DEPARTMENT OF PHILOSOPHY, LINGUISTICS AND THEORY OF SCIENCE

SEEK AND FIND

A retrieval approach to construction search

Anna Ehrlemark

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2016
Supervisors	Richard Johansson and Benjamin Lyngfelt
Examiner	Simon Dobnik
Report number	(number will be provided by the administrators)
Keywords	Construction Grammar, Information Retrieval, Construction detection, Corpus Linguistics, Lexical Semantics, Constructicography

Abstract

Finding occurrences of specific *grammatical constructions* in running text is a central issue to constructionist approaches to linguistics and language processing. Of special concern are partially schematic constructions that cannot be distinguished from unrelated constructions by surface form alone. In order to detect instances of such complex constructions we consider using features that are intended to capture semantic restrictions of particular construction elements. We address this task as an information retrieval (IR) problem, and describe a simple interactive architecture for searching for constructions.

The retrieval system is guided by the user who provides it with a number of positive seed examples (occurrences of the construction) and tailors a ranking function based on a combination of lexical-semantic similarity features (lexicon-based or distributional).

The system was evaluated using standard IR metrics on a new benchmark for construction retrieval in Swedish, and we observed that a lexical-semantic reranker can give significant improvement over a lemma-based baseline, but must be tailored for the construction at hand. The search system is effective even with a small number of positive seed examples, which proves the feasibility of our approach from a user perspective.

Preface

Here are some of my favorite idioms:

As holy as a horse. A horse kiss. A free horse, a hired horse, a good horse of any color. Horse and man. Horse in hand. When the horse is starved you bring him oats.

I would like to thank my supervisors Richard Johansson and Benjamin Lyngfelt for solid feedback and interesting discussions.

Contents

List of Figures

List of Tables

1 Introduction

Constructions are larger than words. Or, put differently, *X is ADJECTIVE-er than Y*. Construction Grammar is a fairly recent development in theoretical linguistics and has revitalized the discussion in fields such as syntax, lexicography, and language acquisition. In particular, construction grammar is useful for describing partially schematic constructions: templatic patterns that exhibit lexical as well as syntactic properties, too specific to be referred to general rules of grammar, but too general to be attributed to specific lexical units (Fillmore et al., 2012). These constructions are notoriously difficult to capture from either side of the language spectrum, but should be of special interest for grammarians and computational linguists alike since they pose interesting theoretical and methodological questions that have been historically overlooked.

Despite the memorable statement by Sag et al. (2001) that multiword expressions are *'a pain in the neck for NLP'*, the constructionist view on linguistic units as form-meaning pairings on all levels of language complexity also carries great potential to improve the language models of NLP systems and influence the way we formulate and solve typical NLP-tasks. Taking the theoretical insights of Construction Grammar seriously means to integrate form, function and meaning into the design and implementation of NLP applications from the beginning.

1.1 Seek

The major impediment to a wide adoption of construction-based approaches is probably the lack of resources of a nontrivial size. Before we can even attempt to analyse language automatically based on a constructionist theory, it will be necessary to list, describe, and exemplify the constructions that are present in a language. Efforts to build inventories of this kind – *constructicons* – are underway for a number of languages, including Swedish. The emerging practice of constructicon development, or *constructicography*, may in short be characterized as a combination of construction grammar and lexicography (Bäckström et al., 2014).

Developing a constructicon is a large investment, just like building a traditional lexicon is, and it is crucial that corpus-based search tools are available that allows constructicographers to find suitable examples (Gries, 2003) or to compute various kinds of statistics on e.g. word–construction cooccurrence (Stefanowitsch & Gries, 2003). But searching for and detecting constructions in corpora is not a trivial task. Available corpora search tools are restricted to linear search strings with syntactic and lexical variables, an obvious methodological holdback for a theory that questions the division between form and meaning in linguistic units. To illustrate the problem, consider the examples in (1), (2) and (3)[1].

(1) Efter [en timme-s försening] kom vi fram till Göteborg.
After one hour-GEN delay.INDEF come.PST we forth to Göteborg.
'After an hour's delay we arrived in Gothenburg.'

(2) Hon började sjunga på barer redan vid [tio år-s ålder].
She start.AUX sing.INF on bars already by ten year-GEN age.INDEF

[1] All example sentences throughout the thesis are authentic occurrences selected from corpora, sometimes shortened to highlight the phenomena we wish to illustrate. The construct instantiating the construction in question is delimited by square brackets.

'She started to sing in bars already at the age of ten.'

(3) Charlie hade [en ängel-s tålamod].
Charlie have.PST an angel-GEN patience.INDEF
'Charlie had the patience of an angel.'

The sentences above contain examples of three different constructions that share the same surface form – a determiner (or number) followed by a noun in the genitive and an indefinite noun; $\text{DET N}_{\text{GEN}}\ \text{N}_{\text{INDEF}}$. Despite the structural similarity they clearly differ in meaning and use. While the example in (1) is a temporal construction that modifies the duration of an activity, (2) is an instance of a scale construction that measures the value on a scale expressed by the head noun, and (3) is a regular genitive phrase indicating possession.

Now here is the catch. Using available corpus search tools, a search for one of these constructions would return sentences containing all three of them (and more unrelated constructions that share the same surface form). A constructicographer interested in analyzing a particular construction would have to manually sort through a possibly very large number of hits to sift out the relevant examples. In order to disambiguate the constructions from each other we clearly have to consider more features than the surface structure.

Enter construction search.

1.2 And find

Partially schematic constructions with open and variable slots are typically defined by restrictions on these construction elements; particular slots are expected to prefer a certain set of lexical items or be restricted to a specific semantic class (Stefanowitsch & Gries, 2003). Guided by the theoretical expectation that semantic restrictions are delimiting features in disambiguating syntactically similar constructions, we would like to test the hypothesis by developing a search system that is capable of ranking search results by semantic similarity.

If the experiment falls out well, it will not only help constucticographers find what they are looking for, it will also give experimental support for the theory that such dependencies between semantic classes and grammatical structures are indeed present. This insight can be extended to other NLP tasks that can benefit from a more holistic view of the grammar-lexicon division.

In this thesis, we cast the task of searching for construction occurrences as an information retrieval (IR) problem. The goal is to rank relevant search hits first, based on user-defined criteria and a simple reranking function that uses different semantic similarity measures.

Treating construction detection as a retrieval task is innovative, and while we are only aware of one previous attempt to use this approach, we claim that the IR perspecive is fruitful in construction-based research and constructicography for a number of reasons. First, while a partially schematic construction can have a surface form that is easy to describe as a structural pattern, its semantics can be hard to capture as a clear-cut binary decision, which makes it natural to re-rank corpus hits according to a semantics-based scoring function. Also, the IR perspective is natural in terms of interaction: a user can pose queries, rank and re-rank repeatedly according to different functions defined on the fly, to get a comprehensive overview of the various uses of a construction in a corpus.

To demonstrate the feasibility of this approach, we will present a system that retrieves occurrences of partially schematic constructions in Swedish corpora, based on flexible user-defined ranking functions using lexicon-based and distributional similarities to describe the slot fillers. The ranking functions are learned from a small number of seed examples. Methodologically, we also take an IR perspective, and we have developed a new benchmark with constructions from the Swedish Constructicon that allows us to compare different ranking functions according to standard IR evaluation protocols.

Our results show that ranking functions based on lexical-semantic features are effective, but that there is considerable diversity among constructions as to which features are most useful. This shows that it is crucial for the ranking functions to be flexible and user-defined.

The rest of the paper is structured as follows. In the next section, 2, we will give a short introduction to the theoretical background of the thesis, including brief excursions into the fields of Construction Grammar, Lexical Semantics and Information Retrieval. We then move on to the design and implementation of the search system in section 3. In section 4 we present a benchmark collection of six different constructions from the Swedish Constructicon. Section 5 contains experiments and evaluations of construction search, and the insights and implications of these results are discussed in section 6.

2 Background

The design and implementation of a semantically guided search system is couched in *Construction Grammar*, for which the abundance of semi-general and partially schematic patterns are a key motivation. The basic principles of constructionist approaches and their potential impact on language technology will be sketched out in the following section, 2.1. In order to implement a constructional search system we are relying on different ways of modelling semantic similarity between lexical units, and both hand-crafted and distributional approaches to *Lexical semantics* will be introduced below in 2.2. Finally, since we have framed the challenge of construction detection as a search problem, the basic principles of *Information Retrieval* and rank evaluation will get a short introduction in 2.3.

2.1 Construction Grammar

As a recent development in linguistic theory Construction Grammar has sparked new life in grammatical discussion by questioning the traditional division between grammar and lexicon. Instead of treating language as a set of lexical units combined in phrases and clauses by grammatical rules, constructionists propose *constructions* as the basic linguistic units in language, "conventional, learned form–function pairings at varying levels of complexity and abstraction' (Goldberg 2013). In this view, the only thing language users must know when they know a language is constructions and how these can be combined together into utterances (Hilpert, 2014).

From the beginning many constructionist contributions focused on examples of constructions that seemed to diverge from established patterns and which traditional generative approaches failed to give satisfying explanations (see for example Jackendoff's investigation of the 'time'-*away* construction (1997) and Kay and Fillmore's adventures with the *What's X doing Y* construction (1999). By pointing out that such exceptions are neither rare nor peripheral to language (Fillmore et al., 1988) construction grammarians built the case for a new, holistic language model

consisting of a network of constructions from the most specific to the most general signs, and everything in between. As Adele Goldberg (2003) put it, "it's constructions all the way down". Importantly, the abundance of constructions should not be thought of as an unstructured set of unrelated chunks, they are rather organized in a hierarchical network, a *constructicon*, where constructions are related to each other via inheritance links. Constructions at the top of the hierarchy are highly schematic patterns that combine with more specific constructions down to lower levels of fully fixed lexical constructions. Constructions can be combined as long as they are not in conflict with each other, and specific constraints can override more general constraints as the puzzle is pieced together. Several different kinds of inheritance links have been proposed to explain how constructions interact and combine into larger expressions (Hilpert, 2014; Goldberg, 2013), but a complete map of the constructicon has yet to be drawn. Some examples of constructions of different degrees of schematicity can be seen below in Table 1.

Sentences, phrases and strings that are concrete instances of a more general construction are called *constructs*, so that for example the Swedish compound *grammatiknörd* 'grammar nerd' is a construct of the construction N-NÖRD 'N-nerd'.

Degree of schematicity	**Example**
Filled and fixed	*Slovenia, them, pencil, all in all, hot potato*
Filled and partially flexible	*Give* someone *the cold shoulder, N-nerd, put up with* something
Partially schematic	[unit] *per* [time], *the Xer the Yer, V* one's *way PP*
Fully Schematic	*Subj V Obj1 Obj2, S (VP NP), Verb*[tense]

Table 1: Examples of constructions at varying levels of schematicity.

Another point where Construction Grammar challenges previous generative models is in opposing the practice of analyzing language structure in terms of transformations and derivations (for example in Chomsky's Minimalist Programme (1995). Instead, the focus is on surface form. When constructionists argue that we should take surface form as the primary target of investigation, they underscore that difference in form is clearly associated with difference in meaning. Studying constructions as surface generalizations reveals relevant distributional and discursive properties that are not shared with their supposedly derived counterparts (Michaelis, 2013).

From a constructionist perspective, the intermingling of linguistics levels is seen as the norm and not the exception, and each construction must be analysed and defined by all semantic, pragmatic and formal properties that are relevant to the construction at hand. Construction Grammar is thus typically usage-based, not only from a methodological perspective where constructions are identified and characterized according to authentic use as perceived in corpora; but also in that most constructionist approaches argue that constructions are entrenched in the mind of language users as generalizations over actual utterances (Bybee, 2010). This usage-based hypothesis brings construction research close to cognitive linguistics and further motivates close study of statistical measures such as token frequency of certain items in a construction; prototypical and rare examples of construction instantiation; and productivity and variation of particular constructions, diachronically and/or synchronically (Bybee, 2013).

To sum up. Most constructionist approaches to grammatical theory agree on the following (adopted from Goldberg (2013):

- **Grammatical constructions**. Constructions are learned form-meaning pairings at varying levels of complexity and abstraction.
- **Surface structure**. What you see is what you get. Difference in surface form is associated with difference in meaning.
- **A network of constructions**. Constructions are organized in a hierarchical network, a constructicon, joined together by inheritance links.
- **Crosslinguistic variability and generalizations**. Languages vary and should be studied independently.
- **Usage-based**. Constructions are learned as generalizations over utterances and are studied by authentic use as perceived in corpora.

2.1.1 Constructicons and constructicography

In recent years constructionist theory has turned to practical resource building as *constructicons* for several different languages are starting to take shape. Here, a constructicon should be understood as a concrete and partial realization of the theoretical constructicon, described above. The emerging practice of constructicon building can best be described as a combination of construction grammar and lexicography, for which the term constructicography has been coined (Bäckström et al., 2014). The first constructicon, the Berkeley English Constructicon (Fillmore et al., 2012) was built as an extension of the Berkeley FrameNet (Baker et al., 1998), and similar projects are now under way for Swedish (Lyngfelt et al., 2012), Brazilian Portuguese (Torrent et al., 2014), Japanese (Ohara, 2013) and German (Boas, 2014) .

The close connection between Construction Grammar and Frame Semantics, and the descriptive resources based on these theories - framenets and constructicons - is due to their historical and conceptual origins in Berkeley, as developed by Charles J. Fillmore and associates (Fillmore, 2008). Both theories are concerned with the intimate relation between language form and meaning. But while framenets were designed to document and quantify the semantic and syntactic valency of individual lexical units, constructicons make room for detailed and multilayered descriptions of more complex grammatical constructions, especially such partially schematic patterns that have previously escaped catalogization in other lexical or grammatical linguistic resources. The similarities in design and description format make the two kinds of resources compatible for interlinking (Ehrlemark, 2014), and the semantic component of construction descriptions is often given a frame semantic representation. In this thesis we will use framenet as a resource to model semantic similarity, and we will get back to frame semantics in section 3.2.3 below.

In a constructicon database each construction is described in a construction post that can be characterized as a mix between a dictionary entry and a formal representation. In order to facilitate vide coverage and broad applicability, the format is simplified compared to more wide construction descriptions and formalisms. Such a constructicon post would minimally include a dictionary style free text description, a simple structure sketch and a handful annotated example sentences of the construction in authentic use. Some constructicon projects have also made attempts to

model inheritance networks, at least locally as sub-trees within the database (Torrent et al., 2014, Lyngfelt et al., forthcoming)

Constructicon building is a laborious process and much of the work consists of identifying and characterizing potential construction candidates before entering them into the database. Some different approaches to finding construction candidates are automatic extraction of hybrid N-grams from large corpora, mining L2 student essays for typical mistakes and collecting potential candidates from secondary sources like grammars, lexicons and academic papers (Sköldberg et al., 2013). Before being entered in the database each construction must be analyzed in terms of variation, productivity and other formal, semantic or functional constraints. This analysis is conducted by studying the construction in authentic use as perceived in corpora. Even though the format is somewhat simplified for the purposes of scalability, descriptive adequacy is a high priority.

Constructicons make available detailed and accessible linguistic descriptions of previously understudied constructions. As the constructicon repositories grow they are expected to be useful for many different purposes such as second language learning, interlingual comparison and language technology applications.

2.1.2 Computational approaches

So far, construction grammar has had no major impact on natural language processing. There are a number of possible reasons for that. While computational methods are often relying on large repositories of annotated data and give priority to scalability and overall performance over theoretical exactness, construction grammarians have shown more interest in linguistic theory building, descriptive accuracy and deep explorations of particularly interesting constructions. Concrete steps to bridge this gap can be seen in cross-disciplinary collaborations between computational and theoretical approaches, like in constructicon building.

