Constructions in Wonderland

Exploring the functionality of constructions through N-grams

Jensen, Kim Ebensgaard; Shibuya, Yoshikata

Publication date:
2014

Citation for published version (APA):
Jensen, K. E., & Shibuya, Y. (2014). *Constructions in Wonderland: Exploring the functionality of constructions through N-grams*. Paper presented at Sprogets Funktionalitet, Aalborg, Denmark.

Constructions in Wonderland: Exploring the functionality of constructions through N-grams

(manuscript – talk given at Sprogets Funktionalitet, Aalborg University, 10 December 2014)

Yoshikata Shibuya, Kyoto University of Foreign Studies
Kim Ebensgaard Jensen, Aalborg Universitet

1. Introduction

We are interested in exploring the extent to which the N-gram text-mining technique can help researchers identify constructions and their functional contributions to the discourses in which they appear. We will address the following three questions. Is there a way to automatically or semi-automatically identify constructions in texts, discourses, and corpora? Could N-gram analysis and N-gram based network analysis be ways to do that? Can (semi-)automatic identification of constructions help us learn about the functional contributions of constructions in discourse, and, if yes, what can we learn. To answer these questions, we have analyzed two literary classics and four political speeches.

In terms of theory, our study is positioned within the framework of construction grammar (e.g. Goldberg 1995; and Croft 2001), which is now considered a central cognitive theory of language. There are also elements of cognitive stylistics (e.g. Stockwell 2002) and cognitive discourse analysis (e.g. Hart 2013), as we address the contributions of constructions as cognitively entrenched functional units to literary texts and political discourse. Methodologically, our main framework is computational corpus linguistics – in particular, text-mining (e.g. Miner et al. 2012) – and, in tandem with the elements of cognitive poetics and cognitive discourse analysis, we also include elements of corpus stylistics (e.g. Semino & Short 2004) and corpus-aided discourse studies (e.g. Baker 2012).

It should be mentioned that our study is purely exploratory, and that our main goal is not necessarily to show that stylistic analysis and discourse analysis need construction grammar; our purpose is to enhance construction grammar empirically and to show that construction grammar needs stylistic analysis and discourse analysis in the sense that, if we want a powerful usage-based construction grammar, we need to address discursive aspects of constructions and look at how constructions are used in discourse from as many angles as possible.

2. Constructions

In construction grammar, a construction is a functional unit that pairs form and semantic and/or discourse-pragmatic function (Goldberg 1995, 2006; Croft 2001, 2005; Hilpert 2014). Here are some examples of constructions that have been documented in constructionist research (we are using the Langackerian (1987) [form]/[function]-representation style):

- [S V IO DO]/[TRANSFER OF POSSESSION] (Goldberg 1995)
- [X BE *so* Y *that* Z]/[SCALAR CAUSATION] (Bergen & Binsted 2004)
- [*you don't want me to* V]/[THREATENING SPEECH ACT] (Martínez 2013)
- [*to begin with*]/[INTRODUCTION OF LIST OF ITEMS] (Lipka & Schmid 1994)
- [PRO$_{acc}$ CL$_{inf}$ (or NP)]/[DISBELIEF TOWARDS PROPOSITION] (Lambrecht 1990)

A central notion in construction grammar is that constructions may be atomic/simple, complex, or something in-between and form a lexicon-syntax continuum (e.g. Goldberg 1995, Croft 2001). In other words, morphemes, lexemes and syntactic structures as well as even structures that exceed the boundaries of the sentence – as long as they can be associated with some conventional function – are constructions. On a related note, constructions may be schematic, substantive (fixed), or something in-between (Fillmore et al. 1988). Rather than being compartmentalized, then, language competence is an inventory of constructions (aka. the construct-i-con) of varying degrees of abstraction which are instantiated in language use (e.g. Goldberg 1995). In most contemporary incarnations of construction grammar, the construct-i-con is usage-based (e.g. Croft 2001).

Constructions are subject to general human cognitive processes and principles, such that language is not a separate, autonomous cognitive faculty. That is, language is not autonomous, and consequently, construction grammar is part of the overall endeavor of cognitive linguistics.

