Towards an English constructicon using patterns and frames

Perek, Florent; Patten, Amanda

DOI:
10.1075/ijcl.00016.per

License:
Other (please specify with Rights Statement)

Document Version
Peer reviewed version

Citation for published version (Harvard):
Perek, F & Patten, A 2019, 'Towards an English constructicon using patterns and frames', *International Journal of Corpus Linguistics*, vol. 24, no. 3, pp. 354–384. https://doi.org/10.1075/ijcl.00016.per

Towards an English Constructicon using patterns and frames*

Florent Perek and Amanda L. Patten
University of Birmingham

Recent research in construction grammar has been marked by increasing efforts to create constructicons: detailed inventories of form-meaning pairs to describe the grammar of a given language, following the principles of construction grammar. This paper describes proposals for building a new constructicon of English, based on the combination of the COBUILD grammar patterns and the semantic frames of FrameNet. In this case study, the valency information from FrameNet was automatically matched to the verb patterns of COBUILD, in order to identify the frames that each pattern is associated with. We find that the automatic procedure must be complemented by a good deal of manual annotation. We examine the "V that" pattern in particular, illustrating how the frame information can be used to describe this pattern in terms of constructions.

Keywords: constructicon, COBUILD, FrameNet, construction grammar, lexicogrammar

1. Introduction

One of the central tenets of constructional approaches to grammar (Fried & Östman, 2004; Goldberg, 1995, 2006) is that grammatical knowledge is better described as a vast structured inventory of direct pairings of form with meaning, called constructions, as opposed to a system of abstract rules strictly separated from the lexical items inserted in them (e.g., Chomsky, 1965). Much of the research that laid the groundwork for construction grammar (e.g. Fillmore et al., 1988; Kay & Fillmore, 1999) was focused on idiosyncratic expressions

* We would like to thank audiences at AAAL 2018 and ICCG 2018, and participants of the "Multilingual Frame Semantics" Workshop in 2018, where earlier versions of this work were presented. We are grateful to two anonymous reviewers for their useful comments on the first draft of this article; any remaining shortcomings are of course our own.

which, with their irregular syntax and/or non-compositional semantics, are typically challenging to earlier approaches to syntax. However, the approach was designed from the start to be able to account not just for the periphery but also 'core' areas of grammar, such as the syntactic realisation of the argument of verbs (Goldberg, 1995). Construction grammarians commit to the view that the *entirety* of grammar consists of a structured inventory of constructions linked by relations of various kinds, i.e., a *constructicon*.

In recent years, the term 'constructicon' has been given a more practical use (alongside its theoretical and psychological sense) to refer to databases of fully described constructions in a given language, typically in electronic form. This new field of 'constructicography', the lexicography of constructions (Lyngfelt et al., 2018), was largely spearheaded by the FrameNet Constructicon for English (Fillmore et al., 2012), soon followed by similar projects in other languages (e.g., Lyngfelt et al., 2012; Ohara, 2013; Torrent et al., 2014). One major appeal of constructicon projects is their potential to bestow construction grammar with wide-scope empirical validation. To date, the construction grammar literature largely consists of individual studies of separate constructions or small families of constructions, with such hallmark examples as the caused-motion construction or the *way*-construction repeatedly cited as evidence for the approach. Comparatively little progress has been made in expanding the empirical coverage of construction grammar beyond isolated pockets of constructions. Equally importantly, the building of constructicons is likely to enable construction grammar to better present itself as a serious alternative to other grammatical frameworks in various areas of applied linguistics, such as designing methods and materials for language teaching, or creating tools for automatic language processing.

Yet, the ideal of describing grammar as constructions *in toto* has been reached to various extents by current constructicon projects. Some of them tend to focus more on phraseology and idiosyncratic constructions than common and fully regular patterns, and/or still have limited coverage. For English in particular, the FrameNet Constructicon only contains 73 entries as of 13 September 2018 (cf. http://www1.icsi.berkeley.edu/~hsato/cxn00/21colorTag/), and fewer still are in a final state of completion. That said, because it was designed as a complement to the FrameNet database, the FrameNet Constructicon is mostly meant to capture aspects of grammatical behaviour that are not covered by the FrameNet lexical entries. As a result, it currently consists of a diverse collection of idiosyncratic constructions (such as "be_recip", e.g. *Sue is good friends with Bob,* cf. Lee-Goldman & Petruck, 2018) and "non-canonical" syntactic structures (e.g.,

ellipsis constructions such as gapping, *I had a salad and Mary a burger*), but it does not cover more general constructions, such as the ditransitive (e.g., *You gave me a book*) or other argument structure constructions.

Against this backdrop, this paper describes and explores a way in which a more comprehensive English Constructicon can be built efficiently; for this case study, we focus in particular on constructions of the verb. Our approach consists in merging two existing corpus-based resources: (i) the COBUILD Grammar Patterns (Francis et al., 1996, 1998, Hunston, this volume), as a source of lexicogrammatical information, and (ii) the FrameNet database (Ruppenhofer et al., 2016), as a framework for semantic description, introduced in Section 2 and Section 3 respectively. In Section 4, we show that these two resources are very complementary, which motivates the idea of combining them to provide the basis for a constructicon of English verbs. In Section 5, we introduce an automatic procedure to match the COBUILD pattern with the FrameNet frames, and we examine the output of this procedure, pointing out that it will have to be supplemented by a good deal of manual annotation. In Section 6, we illustrate how the manually corrected matching of a pattern with semantic frames can be used to describe this pattern in terms of constructions at various level of generality, focusing on the case of the "V that" pattern in particular.

2. The COBUILD Grammar Patterns

The COBUILD project (Collins Birmingham University International Language Database) was a lexicographic enterprise started in the 1980s by John Sinclair at the University of Birmingham, in collaboration with Collins Publishers. COBUILD's innovative aim at the time was to design dictionaries entirely from authentic corpus data. Its main output was the Collins COBUILD English dictionary, first published in 1987 and based on the Bank of English corpus collected for this purpose. The dictionary was soon followed by other reference works, such as dictionaries of phrasal verbs and idioms, and a reference grammar.

One of the new key insights gained by the COBUILD project was that a word is better described not just in terms of a general semantic definition, but more importantly with reference to its typical uses. In particular, the COBUILD entries include the syntactic frames, or "patterns" that each word can occur in. This idea was further taken up by proposals to compile a pattern grammar of English (Francis, 1993; Hunston & Francis, 2000), which gave birth to the COBUILD Grammar Pattern series (Francis et al., 1996, 1998). This two-volume collection catalogues all the patterns mentioned in the COBUILD dictionary entries for verbs

(in Volume 1: Francis et al., 1996), nouns, and adjectives (in Volume 2: Francis et al., 1998), and lists all the lexical items attested in each pattern.

"V n of n" is an example of a verb pattern (Francis et al., 1996: 399-401). As is evident here, the COBUILD patterns are described using a simple, 'flat' notation that is easy for non-expert users to interpret. In more traditional grammatical terms, the "V n of n" pattern correspond to verbs followed by a noun phrase (direct object) and a prepositional phrase headed by *of*. The items in each pattern are further sorted into meaning groups containing semantically similar words; meaning groups are named after one or more of its most typical members. By way of illustration, the pattern "V n of n" has three meaning groups:

(i) The 'rob' and 'free' group, "concerned with taking something away from someone either physically or metaphorically" (Francis et al., 1996: 399), contains 24 verbs, e.g., *cure*, *deprive*, *relieve*;

(ii) The 'inform' group, "concerned with talking or writing, for example giving someone information, warning someone about something, or reminding someone of something" (ibid.), contains 11 verbs, e.g., *assure*, *notify*, *remind*;

(iii) The 'acquit' and 'convict' group, "concerned with declaring or thinking that someone has or has not committed a crime" (ibid.), contains 5 verbs, e.g., *accuse*, *suspect*.

In addition to these groups, Francis et al. (ibid.) also list 11 other miscellaneous verbs that do not seem to share any particular aspect of meaning. According to Hunston & Su (2017), the COBUILD Grammar Patterns comprise about 200 patterns of verbs, nouns, and adjectives, covering an estimated 1,000 meaning groups.

3. FrameNet

Started in 1997, FrameNet (Ruppenhofer et al., 2016) is another lexicographic project that aims to describe the lexicon of English in terms of the theory of frame semantics. According to Fillmore & Atkins (1992: 76-77), "a word meaning can be understood only with reference to a structured background of experience, beliefs or practices, constituting a kind of conceptual prerequisite for understanding the meaning". In frame semantics, word meanings are grounded in conceptual structures called semantic frames, defined by Fillmore (1985:

223) as "some single coherent schematization of experience or knowledge". For example, the word *revenge* refers to a particular action which presupposes a certain amount of background information, yet without directly asserting it: namely, that some wrongdoing has been committed, and that the agent of the revenge acts in retaliation to this wrongdoing.

As of 2nd August 2018, the FrameNet project, hosted at the International Computer Science Institute (ICSI) in Berkeley, lists 1,224 frames describing the meaning of 13,640 nouns, verbs, adjectives, and a few other word types (including some multi-word expressions). Each frame is stored with a list of lexical units, i.e., words that evoke this frame, and a definition of the event or situation that it captures, which makes reference to particular actors and props in the scene, called frame elements (FE). FEs are reminiscent of the traditional notion of semantic roles, with the proviso that they are by definition frame-specific. They can be referred to by sentence constituents that occur with lexical units in all of their uses. For example, the `Lending`[1] frame is defined as referring to an event in which "The Lender gives the Theme to the Borrower with the expectation that the Borrower will return the Theme to the Lender after a Duration of time". In this definition, Lender, Theme, Borrower, and Duration are frame elements. The `Lending` frame contains the lexical units *lend* and *loan* (both as noun and verb). Example (1) below (from FrameNet) illustrates the verb *lend* with the FEs of the `Lending` frame marked in square brackets.

