sciendo
DOI 10.2478/jazcas-2025-0009

MAPPING RECURRENT LEXICO-GRAMMATICAL PATTERNS IN ENGLISH THROUGH SUBTREE FRAGMENTS

ALEKSANDAR TRKLJA
Institute for Translation Studies, University of Innsbruck, Innsbruck, Austria
(ORCID: 0000-0002-7287-5338)

TRKLJA, Aleksandar: Mapping Recurrent Lexico-grammatical Patterns in English Through Subtree Fragments. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 95 – 112.

Abstract: This paper examines subtree fragments (StF) as a corpus-informed method for identifying recurrent lexico-grammatical structures and compares them to two established approaches: collocational frameworks (Sinclair and Renouf 1988) and pattern grammar (Hunston and Francis 2000). StFs differ from these approaches in two major respects. First, they are grounded in a theoretical linguistic assumption that lexical heads project syntactic structures, incorporating part-of-speech categories, phrase structures, and thematic role assignment. Second, StFs are identified semi-automatically from parsed corpora by exploring patterns of grammatical words and syntactic categories, in contrast to the predominantly manual, concordance-based methods of the other two approaches. The findings suggest that StFs provide a productive interface between theory-driven syntactic analysis and data-driven corpus linguistics, allowing for fine-grained mapping between form, meaning, and use while retaining compatibility with probabilistic and statistical perspectives.

Keywords: subtree fragments, collocational frameworks, grammar patterns, thematic roles, argument structure, vector representation

1 INTRODUCTION

This paper proposes an approach to identifying subtree fragments (StFs) in corpora. StFs share similarities with both collocational frameworks (Renouf and Sinclair 1991) and grammar patterns (Hunston and Francis 2000). Developed during the peak of the Cobuild project, collocational frameworks refer to discontinuous sequences of two high-frequency grammatical words with a lexical word in between, such as *a + ? + of* or *too + ? + to*. Warren and Leung (2016) later proposed a broader definition of frameworks not limited to two grammatical words. Renouf and Sinclair regarded frameworks as genuine components of language rather than mere analytical tools although they provided no empirical references to support this claim. The key insight from their study is that grammatical words combine with each other to form regular 'scaffolds' into which certain lexical items fit. These combinations are not random but systematic, frequent, and selective. This systematicity makes frameworks valuable for investigating statistical tendencies, such as the distribution of lexical words within specific grammatical environments, as well as for identifying potential semantic classes. The classification of lexico-grammatical sequences into semantic categories is explored in greater depth in the pattern grammar approach proposed by Hunston and Francis (2000). Grammar patterns defined as "a phraseology frequently associated with (a sense of) a word, particularly in terms of the prepositions, groups, and clauses that follow the word" (Hunston and Francis 2000, p. 3). It is assumed that a pattern, together with all its lexical items, constitutes an extended unit of meaning (building on Sinclair 1996).

StFs proposed in the present paper are akin to both collocational frameworks and grammar patterns in that they concern the association of lexical items with sequences of grammatical words. However, as will be explained in the next section, StFs differ from these two notions in specific ways. I will then demonstrate how StFs can be identified semi-automatically in corpora, how their distribution can be explored, and how they can be classified using word embeddings and the information about thematic structures.

2 SUBTREE FRAGMENTS

2.1 Subtree fragments and their identification in corpora

Two major characteristics of both collocational frameworks and grammar patterns are that

i. they are explored without regard to syntactic structures as they are conventionally defined in theoretical linguistics, and
ii. they are identified through the manual exploration of concordance lines. The former follows from the general scepticism in Sinclairian corpus linguistics towards the notions from theoretical linguistics and from the idea that only minimal assumptions should be made when approaching language (Sinclair 1994; Mahlberg 2005). As Sinclair (1994, p. 25) puts it: "we should trust the text. We should be open to what it may tell us. We should not impose our ideas on it, except perhaps to get started. We should only apply loose and flexible frameworks until we see what the preliminary results are in order to accommodate the new information that will come from the text."