Depending on what angle you look at it from, constructions are either a problem or a potential for natural language processing. In their widely cited article about multiword expressions as a pain in the neck for NLP, Sag et al. also stated that 'modern statistical NLP is crying out for better language models' (2001). The motivation for such a development can perhaps be better understood in terms of performance. Baldwin et al. (2004) have shown that 39% of parse failures on clean BNC data occured on particular constructions. But error detection may be just a short step from error solving. Zhang et al. (2006) similarly used error mining techniques on parse trees to detect unseen multiword expressions and by adding these to the grammar as new lexical entries they managed to increase the coverage of their model by 14.4%. Nivre & Nilsson (2004) have shown that recognizing multiword expressions can improve parser accuracy also for the surrounding syntactic structures.

In recent years a number of automatic methods in constructicography have been presented that mine large corpora for frequent patterns. For instance, the StringNet project (Wible & Tsao, 2010) collected a very large number of hybrid n-grams (that is, combinations of words and part-of-speech tags) and applied standard measures of collocational strength to select n-grams that seem to be recurrent patterns. In an attempt to model the hierarchical dependencies that hold between lexically fixed and schematic constructions, patterns were stored in parent-child relations where an underspecified hybrid n-gram like ***pay*** *[noun] to* is a parent of the fixed child trigram *pay attention to*. StringNet also supports investigating nested constructions, following links

between sub-parts of hybrid n-grams and exploring family resemblances between and among constructions.

The Swedish Constructicon project used similar automatic methods to search for patterns of hybrid n-grams (Forsberg et al., 2014), but also extended the approach by considering patterns containing phrase labels (e.g. *NP-and-NP, neither-AP-nor-AP*). While the StringNet project resulted in a large repository of automatically extracted hybrid n-grams, the Swedish Constructicon experiment followed up the statistical pattern mining with manual evaluation to identify construction candidates for entry into the constructicon. This qualitative approach was motivated since not all patterns were deemed equally relevant to the Swedish Constructicon which is integrated with other language resources in the linguistic resource environment of Språkbanken at Gothenburg University (Lyngfelt et al., 2012). The choice of how much redundancy to allow in pattern mining is thus a design decision, and the manual inspection and analysis of construction candidates finally yielded about 200 actual construction entries in the Swedish experiment.

Computational and quantitative methods have also been employed to increase the descriptive accuracy of construction analysis. O'Donnel and Ellis (2010) exploited the usage-based hypothesis of construction entrenchment as a function of learners' generalizations over relative frequencies of particular lexemes in particular construction slots to develop an inventory of verb-argument constructions. Their study showed trend results attesting particularly 'prototypical' verbs (high frequency) as indicative of the construction meaning as a whole, while expressly 'faithful' verbs (low frequency, high contingency) highlight construction specific properties. Their study attested that verb-argument constructions are not semantically void patterns waiting for insertion of lexical content, instead each construction is clearly associated with a certain class of semantically related verbs, and verbs appearing in a certain construction tend to converge in meaning by cohesion.

Precise methods to conduct construction analysis and measure degree of association between lexemes and grammatical structures have been refined by Stefanowitsch and Gries (2003) and experimental findings in such *collostructural* analysis have shown convincing support for constructional theory by revealing distributional variation and preferences also on a highly schematic level. As an example, Stefanowitsch and Gries (2003) were able to show strong collostructural association between the arguably very abstract past tense construction and particular verbs, such as *be, say and die*. Collostructural analysis is geared to provide precise and adequate grammatical descriptions that capture the semantic restrictions of particular slotfillers, but since many partially schematic constructions are difficult to disambiguate by surface form alone, manual item-to-item inspection is often needed to get to raw frequency counts.

In a pilot study, Hwang, Nielsen and Palmer (2010) showed that assigning meaning on a constructional rather than lexical level can improve automatic semantic interpretation and help generalize over unseen predicates. They used classical machine learning methods to train a classifier to recognize the caused-motion construction exemplified in *'Frank sneezed the napkin off the table"* from non-motion instantiations of the same pattern, like *'Mary kicked the ball to my relief'*. With hand annotated training data they considered a wide range of syntactic, lexical and typological features, and reached the same classification accuracy with unseen as with seen verbs in their evaluation set. While this is of course good news for advocates of constructional NLP, it should be pointed out that their results were highly construction specific (the most useful feature in this case was the preposition) and labor intense, since each subsequent construction classification task would require a substantial amount of new annotated data.

The IR perspective that we employ in this thesis is innovative, but we have found one other example of computational construction research that approaches the problem of construction detection as a non-binary ranking task. Dubremetz & Nivre (2015) used methods from information retrieval to dig out occurrences of the very rare rhetorical *chiasmus*[2] construction from corpora. They argued that ranking is preferred over classification on rare constructions since the complexity of the constructions do not lend themselves to clean classification and that the existence of borderline cases should be embraced instead of ignored. Since Dubremetz & Nivre limited themselves to searching for one particular construction, they also tailored the features and weights used in the ranking function to fit exclusively for this case.

Summing up and moving on to different ways of modeling semantic similarity, constructional NLP is still to a big extent uncharted territory. How to model form and meaning together on all levels of interpretation is definitely a difficult challenge, but the good news is that machinery required to account for patterns with both grammatical and lexical properties should also be able to handle those that are purely grammatical or purely lexical. On the one hand causing errors, on the other hand promising to disentangle those very errors, this is an interesting field of investigation for future explorations in NLP.

2.2 Lexical semantics

Computational lexical semantics is concerned with how to represent word meaning, for example as a way to measure semantic similarity and relatedness between word senses (Cruse, 1986). Semantic *similarity* is concerned with synonymous similarity, that is words that can be used interchangeably in the same contexts, like 'seek' and 'search'. Semantic *relatedness* on the other hand is reserved for topical similarity, that is words that belong in a certain domain, like 'bread' and 'bake'. For the purposes of this thesis we will consider two different approaches to representing word meaning.

- **Hand-crafted lexical resources.**. Defining word meanings as associated to concepts, typically structured hierarchically in a network of relations such as hyponymy and hypernymy.
- **Distributional data-driven models.** Defining word meaning geometrically as a point in a vector space derived statistically by counting the distributional contexts in which a word occurs.

Even though the two approaches are diametrically different they have both been used widely in NLP applications for many different purposes, such as word sense disambiguation, automatic translation, vocabulary expansion and sentiment analysis (Curran, 2003). Despite, or perhaps because of, the different underlying theoretical frameworks, the approaches have been successfully combined in tasks that rely on quantifyable similarity measures (see i.e Johansson (2014) and Agirre et al. (2009). This shows that the approaches are not entirely overlapping, but manage to capture partially different aspects of word meaning.

[2] *Chiasmi* are a family of figures that can be characterized as repeating linguistic elements in reverse order to achieve a rhetoric effect, like the classical example *'One for all, all for one.'*

2.2.1 Hand-crafted resources

Manually constructed lexical resources are carefully compiled by experts to model word sense meaning in a semantic network, as related to other word senses (Curran, 2003). The taxonomic structure formalizes semantic relations grounded in psycholinguistic theories about the mental lexicon and allows users to browse the tree structure conceptually, rather than merely alphabetically like in linearly structured lexicons and dictionaries (Miller et al., 1990). Well established semantic relations include *synonymy* ('interchangeable with x'), *antonymy* ('not-x'), *hyponymy* ('is-a') and *meronymy* ('part-of'). Depending on the theory that motivates the construction of a new lexical resource, other relations can be built into the ontology. Fixing a small set of well defined and non-ambiguous relations that can be applied over the whole network is the lexicographic challenge that decides the quality of the resource. An example slice of a semantic network with relation links can be seen in figure 1.

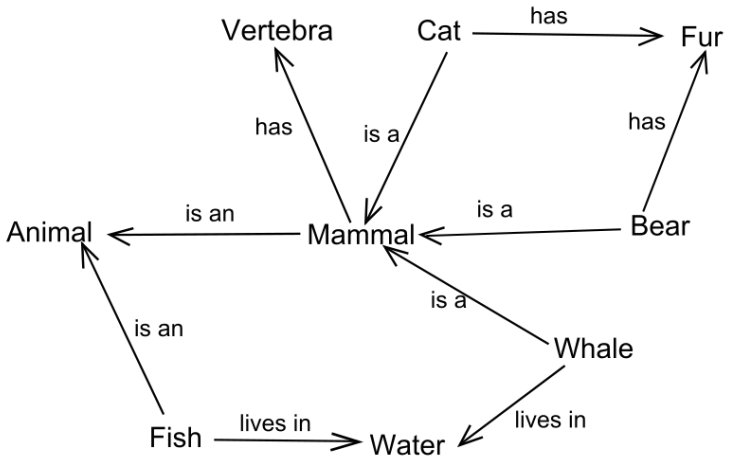


Figure 1: Example of a semantic network.

Given that the word senses in a lexical-semantic resource are organised in a directed graph, the edges linking the senses together can be used to calculate a measure of semantic similarity between different senses. Several ways to analyse semantic similarity mathematically have been proposed (Blanchard et al., 2005), all of them are in one way or another based on distance in the tree, shortest path between two nodes in the network and the depth of the whole tree or common ancestors. These operations serve to computationally quantify the human intuition that for example *bird* and *mammal* are more similar than *bird* and *ashtray*, and also that words close to each other on a lower level in the tree structure, like *parrot* and *eagle* are more similar than words higher up in the hierarchy, like *plant* and *food*. For the latter distinction to play out well in the calculation, the resource must be constructed evenly, with the same level of granularity over the board.

Hand-crafted resources are considered reliable and cognitively grounded sources of word sense meaning and are often used as gold standard references for computational applications. But like all manually constructed linguistic resources they require a significant amount of expertise, time and labor to develop. Consequently, the biggest drawback to most lexical-semantic databases is low coverage, especially for resources that aim to describe language in general, rather than any specified domain (Curran, 2003).

2.2.2 Distributional models

Distributional models of word meaning are based on the *distributional* hypothesis that *similar words appear in similar contexts*. The hypothesis indicates that to compare words is to compare the contexts in which they occur (Clark, 2015; Turney & Pantel, 2010; Lenci, 2008). If the hypothesis is valid, we may represent the meaning of a word by a vector of frequency counts recording the linguistic contexts in which they appear. The advantage of representing word meaning as a vector is that we can apply standard geometric operations to calculate the similarity between two words using simple functions like cosine similarity, where similarity is interpreted as the distance between points in a multidimensional space. Weather to take this distributional similarity as merely correlational and practical for computational applications, or actually causal in shaping the semantic content on a cognitive level, is an empirical question, usually distinguished between as the 'weak' and the 'strong' distributional hypothesis (Lenci, 2008). The greater implications of what semantic vector space models actually model are often glossed over in computational semantics as long as they manage to do the job when applied to different NLP tasks, like word sense disambiguation, information retrieval and question answering. Figure 2 shows an example of word vectors represented as points in a distributional space. Since high dimensional vectors do not lend themselves to simple visualization, the number of dimensions has been scaled down from 512 to 2 in order to be readable on paper.

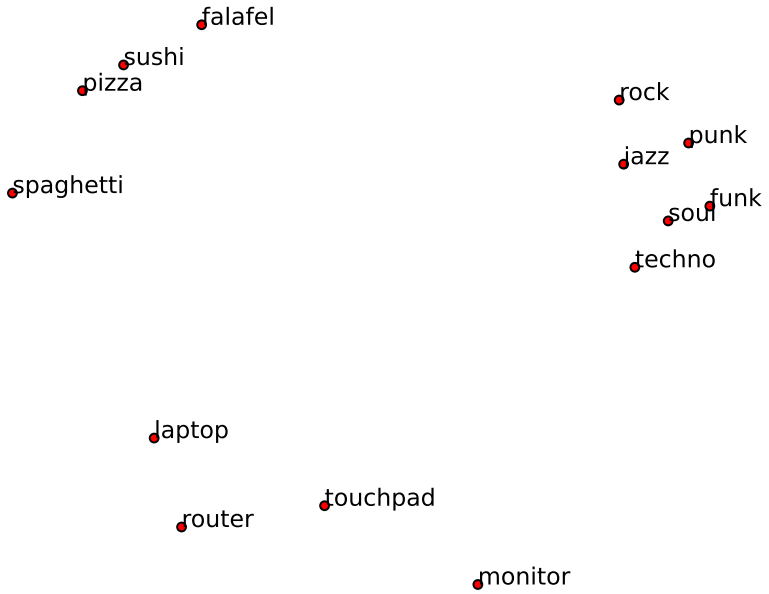


Figure 2: Example of word vectors as points in a distributional space.

A couple of parameters influence the construction and quality of vector-space models.

Corpora. Distributional models are data-driven and data-hungry. The quality of the model is to a large extent dependent on what you feed into it - the size and domains of the corpora used to extract the context counts. The general message is that the more text you can get your hands on, the better the model. The 100 million words of the British National Corpus have been used as a starting point in a lot of experimental research, but the best performing model in the class (beating all other models on standard evaluation tasks) was constructed by researchers with access to Google's computing infrastructure and used a corpus of 1.6 terawords (Clark, 2015; Agirre et al., 2009).

Context. Depending on what semantic relations we wish to capture in the distributional model, different contextual features around the word can be considered. A large context, such as the whole document or sentence in which the target word appears, tends to capture *topical* similarity, while smaller windows of a few words surrounding the target word on either side is normally used to capture *synonymous* similarity (Clark, 2015). Another possibility is to introduce linguistic processing and record syntactic contexts, such as part-of-speach tags or dependency labels of words around the target word (Padó & Lapata, 2007). Using the syntactic context of a target word, expands the distributional hypothesis to something that echoes constructionist theory - *similar words appear in similar syntactic constructions*.

Weighting. To increase the informative value of the linguistic context used to compute the vectors, different weighting functions can be performed on the raw frequency counts. The intuition behind weighting is that some words in the context will be more indicative of the meaning than others, and any collocational statistic can be used to scale the context and improve the quality of the model (Curran, 2003).

Count or predict. In recent years an alternative approach to derive geometric word vectors has outperformed the statistical coocurrence counting methods described above (Mikolov et al., 2013a). Instead of counting and then weighting the counts, the vectors can be derived indirectly as a by-product of classifiers trained to predict the contexts of a target word (Baroni et al., 2014). Since similar words would occur in similar contexts, the system assigns similar vectors to similar words. Mikolov (2013a) has made the predicting word vector models accessible to a wide audience by developing a simpler and computationally effective method, the skip-gram with negative sampling, and released it for public use as the `word2vec` model.

In contrast to hand-crafted lexical resources, distributional models of word meaning are fast and cheap to construct, and can easily cover a much larger portion of the vocabulary than any manually constructed semantic network. The models capture a gradual notion of semantic similarity, but fail to make explicit the nature of the semantic relations (Lenci, 2008). A distributional model will be able to tell us that *eagle* and *bird* are somehow semantically related, but not that the relation is assymmetrical since an *eagle* is-a *bird*. Consequently, the distributional notion of semantic similarity is unspecified.

2.3 Information Retrieval

Information Retrieval is a major independent area in computer science, concerned with separating relevant and useful information from irrelevant and uninteresting ditto (Baeza-Yates & Ribeiro-Neto, 2011; Manning et al., 2008). This is obviously not a trivial task and before going any further one has to ask (1) Relevant to whom? and (2) Relevant in comparison to what? Both questions are highly subjective and helps frame the IR task as a ranking problem, rather than a classification task where the goal is to separate a data set into different predefined classes. Baeza-Yates and Ribeiro-Neto (2011) defines the IR problem in the following way:

> The primary goal of an IR system is to retrieve all the documents that are relevant to a user query while retrieving as few nonrelevant documents as possible.

