Vol..3, Issue 6

ISSN:2349- 0012

I.F. 8.1

UDK:81-133

GRAMMAR IN CORPORA

Kaxorova Nargiza Nusratovna

The teacher of Bukhara State university

Kaxorovanargiza5@gmail.com

Abstract: First, we must operationally define the structure itself in such a way that we (and other researchers) can reliably categorize potential instances as manifesting the structure or not. This may be relatively straightforward in the case of simple grammatical structures that can be characterized based on tangible and stable characteristics, such as particular configurations of grammatical morphemes and/or categories occurring in sequences that reflect hierarchical relations relatively directly. It becomes difficult, if not impossible, with complex structures, especially in frameworks that characterize such structures with recourse to abstract, non-tangible and theory-dependent constructs.

Key words: a query, retrieve, potential, candidates, corpus, case morphologically, marked, relatively simple, grammatical structures, the s-possessive.

Introduction

We must define a query that will allow us to retrieve potential candidates from our corpus in the first place (a problem we discussed in some detail). Again, this is simpler in the case of morphologically marked and relatively simple grammatical structures, for example, the s-possessive (as defined 8 Grammar above) is typically characterized by the sequence ⟨ [pos="noun"] [word="'s"%c] [pos="adjective"]* [pos="noun"] ⟩ in corpora containing texts in standard orthography; it can thus be retrieved from a POS-tagged corpus with a fairly high degree of precision and recall. However, even this simple case is more complex than it seems.

The query will produce false hits: in the sequence just given, ’s may also stand for the verb be (Sam’s head of marketing). The query will also produce false misses: the modified nominal may not always be directly adjacent to the ’s (for example in This office is Sam’s or in Sam’s friends and family), and the s-possessive may be represented by an apostrophe alone (for example in his friends’ families). Other structures may be difficult to retrieve even though they can be characterized straightforwardly: most linguists would agree, for example, that transitive verbs, are verbs that take a direct object. However, this is of very little help in retrieving transitive verbs even from a POS-tagged corpus, since many noun-phrases following a verb will not be direct objects (Sam slept the whole day) and direct objects do not necessarily follow their verb (Sam, I have not seen); in addition, noun phrases themselves are not trivial to retrieve. Yet other structures may be easy to retrieve, but not without retrieving many false hits at the same time. This is the case with ambiguous structures like the of - possessive, which can be retrieved by a query along the lines of ⟨ [pos="noun"] [pos="determiner"]? [pos="adjective"]* [pos="noun"] ⟩, which will also retrieve, among other things, partitive and quantitative uses of the of -construction. Finally, structures characterized with reference to invisible theoretical constructs (“traces”, “zero morphemes”, etc.) are so difficult to retrieve that this, in itself, may be a good reason to avoid such invisible constructs whenever possible when characterizing linguistic phenomena that we plan to investigate empirically. These difficulties do not keep corpus linguists from investigating grammatical structures, including very abstract ones, even though this typically means retrieving the relevant data by mind-numbing and time-consuming manual analysis of the results of very broad searches or even of the corpus itself, if necessary. But they are probably one reason why so much grammatical research in corpus linguistics takes a word-centered approach. A second reason for a word-centered approach is that it allows us to transfer well-established collocational methods to the study of grammar. In the preceding chapter we saw that while collocation research often takes a sequential approach to co-occurrence, where any word within a given span around a node word is counted as a potential collocate, it is not uncommon to see a structure-sensitive approach that considers only those potential collocates that occur in a particular grammatical position relative to each other – for example, adjectives relative to the nouns they modify or vice versa. In this approach, grammatical structure is already present in the design, even though it remains in the background. We can move these types of grammatical structure into the focus of our investigation, giving us a range of research designs where one variable consists of (part of) the lexicon (with values that are individual words) and one variable consists of some aspect of grammatical structure. In these studies, the retrieval becomes somewhat less of a problem, as we can search for lexical items and identify the grammatical structures in our search results afterwards, though identifying these structures reliably remains non-trivial. We will begin with word-centered case studies and then move towards more genuinely grammatical research designs.

An early extension of collocation research to the association between words and grammatical structure is Renouf & Sinclair (1991). The authors introduce a novel construct, the collocational framework, which they define as “a discontinuous sequence of two words, positioned at one word remove from each other”, where the two words in question are always function words – examples are [a __ of], [an __ of], [too __ to] or [many __ of] (note that a and an are treated as constituents of different collocational frameworks, indicating a radically word-form-oriented approach to grammar). Renouf and Sinclair are particularly interested in classes of items that fill the position in the middle of collocational frameworks and they see the fact that these items tend to form semantically coherent classes as evidence that collocational frameworks are relevant items of linguistic structure. The idea behind collocational frameworks was subsequently extended by Hunston & Francis (2000) to more canonical linguistic structures, ranging from very general valency patterns (such as [V NP] or [V NP NP]) to very specific structures like [there + Linking Verb + something Adjective + about NP] (as in There was something masculine about the dark wood dining room (Hunston & Francis 2000: 51, 53, 106)). Their essential insight is similar to Renouf and Sinclair's: that such structures (which they call "grammar patterns") are meaningful and that their meaning manifests itself in the collocates of their central slots:

The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively, if it is dependent on a particular word choice, and if there is a clear meaning associated with it.