This view is understandable given the fact that it stems from lexicographic research, which attempts to provide item-specific descriptive information for practical uses. However, aside from ignoring decades of theoretical and empirical research in syntax, the problem with this view is that it risks treating structural generalisations as irrelevant or even obstructive. By focusing exclusively on surface co-occurrence patterns, such an approach loses explanatory depth since it does not account for why certain combinations are possible or impossible in terms of underlying grammatical relations. It also has limited generalisability, as observations remain tied to attested forms and do not easily extend to potential but unattested structures. Finally, it may lead to misclassification, grouping together formally similar sequences that are structurally distinct (for more details see Trklja, forthcoming).

As for the second common feature, at the time when this research was conducted, the main analytical tool in corpus linguistics was the concordance line, supported by tools for displaying collocations. Since then, both corpus resources and computational tools have developed considerably, making it possible today to automate to a much greater extent the exploration of patterning in corpora.

Subtree fragments (StFs) differ from collocational frameworks and grammar patterns in relation to the features discussed above. First, they are based on the generally accepted assumption in theoretical linguistics that lexical items project syntactic structure – an idea central to the Projection Principle in generative grammar (Jackendoff 1977; Chomsky 1981). In other words, the lexical properties of a head determine the syntactic configuration in which it can appear. As their name suggests, StFs are derived from syntactic trees (see below for more details). These syntactic structures are associated with semantic interpretation and contribute to the construction of thematic structures (theta-grids or argument structures) that encode the roles of participants in events (Williams 1994). Although there is no consensus on whether part-of-speech categories are universal – with Baker (2003) arguing in favour of universality and Croft (2001) arguing against – I will assume here that such categories do exist.

Second, in the present study StFs are identified semi-automatically by analysing the patterning of grammatical words and syntactic categories in corpora, rather than through the manual investigation of concordance lines. This involves using parsed corpora and computational tools capable of extracting and classifying structural configurations according to specified grammatical and lexical criteria. While some manual checking may still be required to ensure accuracy, the reliance on syntactic annotation and automated search distinguishes this approach from the purely concordance-based, manual methods used in the early studies of collocational frameworks and grammar patterns. Crucially, because StFs are grounded in syntactic theory, their identification and interpretation are linked to an explicit model of grammar, rather than to surface-level co-occurrence patterns alone.

What kinds of structures are StFs? The notion of subtrees used here is adopted from Bod (1995) and the following three generalizations define subtrees:

> “A subtree of a tree T is a subgraph t of T such that
> (1) t consists of more than one node
> (2) t is connected
> (3) except for the frontier nodes of t, each node in t has the same daughter nodes as the corresponding node in T” (Bod 1995, p. 36).

Unlike some other approaches that rely on the notion of subtrees or similar concepts (Aravind et al. 1975; Marcus 2001), Bod (1998) explicitly states that subtrees are elements of the speaker's linguistic experience. Bod (1998) argues that grammatical knowledge consists of a "statistical ensemble of language experiences" (Bod 1998). In this view, the corpus is regarded as a representation of the speaker's past language experience, and statistical learning is implicitly assumed as the mechanism through which this experience is encoded. The frequency with which utterances have previously been used influences the probability with which speakers will produce expressions and sentences in the future[1]. In particular,

> "this means that new utterances are constructed by combining fragments that occur in the corpus, while the frequencies of the fragments are used to determine the most probable utterance for a given meaning" (Bod 1998).

This does not mean that speakers are unable to produce novel sentences or expressions, but the proposal emphasises that previous experience contributes to the production of such units.

The level of detail in the representation, in terms of sub-trees derived from corpora, depends on the availability of annotation sets and will therefore vary from language to language. Grammatical information is encoded in corpora using parts-of-speech (POS) tag sets and/or syntactic parsers. For the purposes of this study, I will assume a sparse representation of functional categories using the TreeTagger PoS tagset (Marcus et al. 1993). The focus of the study will be on English verbs for illustrative purposes, making use of the English TreeTagger PoS tagset and the British National Corpus (BNC) (Leech 1992). This tagset contains the grammatical categories which are annotated with basic features. For example, verbs are annotated with information about tense. In the present study only the general grammatical information is included (e.g. V, N, A) with the lexical categories being represented without any grammatical features. Pronouns are included in the category N. No claim is made here that the data are representative of the English language as a whole or that register- and genre-specific differences are irrelevant. The present approach enables the identification of sequences of POS categories with function words (such as *V the N of the N*), as well as the combination of particular lexical words with POS categories and function words, (such as *find the A N)*. I will explore both types of StFs below. The tabular representation of PoS tags from TreeTagger for the expression *arrives at the station* is as follows:

