin, Markus Forsberg, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg & Sofia Tingsell. 2012. Adding a Constructicon to the Swedish resource network of Språkbanken. *Proceedings of KONVENS 2012 (LexSem 2012 workshop)*, 452–461. Vienna, September 2012.

McCarthy, Michael & Felicity O'Dell. 2001. *English vocabulary in use: Upper-intermediate*. Second Edition. Cambridge: Cambridge University Press.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2010. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014). 176–182.

O'Donnell, Matthew Brook. 2011. The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal* 35. 135–169.

Pulvermüller, Friedemann & Andreas Knoblauch. 2009. Discrete combinatorial circuits emerging in neural networks: A mechanism for rules of grammar in the human brain? *Neural Networks* 22. 161–172.

Pulvermüller, Friedemann, Bert Cappelle & Yury Shtyrov. 2013. Brain basis of meaning, words, constructions, and grammar. In Thomas Hoffmann & Graeme Trousdale (eds.), *Oxford handbook of Construction Grammar*, 396–416, Oxford: Oxford University Press.

Robinson, Peter. 2006. Attention, memory, and the "noticing" hypothesis. *Language Learning* 45(2). 283–331.

Römer, Ute. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3(1). 95–119.

Schmidt, Richard W. 1990. The role of consciousness in second language learning. *Applied Linguistics* 11(2). 129–158.

Sinclair, John. 1991. *Corpus, concordance, collocation: Describing English language*. Oxford: Oxford University Press.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Wible, David & Tsao, Nai-Lung. 2010. StringNet as a computational resource for discovering and investigating linguistic constructions. *Proceedings of the NAACL HLT workshop on extracting and using constructions in computational linguistics*, 25–31. Los Angeles, CA: ACL.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wulff, Stefanie & Stefan Th. Gries. 2011. Corpus-driven methods for assessing accuracy in learner production. In Peter Robinson (ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance*, 61–88. Amsterdam: Benjamins.

Zipf, George K. 1935. *The psychobiology of language: An introduction to dynamic philology*. Cambridge, MA.: MIT Press.