Collocational frameworks and especially grammar patterns have an immediate applied relevance: the COBUILD dictionaries included the most frequently found patterns for each word in their entries from 1995 onward, and there is a two-volume descriptive grammar of the patterns of verbs (Francis et al. 1996) and nouns and adjectives (Francis et al. 1998); there were also attempts to identify grammar patterns automatically (cf. Mason & Hunston 2004). Research on collocational frameworks and grammar patterns is mainly descriptive and takes place in applied contexts, but Hunston & Francis (2000) argue very explicitly for a usage-based theory of grammar on the basis of their descriptions (note that the definition quoted above is strongly reminiscent of the way constructions were later defined in construction grammar.

As an example of a collocational framework, consider [a __ of], one of the patterns that Renouf & Sinclair (1991) use to introduce their construct. Renouf and Sinclair use an early version of the Birmingham Collection of English Text, which is no longer accessible. To enable us to look at the methodological issues raised by collocational frameworks more closely, I have therefore replicated their study using the BNC. As far as one can tell from the data presented in Renouf & Sinclair (1991), they simply extracted all words occurring in the framework, without paying attention to part-of-speech tagging, so I did the same; it is unclear whether their query was case-sensitive (I used a case insensitive query for the BNC data). Table 8.1 shows the data from Renouf & Sinclair (1991) and from the BNC. As you can see, the results are roughly comparable, but differ noticeably in some details – two different corpora will never give you quite the same result even with general patterns like the one under investigation. Renouf and Sinclair first present the twenty items occurring most frequently in the collocational framework, shown in the columns labeled (a). These are, roughly speaking, the words most typical for the collocational framework: when we encounter the framework (in a corpus or in real life), these are the words that are most probable to fill the slot between a and of. Renouf and Sinclair then point out that the frequency of these items in the collocational framework does not correspond to their frequency in the corpus as a whole, where, for example, man is the most frequent of their twenty words, and lot is only the ninth-most frequent. The “promotion of lot to the top of the list” in the framework [a of], they argue, shows that it is its “tightest collocation”. As discussed in Chapter 7, association measures are the best way to assess the difference in frequency of an item under a specific condition (here, the presence of the collocational framework) from its general frequency and I will present the strongest collocates as determined by the G statistic below. Renouf and Sinclair choose a different strategy: for each item, they calculate the percentage of all occurrences of that item within the collocational framework. The results are shown in the columns labeled (b) (for example, number occurrs in the BNC a total of 48 806 times, so the 13 799 times that it occurs in the pattern [a of] account for 28.21 percent of its occurrences). As you can see, the order changes slightly, but the basic result appears to remain the same. Broadly speaking, the most strongly associated words in both corpora and by either measure tend to be related to quantities (e.g. lot, number,couple), part-whole relations (e.g. piece, member, group, part), or types (e.g. sort or variety). This kind of semantic coherence is presented by Renouf and Sinclair as evidence that collocational frameworks are relevant units of language.

Reference

Biber, Douglas, Stig Johansson, Geoffrey N. Leech, Susan Conrad & Edward Finegan. 1999.

Longman grammar of spoken and written English. Harlow: Longman. Biber, Douglas & Randy Reppen. 2015.

The Cambridge handbook of English corpus linguistics (Cambridge Handbooks in Language and Linguistics). Cambridge & New York: Cambridge University Press. Bock, J.Kathryn. 1986.

Syntactic persistence in language production. Cognitive Psychology 18(3). 355–387. DOI:10.1016/0010-0285(86)90004-6 Bondi, Marina & Mike Scott (eds.). 2010.

Keyness in texts (Studies in corpus linguistics 41). Amsterdam & Philadelphia: John Benjamins. Borkin, Ann. 1973.

To be and not to be. In Claudia Corum, T. Cedrik Smith-Stark & Ann Weiser (eds.), Papers from the ninth regional meeting of the Chicago Linguistics Society, vol. 9, 44–56. Chicago: Chicago Linguistic Society. Butler, Christopher. 1985.

Statistics in linguistics. Oxford & New York: Blackwell. Caldas-Coulthard, Carmen Rosa. 1993.

From discourse analysis to Critical Discourse Analysis: The differential re-presentation of women and men speaking in written news. In John McHardy Sinclair, Michael Hoey & Gwyneth Fox (eds.), Techniques of description: Spoken and written discourse, 196–208. London: Routledge.