(1) [I $_{\text{Lender}}$] lent [my girlfriend $_{\text{Borrower}}$] [my car $_{\text{Theme}}$] [for the weekend $_{\text{Duration}}$].

A distinction is made in FrameNet between core and non-core frame elements. Core FEs refer to aspects that are central to the frame and obligatorily expressed or implied in all its uses, and are typically realized as major clause elements such as subject, object etc. Non-core FEs correspond to more peripheral and typically optional information, which is often realized as adverbials and modifiers. In the `Lending` frame, LENDER, BORROWER, and THEME are core FEs; DURATION is a non-core FE. The non-core FEs of `Lending` also include other kinds of less central information such as MANNER, PLACE, PURPOSE, and TIME.

In lexicographic terms, the frames of FrameNet do not correspond to definitions as would be found in a traditional dictionary; rather, they are a higher level of lexicographic description that captures shared aspects of the conceptual import of words. Another important way in which FrameNet goes beyond a traditional dictionary is that it contains information about relations between frames. For instance, inheritance relations relate frames in a

taxonomy, e.g., the `Lending` frame inherits from the `Giving` frame (evoked by such lexical units as *give*, *donate*, and *gift*), marking that lending is a kind of giving, albeit with the added notion that the theme is supposed to be returned to the giver at some later point in time. `Supply` (evoked by *equip*, *provide*, and *supply*) is another frame that inherits from `Giving`; it adds the notion that the giving occurs in order for the recipient to fulfil a particular need or purpose. Another relation is "using", which marks that a frame draws on the conceptual content of another frame, without actually being a subtype of that frame (as is the case with the inheritance relation). For instance, the `Offering` frame describes an event in which someone makes something available for someone else to receive; hence, while it presupposes the idea of giving, it is not an actual instance of giving. Therefore, the `Offering` frame uses the `Giving` frame, rather than inherits from it. In some cases, the "using" relation can be seen as a form of partial inheritance, although, as one anonymous reviewer clarified for us, it more generally is a somewhat loose label that covers relations between frames that seem important (enough) to recognize but which cannot be categorized as instances of the other better-defined relation types. Finally, the "perspective" relation marks that a certain frame is a particular construal of another, more generic frame, usually one in which one frame element is made a more focal participant. For instance, the `Giving` frame describes a certain perspective on the more general `Transfer` frame, namely that of the GIVER. The `Receiving` frame offers a different perspective on the same frame: that of the RECIPIENT. Frame-to-frame relations between the frames mentioned above are summarized in Figure 1. The FrameNet database contains a few other types of frame-to-frame relations, but it goes beyond the scope of this paper to describe them all.

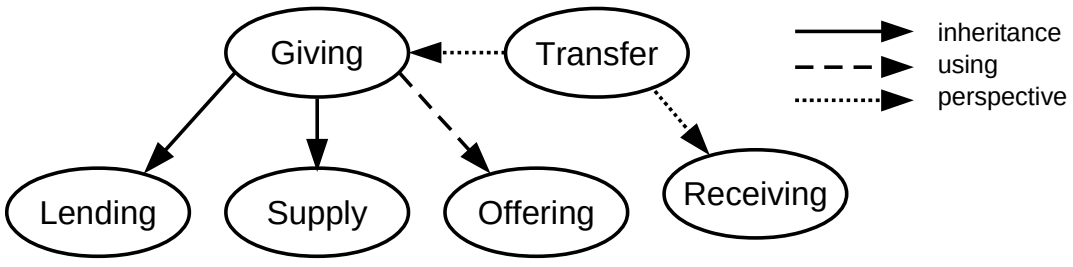

Figure 1: Frame-to-frame relations in FrameNet

FrameNet is less obviously corpus-based than COBUILD, as it largely relies on the semantic intuitions of the compilers instead of directly following from corpus data. However, these

intuitions are largely informed by corpus data, mostly adduced from the BNC. Corpus evidence is used at every step of frame development, notably to identify frame elements (through their linguistic realisations), to distinguish between core and non-core frame elements, and to find the evidence needed to add lexical units to existing frames. Likewise, most examples used as illustration in the frame descriptions come from corpora. More importantly, the lexical entries of lexical units include sets of annotated corpus examples which provide information as to how frame elements are grammatically encoded in actual language use. The annotated examples augment the FrameNet database with information on argument realization: the lexical entries contain lists of frame element configurations, i.e., sets of FEs simultaneously realised in uses of the LU, and information about the syntactic realisations of these FEs in the annotated examples. For example, the lexical entry for the LU *loan* in the `Lending` frame contains the following two examples (among others), taken from the BNC, in which the FEs LENDER, BORROWER, and THEME are realised respectively as an external NP and two object NPs in (2), and as an external NP, a prepositional phrase headed by *to*, and an object NP in (3). From these attestations, the corresponding two valency patterns are added to the lexical entry of *loan*.

(2) They asked [the CIA $_{\text{Lender}}$] to loan [them $_{\text{Borrower}}$] [an agent $_{\text{Theme}}$].

(3) [He $_{\text{Lender}}$] had loaned [five thousand pounds $_{\text{Theme}}$] [to Phillip Wreck $_{\text{Borrower}}$].

This effectively makes FrameNet a source of lexicogrammatical information. Thus, while FrameNet takes a different perspective from the COBUILD Grammar Patterns and although it is still primarily a semantic database, the addition of valency information to FrameNet makes the two resources more similar. It should be acknowledged, however, that the coverage of FrameNet, in terms of the number of lexical items and the range of valency patterns described for each, is still far below that of the COBUILD Grammar Patterns (as will become apparent in Section 5. That said, their different scope and approach make the two resources quite complementary, as argued in the next section.

4. Two complementary resources

The main focus of COBUILD Grammar Patterns is on lexicogrammar: the aim is to document what patterns are available for English verbs, nouns, and adjectives, and what

words can be used in them. Meaning, on the other hand, is secondary. Beyond the indication of the relevant lexical senses in the COBUILD English dictionary (Sinclair et al., 1995), the meaning of lexical items is not characterised. As mentioned earlier, words are sorted into meaning groups, but this is merely offered as an ad-hoc way to organize the pattern entries. While these meaning groups are very useful to readers for them to assimilate the pattern entries and make generalisations from the lexical distribution of patterns, this classification is intuitively inferred and not based on a particular theory of meaning, or any prior characterization of word meaning. The limitations of such an intuitive approach are acknowledged by Hunston & Francis (2000: 86):

> [I]t must be conceded that the division into meaning groups, such as that given above for V *of* n, is not achieved through anything other than the intuition of the person looking at the list. Different researchers or teachers may well come up with a different set of meaning groups, and even the same observer may on different occasions and for different purposes wish to propose different groups. In Francis et al. (1996), for example, the meaning groups given in the section 'V of n' (p. 211–214) are similar to but not the same as the groups suggested above. This is largely because the verbs in the groups are not synonyms of each other, but simply share an aspect of meaning, and different observers would prioritise different aspects. On the other hand, any observer could identify some meaning groups, and it is probable that most observers would arrive at meaning groups that were very similar to each other.

The potential subjectivity of meaning groups also means that they might not be based in the same exact criteria in different patterns. Accordingly, there is no systematic way to relate the meaning groups (for instance, judging how similar they are), other than that offered by intuitions. In sum, meaning is rather secondary in the COBUILD Grammar Patterns.

FrameNet can be seen to have the opposite organization: the main focus is word meaning, i.e. what semantic frames are required to describe the lexicon of English, how they are related, and by what words these frames are evoked. Lexicogrammatical information, however, is merely an addendum to the lexical units of FrameNet: it is derived from the annotated examples, which are only added after a frame is created and lexical units added to it. In fact, annotated examples were not always part of the database: while the oldest frames recorded in FrameNet date back from 2001, annotated examples were only added from 2003 onwards. Valency information in FrameNet is still piecemeal as of 2nd August 2018: only 62% of lexical units are claimed to have lexicographic annotation (cf.

https://framenet.icsi.berkeley.edu/fndrupal/current_status), and 24% (3,339 out of 13,640)[2] have no annotated examples at all (and hence no valency information).[3] In addition, the valency information derived from the annotated examples is necessarily limited to the range of grammatical properties exemplified by these examples. While Ruppenhofer et al. (2016: 9) claim that "the set of examples (approximately 20 per LU) illustrates **all** of the combinatorial possibilities of the lexical unit" (emphasis in the original), this will be limited by the particular corpora used by the project (mostly the BNC), and it is also not clear how many lexical units have indeed reached the stage of full coverage in the corpus, though this is not to deny the impressive annotation work that has already been carried out.

Another way in which FrameNet differs from the COBUILD Grammar Patterns is that there is no systematic inventory of all valency constructions listed in the lexical entries; currently, the only way to find them is to look up the lexical entries one by one. In this way, FrameNet is more comparable to a dictionary that *contains* pattern information in its lexical entries, such as the COBUILD English dictionary (albeit with less coverage).

Although they differ in their scope and approach, FrameNet and COBUILD can thus be seen to be complementary resources. FrameNet is based on sound semantic principles derived from a specific theory of word meaning (frame semantics), while the COBUILD Grammar Patterns lack such a strong semantic foundation. Conversely, the COBUILD Patterns contain a wealth of information about the combinatorial properties of a large number of English nouns, verbs, and adjectives, while FrameNet has not yet approached the question of valency in a comprehensive and systematic way. Hence, the two resources can benefit from each other in many ways. Since both provide valency information as part of their output, it should be possible to systematically compare and match this information, but so far, no attempt has been made at doing so.