[1] From a statistical learning perspective, this proposal aligns with findings in cognitive science and psycholinguistics showing that speakers are sensitive to distributional regularities in their linguistic input (e.g. Ellis 1996, 2002; Armstrong et al. 2017). Thus, high-frequency substructures become entrenched in memory and are more readily retrieved and recombined, whereas low-frequency or novel combinations are less predictable and may require greater processing effort.

Word	Lemma	POS
arrives	Arrive	VV
At	At	PP
The	The	DT
station	Station	NN

Word	Lemma	POS
arrives	arrive	arrive
at	at	at
the	the	the
station	station	NN

Tab. 1. Tabular representation of POS categories in a tagged corpus

I wrote a Perl script to identify StFs by detecting sequences of POS categories and function words. The script can also replace a POS category with the lemma form of a lexical item enabling StFs associated with a lexical word to be identified in a manner similar to the representation in the pattern grammar. In the next step, an n-gram function was used to compile combinations of the actual word within a defined window size. I explored n-grams of three, four and five words. To give an example, one StF associated with for the verb *arrive* is *arrive at the NN*, which occurs 715 times in the BNC. But, the resulting n-grams are not always grammatically complete sequences. Thus, the structure *find the N of* which is generated from the corpus is excluded because it contains a syntactically incomplete prepositional phrase. On the other hand, the structures such as *find the N of the N*, *find the N of a N* or *find the N of the A N* are regarded as StF because they constitute complete VP. At the final stage, all sequences were manually inspected, and only those forming a grammatically complete verb phrase (VP) were included for further analysis. Fig. 1 shows a tree representation of StFs for the verb *find*, with the four types of StFs identified in the BNC.

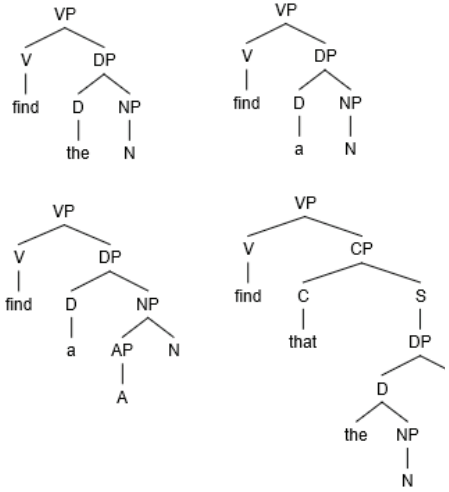
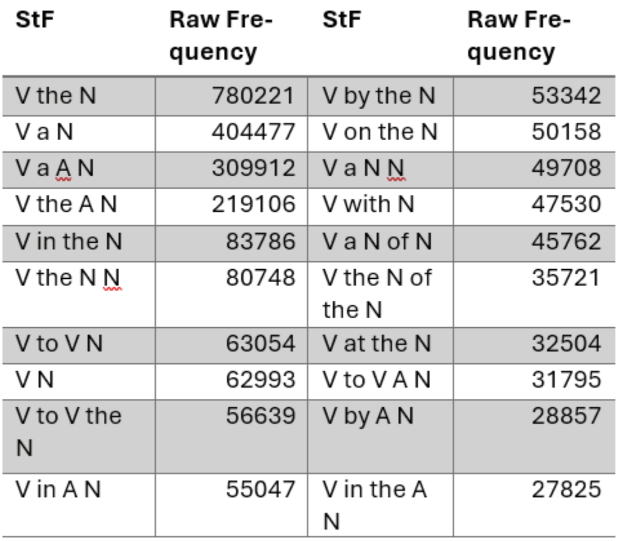


Fig. 1. Four StFs associated with the verb *find* identified in the BNC

2.2 Distribution of general StFs in the BNC

In this subsection I will explain how the distribution of StFs consisting of sequences of POS categories and function words was explored. For illustrative purposes, the present data focuses only on the most representative StFs defined as those that occur at least 1,000 times. In total, 84 StFs that build a VP were identified in this manner in the BNC. The top 20 StFs are presented in Tab. 2 and the comprehensive list can be found in Appendix A.



