Globe: A Journal of Language, Culture and Communication, 2: 23-54 (2015)

Mining for constructions in texts using N-gram and network analysis

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

Abstract: In constructionist theory, constructions are functional entities that pair form and conventionalized semantic and/or discourse-pragmatic function. One of the main tasks of the construction grammarian is thus to identify and document constructions. Seeing that it is unlikely that this can be done satisfactorily via introspection, there is a need for different ways of identifying constructions in language use. In this paper, we will explore the extent to which the N-gram information retrieval technique – which has seen use in phraseological analysis, discourse analysis, register characterization, and corpus stylistics – is applicable in the identification of constructions and their functionality in discourse. An N-gram is a constellation of a specified number (N = number) of entities that frequently (co)occur in a data population. In this paper we will report on an exploratory study in which we apply N-gram analysis to Lewis Carroll's novel *Alice's Adventures in Wonderland* and Mark Twain's novel *The Adventures of Huckleberry Finn* and extrapolate a number of likely constructional phenomena from recurring N-gram patterns in the two texts. In addition to simple N-gram analysis, the following will be applied: comparative N-gram analysis which draws on a slightly adjusted distinctive collexeme analysis, hierarchical agglomerative cluster analysis, and N-gram-based network analysis. The latter is explored as a way to capture different N-gram types, and underlying constructions, in one representation. The main premise is that, if constructions are functional units, then configurations of words that tend to recur together in discourse are likely to have some sort of function that speakers utilize in discourse. Writers of fiction, for instance, may use constructions in characterizations, mind-styles, text-world construction and specification of narrative temporality. In this paper, our special interest lies in the relationship between constructions and the discourse of fiction. As the study reported in this article is exploratory, it serves just as much to test the methods mentioned above as to analyze and characterize the two novels.

Keywords: Constructional functionality, literary language, N-gram analysis, network analysis.

1. Introduction

The construction as a pairing of form and conventionalized function is central in constructionist approaches to language (e.g. Fillmore et al. 1988; Goldberg 1995, 2006; Croft 2001), as it is held to be the basic unit of language. Consequently, constructionist language descriptions do not address combinatorial rules that generate grammatical sentences. On the contrary, construction grammarians seek to describe the constructions of the language in question, addressing their forms, their functions, their symbolic structures, their contextual patterns, and their relations to general human cognition. Thus, an important task is the discovery and documentation of constructions. Language is so diverse and complex that most constructions cannot be documented via introspection, and more empirical/objective and more efficient analysis is called for. There are many ways to do this, but in any case it is required that the analyst be able to identify and quantify recurring patterns and their potential functions in discourse. Text-mining, in a nutshell, covers a set of analytical techniques that can derive patterns from structured and unstructured textual datasets (e.g. Miner et al. 2012). In this article, we suggest that a possible way to identify recurring patterns in discourse that are reflective of constructions could be to apply text-mining techniques.

More specifically, we will use N-gram analysis, which has already seen use in phraseology (Stubbs 2007, 2009) in the discovery of fixed expressions. In this particular study, we apply N-gram analysis to the two classic novels *Alice's Adventures in Wonderland* by Lewis Carroll and *The Adventures of Huckleberry Finn* by Mark Twain to see whether N-gram analysis is useful in identifying constructions in the two texts. Expanding on N-gram analysis, we will further explore the usability of comparative N-gram analyses as well as the more advanced technique of network

analysis, in which inter-word relations are derived automatically from texts and represented as networks. Note that the research reported in this study is first and foremost exploratory, and the purpose has been just as much to experiment with the above-mentioned text-mining techniques in the name of construction grammar as it has been to analyze and describe the two novels. A further aim is to investigate the functionality of the constructions that emerge from these patterns and thus address how interlocutors, in this case writers of fiction, use constructions to convey the discursive contents, in this case narratives and fictional worlds in which they take place.

This article is organized as follows. In section 2, we provide a brief and very basic account of the fundamental principles of construction grammar as such, focusing on the functionality of constructions. In section 3, the data and methodological framework are accounted for. In section 4, we present our N-gram analyses and account for a number of patterns that display constructional behavior; this section also presents our comparative N-gram analysis. Section 5 presents our network analysis and also briefly discusses node centrality (an advanced analytical method within network analysis) in connection with linguistic data.

2. Constructions and functionality

The theoretical framework of the present study is that of construction grammar (e.g. Fillmore et al. 1988; Goldberg 1995, 2006; Croft 2001; Hilpert 2014) in which the construction is a pairing of form and conventionalized meaning and may range in complexity from atomic to complex structures. That is, constructions are held to form a lexicon-syntax continuum. Since the primary unit of grammar is the construction, language competence is an inventory of constructions (sometimes called the *construct-i-con*) of varying degrees of abstraction which are instantiated in language use. In most contemporary incarnations of construction grammar, the construct-i-con is usage-based and thus allows for redundancy in the constructional network if usage-patterns indicate that this is the case (see Barsalou 1992 who suggests from a psycholinguistic perspective that evidence tends to favor redundant representations over nonredundancy). As Croft (2005: 274) points out, a construction may be defined generally as "an entrenched routine …that is generally used in the speech community ... and involves a pairing of form and meaning". In other words, a construction is a functional unit of language within the code adopted by the community in question. Constructional meaning, it should be pointed out, covers conceptual semantics and discourse-functional properties as well as pragmatic properties (Croft 2001: 18). For the sake of illustration, here are some constructions from English:

- [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)
- [V (DO) *until* ADJ]/[INSTRUCTION IN PREPARATION OF INGREDIENTS IN COOKING SCENARIOS] (Jensen 2014)

The first two constructions have primarily semantic functions. The first one is, of course, the ditransitive construction, which serves to express scenarios of TRANSFER OF POSSESSION, while the second sets up a causal relation between a POINT on a SCALE expressed by [*so* ADJ] and a RESULTING SITUATION expressed by the following *that*-clause. Interestingly, the causal relation is implicit, making it an example of conventional implicature (Grice 1975: 44-45). The third construction is primarily a speech act construction, whose function is that of a THREATENING SPEECH ACT. Thus, this construction is functionally primarily pragmatic. The fourth construction serves to INTRODUCE A LIST OF ITEMS IN A TEXT , making it a primarily discourse-functional construction, whose function is of a

meta-discursive, text-structuring nature. The last construction functionally combines semantics and pragmatics. Semantically, it describes the PREPARATION of an INGREDIENTS in a COOKING SCENARIO. Pragmatically, it serves as an instruction in how to prepare said INGREDIENTS, as this construction most frequently appears in recipes.

Constructions are thus symbolic structures, combining form and semantic and/or discourse-pragmatic function, which are entrenched cognitively in speakers. Constructions may be schematic, substantive (fixed), or something in-between (Fillmore et al. 1988). For instance *to begin with* is fully substantive, while the ditransitive construction is fully schematic. The SCALAR CAUSATION and INGREDIENT PREPARATION constructions contain both schematic and substantive elements. Constructions are subject to general human cognitive processes and principles, such that language is not a separate, autonomous cognitive faculty; thus, construction grammar is part of the overall endeavor of cognitive linguistics (e.g. Croft & Cruse 2004; Evans & Green 2006).

Our main premise is that, if constructions are functional units, then configurations of words that tend to recur together in discourse are likely to have some sort of function that speakers utilize in discourse. Moreover, if constructions are functional units (pairings of form and function), then they must contribute to discourse as part of a speaker's linguistic repertoire. Writers of fiction, for example, 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 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 this paper, our special interest lies in the relationship between constructions and the discourse of fiction, and that is why we have chosen as a test ground two literary texts.

3. Data and method

In this exploratory study, we primarily make use of N-gram analysis and network analysis. Our data consist of the following classic novels, both of which were downloaded in text-format from Project Gutenberg's text archives:

- Mark Twain: *The Adventures of Huckleberry Finn* (published 1884/1885), henceforth *HF.*
- Lewis Carroll: *Alice's Adventures in Wonderland* (published 1865), henceforth *AW.*

After removing the Gutenberg metadata and generally cleaning up the files, the two texts were subjected to two word counts each:

Table 1: Word counts

Text	Word count	Tokenized word count
AW	26,679	27,330
HF	111,002	117,299

In the first word count, units between spaces were treated as words. Thus, in this count, *I don't know* consists of three words. In the second word count, the texts were tokenized such that contracted forms were split up into their constituents. In this count, *I don't know* then consists of four words – namely *I, do*, *n't*, and *know.* Note that, following the way they are represented in *R*,

which we used for our statistical analyses, contracted forms, when treated as N-grams, such as *don't*, *didn't*, and *ain't* will be represented as *don t*, *ain t*, and *ain t* in the remainder of this paper; when treated as constructions, they appear in their standard contracted forms. At this point, some might protest that such texts, because they are literary texts and thus not as such representative of more regular discourse, are not suitable if one wants to convincingly show that a given method of analysis works for identification of recurring patterns in discourse. While this criticism is warranted if the purpose is indeed to convince people *that* the methodology works, the purpose of the present study is not to sell the method, as it were, but to test it and see *if* it works and *how* it works when applied to quirky literary discourse. Granted, the method should be tested on a variety of different data, and, elsewhere (Jensen & Shibuya in prep a; in prep b), we do apply it to more regular language data. However, here, our purpose is to experiment with the method in applying it to literary texts known for their stylistic deviance from regular discourse. Here, it should be reiterated that we are applying the method in addressing the functional contributions of constructions to texts in which they appear; this is as relevant to deviant literary texts as it is to regular discourse. Moreover, while perhaps not interesting to those who want to investigate regular language or other everyday discourses which are less deviant, the two texts we have chosen to explore here are stylistically very interesting exactly because they deviate from everyday language, the artistically motivated foregrounding strategy of deviation being a central topic in literary stylistics (Simpson 2004: 50-51; Short & Leech 2007: 39).

Automatic N-gram analysis was applied to the cleaned-up files in conjunction with concordancing as a way to not just identify potential constructions formally, but also to address their discursive behaviors in the texts and thus their functionalities in the two novels.

3.1. N-grams

N-grams are contiguous strings of items, most often words, that appear in a stretch of discourse. Retrieval of N-grams is an automated text-mining technique, which is essentially a quite simple but efficient one. At its core, N-gram analysis consists in retrieving strings of a specified number of words and then quantifying the strings and ranking them in descending order in terms of frequency. For instance, if we are interested in finding all four-word strings in a dataset, this is the procedure:

- Find all instances of **word + word + word + word** combinations in the dataset.
- Calculate frequencies of **word + word + word + word** combinations in the dataset.
- List the **word + word + word + word** combinations in terms of frequency in the dataset.

N-grams are specified by the number of words in the string in question. Thus, the type of N-gram referred to above is called a fourgram. N-grams of two words are called bigrams, while N-grams of three words are called trigrams, and N-grams of five words are called fivegrams and so forth. N-gram analysis and its variants have seen numerous uses in linguistics. In computational linguistics, for instance, it is often used in the generation of linear probabilistic predictive language models, while in corpus-based language and discourse studies, it has been used to identify various characteristics of texts and discourses. Vasquez (2014: 25-56) identifies a number of word strings in the discourse of consumer reviews, using N-gram analysis, and links these up with trends of expression of positive evaluation. Gries & Mukherjee (2010) and Gries et al. (2011) have applied N-gram analysis in the characterization of registers and language varieties. Corpus stylisticians have also made use of N-gram analysis to address aspects of literary language. Notably, Mahlberg (2007a, 2007b) has made use of N-gram analysis to identify word clusters in the writing of Dickens. More generally, Stubbs (2007, 2009) uses N-gram analysis to identify frequent phraseology, or multi-word expressions.

Automatic N-gram analysis is particularly attractive, because it can return clusters of words that the human analyst may not even have considered. Consequently, it allows the analyst to address linguistic phenomena which might have been missed in manual or introspective analysis. In this exploratory study, we are going to apply N-gram analysis in a manner similar to Mahlberg (2007a, 2007b) and Stubbs (2007, 2009). However, we will take it a step further, in the perspective of construction grammar, and use N-grams to identify constructions through a process of bottom-up abstraction in which we identify constructional schemata that emerge from recurring patterns in our N-gram analyses and then address their functionalities from contextualized patterns of usage in the two novels. We will also apply a comparative N-gram analysis, in which the significance of N-grams in the two texts is established.

We will rely on dispersion measures to help us determine which N-grams, and potentially underlying constructions, are spread so evenly throughout the narrative that they could be considered characteristic of the novel. Seeing that, according to Lyne (1985), Juilland's *D* measure is one of the most reliable dispersion measures, we use *D*-scores to measure dispersion in the present study. A *D*-score is a number between 0 and 1: the closer to 1 it is, the more even the dispersion. The starting point of this measure is the division of the text or corpus in question into equally sized parts. *AW* was divided into five equally sized parts and *HF* into ten equally sized parts (this is because *HF* is larger than *AW*). On the basis of this division of the texts into equally sized parts, a *D*-score was calculated, as described in Oakes (1998: 190), for each N-gram discussed in the following sections. These dispersion measures will be supplemented with dispersion plots (e.g. Jockers 2014: 29-31) to visualize the distribution of N-grams throughout the novels. While numeric dispersion measures are more objective than visual representations of dispersion, it may be easier for readers to relate to visual representations. It should be born in mind, of course, that dispersion plots only offer an approximate visual representation and not a totally precise one. That is why we include both numeric and visual representations in this article. The reason why we include dispersion measures in our analysis is that an N-gram may have a high frequency in a text, but if all its tokens occur in the same place in the text, then the N-gram is not likely to be typical of the narrative, but only serves a special purpose in the portion of the narrative where it appears. While N-grams that appear in high-density groups are undeniably also functionally interesting, our focus here is on N-grams, and underlying constructions, that contribute functionally to the text generally.