The present research seeks to mend this gap, focusing in particular on verbs. We propose that FrameNet can serve as a semantic component for the COBUILD Grammar Patterns, while the Patterns can be used to complement the lexicogrammatical information of FrameNet. To achieve this, the verbs listed in the COBUILD Patterns entries must first be matched to the corresponding lexical units in FrameNet. Matching the COBUILD Patterns with FrameNet will form the basis for building the English Constructicon, a database of English constructions in the construction grammar sense (Goldberg, 1995), i.e., pairings of a grammatical form with an abstract meaning that describe how certain sets of words combine with particular syntactic structures. For example, Goldberg (1995) describes the ditransitive construction, which pairs a clause consisting of a subject, a verb, and two post-verbal noun

phrases, with an abstract meaning of transfer, allowing such verbs as *give*, *hand*, and *sell* (among many others) to occur in this syntactic pattern, and explaining why certain other verbs, such as *buy* and *bake*, take on a meaning of intended transfer when they are used in this construction. Importantly, constructions are not limited to abstract syntactic patterns like the ditransitive construction: they aim to capture the grammar of a language in its entirety, including phraseological units of intermediate status between grammar and lexicon, of varying sizes and complexity, such as idiomatic expressions (like *pull one's leg* or *a storm in a teacup*), syntactic idioms ("Verb *one's way* PP", e.g. *He typed his way to a promotion*, "Verb *the hell out of* NP", e.g. *You entertained the hell out of everyone last night*), and semi-fixed phrases and collocations (*give someone a hand*, *take a bath*). In Goldberg's (2006: 18) often-quoted words, "it's constructions *all the way down*" (emphasis added).

At least at first glance, the COBUILD patterns seem very similar to constructions, as also noted by Hunston & Su (2017), since they are conceptualised as single coherent grammatical units posited somewhat independently of the words they combine with, and consist of fixed parts and open slots (see also Hunston, this volume). Hence, the COBUILD patterns can provide the basis for a more comprehensive constructicon of English, focusing in particular on the grammar of verbs, nouns, and adjectives. Such a database would nicely complement the FrameNet Constructicon project (Fillmore et al., 2012), as the latter tends to focus on idiosyncratic constructions rather than the common, regular constructions exemplified by the COBUILD patterns. Contrary to constructions, the identification of patterns is not semantically motivated, and they are not explicitly paired with meaning or semantic role descriptors, although the semantic groupings of verbs found in a pattern's entry do provide an indication of the kind of verbal semantics that the construction tends to convey. FrameNet can be used to provide the semantic component that is missing in patterns, and frame-to-frame relations and similarities between frames can be drawn upon to identify the meaning of constructions at various levels of generality.

In the remainder of this paper, we explore two main questions: (i) how the COBUILD patterns can be matched to FrameNet, (ii) how this matching can be translated into a description of constructions corresponding to each pattern. To answer the former question, we describe and evaluate a method to automatically match the electronic version of the two resources in Section 5. We address the latter question in Section 6, where we focus on the pattern "V that". Using the results of the automatic procedure complemented with manual annotation where necessary, we show how different constructions corresponding to this pattern can be identified and described, drawing on the semantic information from FrameNet.

5. Merging the two resources

Both FrameNet and the COBUILD Grammar Patterns were available to us in machine-readable versions. FrameNet was designed as an electronic resource from the start, and is distributed under a Creative Commons license that allows unrestricted access and use. An XML version of the database can be requested from the FrameNet website (https://framenet.icsi.berkeley.edu/fndrupal/framenet_request_data). HarperCollins Publishers kindly provided us with an electronic copy of the COBUILD Grammar Patterns in XML format corresponding to the revised version published in 2018 on their online dictionaries platform (https://grammar.collinsdictionary.com/grammar). Although the Patterns XML file is not a structured database in the same way that FrameNet is, but rather an XML-ised version of the Grammar Patterns book complete with information about layout and sectioning, it still lends itself well to automatic processing, which enables it to be automatically matched to the FrameNet database. In this section, we describe how FrameNet and the COBUILD Grammar Patterns were combined, using these two electronic resources. For the sake of simplicity and because of time constraints, this initial research is restricted to the patterns of verbs (Francis et al., 1996), but future work is planned to carry out a similar procedure on the patterns of nouns and adjectives (Francis et al., 1998).

We implemented an automatic procedure as a computer program written in Java, which retrieved all verbs listed in each pattern in Francis et al. (1996) from the XML version, and looked up every verb in the FrameNet database. If the verb was found, this returned one or more lexical units, each evoking a different frame. In each lexical unit, the valency patterns (if any) derived from the annotated examples and describing the syntactic realization of FEs were consulted and searched for any match with the COBUILD pattern under consideration. Only core FEs were considered in the matching of lexical units to patterns, because non-core FEs tend to include semantic roles that are traditionally considered adjuncts, such as place, time, and manner, and such roles would normally not be part of the Grammar Patterns of verbs, as they refer to generic and usually optional information that can be added to many verbs, if not all. If the pattern from COBUILD was found among the valency patterns of the core FEs of the lexical unit, the LU was mapped onto the verb entry of that pattern; several LUs can match a single entry if the relevant evidence is found. In addition, information about what frame element is mapped onto each slot of the pattern was retrieved.

Some limitations had to be put on the automatic matching procedure. First, all multi-word entries from COBUILD, such as particle verbs (e.g. *point out*, *pick up*) and combinations of a verb with typically co-occurring words like negative adverbs or modals were ignored in the matching process (e.g. *(never) dreamed* as in *I never dreamed that I would be able to afford a home here*, or *(cannot) bear* as in *I can't bear people who make judgements and label me*; examples are from the online Collins COBUILD dictionary at https://www.collinsdictionary.com/), since they are harder to extract from the annotated examples in FrameNet. The particle of the particle verbs listed in COBUILD is sometimes considered a frame element, sometimes part of the verb, and sometimes annotated differently. For example, for *send out* (e.g., a letter) in the "V n" pattern, the particle *out* is annotated as the FE GOAL in the `Sending` frame, and for *prattle on* in the "V about n", *on* is annotated on a separate "Aspect" layer (cf. Ruppenhofer et al. 2016: 42). Other multi-word elements are simply hard, if possible at all, to reliably parse out from the annotated sentences; identifying such expressions is likely to require human intervention, especially to avoid false hits. Second, some patterns could not be reliably matched to FrameNet due to the annotation scheme, specifically patterns containing a 'dummy' *it* (e.g. "V *it* adj *that*", *I find it surprising that he came*) or existential *there* (e.g., *There remain major differences between the two groups*). Since the words *it* and *there* play a purely grammatical role and do not correspond to any frame element, they are not reported in the valency sets listed with each lexical unit.[4] For the sake of simplicity and because automatic matching is unlikely to produce reliable results for these patterns, we decided to ignore these patterns altogether for the purpose of this case study, although they could be brought back into consideration in future work.

These limitations left us with 78 matchable patterns, listed in Table 1 below. For each pattern, Table 1 mentions the number of distinct verb types listed in the COBUILD verb patterns XML file, as well as how many of these verbs were matched to at least one lexical unit from FrameNet, and the corresponding matching rate.

Table 1: Automatic matching information for each of the COBUILD verb patterns, ordered by matching rate

Pattern	Total verbs	Matched verbs	Matching rate	Pattern	Total verbs	Matched verbs	Matching rate
V amount	58	49	84%	V n to-inf	175	54	31%
V n	447	351	79%	V n *on* n	122	37	30%
V wh-to-inf	41	30	73%	V *against* n	79	23	29%
V	267	193	72%	V *into* n	136	39	29%
V *as if*/*as though*/*like*	10	7	70%	V *through* n	55	16	29%
V inf	3	2	67%	V n *off* n	28	8	29%
V n wh	9	6	67%	V n n	105	29	28%
V n wh-to-inf	9	6	67%	V n *from* n	172	49	28%
V *of* n	30	19	63%	V n *for* n	153	41	27%
V that	255	157	62%	V n *into* n	203	52	26%
V prep/adv	498	306	61%	V *with* n	174	43	25%
V *by* amount	28	16	57%	V adv	89	21	24%
V *with* quote	236	129	55%	V *at* n	164	39	24%
V prep	85	45	53%	V *out of* n	25	6	24%
V *about* n	103	55	53%	V n *at* n	37	9	24%
V to-inf	151	79	52%	V n *to* n	296	71	24%
V n *as* adj	46	24	52%	V n *with* n	256	62	24%
V wh	134	67	50%	V *as* n	34	6	18%
V *so*/*not*	10	5	50%	V n *over* n	17	3	18%
V *after* n	10	5	50%	pl-n V *together*	18	3	17%
V n prep/adv	496	241	49%	V *towards*/*toward* n	24	4	17%
V -ing	92	43	47%	V n *against* n	54	9	17%
V *in* n	104	48	46%	V n *out of* n	72	12	17%
V n that	26	12	46%	V n *into* -ing	78	12	15%
V n *about* n	56	25	45%	V n *onto* n	33	5	15%
V *between* pl-n	18	8	44%	V *as to* wh	14	2	14%
V n *of* n	45	20	44%	V n adj	103	14	14%
V *in favour of* n	10	4	40%	V *onto* n	16	2	13%
V *from* n	112	44	39%	V n *towards*/ *toward* n	8	1	13%
V *over* n	67	26	39%	V n inf	11	1	9%
V *across* n	8	3	38%	V adj	192	16	8%
V *around*/*round* n	21	8	38%	V pl-n *with together*	60	5	8%
V *to* n	210	79	38%	V n *as to* wh	24	2	8%
V pl-n	50	18	36%	V n *between*/ *among* pl-n	13	1	8%
V n *as* n	122	44	36%	V n *by* n	29	2	7%
V *on* n	195	69	35%	V n -ing	58	2	3%
V n *with* adv	24	8	33%	V *as* adj	5	0	0%
V n *in* n	156	52	33%	V *under* n	4	0	0%
V *for* n	188	59	31%	V n *with* quote	6	0	0%

The automatic matching procedure of these 78 patterns with FrameNet met with limited success. Overall, only 40.5% of the verbs listed in the COBUILD Grammar Patterns entries could be matched to at least one lexical unit in FrameNet (3,063 out of 7,572). However, as can be seen in Table 1, the matching rate is highly variable from pattern to pattern, ranging

from a maximum of 84% all the way down to 0% (for three patterns with very few verbs). Figure 2 below visually shows the distribution of matching rates across patterns. The top chart presents the matching rates as a bar plot, with one bar for each pattern, while the bottom chart plots the probability density function, i.e., how likely every value of the matching rate is according to the matching rates found in the distribution. The lower quartile, median, and upper quartile of the distribution are marked by vertical dashed lines on the plot, at 16.9%, 31.1%, and 49.6% respectively. In other words, 25% of the patterns have less than 16.9% matches, only another 25% have 46.6% or more matches, and 50% patterns have between 16.9 and 46.6% matches. Most of the distribution thus occupies the lower range of the scale.

