John Benjamins Publishing Company

This is a contribution from *Constructions and Frames 6:1*
© 2014. John Benjamins Publishing Company

This electronic file may not be altered in any way.
The author(s) of this article is/are permitted to use this PDF file to generate printed copies to be used by way of offprints, for their personal use only.
Permission is granted by the publishers to post this file on a closed server which is accessible to members (students and staff) only of the author's/s' institute, it is not permitted to post this PDF on the open internet.
For any other use of this material prior written permission should be obtained from the publishers or through the Copyright Clearance Center (for USA: www.copyright.com).
Please contact rights@benjamins.nl or consult our website: www.benjamins.com

Tables of Contents, abstracts and guidelines are available at www.benjamins.com

From construction candidates to constructicon entries

An experiment using semi-automatic methods for identifying constructions in corpora

Markus Forsberg, Richard Johansson, Linnéa Bäckström, Lars Borin, Benjamin Lyngfelt, Joel Olofsson and Julia Prentice
University of Gothenburg

We present an experiment where natural language processing tools are used to automatically identify potential constructions in a corpus. The experiment was conducted as part of the ongoing efforts to develop a Swedish constructicon. Using an automatic method to suggest constructions has advantages not only for efficiency but also methodologically: it forces the analyst to look more objectively at the constructions actually occurring in corpora, as opposed to focusing on "interesting" constructions only. As a heuristic for identifying potential constructions, the method has proved successful, yielding about 200 (out of 1,200) highly relevant construction candidates.

Keywords: Constructicon, natural language processing, construction grammar, corpus, Swedish, construction, language technology, association measure, collocation measure

1. Introduction

In this article we present an experiment where natural language processing (NLP) tools are used to automatically identify potential constructions in a corpus. The experiment was conducted as part of the ongoing efforts to develop a Swedish constructicon. Its purpose is twofold. On the one hand, we wished to test the tools as such and use the results from the experiment to improve them. In fact, this experiment builds on an earlier one (Bäckström et al. 2013), and the method employed here incorporates insights from that study. On the other hand, it is meant

Constructions and Frames 6:1 (2014), **114–135**. DOI 10.1075/cf.6.1.07for
ISSN 1876–1933 / E-ISSN 1876–1941 © John Benjamins Publishing Company

to provide new entries for the constructicon, in particular such patterns that might not have been discovered manually.

The Swedish constructicon is a freely available online database of Swedish constructions, developed in relation to the Swedish FrameNet (Borin et al. 2010; Friberg Heppin & Toporowska Gronostaj this volume). At the time of writing it consists of about 200 construction descriptions and is eventually intended to be a large-scale resource for research in linguistics, language technology, lexicography, and language pedagogy, in particular L2 education. The Swedish constructicon is inspired by the English constructicon in Berkeley (Fillmore 2008; Fillmore et al. 2012) and through collaboration with constructicon projects for other languages (e.g. Ohara 2013; Torrent et al. this volume; Boas 2014; Ziem et al. 2014), we seek to establish cross-linguistically applicable constructicon resources (Bäckström et al. this volume).

Language technology is both a means and a goal for the Swedish constructicon project. The experiment presented here is a valuable means for identifying potential constructions; the method provides construction candidates, statistically identified recurring linguistic structures, which are then manually evaluated to filter out material for actual constructicon entries. In a longer perspective, we are also working towards developing methods for automatic identification of particular constructions, which is a desirable research objective in itself, with a potential for improving automatic language analysis systems.

In the following, we first discuss the notion of construction and its delimitations, both from a theoretical perspective and in relation to the Swedish constructicon (Section 2). The method for automatically identifying construction candidates in corpora is presented in Section 3, after which the actual experiment is described in Section 4 and the evaluation process discussed in Section 5. In the concluding Section 6, we address what the results imply for future constructicon development.

2. Constructions in theory and practice

Constructions are typically defined as "conventional, learned form-function pairings at varying levels of complexity and abstraction" (Goldberg 2013: 17). It is a quite versatile concept that displays some variety in both theoretical interpretations and practical application. From the perspective of the Swedish constructicon, any conventionalized form-function pair may in principle be considered a construction (cxn), from the most general grammatical patterns to specific lexical items. That does not, however, mean that they are all of equal priority or even that they are all considered relevant to include in the database.

© 2014. John Benjamins Publishing Company
All rights reserved

In particular, since the constructicon is designed in relation to the Swedish FrameNet, there is clearly no need to account for lexical units (at least not as conventionally construed) in both resources. Hence, only patterns with at least one variable or schematic element are eligible as constructicon entries, whereas lexically fixed cxns are covered as lexical units in the Swedish FrameNet. Other restrictions and priorities are less self-evident, depending on both theoretical and practical concerns. In the following, we will first address issues of redundancy and productivity (Section 2.1), after which we present strategies for identification and selection of cxns in the Swedish constructicon (Section 2.2).