StF	Raw Frequency	StF	Raw Frequency
V the N	780221	V by the N	53342
V a N	404477	V on the N	50158
V a A N	309912	V a N N	49708
V the A N	219106	V with N	47530
V in the N	83786	V a N of N	45762
V the N N	80748	V the N of the N	35721
V to V N	63054	V at the N	32504
V N	62993	V to V A N	31795
V to V the N	56639	V by A N	28857
V in A N	55047	V in the A N	27825

Tab. 2. The 20 most frequent VP StFs identified in the BNC

Unlike grammar patterns, but like collocational frameworks, StFs include not only the obligatory elements of an argument structure but also modifiers. This has both disadvantages and advantages. The disadvantage is that it overlooks the fact that these instances still belong to the same verb phrase. The advantage is that it provides detailed information about the specific kinds of modifiers typically used. Both types of information can be explored further. In the present study, however, I focus on a more general classification. All subtrees were grouped into broader structural types. For example, the sequences *V the N*, *V a N*, *V the A N*, *V a A N*, *V the N of the N* are all classified into the same category: transitive verbs serving as the head of the verb phrase and selecting a determiner phrase (DP) as their complement. The final classification comprises 23 distinct classes (see Appendix B), encompassing a total of 84 individual StFs.

The initial descriptive statistics reveal a clear tendency in the distribution of VP across types. The most frequent structures (Type 1), such as *V the N* or *V a N*, involve direct NP complements typically associated with core arguments (e.g. Theme, Patient). In contrast, more complex or marked structures, such as those involving directional PPs (*V into the N*), resultative phrases (*V the A N to N*), or role-identifying as-phrases (*V as a N*), are markedly less frequent. The data indicate that Type 1 overwhelmingly dominates usage, accounting for approximately 66.2% of all VP subtree occurrences and 17% of all sequence types (Fig. 2). Other types are much less frequent, each contributing between 0.2% and 8.4%. As the second pie chart (Fig. 3) indicates structural diversity of VP is more evenly distributed across types, with many contributing around 2–6% of the total.

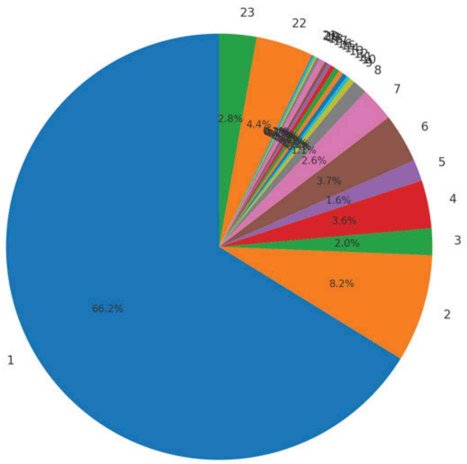


Fig. 2. Total frequency share per VP type

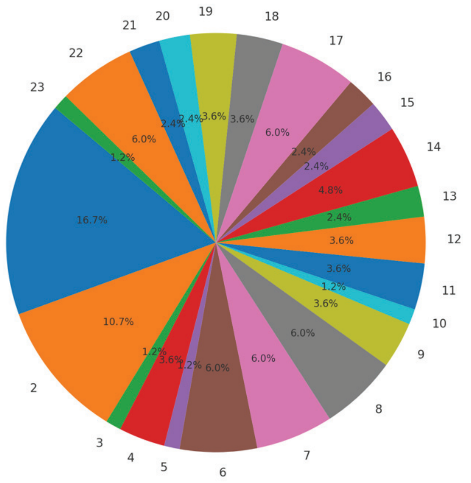


Fig. 3. Distribution of the number of StFs per VP type

To further investigate the data I explored syntactic variety and usage frequency associated with the present set of StFs. I define syntactic variety as the number of distinct StFs grouped under a given VP type (e.g. *V the N*, *V a N*, *V a N of N*), which reflects the degree of formal diversity or grammatical flexibility permitted by a given type. Usage frequency refers to the total number of occurrences of all StFs of a particular VP type.