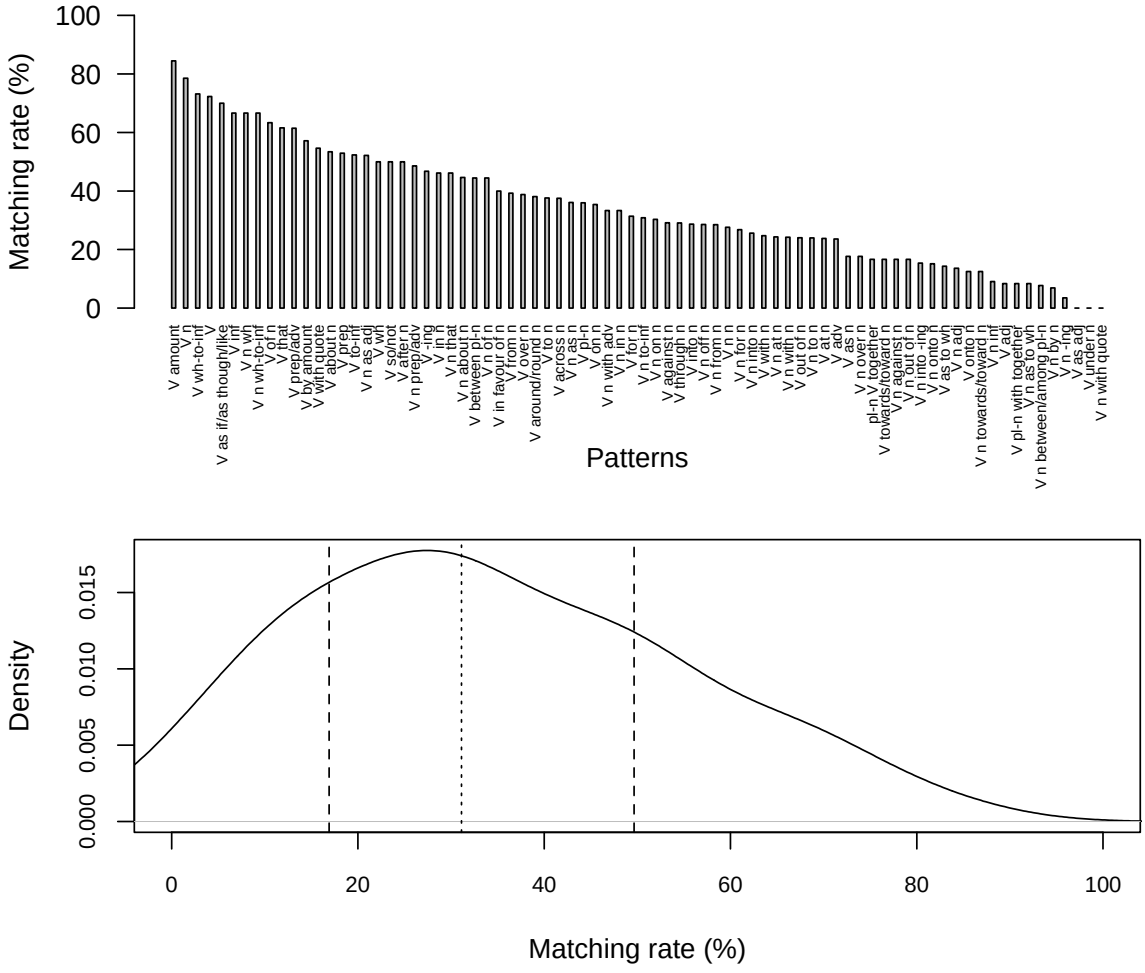

Figure 2: Distribution of matching rates across patterns

In sum, most patterns receive a rather underwhelming matching rate from the automatic procedure. This is primarily explained by a lack of coverage in FrameNet: many verbs remained unmatched because they were not found in FrameNet, or were found but not with

any valency realization information that corresponded to the relevant pattern. Another reason why some patterns were poorly matched is that for some lexical units, one of the positions of the patterns realizes a non-core frame element, but since non-core frame elements are ignored in the annotated examples, they cannot be matched to the pattern. For instance, when the lexical unit *die* in the `Death` frame occurs in the pattern "V *of* n" (e.g., *He died of pneumonia*), the *of*-phrase realizes the frame element EXPLANATION, which is non-core. While it is perfectly understandable, from a semantic point of view, that EXPLANATION is a non-core frame element of `Death`, it is not immediately obvious why ADDRESSEE is a non-core FE of the `Communication` frame and other related frames (e.g., `Communication_noise`, `Statement`). While this is likely the result of FrameNet overt coding criteria for coreness (cf. Ruppenhofer et al., 2016), this prevents LU like *communicate* and *explain* to be matched to such patterns as "V n *to* n", in which the *to*-phrase realizes the ADDRESSEE FE. This might be one reason why the matching rate of "V n to n" is so low (24%), especially if a similar problem occurs in other relevant frames. Thus, it seems that coreness is not a fully reliable criterion for use in the automatic matching procedure, and that applying it actually makes the program miss matches which would require manual intervention to be accurately identified.

It is clear, therefore, that a full, accurate matching of the COBUILD Patterns to FrameNet will necessitate a great deal of manual intervention to check the results of the automatic procedure, and more importantly, to identify lexical units for the entries that failed to be matched. In the simplest case, this involves adding the information that a lexical unit can be used in a pattern even if there is no annotated example to attest this use. If the verb is not listed as any lexical unit in FrameNet, or if none of the available LUs it is listed as correspond to the meaning(s) of the verb as it is used in the pattern, then an appropriate frame will have to be found; this essentially amounts to adding a new LU to a frame. It might be the case that no relevant frame can be found, or that the existing frames only provide a partial and ill-fitting match; in that case, an entirely new frame might have to be created and inserted into the FrameNet network. Interestingly, such manual intervention could thus also contribute information to FrameNet and expand its coverage.

Despite the manual annotation work that it would involve, fully matching the COBUILD Grammar Patterns with FrameNet would create a useful new hybrid resource. By examining the full range of frames associated with each pattern, it should be possible to map out the semantic domain of the pattern and identify different semantic areas that can be

generalized over; these generalisations can in turn be interpreted as the semantic side of one or more constructions. In the next section, we demonstrate this method on the "V that" pattern.

6. Towards the English Constructicon: a case study of the "V that" construction

In this section, we illustrate how the COBUILD Grammar Patterns can be systematically turned into form-meaning pairs, aka constructions in the construction grammar sense, drawing on the semantic information contained in FrameNet when matched to a pattern's lexical entries. To do so, we use the case of the "V that" pattern (Francis et al., 1996: 97-104), consisting of a verb followed by a finite subordinate clause optionally introduced by *that*, as exemplified by (4) and (5) below (from Francis et al., 1996: 97).

(4) I agree that the project has possibilities.

(5) He said the country was unstable.

Our main motivation for choosing this pattern is that it received one of the highest matching rates in the automatic procedure described in the previous section, with 62% of its verb lemmas automatically matched to at least one LU in FrameNet. This limits the amount of manual intervention needed to fully match this pattern to frames. A secondary reason is that this pattern corresponds to a very common construction in English, which has, however, not received much attention in the construction grammar literature. In analysing the "V that" pattern as a form-meaning pair, our case study thus contributes to the literature on constructions, while following an as yet unique methodology in the field.

6.1 From patterns to networks of frames

As discussed earlier, the matching of the COBUILD Grammar Patterns to FrameNet cannot be fully done automatically, and must be complemented by manual annotation. The automatic procedure described in the previous section outputs, for each pattern, a list of all the verbs found in the COBUILD Grammar Patterns XML file, paired with one or more frames from the FrameNet database, if one or more relevant lexical units could be matched to the verb. The manual intervention on this data consists in checking the accuracy of the

automatically found information and manually supplement it with manually selected frames. The COBUILD Grammar Patterns are provided with information about the relevant lexical senses of each entry, as recorded in the *Collins COBUILD English Dictionary 2nd Edition* (Sinclair et al., 1995); indeed, not all senses of a lexical item are attested in a given pattern. When checking the "V that" dataset, we made sure that the lexical senses of each verb listed in the pattern were all appropriately accounted for by the FrameNet frames matched to the verbs.[5]

357 different lexical units were found to correspond to the "V that" pattern after manual annotation.

Table **2** below summarises how these lexical units were identified. 226 were accurately identified by the automatic procedure and kept after manual checking. Four more LUs were wrongly identified due to errors in the FrameNet annotations, and thus removed from the dataset.[6] Three LUs were found not to actually correspond to any of the known senses of the verb and thus also removed.[7] The remaining 131 LUs correspond to lexical entries that were not matched by the automatic procedure, and thus had to be found manually. Of these, 30 are LUs that exist in FrameNet but could not be matched automatically due to the lack of a relevant annotated example (some of these LUs have, in fact, no examples at all). We made sure that the verbs could be used in the "V that" pattern with the relevant sense, mostly relying on our grammatical and semantic intuitions, and occasionally by consulting corpus evidence. Finding these LUs was facilitated by the fact that the automatic procedure also provided a list of all frames evoked by each verb (regardless of the annotated examples). 98 LUs were not found in FrameNet and had to be created by finding one or more frame(s) that appropriately matched the lexical sense(s) of the verb listed in the pattern's entries. While many cases were rather obvious (e.g., *email* in `Communication_means`), for others it was more difficult to identify the right frame. In some cases, it was felt that the presumably best-fitting frame was still not fully adequate, and that a more specific or slightly different frame not found in FrameNet would be preferable. For example, *dream* was assigned to the `Cogitation` frame, although its more specific meaning of "experience thoughts and sensations during sleep" would probably call for a frame of its own. However, we refrained from creating new frames,[8] mostly to maintain the compatibility of our study with the "official" version of FrameNet. Besides, as explained by Ruppenhofer et al. (2016), creating frames is no trivial task, especially compared to adding LUs to an existing frame.