Our main premise is that constructions, if they are functional units (pairings of form and meaning/function), then they logically must contribute to discourse as part of a speaker's linguistic repertoire. Here are two examples that relate to the types of data investigated in this study. Writers of fiction may use constructions in descriptions of actions and happenings. For instance, a writer might use a specific argument structure construction, topicalization construction, or voice construction to perspectivize or construe an event. Writers of fiction may also use constructions in characterizations (Culpeper 2009) and mind styles (Fowler 1977) by having characters use certain constructions in their dialog and narrative, or by using certain constructions in the descriptions of characters or of their actions. Constructions may be used in in setting up the text-world and specifying temporal relations in the narrative, and as ingredients in more general stylistic strategies of foregrounding, deviation, parallelism etc. (e.g. Short & Leech 2007). In political speeches, speaker may use constructions as framing devices, to shape ideologically based representations, to organize topics or issues, and as part of rhetorical strategies.

3. Methodological framework

We use the following data:

- *Alice's Adventures in Wonderland* by Lewis Carroll (1865)
- *Adventures of Huckleberry Finn* by Mark Twain (1884)
- Inaugural speeches by US Presidents.

The two novels were both obtained via the Gutenberg Project. Both novels are known for their quirky styles and we thought they would prove an interesting testing ground for our methodological framework. The presidential speeches were accessed via Bartleby; we chose those because we wanted to apply our analysis to non-fictional discourse as well.

We apply four methods from text-mining – namely, wordclouds, N-grams, and network analysis. We also made use of concordances when we need to explore the immediate contexts of the N-grams in question. The concordances were generated in *AntConc*.

3.1 Wordclouds

A wordcloud is a graphical representations of the lexical texture of a text, based on frequencies. Through very simple means, it indicates frequency through font size, such that frequent words are big and infrequent words are small. That way, a wordcloud is a visual rendering of a frequency list, with the exception that it does not give us any precise information on frequencies. We used the R package 'wordcloud' to generate wordclouds of *Alice's Adventures in Wonderland* and *Adventures of Huckleberry Finn*.

3.2 N-grams

An N-gram (e.g. Stubbs 2009) is a string of words that co-occur frequently in a data set (such as a corpus or a text). In other words, it is a combination of words. N-grams are specified in accordance with the number of words in the string in question (N = number): Monogram (1-gram) = one word, bigram (2-gram) = two words, trigram (3-gram) = three words, four-gram (4-gram) = four words, five-gram (5-gram) = five words etc.

N-grams are digitally identified and the process, roughly, works like this: the analyst asks a computer to find strings of N words, and it returns a list of N-grams ranked in terms of frequency. As an example, one might be interested in finding all 4-grams in the collective body of Shakespeare's plays:

- Find all instances of **word + word + word + word** combinations in the collective body of

Shakespeare's plays.

- Calculate frequencies of **word + word + word + word** combinations in the collective body of Shakespeare's plays
- List the **word + word + word + word** combinations in terms of frequency in the collective body of Shakespeare's plays.

The result of such a search is seen in the PPT presentation (generated in AntConc from the *Tokenized Shakespeare Corpus*).

Our assumption is that is may be possible to extrapolate constructions from recurring patterns across N-grams in a text. The advantage to such an exploratory approach is that we are likely to identify constructions that we would have otherwise not even thought of. We are well aware, however, that it is not possible to identify all types of constructions using this method. For instance, abstract constructions, like argument structure constructions would require that we identify PoS-grams or even what we might theoretically calls SF-grams (syntactic function grams) in a text marked for PoS or syntactic function. Moreover, there is a slim chance of identifying constructions that allow for long-distance discontinuity or which are of a macrosentential nature. We would argue, however, that N-grams can prove effective in, not only finding constructions where we did not even consider looking, but also in finding constructions with substantive elements and perhaps also idiomatic ones.

We used the R package 'tau' and AntConc's N-gram function to generate N-grams. We applied N-gram analysis to *Alice's Adventures in Wonderland*, *Adventures of Huckleberry Finn*, and four Presidential speeches.