One may assume that the two dimensions are positively correlated. In other words, constructions that are structurally more productive – that is, capable of supporting a greater number of grammatical variants – are also expected to occur more frequently in actual language use. In order to test this assumption empirically, I formulated the following hypothesis:

- H_0 (Null Hypothesis): There is no relationship between the syntactic variety of a VP type and its usage frequency in the corpus.
- H_1 (Alternative Hypothesis): There is a positive relationship between syntactic variety and usage frequency in the corpus.

To test this hypothesis, I conducted correlation and regression analyses using VP types classified into 23 categories. Descriptive statistics suggests that VP types with more subtree variants tend to show higher overall frequencies. For instance, Type 1 includes some of the most common VP patterns (e.g. *V the N*, *V a N*, *V the A N*), with 14 distinct subtree structures. This type accounts for 66% of the total frequency across all types (of 2,356,436 occurrences). However, this dominance was not matched by other types with comparable structural diversity. Type 6, which comprises different *with N* structures (e.g. *V with N*, *V with the N*) has a low frequency (of just 116,710 occurrences) Similarly, Type 7, which includes five variants appears 79,800 times in total. To determine whether the observed trend is statistically significant and generalisable, I applied Pearson and Spearman correlation tests. The Pearson test yields a strong linear correlation ($r = 0.830$, $p < .001$), and the Spearman rank correlation also shows a significant monotonic association ($\rho = 0.583$, $p = 0.0047$). These results allow us to reject the null hypothesis, suggesting that VP types with greater syntactic variety do tend to occur more frequently in corpus data. However, further analysis complicates this finding. A linear regression using raw frequency values indicates that syntactic variety explains 69.4% of the variance in frequency ($R^2 = 0.694$). Yet, this model was highly influenced by Type 1, a clear outlier with both high variety and extraordinarily high frequency. A second model using log-transformed frequency values reduces this distortion and explains 57.5% of the variance ($R^2 = 0.575$), showing that the association remains significant, but not uniformly strong across all types. The diminishing returns observed in the log-scale model further suggest that the relationship is not strictly proportional: each additional subtree type adds progressively less to the overall frequency. Taken together, these findings support a partial rejection of the null hypothesis. There is indeed a statistically significant relationship between syntactic variety and usage frequency but the relationship is not linear and is heavily skewed by a small number of functionally entrenched constructions. Type 1, as we saw, is not only syntactically diverse but also highly conventional and semantically general, which likely enhances its functional entrenchment which is a property that cannot be reduced to structural variety alone.

2.3 Investigation of specific StFs in the BNC

At the next stage, it is possible to investigate specific StFs and fine-grained semantic distinctions within a syntactically uniform pattern. I selected for illustrative purposes the StF *V the N* which belongs to Type 1. This subtree instantiates numerous semantically diverse expressions (e.g. *accept the offer*, *cut the cost*, *feel the pain*), making it a good candidate for further analysis. 200 of the most frequent *V the N* expressions was collected in the BNC and passed through the pretrained BERT-based model all-MiniLM-L6-v2 from the sentence-transformers library. This model produces high-dimensional vector representations (384 dimensions) for short texts, encoding rich semantic information learned from large corpora. These vectors were then subjected to KMeans clustering with k = 10 to discover semantically coherent groups. To visualize the structure of these clusters, a t-SNE projection was used to reduce the high-dimensional embeddings to two dimensions. Fig. 4 illustrates the spatial distribution of the clusters where each cluster is related to distinct semantic types. For example, one cluster groups expressions such as *accept the offer*, *assess the situation*, and *address the issue*, which all share a judgmental or evaluative function, with the noun denoting an abstract Theme or Proposition. Another cluster includes verbs like *buy the house*, *cut the cost*, and *cover the expense*, associated with economic transactions or resource manipulation, where the noun represents a Patient or Affected Object.