Finally, in three cases, a LU was automatically found in FrameNet but was changed to another frame that was considered to correspond to the meaning of the verb more closely.[9]

Table 2: Summary of the lexical unit annotations on the "V that" pattern dataset.

Status	**# of LUs**	
LU matched to the pattern automatically and checked manually	226	63.2%
No automatic match; LU found in FrameNet and manually matched to the pattern	98	27.53%
No automatic match; LU not found in FrameNet and manually added	30	8.43%
LU matched to the pattern automatically but reassigned to a different frame	3	0.84%
Total	**357**	

The next step is to use the list of LUs used in the "V that" pattern, and in particular the range of frames they evoke, to describe the pattern in terms of constructions. In a usage-based approach such as most versions of construction grammar, constructional meaning is taken to be an abstraction over the meaning of all tokens of a construction (e.g., Bybee, 2010, 2013; Goldberg, 2006; Langacker, 2000). For verb constructions such as "V that", this is taken to correspond in large part to the meaning of verbs occurring in the construction (or more specifically, the meaning verbs take in this particular grammatical environment), since they contribute a significant share of the overall meaning of the clause (Croft, 2003; Healy & Miller, 1970; Goldberg et al., 2004; Perek & Lemmens, 2010; Perek, 2015). For example, the ditransitive construction is associated with a general meaning of transfer because all of its uses convey the notion of transfer in one form or another, and correspondingly, the construction occurs with verbs that typically refer to transfer, such as *give*, *bring*, and *send*.

Our approach consists in applying this idea to the frame semantic data derived for the "V that" pattern from FrameNet. A construction in this approach is defined as a pairing of a pattern and a generalisation over the semantic frames evoked by verbs occurring in the pattern. Frame-to-frame relations, as mentioned in Section 4, are instrumental in positing generalisations between frames in a systematic way. We thus checked each frame in the dataset for the other frames with which it was related in FrameNet, and used this information to build a network containing as many of the frames in the dataset as possible. Occasionally, we included frames that are not in the dataset if they could serve to provide a link between

frames. The nodes of the network can be taken to correspond to different levels of generalisation in the construction, defining sub-constructions of varying semantic granularity.

At the end of this process, not all frames could be related into a single network; rather, we find sets of related frames, each with their own common semantic denominator, and a few "orphans" not related to any other frame. This indicates that the "V that" pattern likely corresponds to more than one construction,[16] which is in line with a growing body of literature arguing that it might not always be possible to define a semantic generalisation that encompasses all instances of a formal pattern, and that lower-level generalisations are on balance more important than maximally general constructions (Boas, 2003, 2008; Bybee, 2010; Bybee & Eddington, 2006; Iwata, 2008; Langacker, 2000; Perek, 2014, 2015). Alternatively, we could posit a purely formal generalisation as the highest-level "V that" construction, i.e., a construction with a form but with no meaning, as some versions of construction grammar do allow (cf. Fillmore 1999). However, a number of lower-level constructions that do convey meaning are still needed to capture the "V that" pattern in constructional terms.

In the next sections, these different networks of frames, and the constructions corresponding to them, are discussed in turn.

6.2 The `Communication` "V that" construction

We first discuss the largest network, which also covers the highest number of lexical items. There is one frame in particular to which all other frames are ultimately related, through inheritance, using, or (in one case) perspective relations: the `Communication` frame, which thus stand as the highest-level semantic generalisation in the network. Accordingly, the other frames all describe some form or use of communication (mostly, but not exclusively, verbal). These frames and the relations between them are diagrammed in Figure 3. In this and subsequent similar diagrams, frame-to-frame relations are represented by arrows. The numbers in brackets next to the frame names correspond to the number of lexical units of each frame that occur in the "V that" pattern. There is only one frame in Figure 3 that does not have any LU attested in "V that", the `Judgment_communication` frame, which is included in the diagram because it serves to link two other frames.

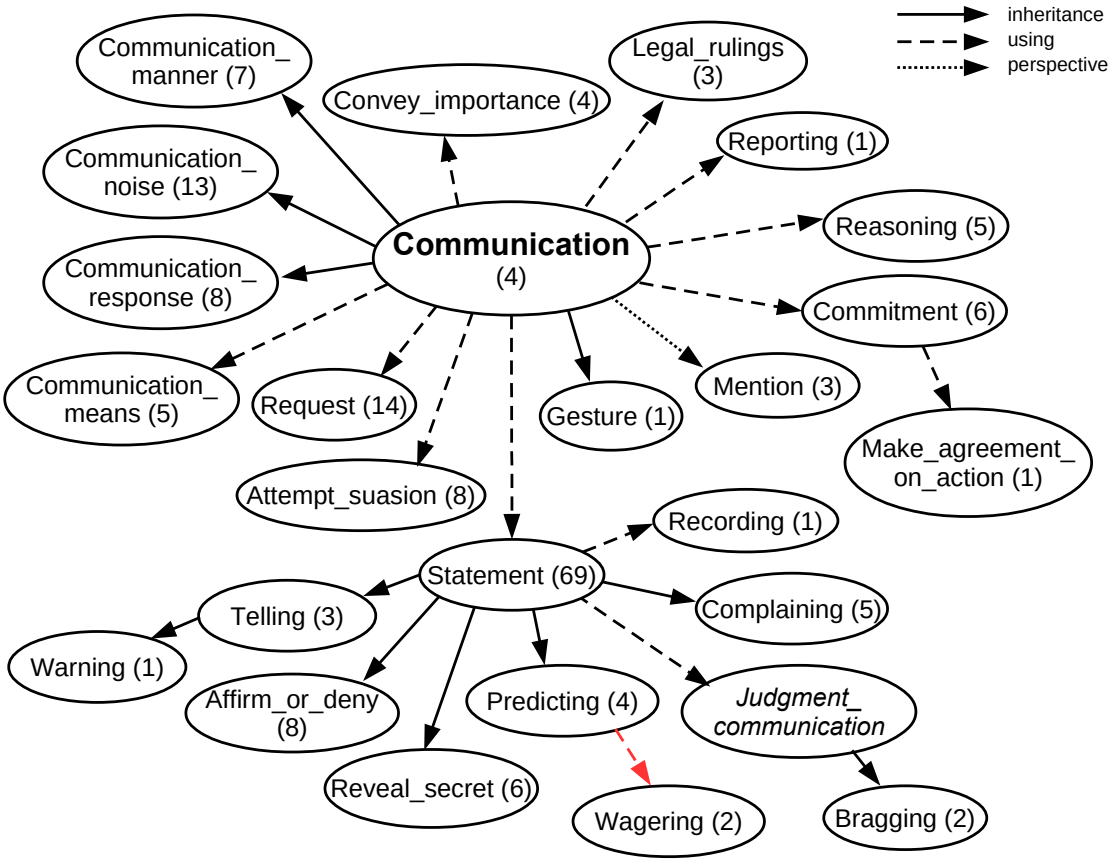

Figure 3: The `Communication` network of frames in the "V that" pattern

All frame-to-frame relations were extracted from FrameNet, except one, marked in red: the using relation between `Predicting` and `Wagering`. We suggest that this relation should be added on the grounds that `Wagering` (e.g., *bet*, *wager*) involves the GAMBLER making and communicating a form of prediction to which they commit an ASSET. This allows `Wagering` to be included in the network, in line with the fact that it is part of the 'say' group in Francis et al. (1996), along with many other verbs in this network.

An important aspect of frame-to-frame relations is that they involve links between the FEs of the frames they relate, marking which FEs are carried over from one frame to the other, and which ones are unique to only one of the frames. All of the frames in Figure 3 share at least two FEs linked by frame-to-frame relations that can ultimately be traced back to the `Communication` frame: COMMUNICATOR, "The sentient entity that uses language in the written or spoken modality to convey a MESSAGE to the ADDRESSEE",[10] and MESSAGE, "a proposition or set of propositions that the COMMUNICATOR wants the ADDRESSEE to believe or take for granted", respectively realised as a subject NP and the *that*-clause of the "V that"

pattern. There are only two exceptions to this claim which are arguably due to gaps in FrameNet. In the Commitment frame (e.g., *guarantee*, *pledge*, *threaten*), the SPEAKER FE should be linked to the COMMUNICATOR FE of Communication, since the agent of a commitment is also the person communicating it. Likewise, in the Legal_rulings frame (e.g., *decree*, *mandate*, *rule*), in which "An AUTHORITY with the power to make decisions hands down a FINDING over a question presented in a formal or informal CASE", the FEs linked to COMMUNICATOR and MESSAGE should be AUTHORITY and FINDING. The frames in Figure 3 also typically contain a MEDIUM FE, defined in the Communication frame as "The physical or abstract setting in which the MESSAGE is conveyed", which can also be encoded as the clause subject instead of the COMMUNICATOR.