3.3 Network analysis

Like N-gram analysis, network analysis in text-mining is based on words that occur next to each other and their frequencies of cooccurrence. Network analysis, however, does not generate a list of word combinations. It treats each word type (as opposed to word token) as a node in a network and sets up network relations between the words based on their frequencies of cooccurrence. In that sense it is an advanced type of N-gram analysis.

We used the R package 'igraph' to perform network analyses of *Alice's Adventures in Wonderland* and *Adventures of Huckleberry Finn*.

4. Wordclouds

The two word clouds can be seen in the PPT presentation. While they do not show us much information in terms of frequencies and interconnections between words let alone in terms of constructions larger than words and morphemes, they are informative to some extent.

For instance, 'said' is the most frequent word in *Alice's Adventures in Wonderland*, suggesting that there is a lot of dialog in that novel. However, one interesting aspect of *Alice's Adventures in Wonderland* is that, with the exception of a handful of frequently used words, its lexical texture consists of many words that are used infrequently in the text. Thus, the style of that novel is lexically varied. In contrast, *Adventures of Huckleberry Finn* consists of fewer words which are used more frequently. Thus, this novel is written in a lexically simplistic style. This makes sense, seeing that it is a first person narrative and that the narrator, Huckleberry Finn, is a child. The lexical simplicity is an ingredient in Huckleberry Finn's childlike and simple mind style.

Wordclouds are visually attractive and to some extent informative, but they do not, as mentioned above, provide much information on frequencies.

5. N-grams

As mentioned above, we believe that it is possible to infer or perhaps induce certain types of constructions from N-grams.

5.1 Simple N-gram analysis

In the PPT presentation, you can see top 20 lists of 2-, 3-, and 4-grams in *Alice's Adventures in Wonderland*. One particularly striking phenomenon in these lists is that N-grams with the 'said the' pattern recurs as in 'said the', 'said the King', 'said the Mock Turtle' and 'said the March Hare'. This could be indicative of an underlying construction at play. Indeed, a concordance of the 2-gram 'said the' (see the PPT presentation) shows that 'said' is often preceded by direct speech and followed by a definite NP with unique reference, whose head is a character designation. This suggests that what we could call the topicalizing reporting clause construction is put to use, serving to organize the dialog in the narrative.

A list of 2-, 3-, 4-, and 5-grams in *Adventures of Huckleberry Finn* (see the PPT presentation) reveals some interesting negation patterns revolving around 'warn't' and 'ain't'. Seeing that Huckleberry Finn is a Southerner, and the narrative takes place along the Mississippi River (and that Twain himself states that the novel features emulations of dialects spoken at the time), this are, of course, indicators of Huckleberry Finn's mind style. A particularly interesting aspect of the 'warn't'-based N-grams is that 'warn't' appears in N-grams after 'it' and 'there'. This raises the question whether 'it warn't' and 'there warn't' behave differently in the novel. The concordances of 'it warn't no' and 'there warn't no', as seen in the PPT presentation, suggest that 'there warn't no' is more productive in terms of what appears after 'no' then 'it warn't no', as 'use' appears often after 'no' in the latter. Given that the two patterns seem to have different productivities, it seems that, in the mind style that Huckleberry Finn has been endowed with, they are treated as instances of different constructions that behave discursively differently.

The Huckleberry Finn lists also feature N-gram patterns that indicate a 'by and by' structure and 'and then' as a 2-gram. While very simplistic in form, both figure as central event-ordering devices in the narrative. The concordance of the 2-gram 'and then' (see the PPT presentation) indicates that it is used as a construction with a event cross-relating function, while the concordance of 'by and by' suggests that its function is to indicate that an even takes place after a short period of time. Thus, both serve to order the temporal structure of the narrative, and their simplistic natures also contribute to the childlike mind style of Huckleberry Finn.