3.2. Networks

Network analysis can be used as a text-mining technique that sets up data points and relations between them, based on the frequency of co-occurrence of the words in the text. Thus, it is essentially an advanced type of N-gram analysis, based on bigrams, which identifies types of word co-occurrences and quantifies the number of tokens of each co-occurrence type. This way, nodes are set up based on words as types, and relations are set up between the nodes based on frequency of co-occurrence. When this is done for every word type, the result is a network of nodes and relations between them. While N-gram analysis presents co-occurring words in ranked lists, network analysis represents them graphically as a network. Network analysis has the advantage over N-gram analysis that it allows one to capture all N-gram types within the same network representation, whereas, in N-gram analysis, the analyst operates across several N-gram lists. Network analysis has been applied in the study of verb-argument constructions by Brook O'Donnell et al. (ms); Römer et al. (fc), Gries & Ellis (2015), and Ellis et al. (2013).

4. N-gram analysis

N-grams allow us to address relations of co-occurrence among words, and, via this, to observe strings of words that may form phraseological units. If we can identify functional patterns of such units (using concordances), then chances are that they may be constructions in the sense of

Goldberg (2006: 5):

> 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.

4.1. N-grams in AW

We generated three N-gram lists from *AW* – namely, a list of bigrams, a list of trigrams, and a list of fourgrams. Below are the top 20s of each type of N-gram:

Table 2: Top 20 bigrams in *AW*

Rank	Bigram	Frequency
1	said the	210
2	of the	133
3	said alice	116
4	in a	97
5	and the	82
6	in the	80
7	it was	76
8	the queen	72
9	to the	69
10	the king	62
11	as she	61
12	don t	61
13	at the	60
14	she had	60
15	a little	59
16	i m	59
17	it s	57
18	mock turtle	56
19	and she	55
20	she was	55

Table 3: Top 20 trigrams in *AW*

Rank	Trigram	Frequency
1	the mock turtle	53
2	i don t	31
3	the march hare	30
4	said the king	29
5	said the hatter	21
6	the white rabbit	21
7	said the mock	19
8	said to herself	19
9	said the caterpillar	18
10	said the gryphon	17
11	she said to	17
12	she went on	17
13	as she could	16
14	i can t	15
15	one of the	15
16	said the duchess	15
17	out of the	14
18	said the cat	14
19	it said the	12
20	minute or two	12

Table 4: Top 20 fourgrams in *AW*

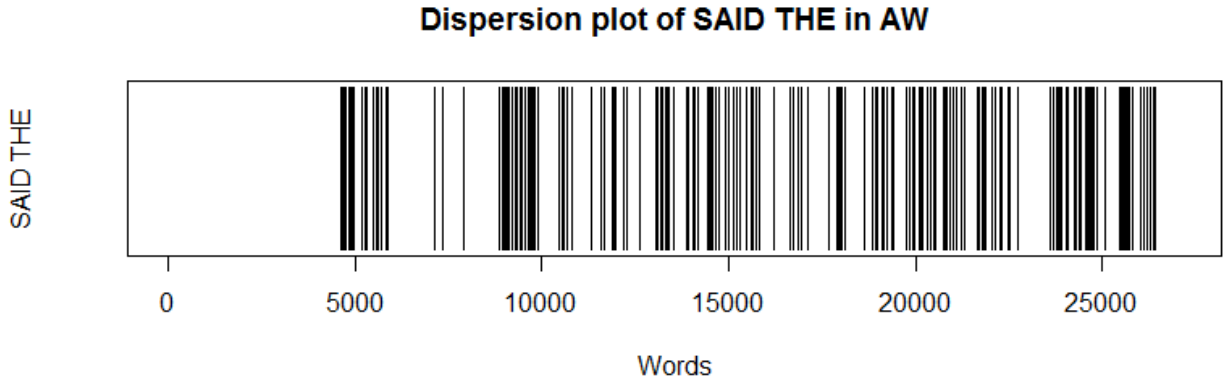
Rank	Fourgram	Frequency
1	said the mock turtle	19
2	she said to herself	16
3	a minute or two	11
4	you won t you	10
5	said the march hare	8
6	will you won t	8
7	i don t know	7
8	said alice in a	7
9	as well as she	6
10	in a great hurry	6
11	in a tone of	6
12	moral of that is	6
13	t you will you	6
14	the moral of that	6
15	well as she could	6
16	won t you will	6
17	and the moral of	5
18	as she said this	5
19	i beg your pardon	5
20	i ve got to	5

Note that in Table 2, *said the* appears in first position, while similar strings appear in Table 3 in the form of *said the king* (ranking 4), *said the hatter* (ranking 5), *said the mock* (ranking 7), *said the caterpillar* (ranking 9), *said the gryphon* (ranking 10), *said the duchess* (ranking 16), and *said the cat* (ranking 18). Likewise, in Table 4, we find *said the mock turtle* (ranking 1) and *said the march hare* (ranking 5). A *D*-score of 0.8103 indicates that the bigram is quite evenly distributed throughout the text. This is reflected in the dispersion plot in Figure 1. This plot shows the distribution of the bigram *said the* throughout *AW* in which each occurrence of the bigram is represented by a black vertical line. The horizontal dimension entitled 'Words' represents the entire novel in a linear fashion; this dimension is based on the location of every word in the novel. Thick vertical lines, then, simply represent multiple instances of *said the* which appear very near each other in the novel. The dispersion plot shows that, apart from in the beginning of the novel,[1] the

1 More specifically, the bigram does not appear in the two first chapters. This may be related to the flow of narrative information throughout the novel. The first *said the X* appears in words number 4526-4528 in the sentence *'Ahem!' said the Mouse with an important air, 'are you all ready?'.* In the first two chapters, however, *said Alice* can be found a few times. As the story goes by, more and more characters are introduced and subsequently referred to in the narrative and hence the X-slot of *said the X* simply becomes more available to those new characters in the story. Moreover, in the first two chapters, Alice does not interact with many characters, but, from the third chapter and onwards, the inventory of characters is considerably expanded, and Alice enters into the type of dialog seen in (6), which is quite characteristic of the novel.

bigram is fairly evenly distributed over the novel:

Figure 1: Distribution of the bigram *said the* in *AW*



A concordance of *said the* was generated and indeed shows a recurring pattern, with only a handful of instances of the bigram deviating from it. The pattern is illustrated by the examples below:

(1) 'Found *what*?' said the Duck.

(2) 'Then you shouldn't talk,' said the Hatter.

(3) 'Hold your tongue!' said the Queen, turning purple.

(4) '*'tis the voice of the sluggard*,' said the Gryphon.

(5) 'There's more evidence to come yet, please your Majesty,' said the White Rabbit, jumping up in a great hurry; 'this paper has just been picked up.'

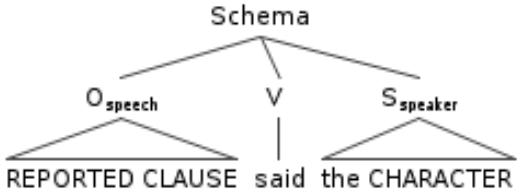
In all examples above, *said the* is preceded by direct speech and followed by a specification of one of the characters in the narrative, allowing us to induce the following schematic generalization:

REPORTED CLAUSE *said the* CHARACTER SPECIFICATION

The function of this particular schema is quite easy to pinpoint. Structurally, it is a reporting clause, and functionally the schema thus serves to assign dialog in the narrative to the character who utters it. More specifically, the character specification is an instance of the definite noun phrase construction, whose function as a presupposition trigger (Huang 2007: 90) is to indicate to the reader that the character is considered GIVEN INFORMATION. At this point, we can thus characterize the schema as a direct speech reporting construction, which we will call the inverted topicalizing reporting clause construction (or the ITRC-construction for short). To anyone who has read literature in English, it should not be a big surprise to find this type of construction in a literary narrative, as novels and short stories typically contain dialog and strategies of assigning dialog to characters within the narrative.[2] If we take a look at the syntactic structure of this particular schema, we see that it involves subject-verb inversion and object fronting:

2 See Short & Leech (2007: 255-270) for a discussion of direct speech and indirect speech in fiction.

Figure 2: Syntactic structure of the schema



In their treatment of inverted direct speech, Short & Leech (2007: 267-268) write that inversion plays a role in connection with direct speech without informing us of the nature of that role. However, later in their discussion of rhetoric and narrative style, they state that "[a]s speakers, we are rarely able to plan the whole of our utterance in advance, so we tend to begin with the thing which is uppermost in our mind, the thing which, from our point of view, is the focal nub of the message" (Short & Leech 2007: 186). This relates to information structure. Bache & Davidsen-Nielsen (1997: 113-114) describe the general principles of information structure in English, reminding us that "[n]ormally the speaker will proceed from what he assumes to be known (the *topic* or *theme*) to what he assumes to be new (the *comment* or *rheme*)" [italics in original] (see also Short & Leech 2007: 170-172). Thus, the schema in Figure 2 involves fronting, or topicalization, of the reported speech and focalization of the character who utters the speech, resulting in a reversal of GIVEN and NEW INFORMATION, in that the character, by virtue of the definite construction, is presented as GIVEN INFORMATION. This suggests that the function of the schema is not only that of assigning dialog to characters, but also topicalize, or highlight, the spoken dialog as particularly salient information. To see whether that is indeed how the schema is used in the narrative, we need to have a look at its discursive behavior. Here is an example:

(6) At this moment the King, who had been for some time busily writing in his note-book, cackled out 'Silence!' and read out from his book, 'Rule Forty-two. *all persons more than a mile high to leave the court*.'
Everybody looked at Alice.
'*I'm* not a mile high,' said Alice.
'You are,' said the King.
'Nearly two miles high,' added the Queen.

Whenever the schema is used, it appears initially in a line with no text preceding it. Contrast the following with the instance of the schema in the sequence in (6):

(7) At this moment the King, who had been for some time busily writing in his note-book, cackled out 'Silence!

(8) The King turned pale, and shut his note-book hastily. 'Consider your verdict,' he said to the jury, in a low, trembling voice.[3]

The schema seems to be used as a type of cohesive device, in that, in fronting speech, it creates a link between the fronted speech and preceding speech, thus highlighting the fronted speech as a reaction to the previous speech. In contrast, (8) breaks with the preceding sequence, as the King addresses the jury rather than responding to Alice. This functional pattern characterizes most of the instances of *said the* in the novel: 90% establish a cohesive link to previous preceding dialog, and

3 There is no subject-verb inversion here so *he* in *he said* has not been focalized.

97% of them appear in the beginning of a paragraph in the novel. While *the X said* does occur in the novel, it only has a frequency of 30, suggesting that, when *said* is used as the reporting verb, *said the X* is the primary dialog-ordering device in the narrative.

From the narrative style emerges a recurring pairing of form and function which serves the purpose of organizing dialog. Its recurrence is such that we can argue that it is used as a construction (recall Goldberg's (2006: 5) definition; see the beginning of Section 4 above). We can now propose a constructional structure in which the form is tied in with a specific functional content:

Figure 3: Form-function structure of *said the X*

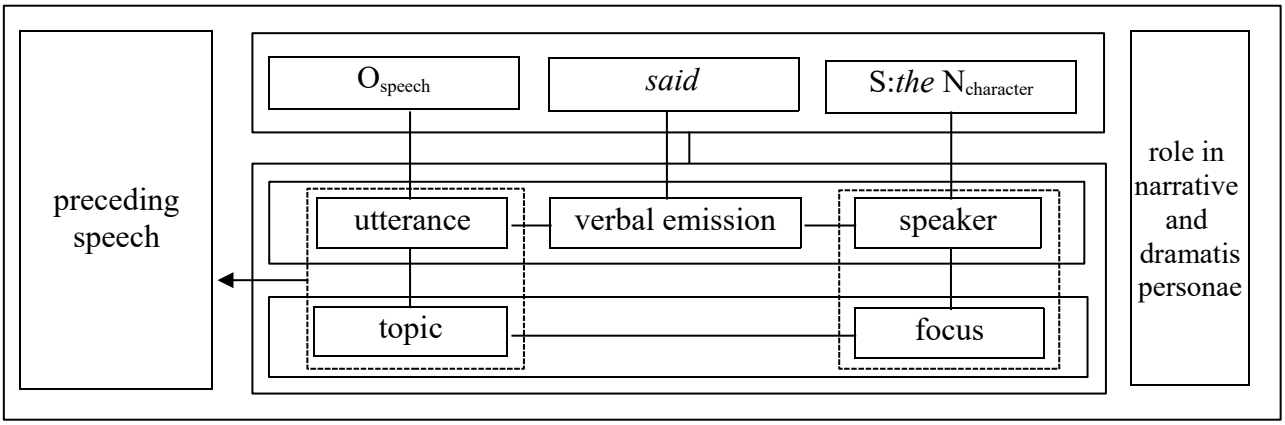


Figure 3 illustrates the construction, using a Croft-style box diagram (Croft 2001). The outer box indicates that this is one construction. The rectangular top box in the middle indicates the form of the construction, and the three boxes within it (entitled 'O_{speech}', '*said*', and 'S:*the* $N_{character}$' respectively) indicate its formal constituents. The big rectangular box underneath represents the functional structure of the construction. It contains two boxes. The one that contains the boxes entitled 'utterance', 'verbal emission', and 'speaker' indicates the semantic structure and essentially represents a semantic frame in the sense of Fillmore (1982), capturing a generalized cognitive model of verbal communication. The links between 'O_{speech}' and 'utterance', '*said*' and 'verbal emission', and 'S:*the* $N_{character}$' and 'speaker' are the symbolic links between the formal elements and semantic components of the construction. The lower box in the function structure represents the information-structural nature of the construction. 'Utterance' links up with 'topic' to indicate topicalization of 'O_{speech}', and 'speaker' links up with 'focus' to indicate focalization of 'S:*the* $N_{character}$'. The punctuated boxes further emphasize that we are dealing with information-structural units. The leftmost box, entitled 'Preceding speech' captures the fact that the construction serves to create a cohesive relation between the reported speech in the construction and preceding speech in the narrative. The arrow from the 'utterance'-'topic' information-structural unit indicates that it is the fronting of 'O_{speech}' which sets up the cohesive relation. At this point, the reader might be puzzled as to why what is essentially mere discursive content is included into the construction. The answer lies in construction grammarians' inclusion of knowledge of contexts in which a construction typically occurs in speakers' language competence (e.g. Fillmore 1988: 36l). Thus, the preceding speech is to be considered a property of the construction. The rightmost box that is entitled 'role in narrative and dramatis personae' is intended to capture such properties of the construction.