To design an IR system one must both figure out a way to define relevance in relation to user intentions, and implement a method to measure and compute it in an informed and efficient way.

These are two distinct fields within IR research.

Since relevance is a fleeting quality and can change over time, context, location (any circumstances, really) it is recognized that no IR system can provide perfect ranking at all times for all users. Thus, systems can always be improved by adjusting to more fine-tuned and user fitted solutions.

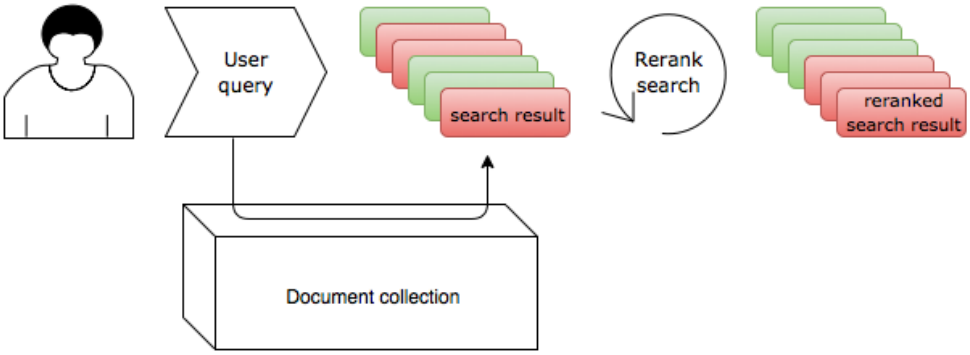


Figure 3: Basic IR system.

The basic architecture of an IR system stays the same, independently of what kind of documents are to be retrieved, what user specified query is considered, and the many different approaches to ranking that exist and continue to evolve (Manning et al., 2008; Baeza-Yates & Ribeiro-Neto, 2011). Figure 3 shows a simple flow chart of a basic IR system. Given a *collection of documents*, a user poses a *query* that reflects their information need; the system continues to parse the query and matches it against the document collection to retrieve a subset of all documents, and finally applies a *reranking function* before returning the retrieved documents to the user in order of relevance.

2.3.1 Evaluation of IR systems

IR systems are evaluated for quality of the result in terms of precision and recall, where precision is the fraction of documents retrieved that are indeed relevant to the user, and recall is the fraction of relevant documents in the collection that were in fact returned by the system, see figure 4.

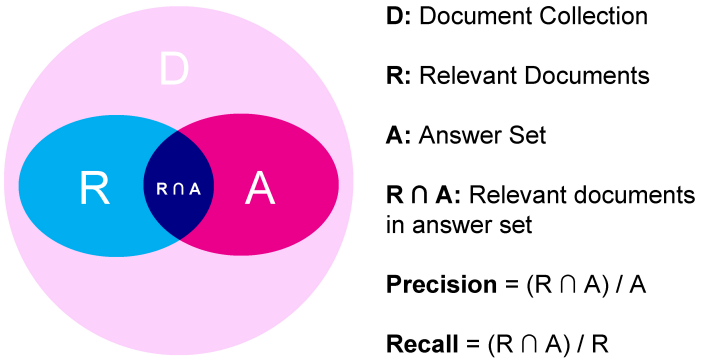


Figure 4: Precision and recall for a given information request.

Simple accuracy, the fraction of correctly retrieved documents over the whole document collection, is not deemed a useful measure for information retrieval problems (Manning et al., 2008).

Since the number of relevant documents compared to the whole collection can be very small, a system tuned to maximize precision could simply judge all documents nonrelevant. A good IR system should attempt to dig out relevant documents even at the cost of returning false positives, thus overall performance is preferably plotted as a curve of precision over recall, gradually increasing the number of retrieved documents until all relevant documents have been found. A single value summary of system performance on a specific query can then be computed as the average precision (AR) of precision over recall, a number between 0 an 1 intuitively corresponding to the area under the curve in the plot. A good retrieval system has a curve that bulges towards the top right corner and a perfect system would score 1 if the average precision equals the average recall so that all relevant documents are retrieved before all nonrelevant ones. Since most IR systems strive for high precision among the top ranked retrieved documents, an additional evaluation score of precision at n-documents (P@n) gives a good indicator of user satisfaction. Figure 5 features a toy example of a precision-recall curve with average precision- and P@n-values for four different retrieval algorithms on the same information query.

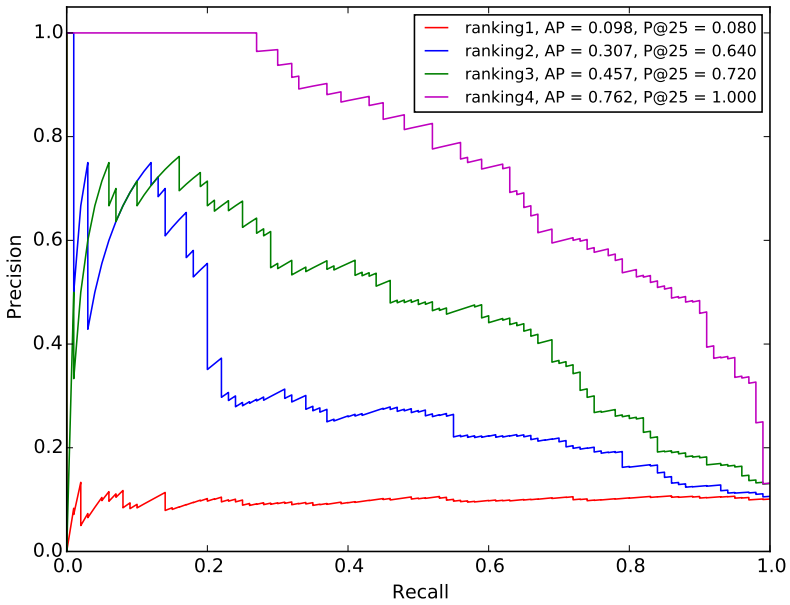


Figure 5: Precision-recall curve, with average precision (AP) and P@n.

Before proceeding with evaluation, however, one has to figure out the standard of relevance to evaluate the results against. A common approach is to compare the results of the system with human relevance judgements on the same set of queries. Such a gold standard is referred to as a *reference collection*, composed of a set of documents related to a specified information need, and judged binary by human experts as either relevant or nonrelevant to the information need. Because of the annotating procedure, reference collections are necessarily quite small and can only give partial indicators of the systems' performance over different document collections and information needs. But they come with the advantage that evaluation can be done quickly and allows comparison over different system versions and ranking functions.

3 Construction search

It is time to introduce the implementation of our construction search system designed to help constructicographers sift out true instances of constructions from corpora. The system has been designed with a certain type of constructions in mind - partially schematic constructions that cannot be distinguished by surface form alone. Searching for such a construction using available corpora search tools will return a (possibly very large) number of hits and depending on the formal properties of the construction, fixed lexical content, variable tokens, or syntactical restrictions on particular slots, this search list will typically be 'contaminated' by hits that are unrelated to the construction we are considering. Separating true instances of the the construction from 'everything else' is the aim of this endeavour.

In this section we will describe the design and implementation of the construction search system, the corpora infrastructure on which we build, and the scoring function used to rerank the original list of hits. We also detail how a user can tailor a scoring function, in particular the various lexical-semantic similarities the user can choose from.

3.1 Guiding principles

The motivation behind construction search comes from practical experience in constructicographic research. Automatic identification of construction occurrences is a desirable objective, both for NLP applications and construction analysis. In this paper we chose to look at construct detection for the purpose of corpus studies, and focus on constructions that are difficult to find using available corpus search tools.

As we have briefly presented above, automatic identification of constructions can be perceived as a two-way problem. Either, following in the footsteps of Wible and Tsao (2010) and Forsberg et al. (2014), we could mine corpora for hybrid n-grams in order to automatically detect and extract frequently occurring construction patterns. Or, as Hwang, Nielsen and Palmer (2010) demonstrated, we could start with an already defined construction and employ different features to detect instances of that construction in running text. In this thesis we are interested in the latter problem, but we expand the task and aim to create a flexible system that can be used on a wide range of constructions without access to previously annotated datasets.

We have formulated the problem of construction detection as a search problem and frame it as an information retrieval task. From the perspective of constructicography, the advantages of this approach are manyfold.

First, it enables flexibility and interaction. We do not know beforehand what constructions the user will be searching for and therefore user interaction should be catered for, allowing users full freedom to formulate and reformulate queries on the fly.

Second, it supports ranking instead of classification. Construction search is not a binary task, making too strict predictions about construction restrictions beforehand is counter-intuitive since we want the search tool to be useful at an early stage of construction analysis. Semantic features cannot be hardwired before the lexical variation of particular slots have been carefully investigated. For the same reasons, maximizing accuracy is not an end goal; from a constructionist perspective borderline cases are of special interest. So instead of custom made solutions, we give the user an additional set of tools to rank and re-rank the search with.

Third, evaluation is meaningful. In analogy with other IR problems, we are only interested in measuring how good the system is at retrieving relevant hits. The rest of the answer set would normally consist of a varied lot of constructions that are difficult to sort into different classes, especially beforehand. The user is expected to know what she is looking for, but does not necessarily need to know what she is *not* looking for.

The guiding principles behind the design of the construction search system are:

- **Work with what we have**. Build the system on top of the existing corpus infrastructure available at the University of Gothenburg. Test and evaluate the system with constructions defined in the Swedish Constructicon. For similarity features, use in-house lexical semantic resources SALDO and Swedish Framenet as well as distributional models constructed from available corpora.
- **Flexibility**. Test and evaluate the system on several constructions with different properties and restrictions to make sure it is flexible enough to handle variation.
- **Short takeoff**. No hand-labeled training set to start with, instead the user is asked to guide the system by selecting a small number of positive examples for every new query.
- **Simple ranking function**. As a starting point, implement a simple and computationally effective reranking function that works with just a few training examples.

3.2 Infrastructure

The construction search system is heavily influenced and inspired by the local environment at the Department of Swedish at the University of Gothenburg. Corpus linguistics, constructicography and linguistic resource building are all profiled research areas at the department, and all the infrastructure needed for a swift takeoff is already in place, under the hood of the research and development unit *Språkbanken*[3], the Swedish language bank (Borin et al., 2012a). In the following we will present all the resources that play a part in the narrative: the corpus infrastructure *Korp*, it's corpora and annotations; the lexical semantic network *SALDO*; the *Swedish Framenet* and the *Swedish Constructicon*.

3.2.1 Korp

We built the construction search tool on top of Språkbanken's corpus infrastructure Korp (Borin et al., 2012b). The infrastructure comprises modules for importing and annotating corpora, web services for searching and retrieving information from the corpora, and a graphical user interface for facilitated search. Korp stores a large collection of Swedish corpora, currently around 10 billion tokens, mostly modern written Swedish. The corpora are automatically processed and annotated with the following types of linguistic information: tokenization, sentence splitting, links to SALDO senses (see section 3.2.2 below), lemmatization, compound analysis, part-of-speach tags, morphosyntactic tags, and syntactic dependency trees.

The backend web service allows users to pose structural queries using the CQP language (Christ, 1994), where a query is expressed as a regular expression over certain conditions, and conditions

[3] `http://spraakbanken.gu.se`

are attribute-value pairs of position attributes from the annotation scheme. Expressions can be concatenated to form complex queries and standard regular operators are available. Below are a few examples of valid CQP expressions, where (4) matches any arbitrary number of repetitions of the word *ha*; (5) matches any sequence of tokens with the part-of-speech tags cardinal number, noun, adjective; (6) matches any token labeled either as a determiner or a cardinal number; (7) matches any token that is both tagged as a noun and with the dependency relation of a direct object; (8) matches the word *sink* if it is not labeled as a noun; (9) matches any token that has the morphosyntactic tag of an indefinite noun in the singular; and (10) introduces a wildcard expression that allows an intervening number (in this case between zero and two) of unconditioned tokens between a determiner and a noun.

(4) [word="ha"]*

(5) [pos = "RG"] [pos = "NN"] [pos = "JJ"]

(6) [(pos = "DT" | pos = "RG")]

(7) [(pos = "NN" & deprel = "OO")]

(8) [(word="sink" & pos =! "NN")]

(9) [(msd = ".*NN.NEU.SIN.IND.NOM.*" | msd = ".*NN.UTR.SIN.IND.NOM.*")]

(10) [pos = "DT"] []{0,2} [pos = "NN"]

3.2.2 SALDO

All lingustic resources under Språkbanken's roof are linked together by one primary lexical resource, the pivot SALDO (Borin et al., 2013). SALDO is a hand-crafted morphological and lexical-semantic lexicon for modern Swedish, with more than 125 000 entries and growing. The lexicon covers all parts of speech (not just open classes) and also multiword expressions, like lexicalized compounds and phrasal verbs.

As a lexical semantic network, the word senses in SALDO are organized hierarchically by association and each word sense is given one or more semantic descriptors which are also entries in the database. The obligatory *primary* descriptor is defined as a more central word in the same semantic neighbourhood as the entry. In this way, each word sense (except a root) is connected to another sense, from the periphery to the core. The descriptors are not labeled with classical semantic relations, but would typically correspond to a synonym or a hypernym. Figure 6 shows a fragment of the SALDO network from the words *vampyr* 'vampire', *papegoja* 'parrot', *pingvin* 'penguin' and *pojke* 'boy' via the primary descriptors up to the root node *PRIM*.

In this study, we use the SALDO network as a way to model semantic similarity between word senses, and introduce it as a feature in our reranking function.

3.2.3 Swedish Framenet

The Swedish framenet, SweFN (Friberg Heppin & Toporowska Gronostaj, 2012), is a lexical-semantic resource based on frame semantic theory and constructed in line with the Berkeley English Framenet (Baker et al., 1998; Fillmore & Baker, 2009). In a framenet, lexical units (LUs) are defined by cognitive frames that they presumably evoke in the mind of a language user. The frames can be described as conceptual scenes populated by frame specific elements (FEs) representing all the participants, props or states of affairs needed to make sense of language in

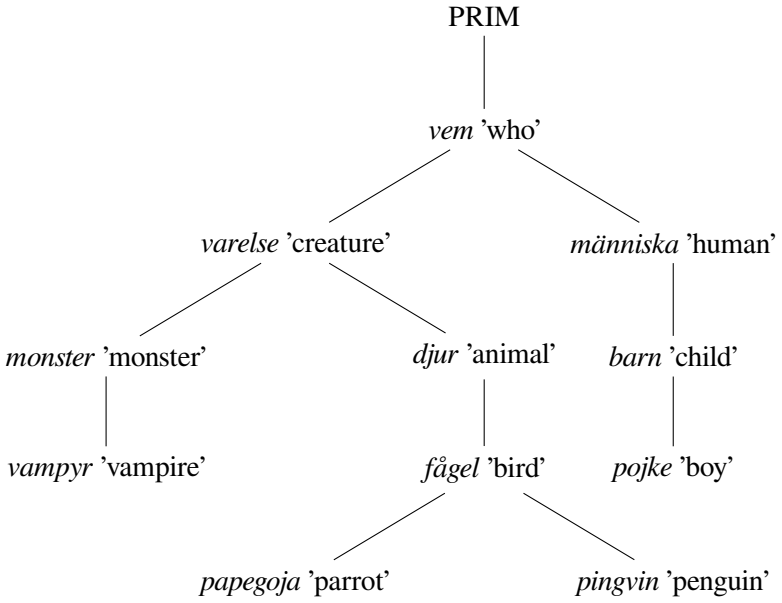


Figure 6: Network example of SALDO association paths.

context. Frames are organized together in a hierarchical semantic network of frame-to-frame relations such as inheritance, causation or perspective.