On the basis of this data, we can posit a Communication "V that" construction, diagrammed in Figure 4 below, with the meaning of the Communication frame. More specifically, the construction imposes a certain "windowing of attention" (Talmy, 1996, 2000) on the frame, which gives prominence to the COMMUNICATOR (or MEDIUM) and MESSAGE FEs, and leaves other FEs in the background (cf. Perek, 2015); this is marked by the list of FEs in brackets after the frame name in Figure 4. Note that the formal component of the construction includes the subject of the clause, contrary to the "V that" pattern and most other patterns in Francis et al. (1996), as their label indicates. This is mostly done to simplify the formal description, since all English sentences have a subject, and thus subjects are usually not a distinguishing feature of patterns and not considered part of them. However, in a construction grammar description, subjects are associated with semantic information that varies from one construction to another, which warrants the inclusion of the subject when patterns are described in terms of constructions.

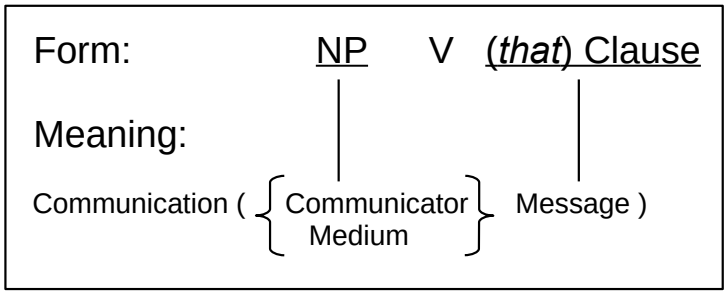

Figure 4: The Communication "V that" construction

The network of frames can also indicate constructional generalisations at intermediate levels of abstraction. In theory, every frame in the network could be taken to correspond to its own construction, with hierarchical relations between sub-constructions matching the frame-to-frame relations. However, it is not clear how useful such a myriad of constructions would be in a constructicon database, especially if it is designed for pedagogical purposes. Yet, frames provide information about some semantic regularities, and thus sub-constructions could add to the description of the construction and its distribution. The `Statement` frame, for instance, seems to occupy a rather central position in the network, with several other frames using it or inheriting from it, and it is evoked by a lion's share of lexical units (69, i.e. 19%, or 101, i.e. 28% including all subframes). It could therefore receive its own sub-construction, and given its prominence, be described as a sort of "prototype" of the `Communication` "V that" construction. Frames that describe specific uses of communication could also deserve their own sub-construction, especially if they are evoked by many LUs. We could thus posit, for instance, a `Request` "V that" construction (14 LUs) and a `Commitment` "V that" construction (7 LUs with related frames), given the specific pragmatic value conveyed by the `Request` and `Commitment` frames. The `Request` "V that" construction is also doubly motivated by a formal regularity: the *that*-clause in this construction should be a mandative clause, i.e., one in which the verb is in the subjunctive form or modified by a suitable modal auxiliary such as *should*, as exemplified by (6) and (7) below from FrameNet.

(6) She did, however, ask that Christine send her a photograph.

(7) Your father ordered that his possessions should be burnt.

6.3 The `Mental_activity` "V that" construction

The second largest network of frames in the "V that" pattern, diagrammed in Figure 5, is centred on the `Mental_activity` frame, in which "a SENTIENT_ENTITY has some activity of the mind operating on a particular CONTENT or about a particular TOPIC". It is a non-lexical frame (marked by a dashed outline in Figure 5), which means that it does not itself contain any lexical units; rather, it is inherited or used by frames with LUs, and captures the commonalities between these frames. As the definition of the `Mental_activity` frame indicates, these frames are all about cognition and cognizers having or processing some

mental content in their mind in some way. The inchoative relation between Awareness and Coming_to_believe (e.g., *realise, guess, deduce*) marks that the former describes the possession of some knowledge in a "static" way, while the latter describes arriving at that knowledge, in a "dynamic", processual way.

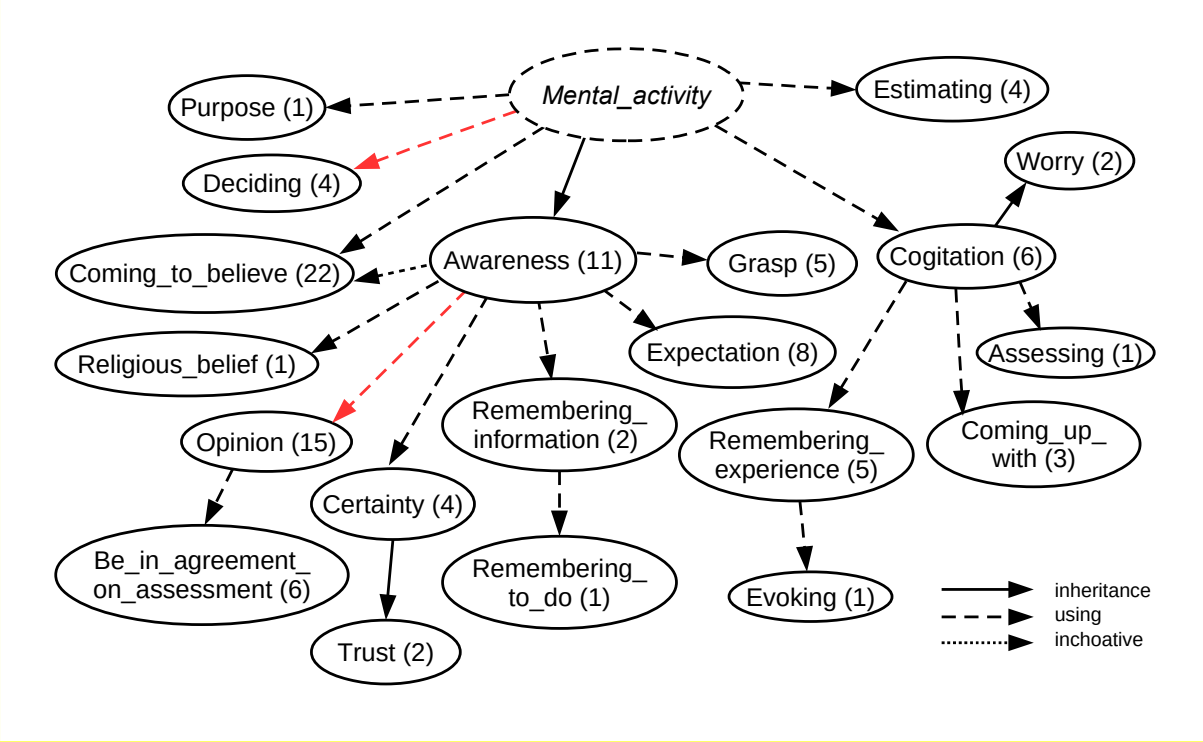

Figure 5: The Mental_activity network of frames in the "V that" pattern

Compared to the Communication network described in the previous section, this network had to be complemented with more changes to the information recorded in FrameNet. First, two using relations were added in order to include frames that qualify as mental activities: between Mental_activity and Deciding and between Awareness and Opinion. While the former is straightforward (making a DECISION involves having that decision as a CONTENT in one's mind), the latter is motivated by the fact that the meaning of Awareness, in which "A COGNIZER has a piece of CONTENT in their model of the world" is at least partially contained in that of Opinion, a subjective version of Awareness, as the name indicates. Second, the Memory frame, which was automatically matched to some verbs (e.g., *recall, remember*), was found to be largely redundant with other frames in the database. The frame is rather loosely defined as "concerned with COGNIZERS remembering and forgetting mental CONTENT", it is not involved in any relations with the other frames in this network, and it substantially overlaps in conceptual content and LUs with the "remember" frames such

as Remembering_information and Remembering_experience. We surmise that Memory is a remnant of very early work on FrameNet (it dates back to the creation of the electronic database in 2001) that was not checked for consistency with other frames created later, which warrants its deletion. The LUs of Memory were reassigned to other frames according to the lexical senses of the corresponding verbs.

Based on this network, the Mental_activity "V that" construction evokes the Mental_activity frame, in which it profiles the SENTIENT_ENTITY and CONTENT FEs, respectively realised as a subject NP and a *that*-clause. Similarly to the Communication network, these two FEs are carried over to all other frames via the frame-to-frame relations. Families of frames using Mental_activity can be identified to posit sub-constructions, such as the Awareness and Cogitation "V that" constructions. Frames within these two constructions illustrate their semantic range; for instance, the subframes of Awareness describe different ways of knowing: personal beliefs (Opinion, Religious_belief), epistemic confidence (Certainty), experience (Remembering_information), inference or prediction (Expectation), or full understanding (Grasp). With its 22 LUs, Coming_to_believe could also deserve its own construction, especially given the way it differs from the other frames. Drawing on the inchoative relation with Awareness, the contrast between the two constructions could also be captured by a horizontal relation between these constructions.

6.4 Other "V that" constructions and relations between constructions

Two other networks are diagrammed in Figure 6 below. For reasons of space we discuss these networks and the constructions corresponding to them in less detail than the first two.