5.2 Comparative N-gram analysis

We have seen that it does appear to be possible to extrapolate constructions from N-grams. Simple N-gram analysis can help us identify and address constructions and their functionality in one text or discourse, as it identifies frequent combinations of words in the text or discourse in question. What simple N-gram analysis does not tell us is whether or not those frequent combination of words can also be found in other text. To obtain a list of N-grams that really delineate a given text (so that we can identify what constructions are characteristically associated with the text), a comparative analysis can be useful.

We made a comparative 2-gram analysis of the two novels and normalized the frequencies to per 10000 words. The comparison was based on Fisher's Exact test. The comparative N-gram analysis (see PPT presentation) confirmed that the topicalizing reporting clause constructions delineates *Alice's Adventures in Wonderland* relative to *Adventures of Huckleberry Finn*, while the latter is characterized by negation constructions and the two even-ordering constructions as well as stereotyping expressions such as 'says I', 'I says', and 'I reckon', all of which are treated by Mark Twain as constructions in the dialect and mind style of Huckleberry Finn.

We applied the comparative 2-gram analysis to four inaugural presidential speeches (George Bush, Bill Clinton, George W. Bush, and Barack Obama as well. This time we normalized the 2-gram frequencies to per 1000 words; we applied fisher's Exact test again. The comparative analysis generated some interesting results (see the PPT presentation). For instance, the 'we will' 2-gram is much more prevalent in George W. Bush's speech than in the other presidents' speeches. This 2-gram is indicative of the [WILL V_{inf}]/[CERTAINTY/FUTURE]-construction. Bush uses the construction as a parallelism as a means to predicate actions that he plans to carry out during his presidency, and it may indeed have the function of establishing a persona characterized by willpower and determination.

6. Network analysis

It seems that comparative N-gram analysis can help us find N-grams that delineate texts or discourses. However, there is a problem. Shorter N-grams are embedded in longer N-grams (2-grams can be found inside some of the 3-grams and 4-grams). Consequently, our N-gram lists contain some redundancy, and there is also redundancy across list. Moreover, our comparative N-gram analyses focused on 2-grams, but texts contain longer strings of words (3-grams, 4-grams, etc) and also shorter N-grams (1-grams). From the perspective of construction grammar, we should also talk about those longer/shorter N-grams, like we did in our isolated N-gram analyses. We can use the methods we have used 2-grams with other N-grams, but is there any simpler way to find both short and long N-grams? That is, can we find 1-grams, 2-grams, 3-grams, 4-grams, etc. all at one time? Is there any way to provide descriptively a more efficient analysis on frequently co-occurring combinations of words (constructions)? Our suggestion here is to use network analysis based on 2-grams. Other N-grams emerge in the network, as a network includes all N-grams.

Our network analysis of *Alice's Adventures in Wonderland* captures N-grams that can be abstracted into the definite NP construction and the topicalized reporting clause construction (see the PPT presentation). Our network analysis of *Adventures of Huckleberry Finn* captures the N-grams discussed above as well as 'and so' which is indicative (as seen in concordance excerpt in PPT presentation) of a causative event cross-relating construction. A closer look at the networks is likely to reveal several more N-grams and underlying constructions.

At the end of the day, short and long N-grams can be found automatically. A list of N-grams can be used to compute *p.*values with Fisher’s exact test, for instance, and thus be part of a comparative N-gram analysis. Then, what’s good about using network analysis? Firstly, it allows us to capture several N-gram types simultaneously without too much redundancy. Secondly, the real advantage that network analysis offers is not just its visual effects, but in fact it tells you a lot about the internal functionality of the network. For instance, it is possible to compute the centrality of nodes in a network. This method provides estimates regarding the relationship between a network and the functionality of nodes in it. Centrality and node functionality translates into centrality and salience in the text in question, allowing us to address more closely the functional contributions of lexical constructions to texts as well as providing us with hints at which patterns to look at in the text, some of which might be indicative of constructions. For instance, we did a betweenness centrality measure (Dehmer & Basak 2012: 70-71) of each of the two novels and found that 'I' was central in *Adventures of Huckleberry Finn*, which is not surprising since it is a first person narrative with numerous self-references (see PPT presentation). Interestingly, the 'n't' negator is also quite central which supports that negation constructions are treated as salient features of Huckleberry Finn's mind style. This would not be possible in a simple N-gram analysis.