To illustrate, the frame Prevarication is described as a situation where a SPEAKER communicates about a TOPIC in such a way as to mislead an ADRESSEE. Lexical units like *bullshit* (v), *kid* (v), *lie* (n) and *fool* (v) evoke the Prevarication frame and for each lexical unit some example sentences are collected from corpora and annotated with frame elements to exemplify different distributional patterns of the LUs, like in table 2[4].

Frame	Prevarication
Frame Elements	SPEAKER, TOPIC, ADRESSEE
Lexical Units	*bullshit, kid, lie, fool*
Example	[We]$_{\text{SPEAKER}}$ might [kid]$_{\text{LU}}$ [ourselves]$_{\text{ADRESSEE}}$ [that life is perfectly OK as it is.]$_{\text{TOPIC}}$

Table 2: Prevarication frame.

The semantic network of frames developed by the Berkeley FrameNet project has been used as a starting point for framenets in other languagues, and the Swedish Framenet has built most of its network, including frame definitions, frame elements and frame-to-frame relations, on the English counterpart. The focus of the Swedish project has instead been on coverage, and with over 34000 lexical units in about 1200 frames it has the largest framenet lexicon to date (Ahlberg et al., 2014).

[4]Example frame taken from Berkeley FrameNet, `https://framenet.icsi.berkeley.edu/fndrupal/home`

In this thesis we use SweFN as a resource to capture semantic similarity between lexical units that belong in the same frame, and introduce it as a feature in our reranking function.

3.2.4 Swedish Constructicon

Finally, let us turn to the lexical resource that plays the lead character in this cast, the Swedish Constructicon, SweCcn (Lyngfelt et al., 2012). The Swedish Constructicon is an online repository of construction descriptions intended for multiple purposes including (second) language pedagogy and language technology. While it was modelled after the Berkeley English Constructicon, it has considerably outgrown its precursor. It currently covers around 400 Swedish constructions, many of which are partially schematic patterns of the kind that is hard to account for from a grammatical or lexical perspective alone.

In SweCcn each construction is described in a construction entry with up to fifteen fields of information, most importantly a *structure sketch*, a free-text *definition* and a number of annotated *example sentences*. The concise format resembles a classical dictionary entry, except that both form and meaning must be accounted for, so the description aims to capture all relevant characteristics of each construction in terms of grammatical structure, semantics, pragmatics and distribution. The free text definition describes the meaning of the construction as a whole as well as semantic roles of individual construction elements. The structure sketch is a simple linear representation of the construction's grammatical form, and includes information about part-of-speech, phrase type, grammatical function and/or fixed lexical items of each element. A translated illustration of a construction entry (including only the fields relevant for our current purposes) can be seen below in table 3, with the example sentence in glossed translation in (11).

Name	N-NÖRD 'N-nerd'
Definition	[Somebody]$_{\text{PROTAGONIST}}$ with a big [special interest]$_{\text{SPECIAL\_INTEREST}}$. Compound construction with two nominal constituents N+N where the prefix describes the special interest and the suffix describes the protagonist.
Structure sketch	N+N
Example	*Bara en vanlig tjej med brinnande intresse för korrekt språk. [[Grammatik]*$_{\text{SPECIAL\_INTEREST}}$*[freak]*$_{\text{PROTAGONIST}}$*]*$_{\text{N-NÖRD}}$ *kan man nog säga.* Just an ordinary girl with a passion for correct language. [[Grammar]$_{\text{SPECIAL\_INTEREST}}$[freak]$_{\text{PROTAGONIST}}$]$_{\text{N-NÖRD}}$ you may call it.

Table 3: SweCcn entry for N-NÖRD.

(11) Bara en vanlig tjej med brinnande intresse för korrekt språk. [Grammatikfreak]
Only an ordinary girl with burning interest for correct language. Grammar freak
kan man nog säga.
can.AUX one probably say.INF
'Just an ordinary girl with a passion for correct language. Grammar freak you may call it.'

The Swedish Constructicon is a living resource under development and the work with identifying and analyzing constructions for entry into the database is firmly usage-based. With its 400 unique entries, the SweCcn offers a wide and relevant selection of constructions to choose from, thus providing a suitable testing ground for the study at hand.

3.3 Design and interaction

Now we will walk through the design of the search system and explain in some detail how the user is expected to interact with the search tool to tailor the ranking function and inspect the results. To illustrate the process, let us reuse the example from the introduction - the KVANTIFIERANDE_GENITIV.TID construction that measures the duration of an activity, as in example (12) below.

(12) Femtusen dollar för [fem sekunder-s arbete].
Five thousand dollar.PL for five second.PL-GEN work.INDEF
'Five thousand dollars for five seconds' work.'

The first step when searching for a particular construction is to translate the formal construction description into a query for the Korp search system. The KVANTIFIERANDE_GENITIV.TID construction has the syntactic description DET N_{GEN} N_{INDEF}, which shows that its surface form consists of a determiner (or number), a noun in the genitive, and finally an indefinite head noun. Since Korp corpora are annotated with dependency labels rather than phrase markers, the CQP query will necessarily have to be a simplified linear search string. Each expression in the CQP query corresponds to a certain slot in the construction. In this case we formulate a CQP query with three slots, first a determiner/number, then a noun in the genitive, followed by an indefinite noun. Below in table 4 is an example sample of a few sentences returned by this initial search query, displayed as KWIC concordances with the the stretch of the sentence corresponding to the search hit displayed as positions in the construction. True instance of the KVANTIFIERANDE_GENITIV.TID construction are highlighted with green.

	1	**2**	**3**	
Han höll cigarettändaren på	*en*	*armlängds*	*avstånd*	*och tände den.*
Kinden trycktes mot	*nån*	*sorts*	*stybb*	*eller grus.*
Vi behövde	*tre*	*fjärdedels*	*majoritet*	*i stortinget.*
Jag hade mer än	*fem*	*års*	*erfarenhet*	*som vice ordförande.*
Nu är det din tur för	*en*	*gångs*	*skull.*	

Table 4: Sample KWIC concordances for initial search result.

As expected for this construction, the initial search query returns a search result 'contaminated' with many other patterns than the time-duration construction we are looking for. Among the concordances in table 4 only the forth hit is in fact an instance of KVANTIFIERANDE_GENITIV.TID - *fem års erfarenhet* 'five year's experience'. The rest of the hits are other noun phrases containing a genitive, including the related scale construction *tre fjärdeldels majoritet* 'three quarters' majority',

en armlängds avstånd 'arm's length', and the partially fixed multiword expressions *nån sorts stybb* 'some kind of coaldust' and *för en gångs skull* 'for once'.

To address this problem we now apply a *reranking* function. The user is asked to provide the system with two additional types of information: (1) a number of positive examples of sentences containing true instances of the construction she is looking for and (2) what semantic features to consider for particular slots in the search string. Typically, the slots that are of interest are variable slots open for content words, where the user expects certain semantic restrictions to be at play. Regardless if these restrictions are already defined, or if the user is only following her intuition, it is enough to point the system in the right direction.

For instance, in our current example, the user could say that the ranking function should consider the distributional similarity function based on the second and third word in the hit, e.g. the time word and the activity word, and then select a number of occurrences such as *an hour's rest, three years' study*. With a carefully designed ranking function and representative seed examples, the system can rank time/activity expressions above other expressions matching that surface pattern.

After the reranking function has been applied the system returns the whole answer set again, ordered by relevance. If the ranking is effective, the top ranked hits for our query could look something like the example in table 5 below, again true instances of the KVANTIFIERANDE_GENITIV.TID construction are highlighted with green.

	1	**2**	**3**	
Åttatusen, för	*två*	*års*	*arbete.*	
Han dömdes till	*femton*	*års*	*fängelse*	
Bosch blev ordinerad	*sex*	*veckors*	*vila*	*i hemmet.*
Vid	*femton*	*års*	*ålder*	*blev stugan mig trång.*
Jag försätter er på fri fot med	*två*	*års*	*prövotid*	*och med Zeus som övervakare.*

Table 5: Sample KWIC concordances for reranked search result.

Of the five sentences above, four are true instances of the time construction, and one is a false positive - *femton års ålder* 'fifteen years of age'. Note that most of the top-ranked sentences contain the same noun in genitive, *år* 'year', e.g. *femton års fängelse* 'fifteen year's prison', *två års arbete* 'two year's work'. Since the ranking function compares new sentences to positive seed examples, this is most likely due to several of the positive sentences containing occurrences with 'year'. From a closer inspection of the concordances we can draw the conclusion that the lemma 'year' (or other time units) is not a strictly delimiting feature to disambiguate the time construction, since years are also units on the scale associated with age (making this false positive an instance of the scale construction described earlier). If the user wishes to proceed from this stage, she could try reranking again, this time only looking at the third position and thereby asking the system to try to predict what different types of activity-nouns fit into this particular slot.

To recapitulate, the complete work flow of the construction search system is illustrated in figure 7. The user starts by posing a corpus search query in CQP format, formulated to capture as many occurences of the particular construction as possible. Each expression in the query corresponds

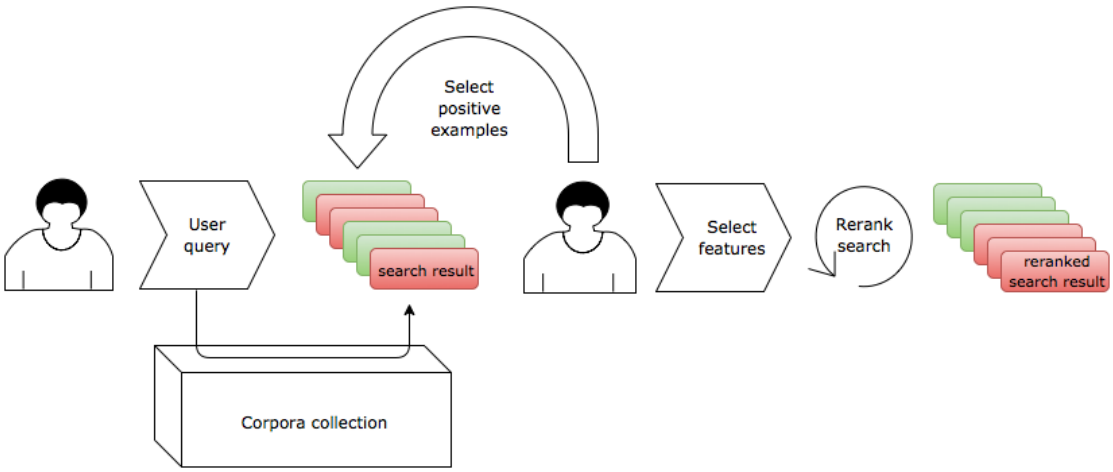


Figure 7: Construction search model.

to a position in the construction. The system matches the query against the corpora collection and returns an answer set with all sentences containing word sequences that fit the conditions. From here the user proceeds to inspect the concordances and select a number of positive example sentences, then indicate what features to consider for particular slots in the construction. Guided by the user input, the system applies a reranking function and returns the reranked search with top-ranked sentences first. The process can be iterated repeatedly to try out different semantic similarity features on different slots in the construction, or to select different positive example sentences and inspect how it affects the outcome.

3.4 Training the reranking function

The ranking process works by applying a scoring function to each hit returned by the corpus system, and then sorting the hits according to this score. In this work we have implemented a ranking function that can be trained simply and quickly, and that can be understood in terms of linguistically meaningful features.

Each hit can be analyzed by a given number of similarity functions selected by the user on the fly by indicating what features and what slots to consider. For each new sentence, the system extracts a feature vector representing the lexical content occupying the relevant slots, and computes the similarity scores in real-time by comparing the vector of the current sentence to the vectors extracted from the positively labeled examples in the hand-picked gold standard. The weight of each feature is set equally so that the final ranking score is computed as the *centroid* of the positively labeled instances.

This learning method is computationally effective and works well with just a few training examples, and as explained above it is intuitive for the users to select a small set of positive examples, while annotating negative examples is less meaningful. The simple scoring function also allows us to use any similarity function, not just numerical values.

3.5 Similarity functions

The corpus search system takes care of basic surface-oriented features (word forms, morphology, grammatical functions), so the central task of the reranker is to use representations of word meaning to go beyond the simple structural information. We investigate different measures: similarities based on hand-crafted lexicons, and distributional similarity computed from corpora.

3.5.1 Network-based similarity

We use the lexical-semantic lexicon SALDO to compute semantic similarity based on network distance.

To measure similarity between two SALDO entries, s_1 and s_2, we use the measure by Wu & Palmer (1994), based on proximity in the tree and the depth of the lowest common ancestor, C. This measure is a number between 1 and 0, where 1 is the score for two identical entries.

$$\text{wup-sim}(s_1, s_2) = \frac{2 \cdot \text{depth}(C)}{\text{depth}(s_1) + \text{depth}(s_2)}$$

A complication is that our corpora lacks sense annotation; however, since the first sense dominates overwhelmingly in corpora for most lemmas (Johansson et al., 2016), we use the first sense to compute the similarities.

3.5.2 Frame-based similarity

An alternative lexicon-based similarity function is based on the Swedish FrameNet. For our current purposes we chose to view frames as semantic classes. Intuitively, two words have a similar meaning if they belong to the same frame; for instance, *timme* 'hour' and *minut* 'minute' are related because they both belong to the frame CALENDRIC_UNIT. Again, we have to deal with the lack of word sense annotation in the corpora, so we define the similarity to be 1 if the two words share at least one frame, and 0 otherwise.

3.5.3 Distributional similarity

As an alternative to hand-crafted semantic resources we also add a semantic similarity feature based on a distributional model.

We trained `word2vec` (Mikolov et al., 2013b) on a 1-billion mixed corpus, preprocessed by lemmatizing the words and splitting compounds. We used the default settings, except the dimensionality which was set to 512. To compute the similarity between two words, we applied the cosine to their lemma vectors.[5] Again, this similarity is at most 1, which happens if the vectors are identical.

4 Benchmark collection

We built a new benchmark for evaluating construction retrival systems. It uses six different constructions taken from the Swedish Constructicon. The constructions are all partially schematic

[5] We pick the first lemma if lemmatization is ambiguous.

and contain at least one lexically fixed element combined with variable slots. They also share the property that the formal description cannot be translated into a CQP query that captures only true instances of the construction in question.

To create the collection, we called the corpus query system with the six respective queries in a corpus of contemporary fiction. Each collection of hits was annotated manually as true or false instances of the construction in question. Table 6 shows the statistics.

Construction	Total hits	True hits
V_*av*_NP	501	169
PROPORTION_*i/om*	1703	205
V_REFL.RÖRELSE	1300	338
KVANTIFIERANDE_GENITIV.TID	945	97
KVANTIFIERANDE_GENITIV.SKALA	945	166
AVGRÄNSAD_AKTION.På	2000	43

Table 6: Statistics for the benchmark.

4.1 Constructions

4.1.1 V_*av*_NP

V_*av*_NP 'V_*of*_NP' is a causal verb phrase construction where the event or state expressed by the verb is caused by the bare noun following the preposition *av* 'of'. The construction includes both literal and metaphorical causative relations such as *stinka av mögel* 'stink of mold' and *dö av skam* 'die of shame' but the formal description V *av* NP_{BARE} also returns search hits with phrasal verbs like *dra av moms* 'subtract sales tax' and passive voice constructions like *läses av flickor* 'read by girls'. The variable construction slots are not restricted to any obvious semantic classes, although emotions and physical states are clearly salient at first glance. Table 7 illustrates the SweCcn entry for V_*av*_NP in English translation (for a glossed translation of the example sentence see (13) below).