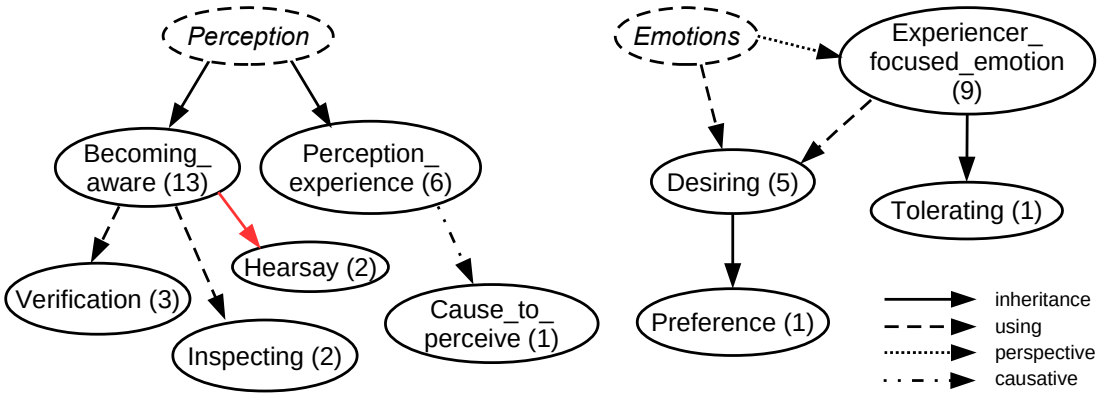

Figure 6: The `Perception` and `Emotions` networks of frames in the "V that" pattern

The first network generalises over the `Perception` non-lexical frame[11] ("A PERCEIVER perceives a PHENOMENON"), but is clearly split into two quite distinct frames: the `Becoming_aware` frame, about "a COGNIZER adding some Phenomenon to their model of the world" (e.g., *discover*, *learn*, *notice*), and the `Perception_experience` frame, which "contains perception words whose PERCEIVERS have perceptual experiences that they do not necessarily intend to" (e.g., *hear*, *see*, *sense*); note that the inheritance relation between `Becoming_aware` and `Hearsay` (e.g., *hear*, *read*) was added to the FrameNet data. We can thus posit a `Becoming_aware` and a `Perception_experience` "V that" constructions, with a potential but abstract and less significant `Perception` "V that" construction. The second network ultimately generalises over another non-lexical frame, `Emotions`, and this is indeed what all frames in the network are concerned with. However, the more specific `Experiencer_focused_emotion` is already a sufficient level of generalisation, as it describes the only kind of verb meaning that is compatible with the "V that" pattern in the semantic domain of emotions. We can thus posit the corresponding `Experiencer_focused_emotion` "V that" construction, evoking this frame.

The network analysis leaves us with eight frames that could not be included in any network. These frames are listed in Table 3, with the number of LUs in the "V that" pattern that evoke them, and up to three examples of verbs for illustration. Only two frame-to-frame relations are found in FrameNet between these frames: `Have_as_requirement` inherits from `Contingency`, and `Sign` uses `Evidence`. It is not immediately obvious if other relations should be added.

Table 3: Frames of the "V that" pattern not included in any network.

Frame	# of LUs	Example verbs
Causation	5	*arrange*, *dictate*, *see*
Contingency	3	*dictate*, *ensure*, *guarantee*
Evidence	16	*confirm*, *imply*, *suggest*
Feigning	1	*pretend*
Have_as_requirement	3	*presume*, *presuppose*, *require*
Prohibiting_or_licensing	1	*provide*
Rite	1	*pray*
Sign	5	*indicate*, *mean*, *signal*

One clear outlier stands out which can be set aside: the Rite frame, which "concerns rituals performed in line with religious beliefs or tradition". It is true that *pray* in its religious sense (as opposed to its Desiring frame sense, similar to *hope*) does qualify as a ritual of sorts, which isolates this LU from the other frames of the "V that" pattern in FrameNet. However, it is not clear how grouping such diverse practices as *pray*, *baptize*, and *sacrifice* into a single frame is useful in understanding how these events actually happen (like for instance the verbal component of *pray*); hence this religious classification could be ignored.[12]

While the other frames are not explicitly related in FrameNet, they can nonetheless be seen to share family resemblances. Contingency and Have_as_requirement on one hand, and Evidence and Sign on the other hand, are frames that establish epistemic relations of conceptual dependency between states of affairs. The Causation frame, which encodes cause-effect relations, also relate states of affairs in a way that can be seen as the concrete counterpart of the Contingency and Evidence frames; in a sense, the latter encode cause-effect relations that are merely imagined for epistemic purposes. Prohibiting_or_licensing describes another kind of conceptual dependency, one imposed by laws or principles, and is similar to Causation in that it relates to enablement and prevention. Even Feigning can be considered a similar frame: in this frame, an agent acts in such a way as to make others believe that a certain state of affairs is true, instead of actually causing it to be true, as would be the case in the Causation frame. In sum, from these frames an abstract "V that" construction encoding relations between states of affairs, or an agent and a state of affair, could be posited; we call this the Relation "V that" construction.

To this point, we have described the "V that" pattern in terms of several constructions, but we have not offered a generalisation over all instances. As we mentioned, such a

generalisation is not strictly required in construction grammar, and in fact recent research gives more importance to more specific constructions than maximally general ones. Given the semantic diversity found in the "V that" pattern, it would be a very abstract generalisation: for instance, the perspective of an agent on a state of affairs realised by the *that*-clause (as suggested by an anonymous reviewer), or something along the lines of Halliday's (2014: 443) notion of projection, i.e., the idea that "the secondary clause [the *that*-clause] is projected through the primary clause [the main clause], which instates it as (a) a locution or (b) an idea". However, because such a meaning might not be particularly helpful in understanding and predicting actual uses of the construction (especially for pedagogical applications), and is likely to overgenerate, it is probably preferable to capture "V that" solely in terms of its subconstructions, and/or allow a purely formal generalisation with no semantic import, as is the case in some versions of construction grammar (cf. Fillmore, 1999). That said, some of the constructions we discussed do share commonalities related to similarities between frames. For instance, `Becoming_aware` (e.g., *discover*, *learn*, *notice*) and `Coming_to_believe` (e.g., *realise*, *guess*, *deduce*) are similar in that they both involve the acquisition of knowledge, the former through sensory perception or observation (in line with the `Perception` frame), the latter through some kind of reasoning or mental processing (in line with the `Mental_activity` frame). This fact which is pointed out by FrameNet itself in the description of the `Becoming_aware` frame: "Words in this frame […] are similar to Coming-to-believe words, except the latter generally involve reasoning from Evidence". An inchoative relation could also be posited between `Becoming_aware` and `Awareness`, as we did with the `Coming_to_believe` frame. By the same token, the frame networks are not as distinct as we described, and there are multiple relations between frames that cut across networks and thus suggest conceptual similarities. `Worry` from the `Mental_activity` network uses the `Emotions` frame; `Predicting` from the `Communication` network uses `Expectation` from the `Mental_activity` network; `Hearsay` from the `Perception` network uses `Communication`, which itself inherits from `Cause_to_perceive`. Not all of these relations link the FEs realised in the "V that" pattern, but they highlight a certain continuity in the semantics of the different constructions and similarities in their argument roles, with some uses straddling the border between them.

By way of summary, a network representation of the constructions that our frame-based study of the "V that" pattern led us to posit is presented in Figure 7; as such, Figure 7 represents a section of our planned English Constructicon. All constructions mentioned

earlier are pictured in box representations typical of construction grammar: the `Communication`, `Mental_activity`, `Perception`, and `Experiencer_focused_emotion` "V that" constructions, as well as the Relation "V that" construction, generalising over various kinds of relations between states of affairs, as described earlier in this section. Three of these constructions have a number of sub-constructions corresponding to more specific frames; only the main sub-constructions mentioned earlier are included, but more sub-constructions could in principle be posited, as tentatively indicated in Figure 7. Inheritance links are represented by arrows; an inchoative relation is posited between the `Awareness` and `Coming_to_believe` "V that" constructions to mark the semantic connection between these constructions. Except for the Relation "V that" construction, these constructions evoke the frame mentioned in their name; the participant roles assigned to the argument slots of the construction correspond to frame elements of this frame, as shown in the bottom line of each box. At the very top of the inheritance hierarchy, the general "V that" construction does not specify any semantic information, as discussed above. Rather, there are two broad kinds of uses: the Relation "V that" construction already mentioned, and an Action "V that" construction generalising over all the other "V that" constructions, in which an animate being (the Actor) acts in some way with respect to some Content.

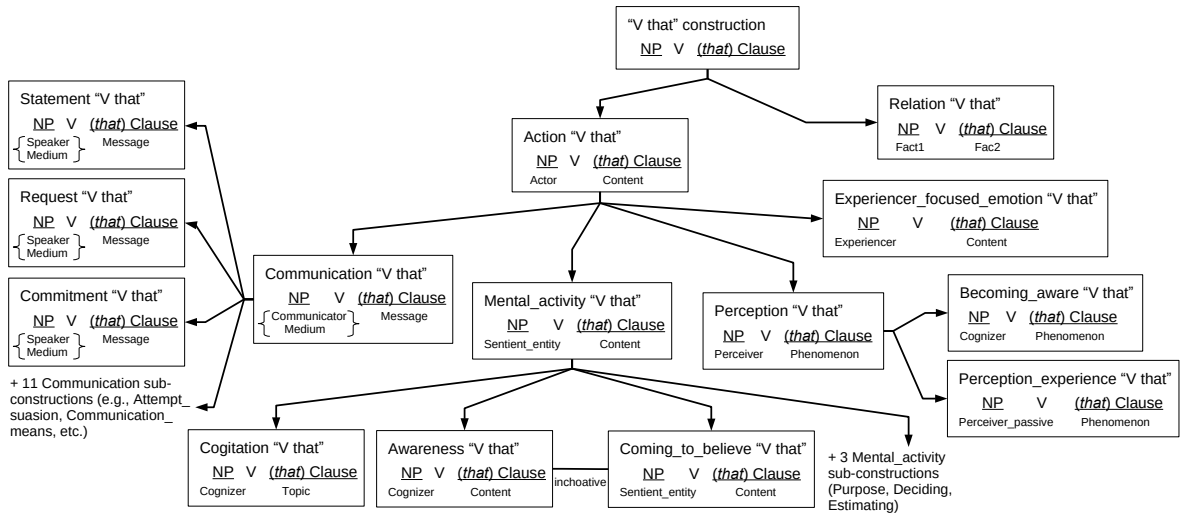

Figure 7: Inheritance network of the "V that" constructions.

7. Summary and conclusion

In this paper, we discussed how two existing corpus-based resources can be combined to form the basis for a new, more comprehensive of database of English constructions: the

English Constructicon. The COBUILD Grammar Patterns resource, on the one hand, documents hundreds of typical grammatical environments, or patterns, in which English verbs, nouns, and adjectives occur, and lists all lexical items attested in them. FrameNet, on the other hand, is a lexical database that aims to describe English words in terms of the theory of frame semantics. The COBUILD patterns are already quite similar to constructions in the construction grammar sense; however, they lack a proper semantic characterization. We suggest that FrameNet can fulfil this role, and that matching the lexical entries of patterns to FrameNet frames can help identify semantic generalizations in patterns robustly and systematically, and turn patterns into a structured set of form-meaning pairs at varying levels of semantic granularity, focussing in particular on constructions of the verb in this case study.