Interestingly, if you look at (6) again, we see the following cases of direct speech, which follow a very similar pattern:

(9) '*I'm* not a mile high,' said Alice.

(10) 'Nearly two miles high,' added the Queen.

In (9), we find the proper noun *Alice* in place of the definite noun phrase. In terms of reference, *Alice* has unique reference which is arguably more closely related to definite reference than to indefinite reference.[4] In (10), we find *added* as the reporting verb in place of *said*. This could suggest that we are dealing with an even more abstract ITRC-construction in which the verb is not lexically fixed and in which the position of the speaker-subject position may be realized by either a definite noun phrase or a proper noun. If we operate with this level of abstraction, the dispersion of the construction generates a D-score of 0.8728 and looks like this in a dispersion plot:

Figure 4: Distribution of the ITRC-construction:

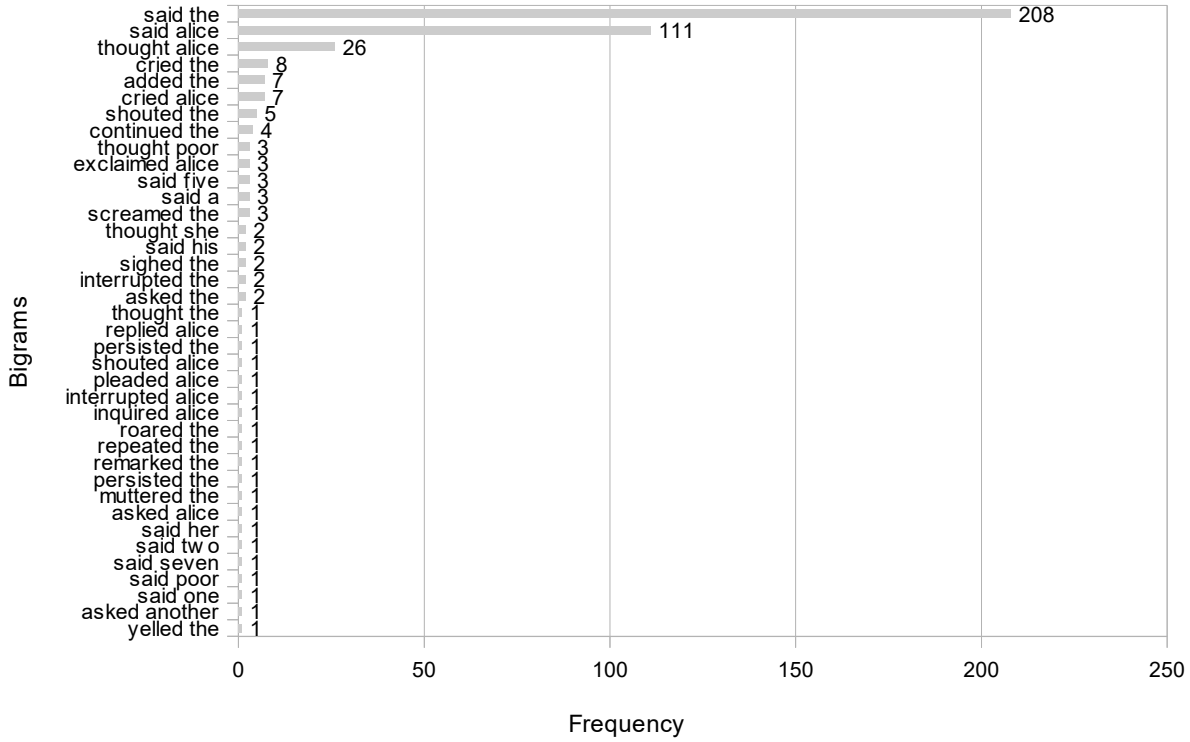
In the dispersion plot above all instances of reporting verbs (including the cognitive reporting verb *think*) followed by speaker-subjects (including definite and indefinite noun phrases and proper nouns) are abstracted into a generalized schema whose occurrences throughout the novel are then tracked.

As Gries & Ellis (2015) point out, constructions are Zipfian in nature (Zipf 1949) – Zipf's law being described by Ferrer i Gancho & Solé (2003: 788) as "a hallmark of human language" and as "required by symbolic systems" (Ferrer i Cancho & Solé 2003: 791) – and it appears to invariably be the case that some instantiations of the construction are more frequent and salient than others.

As the graph in Figure 5 shows, *said the* is the most frequent bigram of all bigrams in the novel that reflect the function. We see that the ITRC-construction displays Zipfian behavior in *AW* and suggests that *said the X* is the most salient realization of the construction. One possible explanation could simply be that *say* is a basic level term for communicative verbal emission in English, while, for instance, *yell*, *mutter*, *persist*, *roar*, and *ask* predicate more specific manner of verbal emission. This suggests that Lewis Carroll specifically draws on *said the* when there is no narrative need for specifying the type of verbal emission involved in characters' utterances, thus using it as a specialized constructional resource in his organization of dialog.

4 *Said* followed by an indefinite noun phrase that refers to a speaker only appears three times in the novel.

Figure 5: Bigrams reflective of the ITRC-construction in *AW*



4.2. N-grams in HF

Having explored N-grams in *AW* and seen how that enabled us to extrapolate a construction and address its functionality as a dialog-ordering strategy, let us turn to *HF*.

Tables 5, 6, 7, and 8 provides are lists of the 30 most frequent bi-, tri-, four-, and fivegrams in the novel. A few interesting patterns occur across the lists above such for instance, *warn t no* (ranking 5 in Table 6) as reflected in *there warn t no* (ranking 1 in Table 7), *it warn t no* (ranking 3 in Table 7), *it warn t no use* (ranking 1 in Table 8), *but it warn t no* (ranking 4 in Table 8), and *there warn t no* (ranking 11 in Table 8), *see it warn t no* (ranking 20 in Table 8), and *but there warn t no* (ranking 28 in Table 8). The pattern is also partially reflected in *warn t* (ranking 8 in Table 5), *it warn t* (ranking 7 in Table 6), *but it warn t* (ranking 12 in Table 7), and *i see it warn t* (ranking 10 in Table 8). Another pattern is *by and by* (ranking 5 in Table 6), which is reflected in *and by and by* (ranking 4 in Table 7), *by and by he* (ranking 22 in Table 7), and *but by and by* (ranking 29 in Table 7). Ranking at 11 in Table 5 we find *and then*, which is also reflected in *and then he* (ranking 25 in Table 6).

In the following sections, we will address the N-grams mentioned above. First we will look at *warn t no*, addressing the possible constructional statuses of *there warn t no* and *it warn t no*. Afterwards, we will turn to *by and by* and *and then*, addressing the functions they have in the narrative.

Table 5: Top 30 bigrams in *HF*

Rank	Bigram	Frequency
1	in the	434
2	it was	370
3	didn t	347
4	don t	340
5	of the	335
6	and the	317
7	ain t	298
8	warn t	293
9	i was	290
10	and i	288
11	and then	250
12	to the	236
13	on the	227
14	it s	226
15	was a	223
16	couldn t	219
17	but i	206
18	he was	204
19	out of	201
20	so i	176
21	wouldn t	176
22	and he	172
23	it and	165
24	i says	163
25	up and	160
26	in a	157
27	t no	153
28	going to	146
29	that s	142
30	got to	141

Table 6: Top 30 trigrams in *HF*

Rank	Trigram	Frequency
1	i didn t	119
2	i couldn t	105
3	i don t	87
4	by and by	85
5	warn t no	71
6	there warn t	70
7	it warn t	69
8	ain t no	67
9	out of the	61
10	it ain t	54
11	was going to	53
12	it was a	50
13	there was a	50
14	all the time	48
15	don t know	48
16	there ain t	48
17	don t you	46
18	the old man	45
19	i warn t	44
20	i wouldn t	43
21	i hain t	40
22	didn t know	38
23	he didn t	38
24	said it was	38
25	and then he	37
26	it s a	35
27	a couple of	34
28	down the river	34
29	i ain t	34
30	it wouldn t	34

Table 7: Top 30 fourgrams in *HF*

Rank	Fourgram	Frequency
1	there warn t no	32
2	i don t know	31
3	it warn t no	30
4	and by and by	24
5	there ain t no	24
6	but i couldn t	22
7	the middle of the	22
8	but i didn t	21
9	i says to myself	21
10	didn t want to	20
11	warn t no use	20
12	but it warn t	19
13	king and the duke	16
14	the king and the	16
15	i didn t want	15
16	it ain t no	15
17	a kind of a	14
18	i didn t know	14
19	in the middle of	14
20	ain t got no	13
21	all the time and	13
22	by and by he	12
23	i couldn t see	12
24	i don t want	12
25	a quarter of a	11
26	ain t going to	11
27	all of a sudden	11
28	and there warn t	11
29	but by and by	11
30	don t want to	11

Table 8: Top 30 fivegrams in *HF*

Rank	Fivegram	Frequency
1	it warn t no use	19
2	the king and the duke	16
3	i didn t want to	11
4	but it warn t no	10
5	ain t a going to	9
6	in the middle of the	9
7	the middle of the river	9
8	a quarter of a mile	8
9	don t make no difference	8
10	i see it warn t	7
11	and there warn t no	6
12	don t know nothing about	6
13	i couldn t help it	6
14	i couldn t see no	6
15	i don t want to	6
16	i never see such a	6
17	it ain t no use	6
18	it don t make no	6
19	made up my mind i	6
20	see it warn t no	6
21	the head of the island	6
22	about a quarter of a	5
23	and one thing or another	5
24	as quick as i could	5
25	at the head of the	5
26	but i couldn t see	5
27	but i didn t see	5
28	but there warn t no	5
29	didn t want to go	5
30	down the lightning rod and	5

4.2.1. *It warn't no* vs. *there warn't no*

Warn t no seems to occur in two constructions: *there warn't no* and *it warn't no* (with the respective frequencies of 32 and 30). This gives rise to the question whether the two have similar or different functions, which, in turns, leads us to the question whether or not they are treated in the narrative as two different constructions. Before going into detail, let us have a look at the distributions of *there warn t no* and *it warn t no* in *HF*. *There warn t no* has a *D*-score of 0.7927 while *it warn t no* has a *D*-score of 0.8208. Thus, both are somewhat evenly dispersed throughout HF, as is also seen in the dispersion plots in Figures 6 and 7:

Figure 6: Distribution of *there warn t no* in *HF*

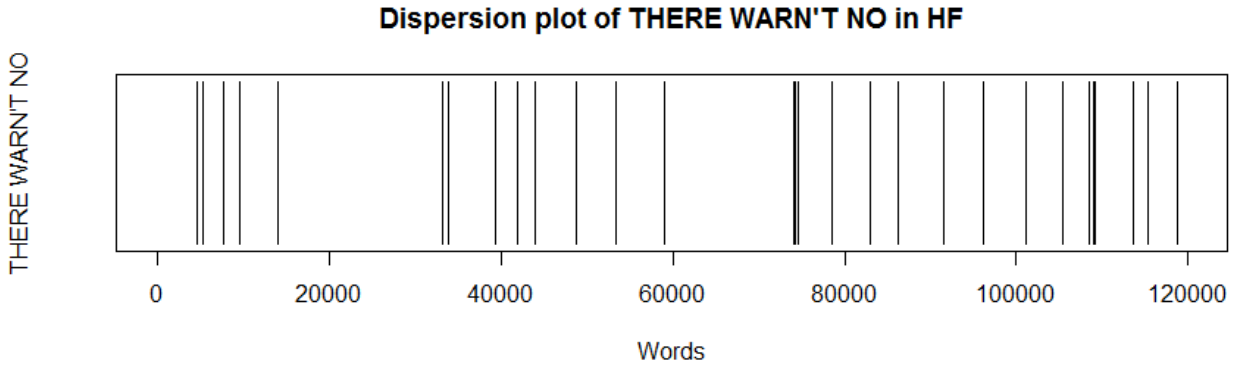
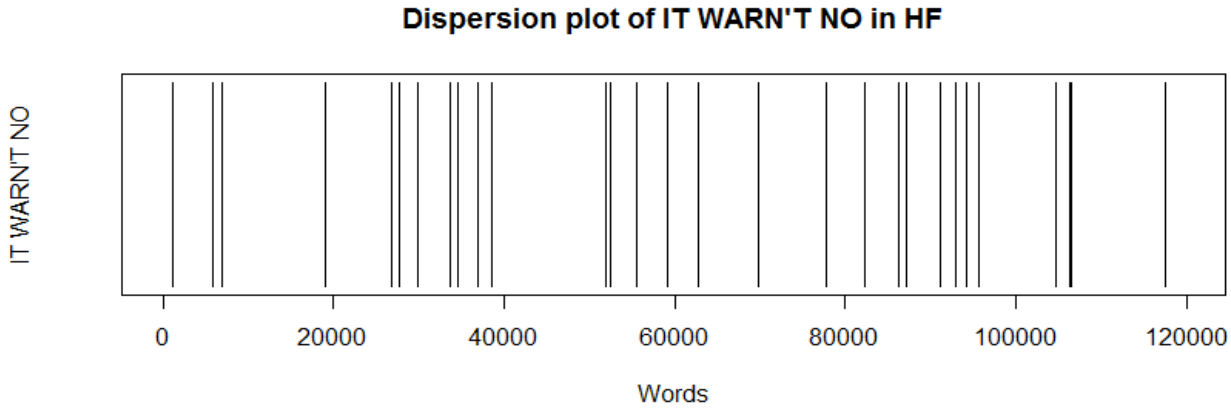


Figure 7: Distribution of *it warn t no* in *HF*



While not extremely frequent, the two expressions nonetheless are more or less evenly distributed over the novel. Thus, we can assume that both, despite their low frequencies, are nonetheless stylistic features of the text and consequently worth investigating further. A concordance was generated for each expression. In Tables 9 and 10, we see excerpts of ten lines from each concordance. It is worth noting that *there warn't no* seems much more productive than *it warn't no*. The following graph, which lists all the lexemes that occur after *no* in both expressions and quantifies their distribution over the two seems to confirm this as seen in Figure 8. As the graph in Figure 8 shows, *it warn't no* occurs with few nouns, with *use* being by far the most frequent. In contrast, *there warn't no* appears with a broader range of lexemes, none of which is particularly frequent. This could suggest that there is a particular affinity between *it warn't no* and *use*.

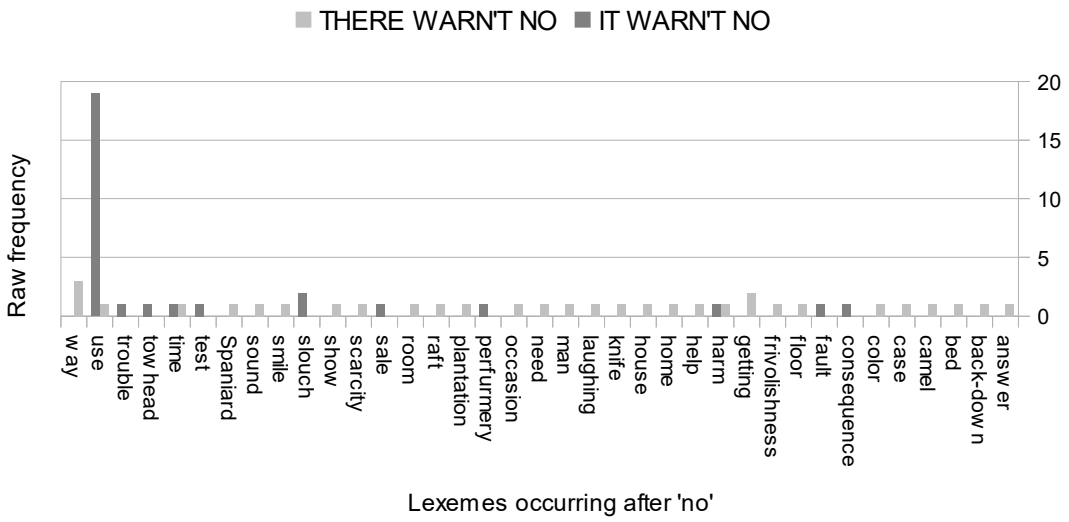
Table 9: Ten lines from the *there warn't no* concordance

to the illinois shore where it was woody and	there warn't no	houses but an old log hut
in the bottom of it with the saw, for	there warn't no	knives and forks on the place
. if he got a notion in his head once,	there warn't no	getting it out again. he was
half a minute it seemed to me and then	there warn't no	raft in sight; you couldn't
't take the raft up the stream, of course.	there warn't no	way but to wait for dark,
we talked about what we better do, and found	there warn't no	way but just to go along
knob to turn, the same as houses in town.	there warn't no	bed in the parlor, nor a
a mahogany cane with a silver head to it.	there warn't no	frivolishness about him, not a bit
jim to get away from the swamp. we said	there warn't no	home like a raft, after all.
and the duke had their legs sprawled around so	there warn't no	show for me; so i laid
he crowd looked mighty sober; nobody stirred, and	there warn't no	more laughing. boggs rode off

Table 10: Ten lines from the *it warn't no* concordance

't run jim off from his rightful owner; but	it warn't no	use, conscience up and says, every
very well i had done wrong, and i see	it warn't no	use for me to try to
duke, and tried to comfort _him_. but he said	it warn't no	use, nothing but to be dead
as it would keep peace in the family; and	it warn't no	use to tell jim, so i
ever put in in the missionarying line. he said	it warn't no	use talking, heathens don't amount
could lock him up and get him sober; but	it warn't no	use -- up the street he would
something muffled up under his coat and i see	it warn't no	perfumery, neither, not by a long
the poor girl's feelings, and all that. but	it warn't no	use; he stormed right along, and
just like the way it was with the niggers	it warn't no	sale, and the niggers will be
't give in _then_! indeed he wouldn't. said	it warn't no	fair test. said his brother william
d that in the woods, whooping and screeching; but	it warn't no	use -- old jim was gone. then

Figure 8: Lexemes occurring with both expressions



Now, the analysis in Figure 8 is based on the raw frequencies of the lexemes occurring after *no*, and hence not the statistically most sophisticated way to determine the differences in productivity, but more sophisticated collostructional analyses will confirm this. Below is the result of a simple collexeme analysis of the lexemes in *it warn't no* in *HF*:[5]

Table 11: Lexemes in *it warn't no*

Rank	Lexeme	Collostruction strength
1	use	256.5564
2	slouch	24.1934
3	test	16.5595
4	perfumery	16.5595
5	consequence	13.7874
6	sale	11.5574
7	fault	9.3610
8	towhead	8.7332
9	harm	8.6283
10	trouble	5.8229
11	time	3.1681

5 Simple collexeme analysis is a type of collostructional analysis (e.g. Stefanowitsch & Gries 2003, 2005; Gries & Stefanowitsch 2004) which statistically measures the degree of attraction of a lexeme to a construction. Its mechanics are as follows. For each lexeme, the following frequencies are specified and entered into a 2x2 table: the frequency of the cooccurrence of item and construction, the frequency of the item in all other constructions, the frequency of the construction with all other constructions, and the frequency of all other items in all other constructions. These are through a Fisher-Yates exact test, which may or may not be log transformed. This results in a *p*-value which is a number that indicates the collostruction strength, or degree of lexeme-construction attraction. The higher the number, the stronger the attraction. The output is a list of lexemes, ranked in accordance with their collostruction strengths. In this study, we used log transformed *p*-values, which allow for more fine-grained distinctions among collostruction strengths. We used Gries (2007) to perform our collostructional analyses. Readers who want to know more about the mechanics, application, and theoretical background of simple collexeme analysis are referred to Stefanowitsch & Gries (2003).

In conjunction with Figure 8 above, Table 11 clearly shows that *it warn't no* attracts *use* very strongly with a collostruction strength of 256.5564 against *slouch*'s collostruction strength of 24.1934. With such a difference between the most and second-most attracted items in a construction, we are not unjustified in concluding that *it warn't no use* has a special status as entrenched in the mind of the narrating character in the novel. Thus, in Mark Twain's writing in *HF*, *it warn't no* is treated as a construction primarily associated with *use* in the vernacular spoken by Huckleberry Finn and thus a trait of his mind-style (Fowler 1977) and other characters in the novel. For the sake of comparison, here is the result of a simple collexeme analysis of *there warn't no*:

Table 12: Lexemes in *there warn't no*

Rank	Lexeme	Collostruction strength	Rank	Lexeme	Collostruction strength	Rank	Lexeme	Collostruction strength
1	getting	16.5794	11	plantation	10.6898	21	show	6.3897
2	back-down	16.4283	12	knife	9.9314	22	use	6.3551
3	frivolishness	16.4283	13	need	9.7316	23	room	6.1910
4	occasion	16.4283	14	laughing	9.3837	24	home	6.0989
5	scarcity	16.4283	15	case	8.8304	25	bed	5.7437
6	way	15.8352	16	harm	8.4978	26	house	5.1669
7	spaniard	13.6562	17	floor	7.6750	27	raft	4.9578
8	camel	12.6103	18	answer	7.5449	28	man	3.2084
9	color	11.9312	19	sound	7.0470	29	time	3.0480
10	smile	11.9312	20	help	6.4607			

Compared to Table 10 we are dealing with much smaller collostruction strengths here, and the differences between them are much smaller (some of them are even identical). Finally, in Table 13 are the results of a distinctive collexeme analysis (Gries & Stefanowitsch 2004), which measures a lexeme's constructional-preference out of a set of two or more constructions.[6] The table confirms that there is a special affinity between *use* and *it warn't no*. It also confirms that more lexemes prefer *there warn't no* than *it warn't n*o which seems to confirm the differences in productivity among the constructions.

This difference in productivity indicates that the two expressions are used as two different constructions in the narrative style of the novel. It is well known that, in *HF*, Mark Twain aimed at emulating the vernaculars spoken in the Mississippi Valley in the early nineteenth century. Indeed, in a prologue to the novel, Twain himself explains this:

> IN this book a number of dialects are used, to wit: the Missouri negro dialect; the extremest form of the backwoods Southwestern dialect; the ordinary "Pike County" dialect; and four modified varieties of this last. The shadings have not been done in a haphazard fashion, or by guesswork; but painstakingly, and with the trustworthy guidance and support of personal familiarity with these several forms of speech.
>
> I make this explanation for the reason that without it many readers would suppose that all these characters were trying to talk alike and not succeeding.

This is where we find the main functional contribution of *it warn't no* and *there warn't no* (in addition to them being *it*- and *there*-constructions).

6 As with simple collexeme analysis, distinctive collexeme analysis that compares two constructions makes use of Fisher-based *p* values for collostruction strengths (in multiple distinctive collexeme analysis, which compares three or more constructions, the statistical mechanics are different). The input frequencies here are: the frequency of the lexical item in construction A, the frequency of the lexical item in construction B, the frequency of all other lexical items in construction A, and the frequency of all other lexical items in construction B. Readers who want to know more about the mechanics, application, and theoretical background of distinctive collexeme analysis are referred to Gries & Stefanowitsch (2004).

Table 13: Patterns of preference among *it warn't no* and *there warn't no*

Lexeme	Preferred construction	Collostruction strength
answer	there warn't no	1.3382
back-down	there warn't no	1.3382
bed	there warn't no	1.3382
camel	there warn't no	1.3382
case	there warn't no	1.3382
color	there warn't no	1.3382
consequence	it warn't no	1.4694
fault	it warn't no	1.4694
floor	there warn't no	1.3382
frivolishness	there warn't no	1.3382
getting	there warn't no	2.7081
harm	it warn't no	0.0022
help	there warn't no	1.3382
home	there warn't no	1.3382
house	there warn't no	1.3382
knife	there warn't no	1.3382
laughing	there warn't no	1.3382
man	there warn't no	1.3382
need	there warn't no	1.3382
occasion	there warn't no	1.3382
perfumery	it warn't no	1.4694
plantation	there warn't no	1.3382
raft	there warn't no	1.3382
room	there warn't no	1.3382
sale	it warn't no	1.4694
scarcity	there warn't no	1.3382
show	there warn't no	1.3382
slouch	it warn't no	2.9749
smile	there warn't no	1.3382
sound	there warn't no	1.3382
spaniard	there warn't no	1.3382
test	it warn't no	1.4694
time	it warn't no	0.0022
towhead	it warn't no	1.4694
trouble	it warn't no	1.4694
use	it warn't no	29.6418
way	there warn't no	4.1113

In constructing, or reconstructing, the vernaculars in question – in particular that spoken by the narrator – Twain quite successfully, in the perspective of a quantitative linguist, manages to imitate in his novel how language is used, to the point of having his characters use constructions in a way that is very compatible with the discoveries about actual language use that construction grammarians, cognitive sociolinguists, usage-based linguists, corpus linguists and other empirically oriented linguists would make in the twentieth century. Twain not only has his characters speak in a way that imitates certain vernaculars. He has them use different constructions at a level of detail that

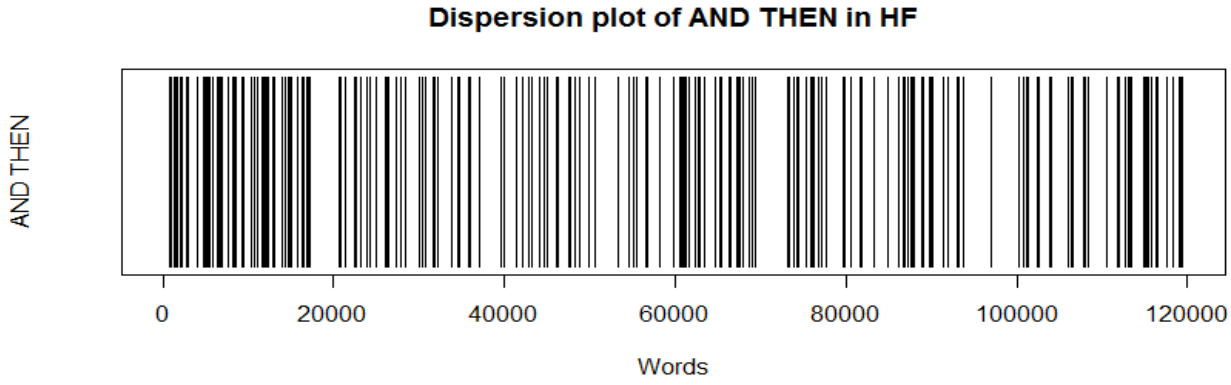
includes differences in productivity and schematicity.