Name	V_*av*_NP 'V_*of*_NP'
Definition	[Somebody]$_{\text{EXPERIENCER}}$ [is affected by an event or enters into a state]$_{\text{EVENT}}$ caused by [something else, expressed by an event noun]$_{\text{CAUSE}}$.
Structure sketch	V *av* NP_{BARE}
Example	*Jag känner att [jag]*$_{\text{EXPERIENCER}}$ *[[rodnar]*$_{\text{EVENT}}$ *[av]*$_{\text{AV}}$ *[ilska]*$_{\text{CAUSE}}$*]*V_*av*_NP· I feel that [I]$_{\text{EXPERIENCER}}$ [[blush]$_{\text{EVENT}}$ [with]$_{\text{AV}}$ [anger]$_{\text{CAUSE}}$]V_*av*_NP·

Table 7: SweCcn entry for V_*av*_NP.

(13) Jag känner att jag [rodnar av ilska].
I feel.PRS that I blush.PRS of anger.INDEF
'I feel that I blush with anger.'

4.1.2 proportion_*i*/*om*

PROPORTION_*i/om* 'proportion_*in/about*' is a rate construction that combines two entities, a numerator and a denominator, joined by the preposition *i* 'in' or *om* 'about'. The construction is restricted to temporal relations such as frequency and speed, but also salary rates fit into the scheme[6]. Search hits for the formal description NP_{INDEF} [I | OM] N_{DEF} captures occurrences like *två gånger om dagen* 'two times a day' and *åttio pesos i månaden* 'eighty pesos per month' as well as false hits like *tio dagar i fängelset* 'ten days in jail'. Table 8 shows the SweCcn entry for PROPORTION_*i/om* with the example sentence in glossed translation in (14) below.

Name	PROPORTION_*i/om* 'proportion_*in/about*'
Definition	The frequency of an iterative [action]$_{EVENT}$ is expressed by the [quantity of a mass]$_{QUANTITATIVE}$ in relation to [a time unit in the definite]$_{CONTEXT}$.
Structure sketch	NP_{INDEF} [I | OM] N_{DEF}
Example	*Han drack sällan mer än [[en flaska]*$_{QUANTITATIVE}$ *[om]*$_{OM}$ *[dagen]*$_{CONTEXT}$*]*$_{PROPORTION\_i/om}$. He rarely drank more than [[one bottle]$_{QUANTITATIVE}$ [a]$_{OM}$ [day]$_{CONTEXT}$]$_{PROPORTION\_i/om}$.

Table 8: SweCcn entry for PROPORTION_*i/om*.

(14) Han drack sällan mer än [en flaska om dagen].
He drink.PST rarely more than one bottle.INDEF about day.DEF
'He rarely drank more than one bottle a day.'

4.1.3 V_refl.rörelse

V_REFL.RÖRELSE 'V_reflexive.motion' is a self-motion construction where an actor expressed with a reflexive traverses a path in a direction from a place or towards a goal. The verb typically describes the means or manner of the motion, while the prepositional/adverbial phrase contributes the direction. The formal description V PN_{REFL} [PP | ADVP] captures prototypical occurences like *sätta sig ned* 'sit down' and *ta sig fram* 'make one's way' as well as instances with verbs that do not usually indicate motion, like *svetta sig igenom* 'sweat one's way through' and *läsa sig bakåt* 'read one's way backwards'.

[6]The related rate construction PROPORTION_*per* can be distinguished from PROPORTION_*i/om* in that it uses the preposition *per* 'per' to combine a numerator with an indefinite denominator. The PROPORTION_*per* construction can also be used for a broader domain of rates, temporal relations as well as relational quantity and price rates. Since rate constructions with 'per' are easy to spot in corpora, we will not include it in this investigation.

The CQP query for this construction cannot be formulated to capture only reflexive pronouns since first and second person reflexives share the same surface form as ditto object pronouns in Swedish. That means that we have no way of indicating that the reflexive should be co-indexed with the agent of the clause.

False hits in the search result include common reflexive verbs, like *känna sig glad* 'feel happy' and *anförtro sig åt* 'confide in' as well as certain lexicalized multiword expressions like *tränga sig på* 'intrude' and *sätta sig på tvären* 'be obstinate'. Since the direction can be expressed either by a preposition or an adverb, we get a few search hits with an intervening manner adverbial further describing the manner of the motion, like the example in (15). These have been annotated as false hits, so as not to confuse the reranking function with contradictory features.

(15) Bosch [satte sig försiktigt] på sängkanten
Bosch sit.PST REFL carefully on bedside.DEF
'Bosch carefully sat down on the bedside'

Table 9 shows the SweCcn entry for V_REFL.RÖRELSE with the example sentence in glossed translation in (16) below.

Name	V_REFL.RÖRELSE 'V_reflexive.motion'
Definition	[An actor]$_{\text{ACTOR}}$, expressed with a [reflexive]$_{\text{REFL}}$, [moves]$_{\text{V}}$ [in a direction, traverses a path, from a place or to a place]$_{\text{LOCATIVE}}$. The verb usually describes the manner of the motion. The construction also encompasses actions that indicate intended motion and non-motion.
Structure sketch	V PN$_{\text{REFL}}$ [PP | AdvP]
Example	*[Vi]*$_{\text{ACTOR}}$ *försökte [[gräva]*$_{\text{V}}$ *[oss]*$_{\text{REFL}}$ *[ut från huset]*$_{\text{LOCATIVE}}$ *]*$_{\text{V\_REFL.RÖRELSE}}$. [We]$_{\text{ACTOR}}$ tried to [[dig]$_{\text{V}}$ [us]$_{\text{REFL}}$ [out of the house]$_{\text{LOCATIVE}}$]$_{\text{V\_REFL.RÖRELSE}}$.

Table 9: SweCcn entry for V_REFL.RÖRELSE.

(16) Vi försökte [gräva oss ut] från huset.
We try.PST dig.INF REFL out from house.DEF
'We tried to dig our way out of the house.'

4.1.4 kvantifierande_genitiv.tid

The KVANTIFIERANDE_GENITIV.TID 'quantifying_genitive.time' construction has already been properly introduced. It is defined as a genitive modifier that specifies the duration of an activity. The formal description Det N$_{\text{GEN}}$ N$_{\text{INDEF}}$ captures true instances of the construction such as *en stunds tystnad* 'a moment's silence', but also the related scale construction *fem meters djup* 'five meter's depth', as well as other genitives like *en människas skugga* 'the shadow of a human'. The SweCcn entry for KVANTIFIERANDE_GENITIV.TID is illustrated in table 10 and the example sentence can be read in glossed translation in (17) below.

Name	KVANTIFIERANDE_GENITIV.TID 'quantifying_genitive.time'
Definition	[The genitive modifier]$_{\text{QUANTITATIVE}}$ specifies the duration in time for [the activity expressed by the head noun]$_{\text{THEME}}$.
Structure sketch	DET N_{GEN} N_{INDEF}
Example	*Efter [[två dagars]*$_{\text{QUANTITATIVE}}$ *[förhör]*$_{\text{THEME}}$ *]*$_{\text{KVANTIFIERANDE\_GENITIV.TID}}$ *har han erkänt att han var med i bilen.* After [[two days']$_{\text{QUANTITATIVE}}$ [interrogation]$_{\text{THEME}}$]$_{\text{KVANTIFIERANDE\_GENITIV.TID}}$ he has confessed that he was present in the car.

Table 10: SweCcn entry for KVANTIFIERANDE_GENITIV.TID.

(17) Efter [två dagar-s förhör] har han erkänt att han var
After two day.PL-GEN interrogation have.AUX he confess.PST that he be.PST
med i bilen
with in car.DEF
'After two days' interrogation he has confessed that he was present in the car.'

4.1.5 kvantifierande_genitiv.skala

The scale construction KVANTIFIERANDE_GENITIV.SKALA 'quantifying_genitive.scale' is the sibling of the time construction described above. Here, the genitive modifier specifies the value on a scale expressed by the noun phrase. It shares the formal description DET N_{GEN} N_{INDEF} with KVANTIFIERANDE_GENITIV.TID but is semantically restricted to scalable measures like height, depth, age, size and other things that can be quantified. Consequently, the search string captures both true instances of the scale construction like *tusen meters höjd* 'thousand meter's height' and *elva månaders hyra* 'eleven month's rent' as well as the aforementioned time construction *tre timmars seglats* 'three hour's sailing trip'. Again, also other genitive phrases like *en kvinnas fot* 'the foot of a woman' end up in the search batch. Table 11 shows the SweCcn entry for KVANTIFIERANDE_GENITIV.SKALA with the example sentence in glossed translation below in (18).

Name	KVANTIFIERANDE_GENITIV.SKALA 'quantifying_genitive.scale'
Definition	[The genitive modifier]$_{\text{QUNATITATIVE}}$ indicates the value on [a scale expressed bu the head noun]$_{\text{UNIT}}$.
Structure sketch	DET N_{GEN} N_{INDEF}
Example	*EES-avtalet antogs med [[tre fjärdedels]*$_{\text{QUANTITATIVE}}$ *[majoritet]*$_{\text{UNIT}}$ *]*$_{\text{KVANTIFIERANDE\_GENITIV.SKALA}}$. The EES-agreement was passed with [[three quarters']$_{\text{QUANTITATIVE}}$ [majority]$_{\text{UNIT}}$]$_{\text{KVANTIFIERANDE\_GENITIV.SKALA}}$.

Table 11: SweCcn entry for KVANTIFIERANDE_GENITIV.SKALA.

(18) EES-avtalet antogs med [tre fjärdedelar-s majoritet].
EES-agreement pass.PST.PASS with three quarter.PL-GEN majority
'The EES-agreement was passed with three quarters' majority.'

4.1.6 avgränsad_aktion.*på*

The AVGRÄNSAD_AKTION.*på* 'bounded_event.*on*' construction is a time expression that modifies the duration in time of a completed action. It is a specific and rather restricted instance of a more general pattern for prepositional time adverbials. The construction can only be used with events of bounded aspect, and thereby specifies the time required to complete the event. For negated events, time adverbials with *på* 'on' can also be used to describe the duration that has passed since an event took place – a separate construction entered in SweCcn as TIDSANGIVELSE.POLARITET 'specified_time.polarity'. The structure sketch *på* NP naturally translates to a search query that captures all prepositional phrases with the preposition *på*, and even though the noun slot is strictly restricted to time expressions it is impossible to delimit it from related time constructions without taking in the context. We include it the collection as an example of a rare construction (only 42 true instances out of 2000 hits in the initial search batch) that can be lifted to the surface of the answer set by clear semantic restrictions, but not fully disambiguated from other time constructions due to the limitations of our approach. Thus, search hits include true instances of the construction such as *rummet tömdes på några sekunder* 'the room was emptied in a few seconds' as well as false relatives like the negative construction *ingen hade samlat ved på ett år* 'nobody had been collecting wood for a year'. However, most of the answer set is full of all kinds of prepositional phrases like *på en stol* 'on a chair'. Table 12 shows the SweCcn entry for PROPORTION_*i/om* with the example sentence in glossed translation in (19) below.

Name	AVGRÄNSAD_AKTION.*på* 'bounded_event.*on*'
Definition	Modifies the [duration]$_{\text{QUNATITATIVE}}$ of a [bounded event]$_{\text{EVENT}}$.
Structure sketch	*på* NP
Example	Hon förvandlades från sju till tre år på ett ögonblick . *Hon [[förvandlades från sju till tre år]*$_{\text{EVENT}}$ *[på]*$_{\text{På}}$ *[ett ögonblick]*$_{\text{QUANTITATIVE}}$ *]*$_{\text{AVGRÄNSAD\_AKTION.}på}$. She [[was transformed from seven to three years old]$_{\text{EVENT}}$ [in]$_{\text{På}}$ [an instant]$_{\text{QUANTITATIVE}}$]$_{\text{AVGRÄNSAD\_AKTION.}på}$.

Table 12: SweCcn entry for AVGRÄNSAD_AKTION.*på*.

(19) Hon förvandlades från sju till tre år [på ett ögonblick].
She transform.PST.PASS from seven to three year.PL on one instant.INDEF
'She was transformed from seven to three years old in an instant.'

5 Experiments

In this section we will inspect and evaluate the effect of the construction search system when tested on the constructions from the benchmark. We start by summarizing the results in section 5.1 and 5.2 and then go on to carry out a closer inspection of each individual construction search

in section 5.3. Here we go beyond ranking scores to analyze how well our approach manages to capture construction deliminating features, and we chose some false positives among the top-ranked sentences to illustrate shortcomings of the ranking function. This qualitative evaluation is telling, since it may also exemplify how a constructicographer would go about defining delimiting characteristics of specific constructions.

Primarily we are interested in two things. First, what similarity features work best for detecting different constructions. Second, how the number of seed examples affects the ranking quality.

Since there are many variable factors that may effect the outcome we will keep some variables stable while performing the experiments. For each evaluation run we apply one similarity feature at a time to the same number of slots for each construction. We also keep the selection of positive seed examples stable by always picking the N first true instances from the benchmark collection.

5.1 Evaluation of features

We first investigated the effect of the choice of lexical similarity. Table 13 shows the average precision scores for three different similarities: Wu–Palmer in SALDO, frame-based, and distributional. As a baseline we include a lemma-based similarity that would correspond to just formulating the search with a number of specified lemmas in the variable construction slots (that is, we get exactly what we asked for and nothing else). As seed examples, the rerankers were trained on the first 15 positively labeled instances in the collection.

Construction	lemma	SALDO	frame	distr
V_*av*_NP	0.69	0.73	0.63	**0.86**
PROPORTION_*i*/*om*	0.64	0.68	**0.95**	0.74
V_REFL.RÖRELSE	0.59	0.53	**0.61**	0.56
KVANTIFIERANDE_GENITIV.TID	0.40	0.48	**0.60**	0.49
KVANTIFIERANDE_GENITIV.SKALA	0.64	0.63	0.52	**0.68**
AVGRÄNSAD_AKTION.PÅ	0.43	0.51	0.36	**0.60**

Table 13: Effect of the lexical representation.

The result clearly shows that reranking based on a lexical-semantic model can give very strong improvement over the lemma-based baseline. However, it should be noted that there is considerable variation in the result. For instance, for PROPORTION_*i*/*om* and KVANTIFIERANDE_GENITIV.TID, the frame-based reranker outperform the others significantly. This is probably because these constructions are clearly restricted to time-related words that nicely correspond to temporal framenet frames such as Calendric_unit and Measure_duration. In the case of V_*av*_NP, KVANTIFIERANDE_GENITIV.SKALA and AVGRÄNSAD_AKTION.PÅ it is instead the distributional model that works best. Considering that these three constructions are very different from each other, we will go on to analyze the effect of the distributional feature on each construction independently. We expect the distributional model to work best when the slotfillers are not restricted to a narrowly defined semantic class, but instead belong to a broader semantic domain, like scalable measures in the case of KVANTIFIERANDE_GENITIV.SKALA or emotional states for V_*av*_NP. When

the distributional similarity achieves the highest average precision, the frame-based reranker is consistently the worst in class.

In the case of V_REFL.RÖRELSE the frame-based reranker just barely beats the lemma-based baseline, and the other similarities perform worse. Here, it seems that none of our features managed to generalize over seen instances in a way that helped distinguish the construction beyond particular lemmas. We will analyse this shortcoming and what it can tell us about the insufficiency of the search system in the following.

The network-based SALDO similarity does better than the baseline in most cases, but never outperforms the other similarity features. It is difficult to speculate about why, but we can at least conclude that framenet frames are better at capturing narrowly defined semantic classes and distributional models do better at generalizing beyond taxonomic similarity scores.

5.2 Evaluation of number of seed examples

We next considered how the number of seed examples affects the ranking quality. Table 14 shows the average precision values for different number of seeds. We used the best reranker from the feature evaluation for each construction.