2.1 What counts as a construction?

A major point of divergence between different views on what should count as a cxn concerns redundancy, particularly whether conventional patterns that are predictable from other cxns should be included or not. The contrast is displayed in the following definition, where both sides agree on the first sentence but not on the second:

> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency (Goldberg 2006: 5)

The standard argument against the second part of the definition is that it would be redundant to store information that can be figured out anyway. It is couched in a long-standing tradition in grammar to avoid redundancy in the system, aiming to generate all and only the grammatical expressions of a language from a minimal number of rules and principles. This view is sometimes defended in terms of parsimony, economy, elegance, or Occam's razor (e.g., Hutchinson 1974: 57–59; Culicover & Jackendoff 2005). However, those who accept the second sentence and maintain that even predictable patterns may be entrenched by common usage also refer to economy, although of another sort. The argument is that deriving utterances compositionally induces a greater processing load than using prefab and semi-prefab expressions (e.g. Wray 2008). Thus, the dispute has to do with the division of cognitive labor between storing and figuring out (see Croft 2003; Jackendoff 2002, ch. 6).

This discussion also relates to the question of what the 'right' level of abstraction is for positing a cxn. While there are corpus methods for comparing the relevance of different levels of abstractions (see Hilpert 2013), these are too time consuming to suit the purposes of the Swedish constructicon. So far, the issue has

© 2014. John Benjamins Publishing Company
All rights reserved

not been much of a problem in practice, but in the present experiment the question is brought to the fore, since several of the auto-generated cxn candidates are similar except for the level of abstraction.

Another 'strict vs. generous' issue has to do with productivity. Construction grammar is constraint based, and from a strictly generative perspective a cxn description should be constrained so as to include all legitimate instantiations of the cxn and exclude anything else. On this view, semi-productive structures are not cxns but so-called patterns of coinage. Such a pattern, taken from Kay (2013: 37–38), is the form [**A as NP**] meaning 'very A', as in *light as a feather, easy as pie, free as a bird*, etc. It is productive in that many different expressions may be generated from it, and if one encounters a previously unfamiliar example, it is still easy to perceive its meaning by the familiar pattern. It is not, however, fully productive; if treated as a (productive) cxn, the above description overgenerates. Kay provides counterexamples such as *strong as a horse* and *heavy as a truck*, which both make perfect sense but are just not idiomatic instances of this pattern. From a more usage-based point of view (e.g. Bybee 2013), cxns do not have to be strictly productive in this sense. Instead, they presumably emerge as generalizations over encountered exemplars, becoming gradually more entrenched with increasingly common usage. A usage-based approach also goes hand in hand with a generous view of redundancy.

From the viewpoint of the Swedish constructicon, these issues, while highly relevant, are more of a practical concern. Filtering out semi-productive patterns would not only be time-consuming; it would also lead to the loss of many cxns that are relevant for the purposes of the resource (see the next section). Likewise, excluding otherwise relevant cxns merely on the grounds of being predictable would be neither desirable nor practically feasible, especially since full predictability is a somewhat utopian ideal (e.g. Svanlund 2002). Predictability is rather a matter of degree, and where to draw the line is not always obvious. On the other hand, excessive redundancy is not an attractive approach either.

Although we do not assume the stricter delimitations above as necessary conditions, they are more useful as sufficient criteria. Clearly, to adhere to them makes for a better case than not to. Hence, Hilpert (2014), while advocating a usage-based view, presents the following list of cxn heuristics:

- Does the expression deviate from canonical patterns? (ex: *by and large, There was cat all over the driveway*)
- Does the expression carry non-compositional meaning? (ex: *call the shots, John broke a finger*; in the latter case, it does not follow from the words alone that the finger is his own)

© 2014. John Benjamins Publishing Company
All rights reserved

- Does the expression have idiosyncratic constraints? (ex: *the dog is asleep* vs. **the asleep dog*, *Mary is a smarter lawyer than John* vs. **Mary is the smarter lawyer than John*)
- Does the expression have collocational preferences? (ex: *will* vs. *be going to* as future auxiliaries showing tendencies to occur with different types of main verbs)

Hilpert's list has been considered for the experiment reported in this article, but the criteria were found too complex for quickly filtering out potential cxn entries from a list of several hundred automatically generated candidates. Nonetheless, they are still relevant for further analysis of the remaining candidates.