At the final stage, one may select a specific verb and analyse its distribution across StF-types. For the present purposes, I selected the verb *find* and explored its distribution across 2,000 concordance lines from the BNC. The results indicate that it occurs in the following StF: *find N*, *find the N*, *find a N*, *find the A N* (Type 1) and *find that S* (Type 23). I have excluded fragments containing the verb *find out* as this is a distinct lexical item. Unlike pattern grammar, this analysis does not indicate semantic interpretation where verbs and patterns are classified into semantic classes. The classification used in pattern grammar is based on intuition and ignores the higher argument structure representation. An alternative approach that I propose here is to explore the thematic structures of the fragments. Let us consider *find* as an example. Thematic roles assigned by find to its complement are typically Theme referring to something located or discovered as in *find the book* or *find a job*. Occasionally, however, the complement receives the role of Patient if it is affected by the action. This occurs in secondary predication constructions (Rothstein 1983, 2004), where *find* takes a DP complement together with an additional predicate that describes a state or property of that DP such as in *find the defendant guilty*, *find the room in a mess* or *find the door locked*. In these complex transitive uses the DP is both the object of *find* and the subject of the secondary predicate and is understood as undergoing or being in the state described and hence its interpretation as a Patient rather than a Theme. This suggests that the syntactically complete fragments in the latter case includes an additional element which can be realised either as a past participle, prepositional phrase or a predicative adjective. In *find that S* constructions, the complement expresses a proposition or a piece of information or a cognitive result: what is found to be true (e.g. *find that it was closed*). This indicates a cognitive or evaluative use of find (semantic overlap with realize or discover).

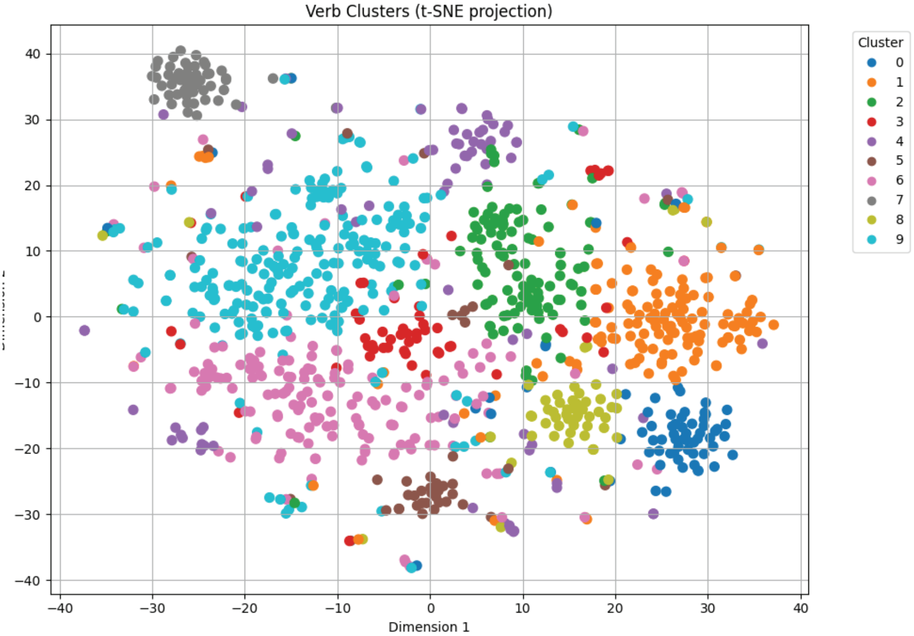


Fig. 4. Clusters of *V the N*-expressions from the BNC

In the current BNC sample, the DP that occur in subject position with the *find*-fragments predominantly fulfils the Experiencer or Cogniser role. But, in addition to its canonical argument structure (Experiencer finds Theme/Proposition), the verb *find* also supports extended argument realizations that introduce Source (*She found the message from John* and *He found her a job*), Beneficiary (*He found a gift for her*), and Means (*They found the solution with a tool*) roles via prepositional phrases or double object constructions.

Formally, this can be represented as:

FIND(x, y, [s], [z], [p])
where
- x = Experiencer (subject NP)
- y = Theme / Patient (Theme if s absent; Patient if s present)
- s = Secondary Predicate (optional; AdjP, Ved, V-ing, PP; makes y = Patient)
- z = Beneficiary (optional; NP or PP)
- p = Source / Asset / Instrument (optional; PP).