Construction	1	5	10	15
V_*av*_NP	0.65	0.81	**0.87**	0.86
PROPORTION_*i*/*om*	0.70	0.82	**0.95**	0.95
V_REFL.RÖRELSE	0.33	**0.61**	0.61	0.61
KVANTIFIERANDE_GENITIV.TID	0.32	0.50	0.58	**0.60**
KVANTIFIERANDE_GENITIV.SKALA	0.48	0.62	0.64	**0.68**
AVGRÄNSAD_AKTION.PÅ	0.31	0.57	**0.60**	0.60

Table 14: Effect of the number of seed examples.

As expected, the scores increase as the number of seeds grows. However, the quality is high even with a small number of seeds, which is important for the usability of an interactive system. It is also worth noting that the steep jump in average precision appears already between 1 and 5 seeds, after that the score stabilizes. Between 10 and 15 seeds the results are just marginally improving for a few of the constructions. It seems that 10 examples are enough to reach about as far as we can get with this way of ranking based on semantic similarity. More seed examples will probably not improve the average precision much, instead we must consider other features or get better at spotting false positives.

5.3 Qualitative evaluation

In this section we conduct a qualitative evaluation of the search system by analyzing each construction search on its own terms. We inspect the reranked concordances with particular focus on false positives – high ranked sentences that are not true instances of the construction we are looking for. False positives are telling indicators of the shortcomings of our approach; since the

reranking function is based solely on positive seed examples we have no way of downranking undesired hits. We will also scroll down the concordances and say something about false negatives – true occurences that for one reason or another end up near the bottom of the reranked search list.

5.3.1 V_*av*_NP

We expected the V_*av*_NP construction to be a hard nut to crack because of its great productivity and high lexical variation, but the distributional similarity function is performing beyond expectations, as can be seen in the precision/recall curve in figure 8. The distributional model succeeds in generalizing beyond seen instances and gives high ranking scores to a wide array of new and creative occurrences of the V_*av*_NP construction.

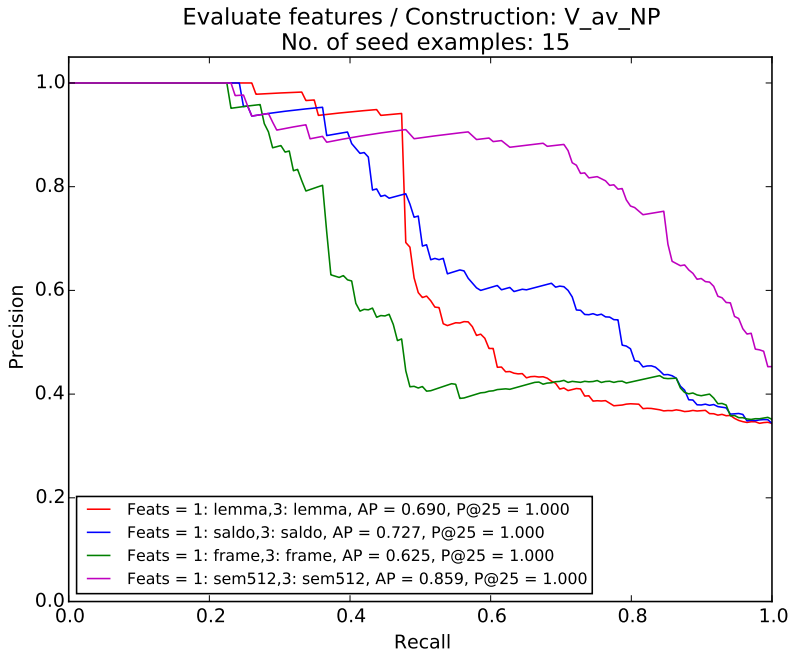


Figure 8: Precision/recall curve for retrieving the V_*av*_NP construction. 15 training instances are used, and different types of lexical-semantic features.

Although there are no distinct semantic restrictions on the variable slots of the V_*av*_NP constructions, the state or event caused by the noun is typically physical or emotional. Representative verbs include *darra* 'shiver', *lida* 'suffer', *dö* 'die', *gråta* 'cry', *rodna* 'blush', *skälva* 'quiver', *kvida* 'whimper', *flåsa* 'pant' and *skaka* 'shake'. The noun slot is typically occupied by event nouns such as *raseri* 'rage', *smärta* 'pain', *ansträngning* 'exertion', *ängslan* 'anxiety', *migrän* 'migraine', *förälskelse* 'infatuation' and *upphetsning* 'excitement'. The distributional model excels in quantifying the common denominator between these words, where the frame-based similarity falls short. The slot fillers display so much lexical variation that semantic frames manage to capture just a fraction of them.

A closer inspection of the reranked list of concordances for the best performing distributional model reveals that top ranked false positives belong to semantically very similar passive constructions, like the one in example (20), which is arguably displaying many of the same properties as

the true occurence in (21). This is not unreasonable, since the agentive adverbial of passive voice constructions can also indicate causation, and the two examples share the bare noun feature that makes this construction interesting. Moreover, if we wanted to exclude the passive examples we could easily introduce a morphosyntactic feature and get rid of them at a structural level.

(20) Mutt-s ansikte [förvreds av raseri].
Mutt-GEN face contort.PST.PASS of rage.INDEF
'Mutt's face was contorted by rage.'

(21) Hennes röst [skakar av ilska], trötthet och rädsla .
Her voice tremble.PRS of anger.INDEF, fatigue.INDEF and fear.INDEF
'Her voice is trembling with anger, fatigue and fear.'

Less prototypical instances of the construction where the cause and effect are not limited to human experiences receive lower ranking scores, both examples in (22) and (23) are ranked somewhere in the middle of the search batch, with many unrelated hits preceding them. Arguably neither *lök* 'onion' nor *stjärnor* 'stars' could be described as event nouns, nor are they semantically related in a way that could help our reranker identify other instances with untypical causative nouns. Simply looking at semantic similarity will not help us find such constructs.

(22) (...) en mun som [stank av lök] (...)
(...) a mouth that stink.PST of onion (...)
'a mounth that stank of onion'

(23) Nätterna [gnistrar av stjärnor].
Night.PL twinkle.PRS of star.PL
'The nights twinkled of stars.'

5.3.2 proportion_*i*/*om*

It is hardly surprising that our construction search system is good at detecting the PROPORTION_*i*/*om* construction or that the best performing similarity feature in this case is frame-based. The rate construction is strictly restricted to temporal relations, so the denominator will always be a time-related word belonging in a few well defined frames. Figure 9 shows to what extent the frame-based similarity outperforms all other ranking functions on this construction. The precision only drops in the end when the frame-based scoring function hands out identical ranking scores to a mixed set of sentences.

The decline in precision at the end of the curve neatly illustrates that the frame-based scoring function has an achilles heel – it is too coarse to detect less obvious cases of the construction. When many sentences receive the same ranking score they are sorted after the index number in the original search batch, which corresponds to no ranking at all. That is why sentences like (24) can get ranked higher than the true hit in (25).

(24) Vi har [några glasspinnar i frysen].
We have.PRS some ice lolly.PL in freezer.DEF

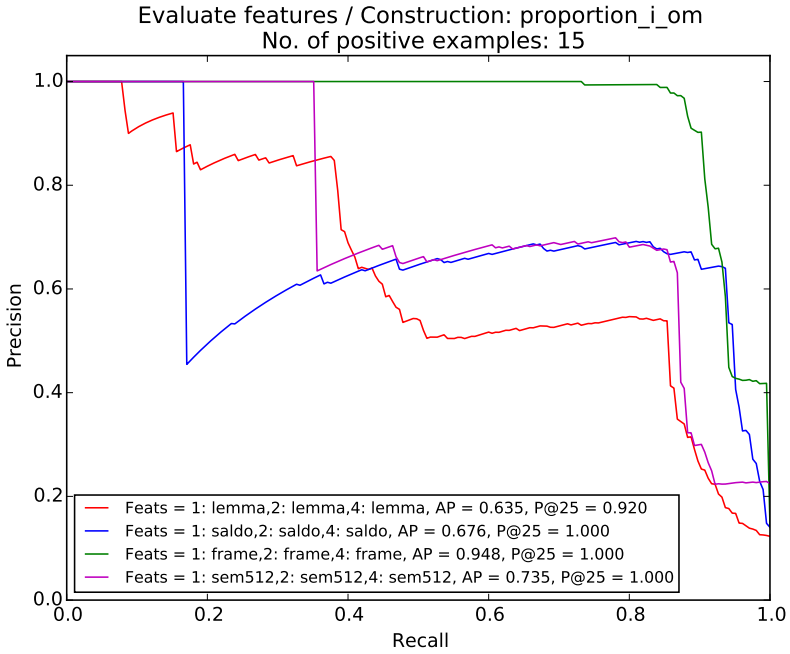


Figure 9: Precision/recall curve for retrieving the PROPORTION_*i*/*om* construction. 15 training instances are used, and different types of lexical-semantic features.

'We have some ice lollies in the freezer.'

(25) [Trettio ord i minuten] räcker.
Thirty word.PL in minute.DEF suffice.PRS
'Thirty words a minute is enough'

Among the false positives we also find a number of instances of a particular lexicalized multiword expression (illustrated in example (26) that fits the construction pattern but has to be considered a construction on its own - *en gång i tiden* 'once upon a time'. If we had a way of excluding such frequent phrases, it would be a quick-fix to improve the precision of the ranking function.

(26) [En gång i tiden] var han en motbjudande moralist.
One time in time.DEF be.PST he an obnoxious moralist.
'Once upon a time he was an obnoxious moralist.'

5.3.3 V_refl.rörelse

The evaluation results for V_REFL.RÖRELSE are quite disheartening. None of the ranking functions are particularly good at detecting the V_REFL.RÖRELSE construction, as can be seen by inspecting the curves in figure 10.

In this evaluation, the best performing frame-based reranker just barely beats the lemma-baseline and a closer look at the concordances for that reranked answer set shows a very confused outcome. The randomness of the result is in part caused by the lack of granularity in the frame-based

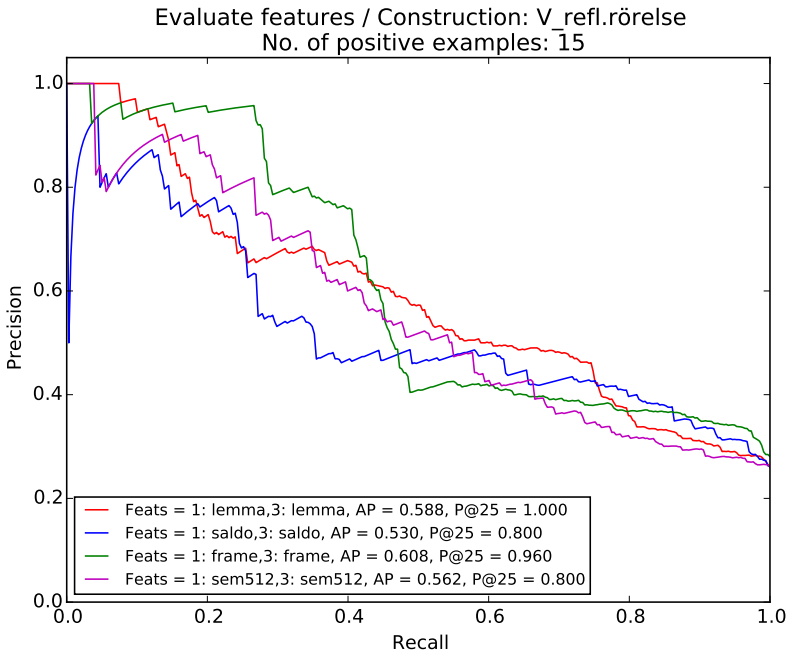


Figure 10: Precision/recall curve for retrieving the V_REFL.RÖRELSE construction. 15 training instances are used, and different types of lexical-semantic features.

scoring function, in part by a quite blatant simplification in the reranking function. Until now we have proceeded under the assumption that the similarity features for different construction slots are independent from each other, so that the task of the reranker is to find the common denominator for lexical items occupying each individual slot. In the case of the V_REFL.RÖRELSE this premise is simply not true. Most of the verbs occupying the verb slot combine with a few specific prepositions to form the motion construction. But this relation is lost when the features are extracted from the positive seed examples, verbs and prepositions are treated independently and the result is quite hap hazard. Another complication is that the goal of the motion is not included in the search string, and that makes it impossible to tell true occurrences like (27) from the false instance in (28).

(27) De [satte sig på] marken.
They sit.PST REFL on ground.DEF
'They sat down on the ground.'

(28) Hon [satte sig i] stället med huvudet i händerna.
She sit.PST REFL in stead.DEF with head.DEF in hand.PL.DEF
'She sat down with her head in her hands instead.'

The reranking function is further confused by the fact that even identical strings can be labeled true or false instances of the construction in question, depending on the context. In example (29) we see the phrase *ta sig till* 'make one's way' used to express directional motion, while example (30) shows the same string as a lexicalized multiword expression with the approximate meaning 'do'. There are several other lexicalized phrases like that in the answer set, like *sätta sig på tvären*

'be obstinate' (lit. 'sit the wrong way') opposed to the true instance *sätta sig på soffan* 'sit down on the couch'; or *tränga sig på* 'intrude' (lit. 'force oneself onto') opposed to the true instance *tränga sig in* 'force one's way in'. Such cases of transferred meaning could indicate a high productivity of the motion metaphor. In any case, our reranking function has no way of disambiguating them without taking in contextual features.

(29) Jag tror att du [tog dig till] båten för att ställa din man
I believe.PRS that you take.PST REFL to boat.DEF for to put.INF your husband
till svar-s.
to answer-GEN
'I believe that you made your way to the boat to confront your husband.'

(30) Nu när han är sysslolös vet han inte vad han ska [ta
Now when he be.PRS unoccupied know.PRS he not what he be.AUX take.INF
sig till].
REFL to.
'Now when he is unuccupied he doesn't know what to do.'

On a lighter note, the ranking result can be tailored by selecting more indicative seed examples. If we are particularly interested in occurences of the V_REFL.RÖRELSE construction where the verb has a more specific manner meaning, we can handpick only that kind of positive examples for the gold standard. By doing so, sentences like (31) and (32) rise to the surface of the answer set. This means that even if the average precision is low over the board, the user can still use the search system to find a certain subset of the construction.

(31) Jag tror att han skulle ha [sovit sig igenom] hela
I think.PRS that he would.AUX have.AUX sleep.PRF REFL through hole
eländet.
misery.DEF
'I think that he would have slept through the whole ordeal.'

(32) Floden [slingrade sig förbi] grusbankar och öar.
River wind.PST REFL past gravel bank.PL and island.PL
'The river winds it's way past gravel banks and islands.'

5.3.4 kvantifierande_genitiv.tid

The evaluation results for the time construction KVANTIFIERANDE_GENITIV.TID that has been following us throughout the thesis can be seen in figure 11. The frame-based reranker is clearly outperforming the other similarity functions in this case since the construction is restricted to time expressions that we have already seen fit neatly into a few particular frames.

Since time expressions are such a strong feature for detecting the KVANTIFIERANDE_GENITIV.TID construction, it comes as no surprise that false positives near the top of the ranked list contain time words as well. Both example (33) and (34) are false hits that are in fact instances of the KVANTIFIERANDE_GENITIV.SKALA construction.