4.2.2. Cross-event structuring constructions

In this section, we are going to have a look at *and then* and *by and by* as well as *and so*. The latter does not appear in the top 30 of bigrams in Table 5. However, ranking 34 with a frequency of 136, *and so* is still among the dominant bigrams in the text. Moreover, it is functionally related to the two other N-grams discussed in this section.

Starting with *and then*, a *D*-score of 0.9136 shows that it is very evenly distributed throughout the novel, which is echoed in the dispersion plot below:

Figure 9: Distribution of *and then*



A concordance was generated, yielding examples like these:

(11) He worked me middling hard for about an hour, and then the widow made her ease up.

(12) And if anybody that belonged to the band told the secrets, he must have his throat cut, and then have his carcass burnt up and the ashes scattered all around, and his name blotted off of the list with blood and never mentioned again by the gang, but have a curse put on it and be forgot forever.

(13) Next day he was drunk, and he went to Judge Thatcher's and bullyragged him, and tried to make him give up the money; but he couldn't, and then he swore he'd make the law force him.

(14) I got the things all up to the cabin, and then it was about dark.

(15) Then I took up the pig and held him to my breast with my jacket (so he couldn't drip) till I got a good piece below the house and then dumped him into the river.

In all examples above, *and then* serves to link one clause to another, and, thus, at a functional level, it creates a cross-event relation between the event or scenario expressed by the clause that precedes *and then* and that expressed by the clause that follows *and then*. Thus, it appears that the bigram *and then* reflects a simplistic cross-event-relating construction (Talmy 2000: 345) that we could call the *X and then Y*-construction. At this point, while he does not take a constructionist perspective, it is worth referring to Bache's (2014, 2015) work on the narrative function of *when* in English, as he demonstrates that, in its narrative function, *when* sets up a cross-event relation between two events,

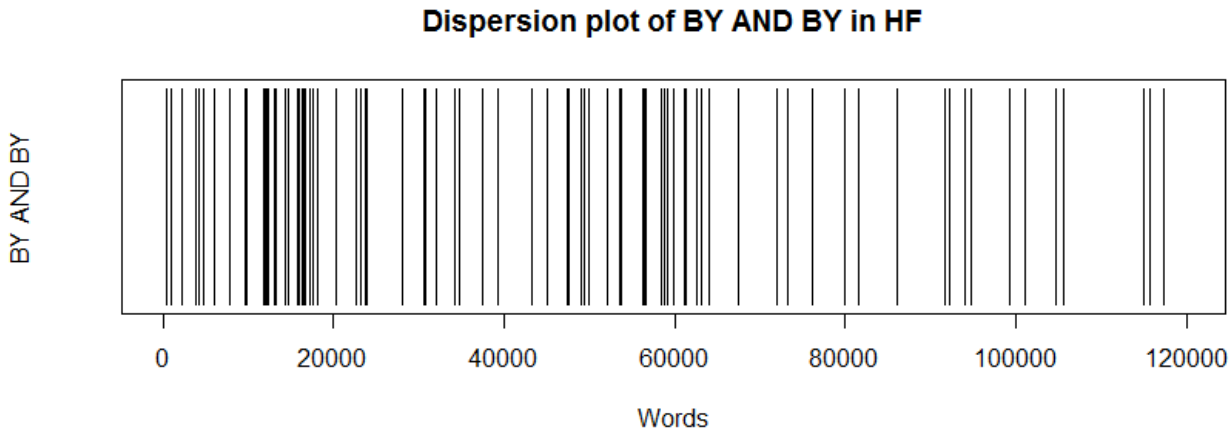
such that one proposition event serves as the background for the other event. The latter event is presented as an important new situation that takes place against the backdrop of the background event. Moreover, the relation between the two cross-related events is characterized by what Bache (2014) calls a narratively intense effect (see also Quirk et al. 1972: 745). This is illustrated by the example below:

(16) I was enjoying the music, *when* suddenly I felt sick.

Bache (2014, 2015) clearly shows that grammatical phenomena, such as *when* can have conventional cross-event relating narrative functions, which can be utilized by speakers and writers in constructing narratives. The cross-event relation in (11)-(15) is one of CHRONOLOGICAL SEQUENCING in which one event follows in a temporal sequence after the other. This applies to 90% of the occurrences of the bigram (the rest are not instances of the construction). Interestingly, Declerck (1997: 212) and Couper-Kuhlen (1989: 20) both suggest that *and then* and narrative *when* are interchangeable. Bache (2014, 2015) points out that this is not quite the case, as the former is mainly a sequentializing expression while the latter adds a sense of narrative intensity to the relation between the cross-related events. In terms of its contribution to the narrative style of the novel, then, the construction serves to organize the events that make up the narrative. Another important contribution by this construction is its simplicity. *HF* is a first person narrative told by the novel's titular character. Huckleberry Finn is a child, and the overall style of the narrative captures the simplicity with which a child would perceive the world. Thus, the simplistic nature of the *X and then Y*-construction not only contributes to the event-structure of the narrative, but also to the naive, simple, and childish mind-style of the character.[7]

Turning to *by and by*, a *D*-score of 0.7698 indicates that this trigram is somewhat evenly dispersed throughout the text. A dispersion plot shows that, while more frequent in the first half of the text, the expression does recur in the novel as such, arguably warranting the generation of a concordance:

Figure 10: Distribution of *by and by*



7 Interestingly, Bache (2015) writes that a group informants who are native speakers of present-day English prefer *and then* over narrative *when*, pointing out that the latter comes across bookish while the former is more suitable for spoken communication. The narrative intensity of the latter, Bache suggests, can be salvaged by adding paralinguistic and prosodic features to the utterance that contains the former. This seems to also have been that case at the time of Mark Twain, and thus it would make much sense for him to bestow Huckleberry Finn with a mind-style that emulates the language of speech rather than that of writing.

A pattern, captured by the following examples, emerges from the concordance in which it is quite clear that the trigram has an adverbial function:

(17) I judged the old man would turn up again by and by, though I wished he wouldn't.

(18) The widow she found out where I was by and by, and she sent a man over to try to get hold of me...

In both cases, *by and by* seems to have the function of a time adverbial. In (17), it seems to express the eventual happening of an event at some point in the future, and, in (18), it specifies that an event took place after a limited period of time.[8] While we are not going to go into any detail regarding which function is primary, we will note that both functions involve the specification of A PERIOD OF TIME. Indeed, one could argue that the future-indicating function logically draws on the notion of a period of time seeing that A PERIOD OF TIME is bound to separate the FUTURE POINT at which the EVENT will happen from the PRESENT MOMENT. Now, this temporal-adverbial function of *by and by* ends itself well for cross-event relation in the sense that it can allow language users to sequentialize events such that one is set up as following the other after a limited period of time. Indeed, we see this in *HF*, as seen in the following examples:

(19) At first I hated the school, but by and by I got so I could stand it.

(20) After supper she got out her book and learned me about Moses and the bulrushers, and I was in a sweat to find out all about him; but by and by she let it out that moses had been dead a considerable long time; so then I didn't care no more about him, because I don't take no stock in dead people.

(21) Being Tom Sawyer was easy and comfortable, and it stayed easy and comfortable till by and by I hear a steamboat coughing along down the river.

In examples (19) and (20), *by and by* appears in structures where clauses are coordinated, thus specifying the sequentiality and temporal relation between the events expressed by the clauses. In example (21), it sets up the same cross-event-relation between a main clause and a subclause. There is also a variant in the novel where an extrasentential cross-event relation is set up, as seen in the following example (in 75% of its occurrences in the novel *by and by* is used to express cross-event sequentiality, and 59% of those occurrences set up an extrasentential cross-event relation, while 41% set up an intrasentential one):

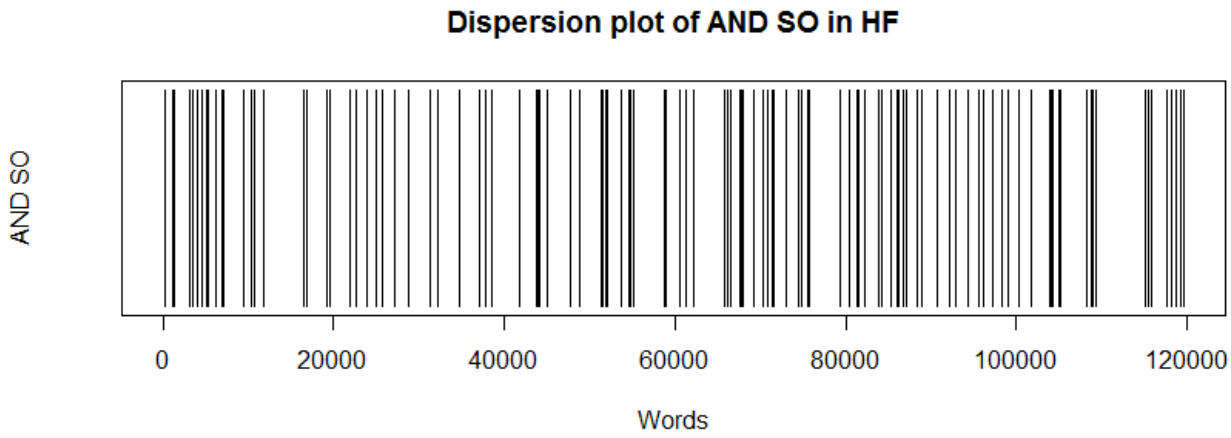
(22) I went to looking out sharp for a light, and sort of singing to myself. By and by one showed.

As with *and then*, this is a very simplistic way to structure events in a narrative which seems perfectly compatible with the simple and childish mind-style of Huckleberry Finn. The difference between *by and by* and *and then* is, of course, that the former expresses SEQUENTIALITY OF EVENTS and specifies that A LIMITED PERIOD OF TIME separates the events, while the latter expresses SEQUENTIALITY, but does not encode a temporal separation of the events.

Lastly, let us turn to *and so*, which has a *D*-score of 0.9247. It is thus very evenly distributed throughout *HF*, as reflected in the following dispersion plot:

8 These functions are corroborated by a number of dictionary entries for *by and by* which list these two meanings, such as *thefreedictionary.com*, *Merriam-Webster*, and *Cambridge Dictionaries Online*.

Figure 11: Distribution of *and so*



A concordance was generated, yielding examples like the ones in (23)-(25) below. In all three examples *and so* has a sequentializing cross-event relating function akin to that of *and then*: Around 83% of occurrences of *and so* are instances of the cross-event relating construction; the remaining portion comprises instances of *and so on* and *and so forth* as well as the combination of *and* and the proform *so*, as in *and so did his leg*.

(23) … but it was rough living in the house all the time, considering how dismal regular and decent the widow was in all her ways; and so when I couldn't stand it no longer I lit out.

(24) He said he would split open a raw Irish potato and stick the quarter in between and keep it there all night, and next morning you couldn't see no brass, and it wouldn't feel greasy no more, and so anybody in town would take it in a minute, let alone a hair-ball.

(25) We didn't have no dog, and so we had to chase him all over the country till we tired him out.

The reader will have noticed that, while SEQUENTIALITY seems to be a function of *and so* in the text, it has an additional cross-event relating function which is perhaps best described as a type of loose causality in which the event expressed by the clause after *so* follows as a consequence from that of the clause before *so*. This is perhaps clearest in (25) where the chasing of a person is presented as the consequents of the people chasing after him not having a dog to help them. Again, this is a quite simplistic way to express such causality, which suits the mind-style of Huckleberry Finn very well.