2.2 Constructions in the Swedish constructicon

The focus of the Swedish constructicon is both wide and narrow. Wide, in that any cxn is in principle relevant and we have therefore treated quite different kinds of cxns in order to ensure that the description format can handle a large variety. Narrow, in that the majority of cxns accounted for so far are semi-general patterns of types that are somewhat peripheral from both a purely lexical and a purely grammatical viewpoint. An obvious reason for this is that more typically lexical or grammatical structures are for the most part covered in dictionaries and grammars, respectively, and a high priority is therefore given to the domains where empirical coverage is lacking. Without abandoning this concern, our aim to build a coherent cxn network also requires that more attention be paid to more general grammatical patterns in the near future. Furthermore, even among linguistic phenomena extensively treated from other perspectives, e.g. word order patterns and definiteness marking in Swedish, many of the known facts suggest that a constructionist account should contribute to a better understanding of the problems at hand.

A category that has been specifically targeted initially, and will remain a central priority, are partially schematic cxns, that is, patterns where at least one element is lexically specific and at least one element is schematic. Consider, for example, the English *way*-cxn — as in "*snoozed her way* through high school" or "*killed his way* from obscurity to infamy"– which is partially schematic in that the verb and the possessive pronoun are schematic units whereas *way* is lexically specific. A Swedish example is "*i* ADJEKTIV-aste *laget*" (*in* ADJECTIVE-superlative *measure*-definite), roughly meaning 'too much of the quality expressed by the adjective': *i hetaste laget* 'too hot for comfort', *i minsta laget* 'a bit on the small side' and *i senaste laget* 'at the last moment'. This pattern consists of two lexically specific units, the preposition *i* 'in' and the definite noun *laget* 'the measure', flanking

© 2014. John Benjamins Publishing Company
All rights reserved

the schematic (superlative) adjective slot. Since such cxns have both lexical and grammatical properties, they are hard to account for from either perspective but constitute a natural object of study for a constructionist approach.

Other concerns follow from the intended applications of the resource, where two main areas are (second) language pedagogy and language technology. Both go well with the focus on semi-general cxns, which are problematic for them both, partly because of the lack of descriptive coverage. While such different areas of application clearly place different — and to some extent conflicting — demands on the cxn descriptions, our main objective is to build a resource of wide applicability; particular adaptations should not restrict the general properties of the database. In order to facilitate large-scale coverage, we aim for as simple a description format as possible, although some degree of complexity is unavoidable. The core of a cxn entry consists of a free text definition of dictionary type and a simple structure sketch representing constituent structure and grammatical relations. In addition, cxn entries include, e.g., FrameNet-style annotated examples, lists of commonly occurring words, inheritance relations between cxns, and links to corresponding frames (where applicable).

In identifying potential constructicon entries, a number of different methods are used. One is to start out from existing cxn analyses. Although these are by definition already accounted for elsewhere, they are not previously collected in one place. Each of them also provides a reference point from which we may identify related cxns, and the analyses as such are useful for improving the description format. For example, by working through valence descriptions and usage examples in general dictionaries, we see patterns that have been noted in relation to particular words but in many cases are actually more productive. In addition, we consider cxn descriptions for other languages, seeking to establish corresponding Swedish cxns; this work is also valuable for future interlingual constructicon applications (see Bäckström et al. this volume).

To identify cxns we are not yet aware of, we employ corpus data. For instance, L2 corpora are investigated to discover cxns that are problematic for learners of Swedish. A number of corpora have also been analyzed with computational methods — such as the experiment presented below.

3. A corpus-based method to extract and rank potential constructions

Focusing on schematic and partly schematic cxns, our method for suggesting potential cxns consists of two steps: (1) extracting and counting syntactic patterns occurring in a large corpus; (2) ranking those patterns by a relevance measure based on our ideas about what characterizes a cxn from a statistical point of view.

© 2014. John Benjamins Publishing Company
All rights reserved

The most highly ranked patterns are then presented to the evaluators as described in Section 4.

3.1 Counting syntactic patterns in a corpus and estimating their occurrence probabilities

In this work, we aim to capture cxns where the surface side can be described as a sequential pattern — a so-called *n-gram* — of *n* adjacent units: specific words, words with a particular part-of-speech (POS) tag, or phrases. This heuristic covers many important grammatical cxns, in particular the most common basic phrases. For instance, the combination of a preposition and a noun phrase is the surface side of a very frequent cxn — a prepositional phrase. There are obviously other cxns involving long-distance interactions that are too complex to be captured using heuristics such as this one, but in order to introduce statistical measures we need something that can be objectively observed and counted in large corpora.

Our method goes through all text in a large corpus. When a word sequence is observed, the method first generates patterns of which that sequence is an instance and then counts those patterns. For instance, if we consider the words *in London*, there are five such generalizations. First, there are three ways to form a pattern by replacing words by POS tags: ***in*** [**PROPER NOUN**], [**PREPOSITION**] ***London***, and [**PREPOSITION**] [**PROPER NOUN**]. Furthermore, since *London* is an NP, there are two more generalizations in terms of that phrase: ***in*** [**NP**] and [**PREPOSITION**] [**NP**]. The patterns are similar to the "hybrid n-grams" used by Wible & Tsao (2010) but more general since our patterns can involve phrases.