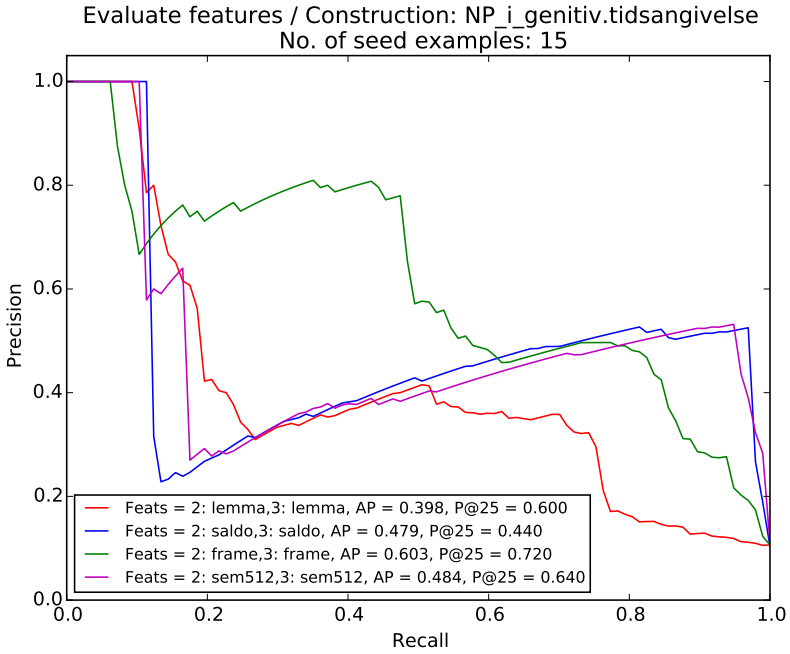


Figure 11: Precision/recall curve for retrieving the NP_I_GENITIV.TIDSANGIVELSE construction. 15 training instances are used, and different types of lexical-semantic features.

(33) Det var [femton år-s åldersskillnad] mellan oss.
It is.PST fifteen year.PL-GEN age difference.INDEF between us.
'It was fifteen years of age difference between us.'

(34) [Två dagar-s post] hade samlats i hallen.
Two day.PL-GEN mail.INDEF have.AUX gather.PST.PASS in hallway.
'Two day's worth of mail had gathered in the hallway.'

On the other side of the ranked search result, near the bottom, we find less prominent examples of KVANTIFIERANDE_GENITIV.TID. The fact that sentence (35) receives a low ranking score is probably due to the fact that the lexical unit *kvart* 'quarter-hour' is only listed under the sense 'quarter' in the frame Part_whole, which is not a time related frame.

(35) Hon annonserar i porrtidningar och tar fyrtio
She advertise.PRS in porn magazine.PL.INDEF and take.PRS forty
dollar för [en kvart-s samtal].
dollar.PL.INDEF for one quarter-GEN conversation.INDEF
'She advertises in porn magazines and charges forty dollars for a quarter-hour's conversation.'

5.3.5 kvantifierande_genitiv.skala

The evaluation results for the sibling KVANTIFIERANDE_GENITIV.SKALA can be inspected in figure 12. At first glance we can conclude that while the frame-based feature worked best for the time

version of this construction, it receives the lowest average precision in this case. The scalable measures that turn up in KVANTIFIERANDE_GENITIV.SKALA are diverse and can not easily be contained within strictly defined semantic frames.

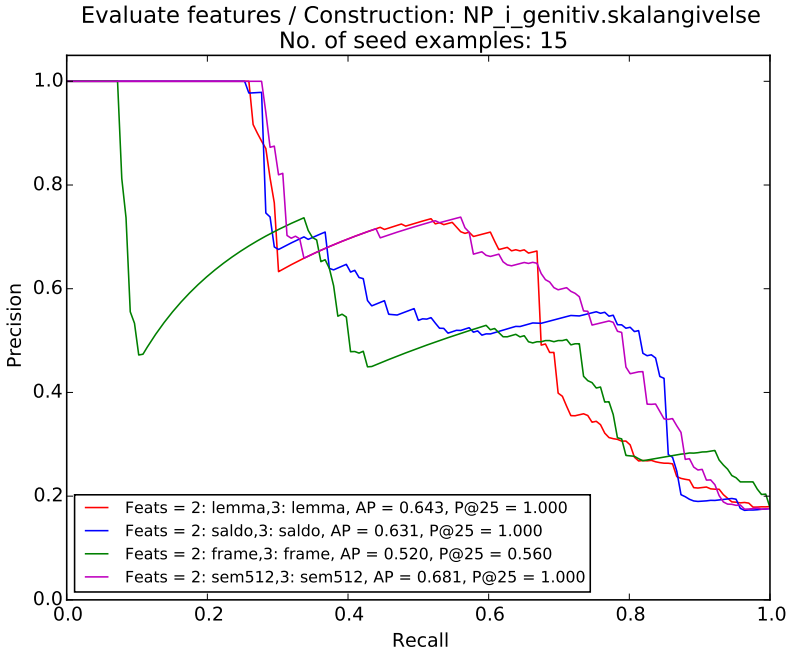


Figure 12: Precision/recall curve for retrieving the KVANTIFIERANDE_GENITIV.SKALA construction. 15 training instances are used, and different types of lexical-semantic features.

The distributional similarity does best at generalizing beyond seen examples, and quite impressively hands out high ranking scores to constructs as diverse as *50 procents chans* '50 percent's chance', *två veckors skäggstubb* 'two weeks of stubble' and *tre spalters bredd* 'three columns' width'. The precision starts to drop with false hits from the KVANTIFIERANDE_GENITIV.TID construction and quite a few instances of the specific construction *någon slags/sorts X* 'some kind of X', as in examples (36) and (37) below. Again, it would be quite useful if we could hand negative example seeds to the ranker, or if we could teach it to differentiate between two particular constructions like in this case.

(36) Kanske står jag för [nåt slag-s hjälp].
Perhaps stand.PRS I for some kind-GEN help.INDEF.
'I might represent some kind of help.'

(37) Det måste vara [nån sort-s rekord] här.
EXPL must.AUX be.INF some kind-GEN record.INDEF here.
'This must be some kind of record.'

Near the bottom of the reranked search batch we find constructs with some unusual units that would only exceptionally be used as scalables and thus do not score high in distributional similarity when compared to words like 'height', 'weight' or 'age', as illustrated in example (38) and (39)

below. The scalables in these examples are compounds so rare that they are probably not even represented in the distributional model.

(38) Wes hade alltid haft [ett tolvmånader-s reservförråd] av
Wes have.AUX always have.PPFV a twelve month-GEN reserve supply.INDEF of
självförtroende .
self confidence.INDEF
'Wes had always had a twelve months' reserve supply of self confidence.'

(39) Gro kom hem en kväll och fann [en 25-liter-s saftbehållare]
Gro come.PST home one night and find.PST a 25-liter-GEN juice container.INDEF
i köket .
in kitchen.DEF
'Gro came home one night and found a 25-liters' juice container in the kitchen.'

5.3.6 avgränsad_aktion.*på*

Finally, let us take a closer look at the precision/recall curves for retrieving the time expression AVGRÄNSAD_AKTION.*på* in figure 13. Even though this construction is strictly restricted to time related words, the frame-based reranker performs worst in the evaluation. The explanation is quite straight forward; just spotting the time word is not enough to disambiguate AVGRÄNSAD_AKTION.*på* from related time constructions, most of the delimiting information can actually be found outside of the hit. Remember, AVGRÄNSAD_AKTION.*på* only occurs with events of bounded lexical aspect.

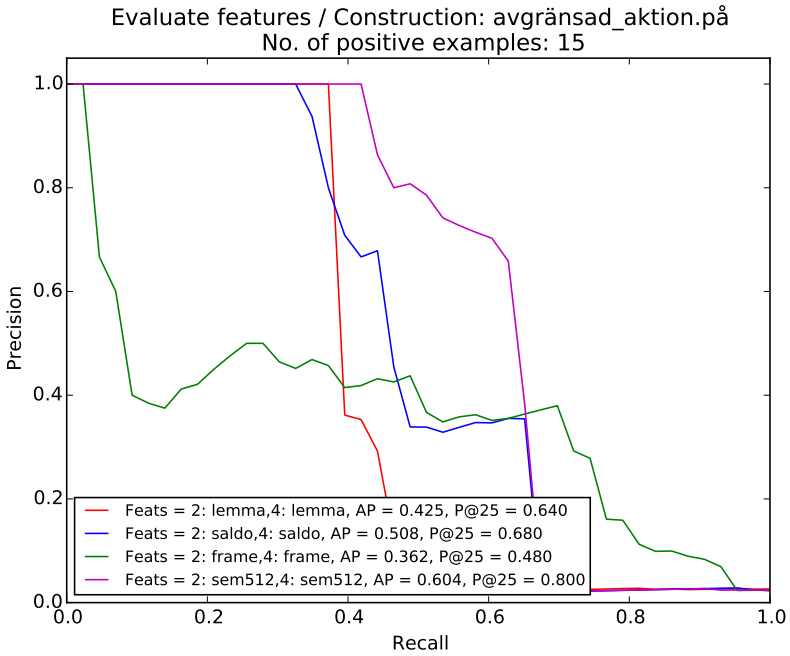


Figure 13: Precision/recall curve for retrieving the AVGRÄNSAD_AKTION.*på* construction. 15 training instances are used, and different types of lexical-semantic features.

Since we had little hope that our reranking function would be effective for such a contextually dependent construction as AVGRÄNSAD_AKTION.*på*, it is a pleasant surprise to see that the distributional similarity is doing significantly better than the lemma-baseline. The possible explanation is that there are some lexical preferences at play, beyond the more general time restriction. Short time spans like *på ett ögonblick* 'in an instant' or *på en halv minut* 'in half a minute' seem to be more likely to be instances of AVGRÄNSAD_AKTION.*på* than for example calendric units like *på en söndag* 'on a sunday'. Even so, also the distributional similarity is incapable of distinguishing between the true instance in (40) and the related negative construction in (41).

(40) I början var jag glad om jag kunde få ihop så här
In beginning.DEF be.PST I happy if I could.AUX get.INF together so here
mycket [på en hel vecka].
much on one whole week.INDEF
'In the beginning I was happy if I could earn this much in a whole week.'

(41) Antagligen hade han inte fått nån riktig nattsömn [på en vecka].
Probably had.AUX he not get.PST any real nigh sleep on one week.
'Probably he had not had any real night sleep for a week.'

Among the top ranked false positives we also find quite a few hits with the lexicalized phrase *på en gång* 'at once', as well as hits with verbs that take complements with the preposition *på* like in the example (41) below.

(42) Han grubblade [på det en stund].
He ponder.PST on it a while.
'He pondered about it for a while.'

Introducing contextual features seems like an obvious way to develop the search system further. However, in this particular case it is not entirely clear how to judge the lexical aspect of the event in an automated fashion. A more straightforward feature to introduce would be negations that can be relatively easily spotted by using a list of negative polarity items.

5.4 Evaluation of seed influence

As a final note, let us say something about the mean and standard deviation of the evaluation results we have inspected so far. It is worth to underscore that the performance of the construction search system is to a large extent hingeing on the selection of positive seed examples. The selection does not only effect the average precision-scores for different similarity functions, it also has a quite dramatic influence on the ranking order of all sentences in the answer set. This flexibility and variation is one of the advantages of our approach, since it allows the user to discover different uses of the same construction depending on the way she tailors the gold standard.

With this in mind, we decided to do a sample evaluation of the probability distribution for the average precision score on each search query. We used the best reranker from the feature evaluation for each construction. This time we stabilized the experiment by fixating all search variables except the selection of seed examples. Instead of picking the N first true instances from the

benchmark collection, we now chose a *random* selection of N true hits from the benchmark every time we run the evaluation. We repeated the experiment 10 000 times and plotted the range of outcomes in a histogram that shows the probability distribution for each search query. We also computed the mean and standard deviation for each retrieval experiment. The plots can be inspected in figure 14.

Visibly, the seed selection has a big influence on the average precision score for every construction, except PROPORTION_*i/om* in plot 14(b). The rate construction is the only construction for which the frame-based reranker so confidently succeeds to capture the temporal restrictions of the variable slots, without simultaneously scooping up false positives. The confidence has less to do with the reranking function, and more to do with the properties of the construction itself. The most uncertain ranking can be found in plot 14(f) for AVGRÄNSAD_AKTION.på. Here, the selection of seed examples dramatically influences the average precision score, with a standard deviation of 0.07. We already know that the AVGRÄNSAD_AKTION.på construction is not easy to detect among false positives from other time constructions, but carefully tailoring the seed selection may improve the scores (for example picking positive examples with a short time span). The remaining four constructions all have a standard deviation fluctuating between 0.03 and 0.05, regardless of how high the mean average precision score is.

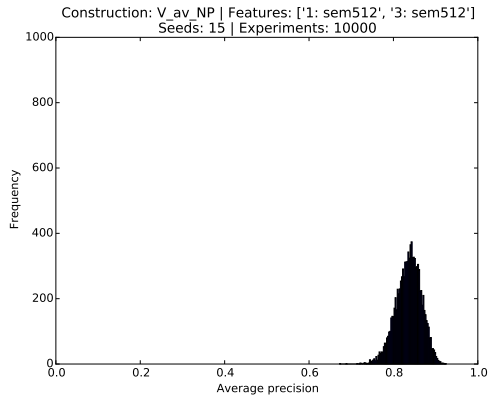
6 Discussion

We have considered the problem of searching for occurrences of partially schematic constructions as a retrieval problem. As a proof of concept, we presented a simple interactive architecture for searching for constructions, where a user provides a number of positive examples (occurrences of the construction) and tailors a ranking function based on a user-defined combination of features.

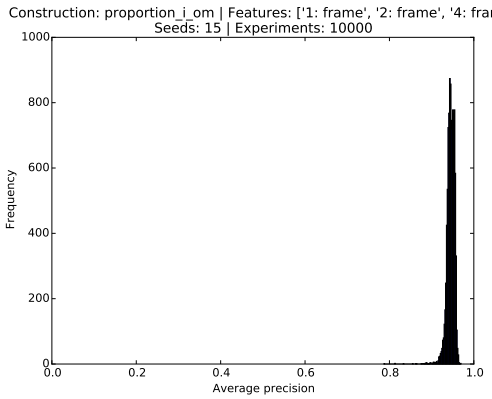
In order to test our system, we annotated a benchmark collection of various constructions from the Swedish Constructicon for evaluation of construction-based retrieval systems. The constructions in the benchmark are all partially schematic and share the property that they cannot be distinguished from unrelated constructions by surface form alone.

For each construction in the benchmark, we evaluated a number of ranking functions based on different feature sets. As expected, searching for construction occurrences is a highly diverse problem for which the ranking function must be tailored for each construction. The results showed that reranking based on a lexical-semantic model can give strong improvement over a lemma-based baseline, but exactly which lexical-semantical similarity – lexicon-based or distributional – is most effective depends on the construction. An important insight is that the precision of the reranker is determined by the construction definition and to which extent the construction elements are in fact semantically restricted. The frame-based reranker tended to work best for constructions that contain variable slots restricted to narrowly defined semantic classes, while the distributional model was better at pinning down the common denominator between slot fillers from a broader semantic domain.