4.3. Comparative N-gram analysis

We have seen that it is possible to extrapolate constructions from N-grams and to address their functional contributions to the texts they appear in. Simple N-gram analysis, like we have seen in sections 4.1. and 4.2., can help us identify and address constructions and their functional contributions in one text or discourse. What simple N-gram analysis does not tell us is whether those frequent combinations of words can also be found in other texts and whether they are particularly frequent in one text, thus delineating it from one or more other texts. To obtain a list of N-grams that really delineate a given text (so that we can identify what N-grams and, at a deeper level, constructions are characteristically associated with the text), a comparative analysis can be useful. A comparative N-gram analysis entails the comparison of frequencies of N-grams across two or more texts or corpora in order to find N-grams that delineate the characteristics of the texts or

corpora in question.

The comparative N-gram analysis is based on the measure for distinctive collexemes (Gries & Stefanowitsch 2004). For each bigram, we entered the following input into a 2x2 table and ran it through a distinctive collexeme analysis:

- the frequency of the bigrams in *AW*
- the frequency of the bigrams in *HF*
- the frequency of all other bigrams in *AW*
- the frequency of all other bigrams in *HF*

The table below summarizes the results. Note that the column named 'collostruction strength', which normally is read as referring to 'degree of lexeme-construction attraction' should in our case be read as 'degree of bigram-text attraction'.

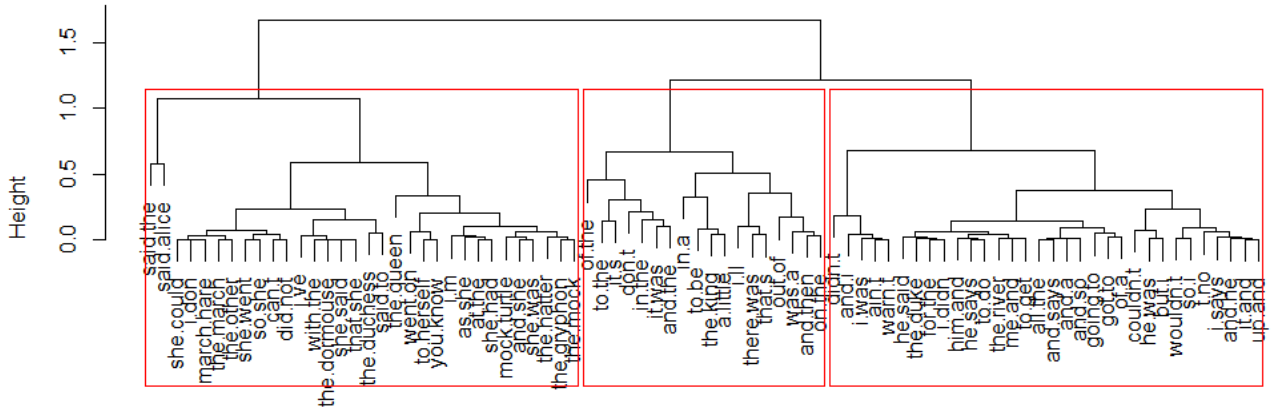
Table 14: Collostruction-based analysis of bigram-text attraction

Bigrams that prefer *AW* (top 20)			Bigrams that prefer *HF* (top 20)		
Rank	Bigram	Collostruction strength	Rank	Bigram	Collostruction strength
1	said the	127.4458	1	ain t	27.046
2	said alice	83.4121	2	warn t	26.9536
3	the queen	51.7526	3	didn t	16.8702
4	mock turtle	40.2463	4	i was	16.4188
5	as she	38.3725	5	i says	15.0384
6	the gryphon	38.0892	6	t no	14.1153
7	the mock	38.0892	7	so i	12.0472
8	the hatter	37.3702	8	and says	11.9
9	to herself	29.4834	9	the duke	11.254
10	the duchess	29.4621	10	couldn t	10.7709
11	she had	28.5327	11	the river	10.7002
12	said to	25.2896	12	and i	10.5681
13	the dormouse	25.1492	13	he was	10.391
14	march hare	22.2742	14	me and	9.5121
15	the march	21.5555	15	by and	9.1315
16	that she	21.4908	16	says i	8.8547
17	went on	20.2767	17	the old	8.3933
18	the mouse	20.1181	18	i reckon	8.1165
19	did not	18.7443	19	he says	8.0106
20	the caterpillar	18.6807	20	done it	7.932

Table 14 shows that the bigram *said the* has the strongest attraction to *AW*, while *ain t* is the most strongly attracted bigram to *HF*. Overall, Table 14 confirms that *AW* is strongly associated with the ITRC-construction, while *HF* is associated with negatives (e.g. *ain t*, *warn t*, *didn t*). It is also important to note that the double-negative marker *t no* as in *warn t no* can also be found in sixth place with a collostruction strength of 14.1153 in *HF*, while *AW* does not have any bigrams that are associated with negatives.

A hierarchical agglomerative cluster analysis was then applied to measure similarities and distances between bigrams, based on their frequencies of occurrence, normalized to per 10,000 words in the two texts. The analysis is summarized in the dendrogram below:

Figure 12: Cluster analysis of bigrams in *AW* and *HF* (Canberra, McQuitty):



The bigrams fall into three clusters: one which contains bigrams exclusive to HF (such as *warn t*, *ain t*, and *i says*), one that contains bigrams exclusive to AW (such as *said the* and *said alice*), and one that contains bigrams that appear in both texts (such as *and then*).

The combination of distinctive collexeme-based comparative N-gram analysis and hierarchical cluster analysis show that there are indeed several bigrams that delimit the two texts, but it also reveals that, while the *and then*-construction is a prominent feature of *HF*, it does not necessarily serve to delineate *HF* from *AW*. In future studies the contrast between stylistic prominence and delimitation is worth exploring further.

5. Network analysis

Network analysis provides a methodology to represent the structure of an object by means of a graph (or network) where a relational structure is represented. A directed graph represents a graph with directed edges between vertices, whereas an undirected graph represents a graph with unordered pairs of vertices. Network analysis is used in a wide range of scientific fields, including biology (e.g. bioinformatics, molecular and systems biology), theoretical physics, and chemistry, as well as computer science and engineering (for a series of informative articles on statistical and machine learning approaches using network analysis, see Dehmer & Basak 2012). As will be outlined below, network analysis allows one to characterize the properties of a system in the way that greatly helps one to investigate the system's structure and function. In biology, for example, network analysis has played an important role in characterizing genomic and genetic mechanisms (Barabási & Zoltán 2004; Barabási et al. 2011). Language can also be seen as a system consisting of structure and function, and hence it seems useful to apply network-based methods to its study. Presentation of a full application of network analysis is beyond the scope of the present paper (for a more active application of network analysis in the context of grammatical constructions, see Jensen & Shibuya (in prep. a, b) as well as Brook O'Donnell et al. (ms); Römer et al. (fc), Gries & Ellis (2015), and Ellis et al. (2013)). Instead, in what follows, we will keep to the minimum necessary to introduce the fundamentals of the methods, and then turn to discussing some of the results yielded by an application of network analysis to our sample.

5.1. Network analysis applied in linguistics

The application of network analysis in linguistics is currently seeing use within cognitive linguistics and corpus linguistics. In the work of Ellis and colleagues, such as Brook O'Donnell et al. (ms); Römer et al. (fc), Gries & Ellis (2015), and Ellis et al. (2013), network analysis is applied to

identify semantic networks in verb-argument constructions. For instance, Gries & Ellis (2015) apply network analysis at the level of semantics to verb-argument constructions and address the prototypicality of verbs in such constructions, the semantic cohesion of verbs in such constructions, and patterns of semantic prototypicality. Thus, they set up a network of verbs in the English *into*-construction and identify several communities of semantically related verbs such as for instance a deceive community (*deceive*, *fool*, *delude*, *dupe*, *kid*, *trick*, *hoodwink*), a force community (*force*, *push*, *coerce*, *incorporate*, *integrate*, *pressure*), and a persuade community (*persuade*, *tease*, *badger*, *convert*, *convince*, *brainwash*, *coax*, *manipulate*) and are able to address degrees of connectivity between members of such communities.

Our application of network analysis, while applying the same measures, differs from the work of Ellis and colleagues in that we apply network science at the *textual* level, and we base it on observed N-grammic relations. That is, while they apply it at the level of verbs, basing it on lexical relations, in particular verb-argument constructions and set up *semantic* networks, we treat the entire text[9] as a network in which every word in the text is a node. On the basis of the connectivity between those nodes, we can identify relations similar to those between words in N-grams, but transcending the limits of specific N-gram types.

Although they do not address constructions, our work is more akin to Brezina et al.'s (2015) approach to collocations in texts and corpora, in which texts and corpora are treated as networks of collocations than it is to Ellis and colleagues' application of network analysis. A difference between Brezina et al. (2015) and the analyses presented here is, of course, that our work takes its starting point in N-grams while theirs as a type of advanced and sophisticated collocational analysis. Note that, while we use packages in *R*, Brezina et al (2015) use a specialized piece of software called *CollGraph* which was still under development while the analyses presented here were being carried out. That is why, although *CollGraph* may well be applicable in the type of analysis we are interested in, we did not use it for this particular study.

5.2. Network analysis as an extension of comparative N-gram analysis

We have so far presented a comparative N-gram analysis, where N-grams were first identified in the texts of *AW* and *HF*, and significant N-grams that are characteristic of each of these texts were captured and discussed with respect to their functionality. As with many other methods, N-gram extraction as well as a comparative N-gram analysis has merits and demerits. N-grams can help us identify and address constructions and their functionality in one text or discourse. Comparative N-gram analysis can help us find N-grams that delineate texts or discourses. A problem, however, is that shorter N-grams are embedded in longer N-grams. Bigrams can be found inside some of the trigrams and fourgrams. Note, for example, *said the* can be found inside *said the mock turtle*. That is to say, as a result, our N-gram lists as presented in Tables 3 and 6 contain some redundancy. In the comparative N-gram analyses so far presented, we have mainly focused on bigrams. However, since texts contain both shorter N-grams (unigrams) and longer N-grams (trigrams, fourgrams, etc), it is preferable if we discuss shorter *and* longer N-grams. One way to overcome this type of problem if one is not interested in abstracting from N-grams to more schematic structures is Brook O'Donnell's (2011) adjusted frequency list approach in which the frequencies of larger N-grams that entail shorter N-grams are subtracted from the frequencies of the embedded N-grams. This approach is extremely useful with frequency lists that distinguish between fully fixed phraseological strings and lexemes, but in a study such as this one in which we generalize over certain units in the string, it is not applicable. This is the case of *said the X* in which we generalized over the elements that appear in the *X*-position. In fact, if we subtract the frequencies of larger N-grams that contain

9 In cases where a full corpus is used, network analysis can be applied at corpus level. In such a case, the entire corpus is represented as a network.

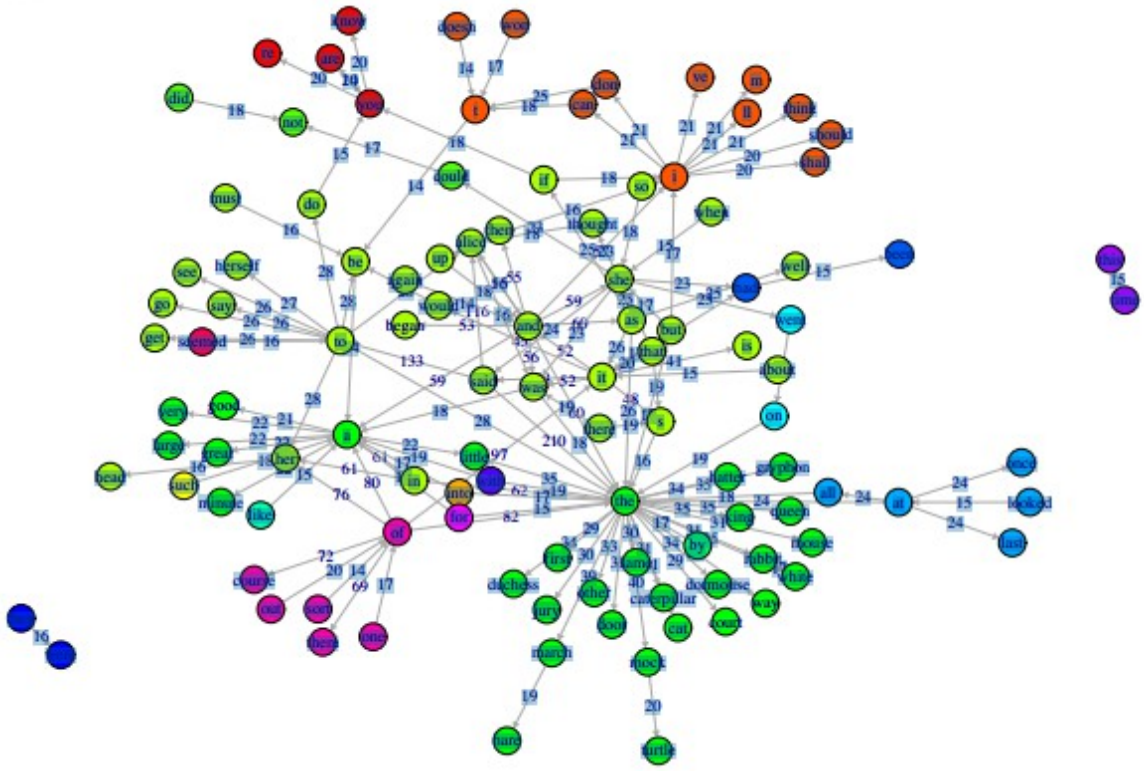
said the from the frequency of the bigram *said the*, the result would be a frequency of 0 for *said the*. The network analysis as illustrated below is an alternative way of handling the descriptive demand of addressing short and long N-grams within the same representational frame.