We offered proof of concept for the merger of the two resources, first in the form of a procedure that automatically matches the verbs of the COBUILD patterns to lexical units of the FrameNet frames on the basis of the examples annotated for valency information found in the electronic version of the database. The results of this procedure are mixed, as many pattern entries could not be matched to a frame, because the relevant information is currently absent from FrameNet. This means that a good deal of manual annotation is necessary to provide the information that cannot be gathered by the automatic procedure and fully match the patterns to frames, even though the automatic matching admittedly does give a significant head start in this process. While the current coverage of FrameNet limits its use for the present purpose, the semantic information it contains is very useful for such a project. We focused on the case of the "V that" pattern, for which we manually identified all the missing frames, which gives a glimpse at the nature of the annotation work that building the English Constructicon with this method would involve. Drawing in particular on frame-to-frame relations, we used this information to describe the "V that" pattern in terms of constructions at different levels of generality, thus illustrating how patterns can be turned into constructions when paired with semantic frames.

In conclusion, this case study demonstrates the potential of using the COBUILD Grammar Patterns and FrameNet to build a large-scale contructicon. As mentioned earlier, there are about 200 patterns listed in Francis et al. (1996, 1998). Considering that most of these patterns are likely to correspond to more than one, possibly many constructions, such a constructicon would be unmatched in terms of size. Given the nature of the COBUILD patterns, there are admittedly certain types of constructions that would not be covered by this work alone, such as idioms, clausal constructions, and constructions related to other parts of speech. Yet, building the English Constructicon from patterns and frames seems like a

promising endeavour that would go a long way to achieve the commitment of construction grammar to describe the entirety of grammar in terms of constructions.

Notes

1. In line with formatting conventions in the field, we use `Courier` font for frame names, and SMALL CAPS for names of frame elements.

2. This figure was calculated from the XML version of FrameNet.

3. It is fair to point out that many of these cases arise because of work on full-text annotation, i.e., entire texts annotated for frames and frame elements, which often requires the addition of frames and LUs according to the requirements of the text, that subsequently do not receive the full lexicographic treatment.

4. As one anonymous reviewer points out, such elements are actually signaled on another layer of annotation, so it should be possible in principle to at least detect their presence in the FrameNet examples, if not match the patterns containing *it* and *there*. However, this would require an ad-hoc procedure which was not included in the simple computer program developed for this case study. In future work, it would be worth considering how extracting information from that annotation layer as well could help to include patterns with *it* and *there* in the automatic matching.

5. The verb *remain* was removed from the distribution of the pattern, since it only occurs in the expression "the fact remains that …". Since this does not follow the same generalisation as the construction(s) of the "V that" pattern but rather instantiates a kind of extraposition construction, this verb is excluded from the present study.

6. Namely, *ask* in the `Questioning` frame, *feel* in the `Give_impression` frame, *find* in the `Locating` frame, and *wonder* in the `Cogitation` frame.

7. Namely, *say* in the `Text_creation` frame, *attest* in the `Statement` frame, and *recall* in the `Memory` frame. The latter two were manually matched to two existing LUs (in the `Affirm_or_deny` and `Remembering_experience` frames), hence these LUs were judged redundant (also because the `Memory` frame was deleted, cf. Section 6.3).

8. As we will see in the next section, we were less scrupulous regarding frame-to-frame relations.

9. Namely, *aver* was changed from `Statement` to `Affirm_or_deny`, *preach* was changed from `Statement` to `Attempt_suasion`, and *remember* was changed from `Memory` to `Remembering_experience`, since the `Memory` frame was judged redundant and deleted (cf. Section 6.3).

10. All definitions of frames and frame elements cited in this and subsequent sections come from FrameNet.

11. The `Perception` network was simplified for reasons of space: `Verification` and `Inspecting` were made to inherit directly from `Becoming_aware`, while there are actually intervening frames in each case that do not contain any LUs of the "V that" pattern (namely `Scrutiny` and `Scrutinizing_for`).

12. According to one anonymous reviewer, this is more specifically because `Rite` is an example of a non-perspectivalized frame. In a complete description, the LUs listed in it should indeed in most cases be re-assigned to more specific frames.

References

Boas, H. C. (2003). *A constructional approach to resultatives*. Stanford: CSLI Publications.

Boas, H. C. (2008). Determining the structure of lexical entries and grammatical constructions in Construction Grammar. *Annual Review of Cognitive Linguistics*, 6, 113–144.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Bybee, J. (2013). Usage-based theory and exemplar representations of constructions. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 49–69). Oxford: Oxford University Press.

Bybee, J., & Eddington, D. (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language*, 82(2), 323–355.

Croft, W. (2003). Lexical rules vs. constructions: A false dichotomy. In H. Cuyckens, T. Berg, R. Dirven, & K.-U. Panther (Eds.), *Motivation in Language: Studies in honour of Günter Radden* (pp. 49–68). Amsterdam: John Benjamins.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, VI(2), 222–254.

Fillmore, C. J. (1999). Inversion and constructional inheritance. In G. Webelhuth, J.-P. Koenig, & A. Kathol (Eds.), *Lexical and Constructional Aspects of Linguistic Explanation* (pp. 113–128). Stanford: CSLI Publications.

Fillmore, C. J., & Atkins, B. T. (1992). Towards a frame-based Lexicon: The semantics of RISK and its neighbors. In A. Lehrer & E. Kittay (Eds.), *Frames, Fields and Contrasts: New essays in semantic and lexical organization* (pp. 75–102). Hillsdale: Erlbaum.

Fillmore, C. J., Lee-Goldman, R. R., & Rhomieux, R. (2012). The FrameNet Constructicon. In I. A. Sag, & H. C. Boas (Eds.), *Sign-Based Construction Grammar* (pp. 283–322). Stanford: CSLI.

Francis, G. (1993). A corpus-driven approach to grammar – principles, methods and examples. In M. Baker, G. Francis, & E. Tognini-Bonelli, E. (Eds), *Text and Technology: In honour of John Sinclair* (pp. 137–156). Amsterdam: Benjamins.

Francis, G., Hunston, S. & Manning, E. (1996). *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins.

Francis, G., Hunston, S. & Manning, E. (1998). Col*lins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

Fried, M., & Östman, J.-O. (2004). Construction grammar: A thumbnail sketch. In M. Fried & J.-O. Östman (Eds.), *Construction Grammar in a Cross-language Perspective* (pp. 11–86). Amsterdam: John Benjamins.

Halliday, M.A.K & Matthiessen, C. (2014) [1985]. *Halliday's Introduction to Functional Grammar* (4th Edition). London & New York: Routledge.

Healy, A. & Miller, G. (1970). The verb as the main determinant of sentence meaning. *Psychonomic Science*, 20(6), 372.

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

Hunston, S. & Su, H. (2017). Patterns, Constructions, and Local Grammar: A Case Study of 'Evaluation'. *Applied Linguistics* (online). URL: https://doi.org/10.1093/applin/amx046

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at Work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, 15(3), 289–316.

Iwata, S. (2008). *Locative Alternation: A lexical-constructional approach*. Amsterdam: John Benjamins.

Jackendoff, R. (1990). *Semantic Structures*. Cambridge, MA: MIT Press.

Langacker, R.W. (2000). A dynamic usage-based model. In M. Barlow & S. Kemmer (Eds.), *Usage-based Models of Language* (pp. 1–63). Stanford: CSLI Publications.

Levin, B., & Rappaport Hovav, M. (2005). *Argument Realization*. Cambridge: Cambridge University Press.

Lyngfelt, B., Borin, L., Forsberg, M., Prentice, J., Rydstedt, R., Sköldberg, E., & Tingsell, S. (2012). Adding a Constructicon to the Swedish resource network of Språkbanken. In *Proceedings of KONVENS 2012 (LexSem 2012 workshop)* (pp. 452–461). Vienna.

Lyngfelt, B., Borin, L., Ohara, K., & Torrent, T. T. (Eds.). (2018). *Constructicography: Constructicon development across languages*. Amsterdam: John Benjamins.

Ohara, K. H. (2013). Toward Constructicon Building for Japanese in Japanese FrameNet. *Veredas*, 17(1), 11–27.

Perek, F. (2014). Rethinking constructional polysemy: The case of the English conative construction. In D. Glynn & J. Robinson (Eds.), *Polysemy and Synonymy. Corpus methods and applications in cognitive linguistics*. Amsterdam: John Benjamins.

Perek, F. (2015). *Argument Structure in Usage-based Construction Grammar: Experimental and corpus-based perspectives*. Amsterdam: John Benjamins.

Perek, F., & Lemmens, M. (2010). Getting at the meaning of the English *at*-construction: The case of a constructional split. *CogniTextes*, 5. Retrieved from http://cognitextes.revues.org/331

Pinker, S. (1989). *Learnability and Cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press/Bradford Books.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., & Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice*. Berkeley: ICSI. Retrieved from https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf

Sinclair et al. (1995). *Collins COBUILD English Dictionary 2nd Edition*. London: HarperCollins.

Talmy, L. (1996). The windowing of attention in language. In M. Shibatani & S. A. Thompson (Eds.), *Grammatical Constructions: Their form and meaning* (pp. 235–287). Oxford: Oxford University Press.

Talmy, L. (2000). *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.

Torrent, T. T., Lage, L. M., Sampaio, T. F., Tavares, T. S., & Matos, E. E. S. (2014). Revisiting border conflicts between FrameNet and Construction Grammar: Annotation policies for the Brazilian Portuguese Constructicon. *Constructions and Frames*, 6(1), 34–51.