3 CONCLUSION

This study proposed the use of StFs as an analytical tool to explore lexical and syntactic patterns in corpora. The aim was to clarify the theoretical assumptions, methodological procedures and potential advantages of the StF approach in relation to existing corpus linguistic approaches, while situating it within the broader corpus linguistic and syntactic theoretical landscape. Unlike collocational frameworks and pattern grammar, which do not commit to syntactic categories beyond those minimally required for corpus annotation, StFs draw directly on syntactic structure and its semantic interpretation. This includes the assignment of thematic roles and the representation of argument structure. Secondly, StFs are identified using a semi-automatic method involving parsed corpora and the computational extraction of patterns defined over syntactic and lexical categories. This method relies less on manual inspection of concordance lines than the other two approaches. Overall, the StF method should offer a bridge between theory-driven and data-driven approaches. This combination enables a more precise mapping of form, meaning, and use than is possible with purely surface-based methods while accommodating probabilistic and statistical insights from corpus linguistics. The findings suggest that incorporating syntactic structure into corpus pattern analysis can enrich theoretical and applied descriptions of language, particularly in contexts where thematic role distinctions and variation in argument structure are important.

4 APPENDIX A: DISTRIBUTION OF THE MOST FREQUENT STFS IN THE BNC

Subtree fragments	Frequency of StFs in the BNC
V the N	780221
V a N	404477
V a A N	309912

V the A N	219106
V in the N	83786
V the N N	80748
V to V N	63054
V N	62993
V to V the N	56639
V in A N	55047
V by the N	53342
V on the N	50158
V a N N	49708
V with N	47530
V a N of N	45762
V the N of the N	35721
V at the N	32504
V to V A N	31795
V by A N	28857
V in the A N	27825
V with the N	26183
V with A N	25867
V to V a N	24230
V by N N	24191
V N of N	22543
V for the N	22028
V into N	21027
V in a N	20181
V into the N	18009
V by the A N	16696
V as a N	15752
V in a A N	13750
V N of the N	12218
V by a N	12199
V for a N	11465
V the A N of the A N	10385

V the A N of the N	10370
V with a N	9873
V through the N	9296
V with the A N	8335
V at the A N	8002
V over the N	7846
V for the A N	7787
V about the N	7768
V as a A N	7612
V into A N	7557
V out of the N	7013
V a N of the N	6261
V with a A N	6257
V as a A N	6104
V for a A N	6016
V into a N	5320
V off the N	5316
V N from N	4305
V under the N	4250
V N for the N	4228
V the N in the N	4152
V against the N	3953
V among the N	3953
V N to N	3712
V across the N	3584
V the N to N	3288
V N from the N	3223
V the N to the N	3060
V as the A N	2665
V after the N	2650
V N as a N	2553
V through the A N	2355
V N to the N	1983

V a N to N	1972
V about the A N	1895
V between N and N	1895
V over the A N	1800
V the N from the N	1771
V through a N	1729
V N as a N	1434
V over a N	1266
V about a N	1220
V N into N	1203
V the N in the A N	1184
V N as N	1135
V a N from the N	1042
V N into the N	1022
V after a N	1016
V the A N to N	1003

5 APPENDIX B: DISTRIBUTION OF THE STF-TYPES

Subtree fragments	**Frequency of StFs in the BNC**	**Type**
V the N	780221	1
V a N	404477	1
V a A N	309912	1
V the A N	219106	1
V the N N	80748	1
V N	62993	1
V a N N	49708	1
V a N of N	45762	1
V the N of the N	35721	1
V N of N	22543	1
V N of the N	12218	1
V the A N of the A N	10385	1
V the A N of the N	10370	1

V a N of the N	6261	1
V in the N	83786	2
V in A N	55047	2
V in the A N	27825	2
V into N	21027	2
V in a N	20181	2
V into the N	18009	2
V in a A N	13750	2
V into A N	7557	2
V into a N	5320	2
V to V N	63054	3
V to V the N	56639	4
V to V A N	31795	4
V to V a N	24230	4
V on the N	50158	5
V with N	47530	6
V with the N	26183	6
V with A N	25867	6
V with a N	9873	6
V with a A N	6257	6
V at the N	32504	7
V for the N	22028	7
V for a N	11465	7
V for the A N	7787	7
V for a A N	6016	7
V as a N	15752	8
V as a A N	7612	8
V as a A N	6104	8
V as the A N	2665	8
V N as a N	2553	8
V through the N	9296	9
V through the A N	2355	9
V through a N	1729	9