5.3. Representing constructions in networks[10]

5.3.1. Network of N-grams (and underlying constructions) in *AW*

Figure 13 below is a network analysis representation for *AW* (96 most frequent bigrams):

Figure 13: Global network of *AW*

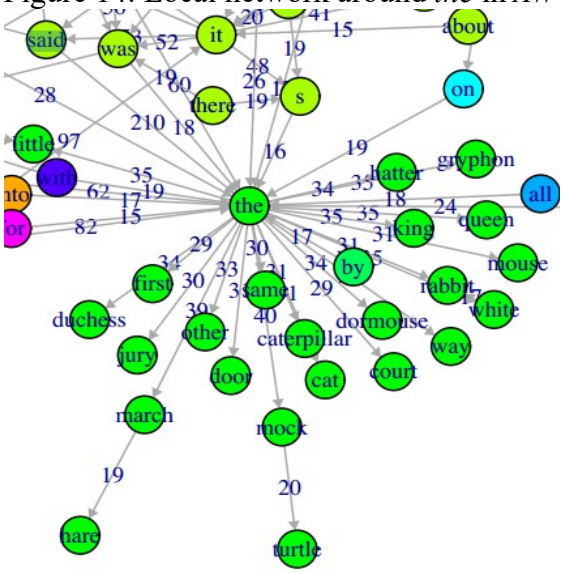


The number between two nodes indicates the frequency of the connected nodes. The color of nodes indicates the connecting edges (community) clustered together based on their "edge betweenness". A network (or graph) consists of nodes that as a whole constitute a global community. A network, however, often forms a nested structure, consisting of several subnetworks (or communities). A subnetwork (or community) is structured such that the nodes included in it are connected often by a number of edges. That is, there is in general a high edge density inside a community. On the other hand, the edge density is low between communities. Each node constitutes a minimal community. A company, for example, is an organization as a whole, consisting of subnetworks called departments or units which ultimately consist of each individual. A way of extracting subnetworks (or communities) in a network is through calculating the edge betweenness of the graph, and this is what is implemented in this figure. For convenience of explanation, consider Figure 14 which zooms in a local network around *the*. First, notice that the node *the* is connected with its co-occurring nouns. The direction of arrows indicates the directionality of word combinations (i.e. *the* and the nouns). As mentioned above, the color of nodes indicates communities in the network. The green nodes, which have been clustered as forming a community in the network, consisting of *the* and the nouns that it determines instantiate the construction [*the* N]/[DEFINITE NOMINAL

10 As an input for the networks discussed here, the bigrams identified in section 5 were used.

REFERENCE].

Figure 14: Local network around *the* in *AW*



Now, notice next that the node *the* is also connected with another important bigram – namely, *said the*. Recall that *said the* was identified in our N-gram analysis as constituting the most important and frequent bigram in *AW*. Notice yet another important fact in the graph that starting from the node *said* it is possible to find longer strings of words (trigrams) such as *said the king*, *said the caterpillar*, *said the cat*, as well as fourgrams such as *said the march hare* and *said the mock turtle*. As illustrated here, the network analysis based on the identified bigrams thus offers a simple but powerful method for representing both short and long N-grams (and underlying constructions) within the same representational framework. The method lists unigrams, bigrams, trigrams, fourgrams, etc. all at one time, and may thus be considered to provide descriptively an efficient analysis on frequently co-occurring combinations of words (and underlying constructions).

There are many more important aspects to be examined concerning the global network given in Figure 13, but since our main concern is to show the usefulness of network analysis for discovering N-grams (and underlying constructions), we will not further explore the graph. Instead, we now turn to the network of N-grams in *HF*.

5.3.2. Network of N-grams (and underlying constructions) in *HF*

Figure 15 shows the bigram network of *HF* (99 most frequent bigrams). As with *AW*, for convenience of explanation, we will focus here on some local networks in the figure that seem worth a special attention. Figure 16 below represents a local network capturing the auxiliary-with-a-negator construction (or negation construction) consisting of instances such as *couldn-t*, *don-t*, *didn-t*, etc. Notice that nodes instantiating double negation are also represented in the figure. In Figure 16, Consider the circled nodes of *warn-t*, *ain-t*, and *t-no* In the global network presented in Figure 15, it is possible to observe, as illustrated in Figure 17, a few interesting bigrams concerning the first person pronoun *I*: *i reckon*, *says i*, and *i says*. There are also some N-grams of *and* consisting of *and then*, *by and by*, and *and so*, as seen in the local network in Figure 18.

Figure 15: Global network of *HF*

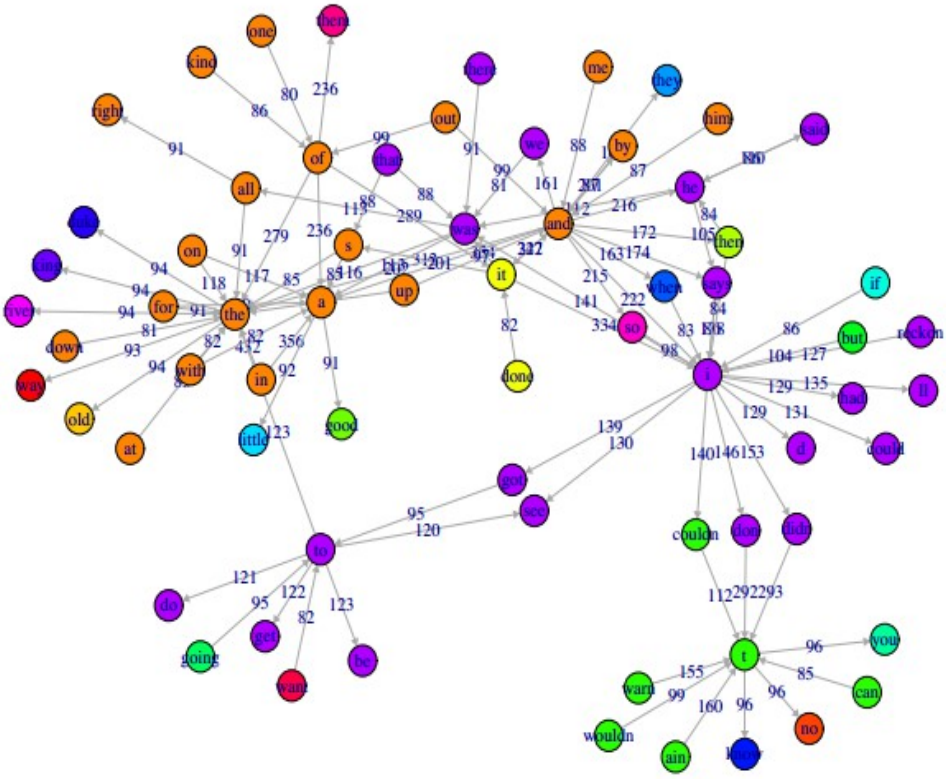


Figure 16: Local network reflective of a negator construction in *HF*

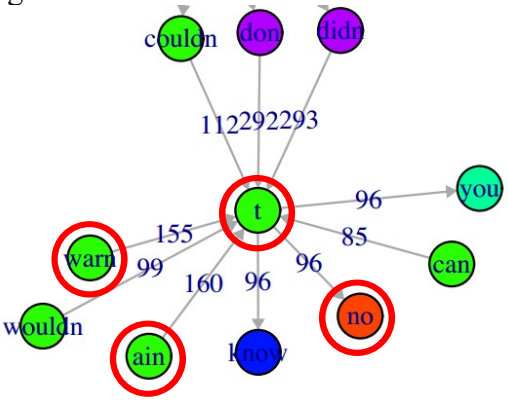
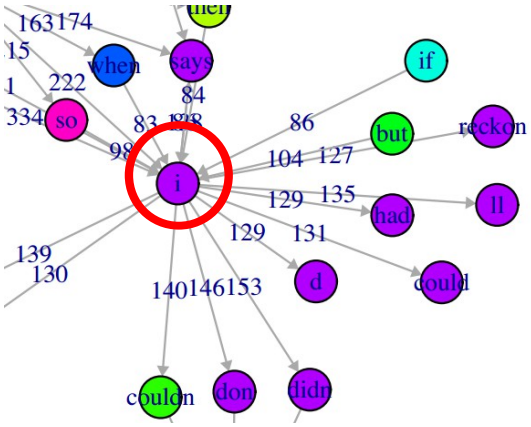


Figure 17: Local network reflective of bi-*grams containing I in HF*



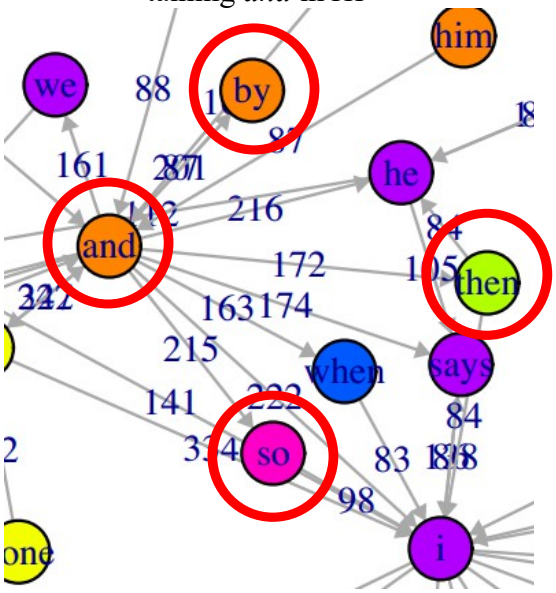
5.4. Nodes and centrality

In network analysis, a set of indices is used to characterize the structural properties of networks. Such indices include density, transitivity, reciprocity/mutuality, and centrality. Centrality is among the most frequently used indices, and here we restrict ourselves to this index.

5.4.1. Introducing the notion of centrality

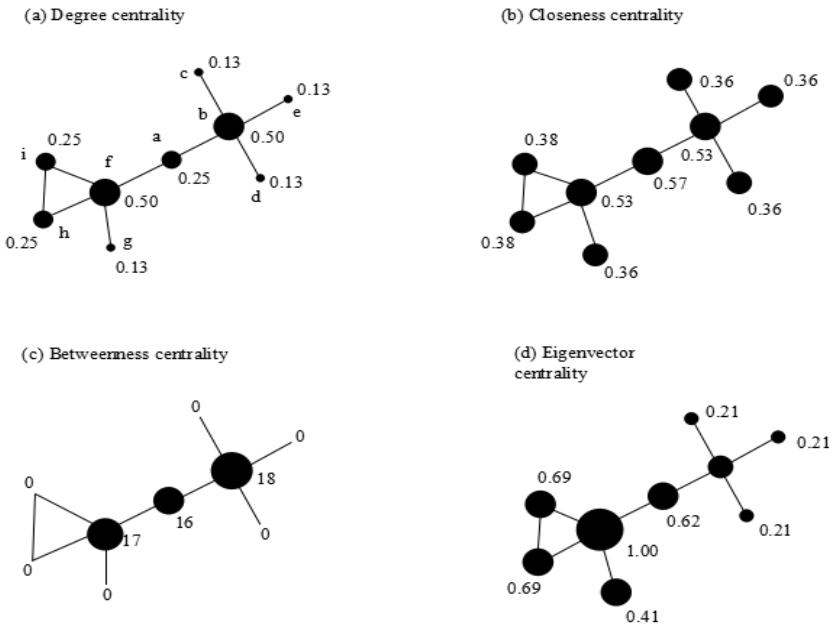
Centrality (commonly called point/node centrality) shows how central each of the vertices (or nodes) in the network is. It is an index used to estimate or compare the importance of each vertex (or node). Several methods have been proposed to evaluate centrality of vertices. One is degree centrality.

Figure 18: Local network of N-grams containing *and* in *HF*



Degree of centrality is the simplest centrality measure among others. It is used to calculate the number of ties that a vertex has in a network. Another centrality measure is called closeness centrality. Closeness centrality measures how many steps are required in order to access every other vertex from a given vertex. A third centrality measure is betweenness centrality. It calculates vertex betweenness. It measures the centrality of a vertex in a network. Its calculation is based on the shortest path between vertices. Yet another centrality measure is eigenvector centrality. This is a higher version of degree centrality in that while degree centrality is measured on the basis of the number of neighbors, the eigenvector centrality measure considers the centralities of neighbors. Figure 19 illustrates the aforementioned centrality measures:

Figure 19: Centrality measures (adapted from Dehmer & Basak 2012: 71)



In the figure, the size of a vertex expresses the centrality value. Centrality values and node identifiers are indicated by the numerical values and lowercase letters, respectively. In (a), nodes *b* and *f* have the highest centrality in the network. This is obvious, because degree centrality reflects the node degree. Note that node *a* has high centrality in (b) and (c). The high centrality of node in closeness centrality and betweenness centrality is due to the fact that this node functions as an intersection between two subnetworks consisting of node sets of {*b*, *c*, *d*, *e*} and {*f*, *g*, *h*, *i*}, respectively. That is to say that node *a*, by constituting an intersection between these two subnetworks, can be interpreted as a central node. Closeness centrality and betweenness centrality are both measures based on the shortest path analysis, and hence they can find a central node. As mentioned above, eigenvector centrality is an extended degree centrality, and this is why the results for these two measures are similar in the figure. The fact that the nodes in the triangle consisting of *f*, *h*, and *i* have high centralities is because eigenvector centrality is based on the centralities of neighbors. As is apparent from this brief description of centrality measures, different centrality measures yield different interpretations. Thus, it is important to choose centrality measures with care.