7. Concluding remarks

Can N-grams be used as means of identifying constructions? Yes, partially. We have seen that N-grams can guide us in terms of where to look for patterns that can be indicative of constructions and apply sophisticated means of analysis. For instance, 'it warn't no' and 'there warn't no' could be subjected to collostructional analysis as a means of addressing their differences in productivity in *Adventures of Huckleberry Finn* to see whether they are indeed treated as two different constructions in the novel. Moreover, as we saw, it is possible to do a comparative N-gram analysis, which, again can guide us in terms of which patterns to investigate; comparative N-gram analysis can also give us insights into the extent to which certain patterns delineate texts.

Is network analysis useful in the identification of constructions? Yes, partially. It as the same advantages as simple N-gram analysis, but it reduces redundancy in N-grams by capturing all N-gram types at the same time within the same representation. Also, it allows us to measure centrality.

However, neither N-gram analysis nor network analysis will enable us to identify all possible types of constructions and they are likely to be most useful in relation to constructions with substantive elements. It is possible that network analysis is able to capture a broader range of constructions than N-gram analysis is, but that is something that needs to be tested in the future.

What about functionality? Neither N-grams nor N-gram based networks tell us much about functionality, as they show us purely formal relations. However, they guide us in terms of connections between words that are salient in a given text and may be indicative of constructional as functional units. We can then look at the discursive behavior of such N-grams and extrapolate constructions and their functionality in the text or discourse (and, depending on the corpus, in general).

We would argue that this 'test drive' of the text-mining methods of N-gram analysis and network analysis shows their potential, and that they are worth exploring further in relation to the identification of constructions.

References

Baker, P. (2012). Acceptable bias? Using corpus linguistic methods withcritical discourse analysis. *Discourse Studies*, 9(3): 247-256.

Bergen, B. & K. Binsted (2004). The cognitive linguistics of scalar humor. In M. Achard & S. Kemmer (eds.). *Language, Culture, and Mind. Stanford*, CA: CSLI. 79-91.

Croft, W. A. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Culpeper, J. (2009). In G. Brône & J. Vandaele (eds). *Cognitive Poetics: Goals, Gains and Gaps*. Berlin: Mouton de Gruyter.

Dehmer, M. & S. C. Basak (2012). *Statistical and Machine Learning Approaches for Network Analysis*. Chichester: Wiley-Blackwell.

Fillmore, C, P. Kay and M. C. O'Connor (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64: 501–38.

Fowler, R. (1977). *Linguistics and the Novel*. London: Methuen.

Hart, C. (2013). Constructing contexts through grammar: Cognitive models and conceptualisation in British newspaper reports on political protests. In J. Flowerdew (ed.), *Discourse in Context*. London: Bloomsbury. 159-183.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.

Lambrect, K. (1990). "What, me worry?": Mad Magazine sentences revisited. *BLS*, 16: 215-228.

Langacker, R. (1987). *Foundations of Cognitive Grammar. Vol. 1: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Lipka, L. & H.-J. Schmid (1994). *To begin with*: Degrees of idiomaticity, textual functions and pragmatic exploitations of a fixed expression. *ZAA*, 42: 6-15

Martínez, N. D. C. (2013). *Illocutionary Construtions in English: Cognitive Motivation and Linguistic Realization*. Bern: Peter Lang.

Miner, G., J. Elder, T. Hill, R. Nisbet, D. Delen & A. Fast (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Elsevier Academic Press.

Semino, E. & M. Short (2004). *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.

Short M. & G. Leech (2007). *A Linguistic Introduction to English Fictional Prose* (2nd ed.). Harlow: Pearson Longman.

Stockwell, P. (2002). *Cognitive Poetics*. London: Routledge.

Stubbs, M. (2009). Technology and phraseology. In U. Römer & R. Schulze (eds.), *Exploring the Lexis-Grammar Interface*. Amsterdam: John Benjamins. 15-31.