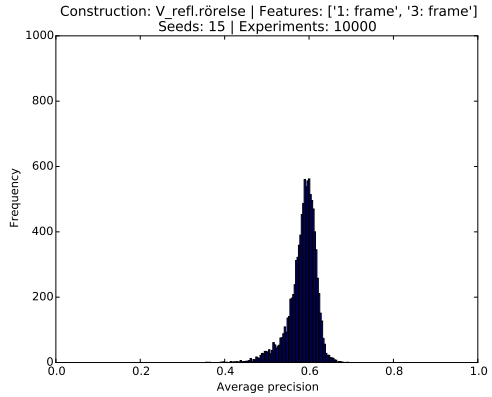
Furthermore, we have also shown that the search system is effective even with a small number of positive seed examples, which proves the feasability of our approach from a user perspective. For our benchmark collection, as few as five positive seeds were enough to reach close to maximum avarege precision for each search query. This indicates that increasing the size of the gold standard would not improve system performance at this point. Instead, it would be useful to extend



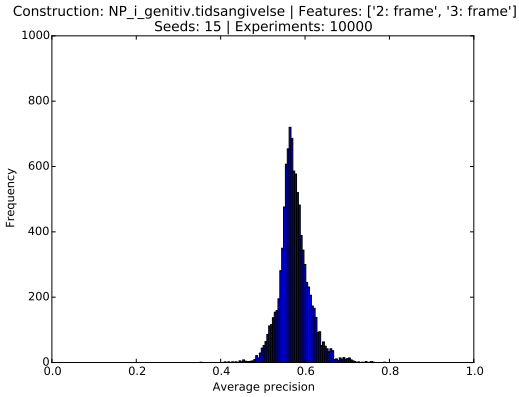
(a) V_*av*_NP
Mean: 0.84 | Standard deviation: 0.03



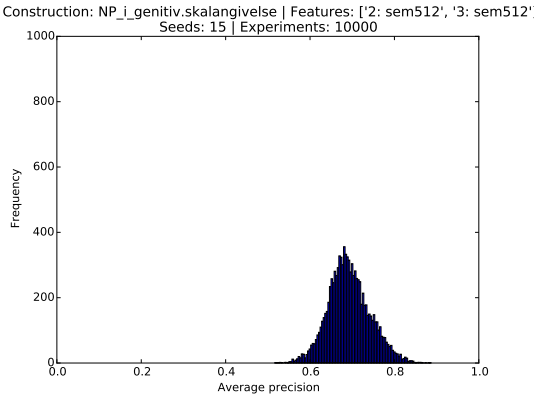
(b) PROPORTION_*i*/*om*
Mean: 0.95 | Standard deviation: 0.01



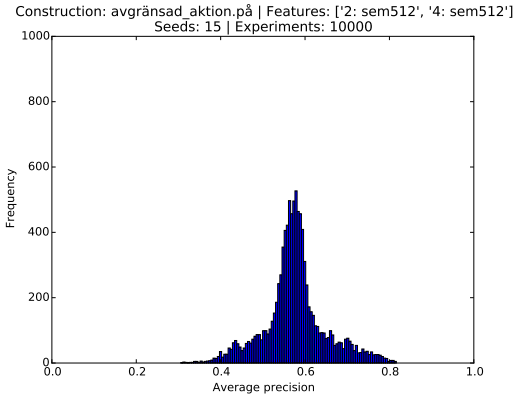
(c) V_REFL.RÖRELSE
Mean: 0.59 | Standard deviation: 0.03



(d) KVANTIFIERANDE_GENITIV.TID
Mean: 0.57 Standard deviation: 0.04



(e) KVANTIFIERANDE_GENITIV.SKALA
Mean: 0.69 | Standard deviation: 0.05



(f) AVGRÄNSAD_AKTION.PÅ
Mean: 0.58 | Standard deviation: 0.07

Figure 14: Probability distributions of the average precision score for retrieving all six constructions. Experiment repeated 10 000 times with 15 randomly selected seed examples each run.

the list of features and improve the weighting scheme of the ranking function.

Lastly, we have demonstrated that the selection of seed examples has significant influence on the reranking results. Even though this makes the average precision scores less reliable, it also illustrates how the user can influence the output of the ranking function by tailoring the gold standard to her satisfaction. The user can pose queries, rank and rerank repeatedly according to different seeds and thus get a comprehensive overview of the various uses of a construction in a corpora.

The user perspective, flexibility and short takeoff is what sets our experiment apart from previous attempts to detect occurrences of particular constructions in running text. We have demonstrated a system that works for a wide range of constructions and that does not require large hand-labeled datasets for training before taking off.

The construction search system is primarily a useful tool for constructicographers in their work with characterizing and defining new construction candidates. However, we have also given strong support that semantic features can be used to disambiguate grammatical constructions, and this insight can be incorporated into other NLP tasks. Since previous studies have shown that construction detection is a key to improve the accuracy of for example parse algorithms (Nivre & Nilsson, 2004; Baldwin et al., 2004; Zhang et al., 2006) we expect that constructionist approaches to natural language processing will attract more attention in future research. The construction search system could for example readily be employed to collect large corpora of authentic construct examples, as an intermediate step before applying more fine tuned machine learning methods for classification.

The qualitative evaluation of the reranked search lists have given us valuable insights about the limits of our approach, and we will now discuss these shortcomings and how we could get around them.

As a point of departure, let us consider an issue that has so far been glossed over – the binary annotation of constructs as true or false instances of the construction we are looking for. Construction Grammar embraces the fact that constructions cannot always be easily classified and recognizes the fuzzy borders of categorization. We have already stated this as a underlying motivation to treat construct detection as a non-binary ranking problem. For the purposes of our current study however, we have treated the annotation of the benchmark collection as an unproblematic procedure before moving on to evaluation. In future work it would certainly be desirable to take this issue more seriously, let different annotators label the search hits independently and measure inter-annotator agreement. It is very likely that even skilled linguists would have difficulties making clean-cut binary decisions about the correct classification of certain search hits. It would also be possible to introduce a graded notion of relevance and let annotators label hits as more or less prototypical instances of the construction at hand. There are standard IR-metrics for evaluating ranking according to non-binary relevance judgements (i.e. *normalized discounted cumulative gain*, see Manning et al. (2008), but introducing more labels does not necessarily make the annotation process less difficult, on the contrary.

We have also learned that peripheral instances of constructions are much harder for our search system to detect than more prototypical occurrences. This is not particularly surprising, since the ranking function works by comparing new hits with already seen examples, but it is nonetheless a shortcoming worth noting. Peripheral cases are harder to detect because they are more unique than typical instances of the construction. Allow us to quote Tolstoy on this one – 'All happy

families are alike; each unhappy family is unhappy in its own way.'

Another problem that we have already mentioned in the qualitative evaluation is that our system treats the features extracted from each position in the construction independently, an underlying assumption that distorts the reranking results when the relation between slot fillers is meaningful. In Construction Grammar, constructions are defined as non-compositional, that is that the meaning of the whole cannot be computed by summing up the parts (Hilpert, 2014). Yet, we have designed our search system in a completely compositional way. It would therefore perhaps be more fair to call it 'semantic search' instead of 'construction search' until we have improved the feature design of the search system. In future implementations we must correct this flaw and account for slot relations as a way to improve precision.

It has also become evident that only looking at the construction elements and the lexical content occupying these slots is often not enough to delimit certain constructions from unrelated hits. We need to consider contextual features as well, such as preceding and following constituents, grammatical dependency labels, and scanning the sentence for negations. To keep the system user-governed, however, it is crucial that these features are kept simple and meaningful, at least to the extent that a linguist or language researcher can make sense of them.

We had two good reasons when we decided to build the reranking function exclusively on positive seed examples. First, the less work for the user, the more user-friendly the system. Second, we expected the rest of the search batch to contain a varied lot of constructions that would be hard to treat as a homogeneous group, labeled 'not it'. However, from the qualitative evaluation we have learned that the average precision of a search query could be quite tangibly improved by allowing the possibility to point out certain lexicalized phrases as 'not it' and get rid of them from the search batch entirely. This possibility should be implemented into future versions of the system.

The ranking function we have used in this experiment is simple and computationally effective, but in future work it would be interesting to investigate more complex learning methods for our scenario, such as the one-class SVM (Manevitz & Yousef, 2001). It would also be possible to consider weights being set manually by the users of the retrieval system.

There are several ways in which this work could be extended. We have now described the machinery of retrieval of construction occurrences, but in future work, it would be interesting to consider the usability and interaction aspect as well. In order for the system to be useful in real life we have to test it on users and make sure that the intermediate steps between search and ranking are simple and meaningful also for the target group.

Another possible extension would be to create a work-bench tool for collostructural analysis (Gries, 2003) that would incorporate the construction search system as a facilitated way to get to raw frequency counts of complex constructions.

The title of this thesis has two different interpretations. To wrap up the paper let us spell out the one that suits our conclusions best. While the bible quote promises that the one who seeks shall find (Matthew 7:7) the children's game is perhaps more to the point when it comes to language: Seek and you shall find many things, interlinked, hard to classify, and beautiful to the eye.

References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09 (pp. 19–27). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ahlberg, M., Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., Friberg Heppin, K., Johansson, R., Kokkinakis, D., Olsson, L.-J., & Uppström, J. (2014). Swedish framenet++ the beginning of the end and the end of the beginning.

Bäckström, L., Lyngfelt, B., & Sköldberg, E. (2014). Towards interlingual constructicography. on correspondence between constructicon resources for english and swedish. *Constructions and Frames*, 6(1), 9–33.

Baeza-Yates, R. A. & Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1* (pp. 86–90). Montréal, Canada.

Baldwin, T., Bender, E. M., Flickinger, D., Kim, A., & Oepen, S. (2004). Road-testing the english resource grammar over the british national corpus. In *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004* (pp. 2047–2050).

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238–247). Baltimore, Maryland: Association for Computational Linguistics.

Blanchard, E., Harzallah, M., Briand, H., & Kuntz, P. (2005). A typology of ontology-based semantic measures. In *EMOI - INTEROP'05, Enterprise Modelling and Ontologies for Interoperability, Proceedings of the Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability, Co-located with CAiSE'05 Conference, Porto (Portugal), 13th-14th June 2005*.

Boas, H. C. (2014). Zur architektur einer konstruktionsbasierten grammatik des deutschen. In A. Ziem & A. Lasch (Eds.), *Grammatik als Netzwerk von Konstruktionen? Sprachliches Wissen im Fokus der Konstruktionsgrammatik* (pp. 37–63). Berlin & New York: de Gruyter.

Borin, L., Forsberg, M., & Lönngren, L. (2013). SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4), 1191–1211.

Borin, L., Forsberg, M., Olsson, L.-J., & Uppström, J. (2012a). The open lexical infrastructure of språkbanken. In *Proceedings of the 8th International Conference on Language Resources and Evaluation : May 23-25, 2012 / eds. Nicoletta Calzolari* (pp. 3598–3602).

Borin, L., Forsberg, M., & Roxendal, J. (2012b). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC-2012)* (pp. 474–478). Istanbul, Turkey.

Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge University Press.

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

Chomsky, N. (1995). *The Minimalist Program*. Current studies in linguistics series. MIT Press.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research* (pp. 23–32). Budapest, Hungary.

Clark, S. (2015). Vector space models of lexical meaning. In S. Lappin & C. Fox (Eds.), *Handbook of Contemporary Semantic Theory*. Chichester, UK.: John Wiley & Sons, Ltd.

Cruse, D. (1986). *Lexical semantics*. Cambridge University Press.

Curran, J. R. (2003). *From Distributional to Semantic Similarity*. PhD thesis, University of Edinburgh.

Dubremetz, M. & Nivre, J. (2015). Rhetorical figure detection: the case of chiasmus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature* (pp. 23–31). Denver, Colorado, USA: Association for Computational Linguistics.

Ehrlemark, A. (2014). *Ramar och konstruktioner – en kärlekshistoria*. Technical Report GU-ISS 2014-01, Institutionen för svenska språket, Göteborgs universitet.

Fillmore, C. J. (2008). Border conflicts: Framenet meets construction grammar. In *Proceedings of the XIII EURALEX International Congress* Barcelona, Spain: Universitat Pompeu Fabra.

Fillmore, C. J. & Baker, C. (2009). A frames approach to semantic analysis. In B. Heine & H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis* (pp. 313–340). Oxford: OUP.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions. the case of *let alone*. *Language*, 64, 501–538.

Fillmore, C. J., Lee-Goldman, R., & Rhomieux, R. (2012). The FrameNet Constructicon. In H. C. Boas & I. A. Sag (Eds.), *Sign-Based Construction Grammar* (pp. 309–372). Stanford: CSLI Publications.

Forsberg, M., Johansson, R., Bäckström, L., Borin, L., Lyngfelt, B., Olofsson, J., & Prentice, J. (2014). From construction candidates to constructicon entries. an experiment using semi-automatic methods for identifying constructions in corpora. *Constructions and Frames*, 6(1), 114–135.

Friberg Heppin, K. & Toporowska Gronostaj, M. (2012). The rocky road towards a Swedish FrameNet – creating SweFN. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC-2012)* (pp. 256–261). Istanbul, Turkey.

Goldberg, A. (2013). Constructionist approaches. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 15–31). Oxford: OUP.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5), 219–224.

Gries, S. (2003). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, 1, 1–27.

Hilpert, M. (2014). *Construction Grammar and Its Application to English*. Edinburgh University Press Series. Edinburgh University Press.

Hwang, J. D., Nielsen, R. D., & Palmer, M. (2010). Towards a domain independent semantics: Enhancing semantic representation with construction grammar. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics* (pp. 1–8).: Association for Computational Linguistics.

Jackendoff, R. (1997). Twistin' the night away. *Language*, 73(3), 534–559.

Johansson, R. (2014). Automatic expansion of the swedish framenet lexicon: Comparing and combining lexicon-based and corpus-based methods. *Constructions and Frames*, 6(1), 92–113.

Johansson, R., Adesam, Y., Bouma, G., & Hedberg, K. (2016). A multi-domain corpus of Swedish word sense annotation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)* Portorož, Slovenia.

Kay, P. & Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: The '*What's X doing Y?*' construction. *Language*, 75, 1–33.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1), 1–31.

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 *11th Conference on Natural Language Processing (KONVENS) Proceedings*.

Lyngfelt, B., Bäckström, L., Borin, L., Ehrlemark, A., & Rydstedt, R. (forthcoming).

Manevitz, L. M. & Yousef, M. (2001). One-class SVMs for document classification. *Journal of Machine Learning Research*, 2, 139–154.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Michaelis, L. A. (2013). Sign-based construction grammar. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 133–152). Oxford: OUP.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations, Workshop Track* Scottsdale, USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3, 235–244.

Nivre, J. & Nilsson, J. (2004). Multiword units in syntactic parsing. In *In Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*.

O'Donnell, M. B. & Ellis, N. (2010). Towards an inventory of english verb argument constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, EUCCL '10 (pp. 9–16). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ohara, K. H. (2013). Toward constructicon building for japanese in japanese framenet. *Veredas*, 17(1), 11–27.

Padó, S. & Lapata, M. (2007). Dependency-based construction of semantic space models. *Comput. Linguist.*, 33(2), 161–199.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2001). Multiword expressions: A pain in the neck for NLP. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002* (pp. 1–15).

Sköldberg, E., Bäckström, L., Borin, L., Forsberg, M., Lyngfelt, B., Olsson, L.-J., Prentice, J., Rydstedt, R., Tingsell, S., & Uppström, J. (2013). Between grammars and dictionaries: a Swedish Constructicon. In *Proceedings of eLex* (pp. 310–327). Tallinn, Estonia.

Stefanowitsch, A. & Gries, S. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243.

Torrent, T. T., Lage, L. M., Sampaio, T. F., da Silva Tavares, T., & da Silva Matos, E. E. (2014). Revisiting border conflicts between framenet and construction grammar: Annotation policies for the brazilian portuguese constructicon. *Constructions and Frames*, 6(1), 33–50.

Turney, P. D. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188.

Wible, D. & Tsao, N.-L. (2010). StringNet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics* (pp. 25–31). Los Angeles, United States.

Wu, Z. & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics* (pp. 133–138). Las Cruces, United States.

Zhang, Y., Kordoni, V., Villavicencio, A., & Idiart, M. (2006). Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, MWE '06 (pp. 36–44). Stroudsburg, PA, USA: Association for Computational Linguistics.