V at the A N	8002	10
V over the N	7846	11
V over the A N	1800	11
V over a N	1266	11
V about the N	7768	12
V about the A N	1895	12
V about a N	1220	12
V out of the N	7013	13
V off the N	5316	13
V N from N	4305	14
V N from the N	3223	14
V the N from the N	1771	14
V a N from the N	1042	14
V under the N	4250	15
V against the N	3953	15
V N for the N	4228	16
V among the N	3953	16
V N to N	3712	17
V the N to N	3288	17
V the N to the N	3060	17
V N to the N	1983	17
V a N to N	1972	17
V across the N	3584	18
V after the N	2650	18
V after a N	1016	18
V between N and N	1895	19
V N as a N	1434	19
V N as N	1135	19
V the N in the N	4152	20
V the N in the A N	1184	20
V N into N	1203	21
V the A N to N	1003	21
V by the N	53342	22

V by A N	28857	22
V by N N	24191	22
V by the A N	16696	22
V by a N	12199	22
V that S	87245	23

References

Armstrong, B. C., Frost, R., and Christiansen, M. H. (2017). ‘The long road of statistical learning research: past, present and future.’ Philos. Trans. R. Soc. B Biol. Sci. 372(1711).

Baker, M. (2003). Lexical Categories: Verbs, Nouns and Adjectives. Cambridge: Cambridge University Press.

Bod, R. (1998). Beyond Grammar: An experience-based theory of language. Stanford, CA: Center for the Study of Language and Information.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MITPress.

Croft, W. (2001). Radical construction grammar: Syntactic theory in typological perspective. Oxford: Oxford University Press.

Ellis, N. C. (1996). Sequencing in SLA: phonological memory, chunking and points of order. Studies in Second Language Acquisition, 18, pp. 91–126.

Ellis, N. C. (2002). Frequency effects in language processing. Studies in Second Language Acquisition, 24(2), pp. 143–188.

Jackendoff, R. (1977). X-bar Syntax: A Study of Phrase Structure. Cambridge, MA: MIT Press.

Hunston, S., and Francis, G. (1999). Pattern grammar. A corpus-driven approach to the lexical grammar of English. Amsterdam and Philadelphia: John Benjamins.

Leech, G. (1992). 100 million words of English: The British National Corpus (BNC). Language Research, 28(1), pp. 1–13.

Mahlberg, M. (2005). English General Nouns: A Corpus Theoretical Approach. Amsterdam/Philadelphia: John Benjamins.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2) pp. 313–330.

Marcus, G. F. (2001). The Algebraic Mind: Integrating Connectionism and Cognitive Science. Cambridge: MIT Press.

Renouf, A., and Sinclair, J. McH. (1991). Collocational frameworks in English. In: K. Aijmer – B. Altenberg (eds.): English corpus linguistics: Studies in the honour of Jan Svartvik, pp. 128–143. London: Longman.

Rothstein, S. (1983). The syntactic forms of predication. Cambridge, MA: MIT.

Rothstein, S. (2004). Structuring events: A study in the semantics of lexical aspect. Malden, MA: Blackwel.

Sinclair, J. M. (1994). Trust The Text. In: M. Coulthard (ed.), pp. 12–25.

Sinclair, J. (1996). The search for units of meaning. Textus, 9(1), pp. 75–106.

Trklja, A. (forthcoming). ‘Distributional properties of near synonyms in lexical domains: A formal and metric-based approach.’ Corpora.

Warren, M., and Leung, M. (2016). Do Collocational Frameworks have Local Grammars?, International. Journal of Corpus Linguistics, 21(1), pp. 1–27.

Williams, E. (1994). Thematic structure in syntax. Linguistic inquiry monographs, Vol. 23. Cambridge, MA: MIT press.