Having outlined the notion of centrality, we can now turn to analyzing the sample using the index.

5.4.2. Measuring centrality of nodes

Here, we measure the nodes centrality by computing the betweenness centrality. As outlined above, betweenness centrality is a measure concerning the number of shortest paths going through a vertex or an edge. In network analysis, a node with a high degree of betweenness centrality is assumed to play an influential role in the network, because the particular node is connected with other nodes with the shortest paths.

The table below shows the top 15 in *AW* and *HF*, respectively:

Table 15: Top 15 in *AW* and *HF*

Rank	*AW*		*HF*	
	Node	Frequency	Node	Frequency
1	the	750	i	411
2	and	512	and	243
3	it	350	the	180
4	she	330	it	147
5	said	261	was	145
6	to	250	to	90
7	of	231	of	89
8	alice	226	t	80
9	a	211	couldn	80
10	i	135	got	80
11	was	135	a	58
12	as	117	he	57
13	be	77	says	38
14	you	68	s	27
15	t	52	all	24

In between centrality, the larger the value is, the higher the centrality of the node is. It is shown in the table that a number of same words can be found both in *AW* and *HF*, which suggests that their

betweenness centrality is perhaps a general characteristic of the English language. At least, that seems to be the case in written English. Despite this similarity, it is important to note that betweenness centrality also shows us that the words in the table are not listed in the same order in the two texts. Starting from the top of the table, for example, notice that *the* is listed as Number 1 in *AW*, while in *HF i* fills that position. This suggests that the community consisting of *the* and its head nouns constitute the highest betweenness centrality in *AW*, while the community consisting of *i* and its co-occurring words as briefly discussed above constitute the highest betweenness centrality in *HF*. Betweenness centrality thus allows one to quantify significant nodes in a network, which in turn serves to characterize the texts under investigation.

5.5. Motivation for network analysis

After all, short and long N-grams can be identified without network analysis (recall Section 4). Then, what is good about using network analysis? We suggest that there are mainly two points to argue for taking a network analysis approach. 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, as discussed in the preceding section, 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. A simple N-gram analysis does not provide answers to these issues.

6. Concluding remarks

Can N-grams and the more advanced N-gram-based network analysis be used to identify constructions? We have seen that both techniques help to identify recurring strings of recurring words, one difference being that simple N-gram analysis requires the analyst to operate with several lists of N-gram types and make cross references across the lists while the latter enables the analyst to capture all N-grams, regardless of their size, in the same representational network. While the latter has an advantage over the former, both have the distinct advantage that they can be useful for identifying recurring phraseological phenomena in texts or corpora in a fashion that would be impossible for human analysts. Further, since both methods provide frequencies, the analyst is enabled to compare N-gram occurrences across texts or corpora, such that, by applying distinctive collexeme analysis for instance, it is possible to see whether or not the N-gram in question delineates one text or corpus.

What about functionality? Neither N-grams nor N-gram-based networks tell us much about functionality, as they show us purely formal relations. That is, they automatically identify phraseological phenomena and quantify them, but they do not show how the N-grams in question are actually used. However, in automatically identifying recurring strings of words, they guide the analyst in terms of connections between words that are salient in a given text and may be indicative of constructions as functional units. The analyst can then manually, according to their theoretical orientation, investigate 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 saw this in our exploratory analyses of *Alice's Adventures in Wonderland* and *The Adventures of Huckleberry Finn.* In the former, in our simple N-gram analysis, returned several N-grams of the *said the* type. In a concordance, we analyzed all instances of *said the* and found a recurring discursive pattern in which *said the* is reflective of a dialog-ordering construction in which the dialog is topicalized and the speaking character is focalized. We were further able to abstract even further, via a list of bigrams, up to a more general constructional level where other reporting verbs occur in the construction. Similarly, a number of N-grams were identified in *The Adventures of Huckleberry Finn* which displayed discursive patterns reflective of communicative

functions. For instance, the *warn t no*-type N-gram captured two entities that are used as separate constructions in the narrative style – namely, *it warn't no X* and *there warn't no X*. The collostructional analyses confirmed that the two are treated as different constructions, as they display rather different degrees of productivity. Their main functional contribution, however, is constructed by Mark Twain, as he captures the typical discursive behavior of constructions (at least in the perspective of usage-based construction grammar) and imbues the mind-style of Huckleberry Finn with a sense of authenticity. We also found a number of N-grams – namely, the N-grams that capture *and then*, *by and by*, and *and so*, all of which are used in the narrative to organize events in the narrative, and to contribute to the simple and childlike mind-style of the narrator.

The methods presented here need to be applied to further data capturing various types of discourses, and it is very possible that they will have to be modified in a number of ways. However, this initial exploratory study does indicate the usability of N-gram-based analyses (including two comparative N-gram analyses and N-gram-based network analysis) in exploring constructions in an objective and efficient way, which ultimately could contribute to the development of constructionist approaches to language.

Bibliography

Agresti, Alan. (2002) *Categorical Data Analysis*. Second edition. New York: Wiley.

Bache, Carl (2014). 'Den narrative anvendelse af *when* i engelsk'. *Ny Forskning i Grammatik*, 21: 5-19.

Bache, Carl (2015). 'The narrative 'when' enigma'. In Claus Schatz-Jakobsen, Peter Simonsen & Tom Pettitt (eds.), *The Book out of Bonds: Essays Presented to Lars Ole Sauerberg*. Odense: Institut for Kulturvidenskaber. 7-21.

Barabási, Albert-László & Zoltán N. Oltvai (2004). 'Network biology: Understanding the cell's functional organization'. *Nature Reviews Genetics*, 5: 101-113.

Barabási, Albert-László, Natali Gulbahce & Joseph Loscalzo (2011). 'Network medicine: A network-based approach to human disease'. *Nature Reviews Genetics*, 12: 56-68.

Barsalou, Lawrence R. (1992). *Cognitive Psychology: An Overview for Cognitive Scientists*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bergen, Benjamin & Kim Binsted (2004). 'The cognitive linguistics of scalar humor'. In Michel Achard & Suzanne Kemmer (eds.). *Language, Culture, and Mind*. Stanford, CA: CSLI. 79-91.

Bondy, John Adrian & Murty, U.S.R. (2008). *Graph Theory*. New York: Springer.

Brezina, Vlacav, Tony McEnery & Stephen Wattam (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2): 139–173

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

Brook O'Donnell, Matthew, Nick Ellis, Gin Corden, Liam Considine & Ute Römer. (ms). 'Using network science algorithms to explore the semantics of verb argument constructions in language usage, processing, and acquisition'.

Cho, Dong-Yeon, Yoo-Ah Kim & Teresa M. Przytycka. (2012). 'Chapter 5: Network biology approach to complex diseases'. *PLoS Computational Biology*, 8(12): e1002820.

Couper-Kuhlen, E. (1989b). 'Foregrounding and temporal relations in narrative discourse'. In A. Schopf (ed) *Essays on Tensing in English, Vol II: Time, Text and Modality*. Tübingen: Niemeyer. 7-30.

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

Croft, William A. (2005). 'Logical and typological arguments for Radical Construction Grammar'.

In Jan-Ola Östman (ed.), *Construction Grammars: Cognitive Grounding and Theoretical Extensions,* Amsterdam: John Benjamins, 273-314.

Croft, William A. & D. A. Cruse (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Culpeper, Jonathan (2009). 'Reflections on a cognitive stylistic approach to characterization'. In Geert Brône & Jeroen Vandaele (eds). *Cognitive Poetics: Goals, Gains and Gaps*. Berlin: Mouton de Gruyter: 125-159.

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

Declerck, Renaat H. C. (1997). *When-clauses and Temporal Structure*. London: Routledge.

Ellis, Nick, Matthew Brook O'Donnell & Ute Römer (2014). 'Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*, 4 (4): 405-431.

Evans, Vyvyan & Melanie Green (2006). *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Ferrer i Cancho, Ramon & Ricard V. Solé (2003). 'Least effort and the origins of scaling in human language'. *PNAS*, 100: 788–791.

Fillmore, Charles J. (1982). 'Frame semantics'. In The Linguistic Society of Korea (Eds.), *Linguistics in the Morning Calm*. Seoul: Hanshin: 11-137.

Fillmore, Charles J. (1988). 'The mechanics of "Construction Grammar"'. *BLS*, 14: 35-55.

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

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

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

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

Grice, Paul (1975). 'Logic and conversation'. In Peter Cole & Jerry Morgan (eds.), *Syntax and Semantics, 3: Speech Acts*, New York: Academic Press. 41-58.

Gries, Stefan Th. (2007). Coll.analysis 3.2: A program for R for Windows 2.x

Gries, Stefan Th. & Anatol Stefanowitsch (2004). 'Extending collostructional analysis: A corpus-based survey on "alternations"'. *International Journal of Corpus Linguistics*, 9(1), 97-129.

Gries, Stefan Th., John Newman & Cyrus Shaoul (2011). 'N-grams and the clustering of registers'. *ELR Journal*, 5(1). URL: http://ejournals.org.uk/ELR/article/2011/1. Retrieved November 14, 2014.

Gries, Stefan Th. & Joybrato Mukherjee (2011). 'Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes'. *International Journal of Corpus Linguistics*, 15(4): 520-548.

Gries, Stefan Th., & Nick Ellis (2015). 'Statistical measures for usage-based linguistics'. *Currents in Language Learning*, 2: 228-255.

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

Huang, Yan (2007). *Pragmatics*. Oxford: Oxford University Press.

Jockers, Mathew L. (2014). *Text Analysis with R for Students of Literature*. New York: Springer.

Jensen, Kim Ebensgaard (2014). 'Performance and competence in usage-based construction grammar'. In Rita Cancino, & Lotte Dam (eds.), *Towards a Multidisciplinary Perspective on Language Competence*. Aalborg: Aalborg University Press: 157-188

Jensen, Kim Ebensgaard & Yoshikata Shibuya (in prep a). 'Exploring inaugural presidential speeches with network analysis' [working title].

Jensen, Kim Ebensgaard & Yoshikata Shibuya (in prep b). 'Delineating travel guides in the American National Corpus' [working title].

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

Lyne, Anthony A. (1985). *The Vocabulary of French Business Correspondence*. Geneva: Slatkine-Champion.

Mahlberg, Michaela (2007a). 'A corpus stylistic perspective on Dickens' Great Expectations'. In Marina Lambrou and Peter Stockwell (eds.), *Contemporary Stylistics*. London: Continuum. 19-31.

Mahlberg, Michaela (2007b). 'Clusters, key clusters and local textual functions in Dickens'. *Corpora*, 2(1): 1-31.

Martínez, Nuria Del Campo (2013). *Illocutionary Constructions in English: Cognitive Motivation and Linguistic Realization*. Bern: Peter Lang.

Miner, Gary, John Elder, Thomas Hill, Robert Nisbet, Dursun Delen & Andrew Fast (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Oxford: Elsevier Academic Press.

Newman, Mark. (2010). *Networks: An Introduction*. Oxford University Press.

Oakes, Michael P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, & Jan Svartvik (1972). *A Grammar of Contemporary English*. London: Longman.

Römer, Ute, Matthews Brook O'Donnell & Nick Ellis, N. C. (fc). 'Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions: Exploring corpus data and speaker knowledge'. In Nicholas Groom, Maggie Charled & Suganthi John (eds.), *Corpora, Grammar, Text and Discourse: In Honour of Susan Hunston*. Amsterdam: John Benjamins.

Schönefeld, Doris (2013). '*It is ... quite common for theoretical predictions to go untested* (BNC_CMH). A register-specific analysis of the English go un-V-en construction'. *Journal of Pragmatics*, 52: 17-33.

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

Simpson, Paul (2004). *Stylistics: A Resource Book for Students*. London: Routledge.

Stefanowitsch, Anatol & Stefan Th. Gries (2003). 'Collostructions: Investigating the interaction between words and constructions'. *International Journal of Corpus Linguistics*, 8(2), 2-43.

Stefanowitsch, Anatol & Stefan Th. Gries (2005). 'Covarying collexemes'. *Corpus Linguistics and Linguistic Theory*, 1(1), 1-43.

Stubbs, Michael (2007). 'An example of frequent English: phraseology: Distributions, structures and functions'. In Roberta Facchinetti (ed.), *Corpus Linguistics 25 Years On*. Amsterdam: Rodopi. 89-105.

Stubbs, Michael (2009). 'Technology and phraseology'. In Ute Römer & Rainer. Schulze (eds.), *Exploring the Lexis-Grammar Interface*. Amsterdam: John Benjamins. 15-31.

Talmy, Leonard (2000). *Toward a Cognitive Semantics. Vol. 1: Concept Structuring Systems*. Cambridge, MA: MIT Press.

Vasquez, Camilla (2014). *The Discourse of Online Consumer Reviews*. London: Bloomsbury.

Wasserman, Stanley and Katherine Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology.* Cambridge, MA: Addison-Wesley.