We applied automatic NLP tools to determine POS tags and to extract phrases: HunPos (Halácsy et al. 2007) trained on the Stockholm–Umeå corpus (Gustafson-Čapková & Hartmann 2006) for the POS tags, and MaltParser (Nivre et al. 2007) trained on the Swedish Treebank (Nivre et al. 2008) for the phrases. MaltParser outputs dependency structures, so we defined heuristics to convert dependencies into phrase structures, e.g. a dependency subtree dominated by a common or proper noun or a pronoun becomes an NP.

The occurrence and co-occurrence probabilities of the patterns were estimated using maximum likelihood estimation from the raw frequency counts observed in the corpus. For example, when estimating the probability of a preposition followed by an NP, we divide the frequency count of that pattern by the total frequency of all token–phrase patterns:

$$p_{MLE}(\text{preposition, NP}) = \frac{\text{count (preposition, NP)}}{\text{count (any token, any phrase)}}$$

© 2014. John Benjamins Publishing Company
All rights reserved

Since our corpora are large we ended up with very large frequency tables, and to reduce memory consumption we applied a pruning heuristic: we counted a pattern only if it had the same start and end point as some phrase. Although this heuristic makes us more dependent on the automatic parser, it reduces the number of fragmentary patterns.

3.2 A statistical measure for ranking syntactic patterns

In Section 2 we have discussed the notion of linguistic cxn, but what characterizes a cxn from a statistical point of view? In particular, what could be said about the cxns that are more general than those found in a lexicon? While it would be difficult to directly operationalize the criteria in Section 2 as statistical measures, they may be reflected in surface properties that can be measured quantitatively: First, the parts of the pattern should be strongly associated, i.e. they should occur together more often than expected if they were independent. Secondly, the pattern should be productive, meaning that it occurs in many variations, as opposed to fixed, formulaic expressions.

To implement the first of these intuitions, we can rely on the established practice of using linguistic association measures, similar to what has often been used in tasks such as collocation extraction and lexicography; see e.g. Evert (2005) or Pecina (2010) for a detailed overview. Pointwise mutual information[1] (Fano 1961; Church & Hanks 1990) is one of the most widely used methods to measure the strength of association. It directly encodes the idea of comparing the estimated probability of cooccurrence of two terms to what would be expected if they were independent:

$$\mathrm{PMI}\,(x_1, x_2) = \log_2 \frac{p(x_1, x_2)}{p(x_1)\,p(x_2)}$$

PMI and other association measures are bivariate, and there are several ways to generalize them to the multivariate case. The most obvious multivariate generalization of PMI is probably what Van de Cruys (2011) terms *specific correlation* (SI_2):

$$\mathrm{SI}_2\,(x_1, \ldots, x_n) = \log_2 \frac{p(x_1, \ldots, x_n)}{\prod p(x_i)}$$

1. Early literature often used the term *mutual information* (MI) to refer to PMI, which is inconsistent with its use in the statistical and information-theoretical literature: MI measures the association between variables, PMI between values.

© 2014. John Benjamins Publishing Company
All rights reserved

PMI, as well as its multivariate generalization, is positive if its parts are positively associated, zero if they are independent, and negative if they are negatively associated.

It has frequently been observed that PMI tends to overemphasize rare patterns, and a number of rules of thumb have been proposed for addressing this problem. For instance, common patterns can be promoted by multiplying the PMI by the co-occurrence frequency or its logarithm (Evert 2005; Kilgarriff & Tugwell 2002).

In our case, scaling the PMI by frequency is not necessarily useful, since this will assign high scores to frequently occurring fixed expressions. The idea of rescaling PMI leads us back to the requirement of linguistic productivity of cxns: Unlike the typical scenario in collocation extraction, we would like to find patterns with a high degree of variation. This intuition makes it natural to scale the PMI by the unique instance frequency (UIF), the number of unique word sequences matched by the pattern in the corpus:

$$\text{UIF-PMI}(x_1, \ldots, x_n) = \text{UIF}(x_1, \ldots, x_n) \log_2 \frac{p(x_1, \ldots, x_n)}{\prod p(x_i)}$$

The experiments described in the following sections are all based on the UIF-PMI measure. We used n-gram lengths of up to 4 only, since data sparsity may be an issue as n grows larger. Before they are presented to the evaluators, we process the ranked lists applying a heuristic similar to the "vertical pruning" method introduced by Wible & Tsao (2010): if a more general pattern (e.g. [**PREPOSITION**] [**NP**]) is ranked higher than a more specific pattern (e.g. ***in*** [**NP**]), then the specific pattern is removed.

4. Experiment

For the evaluation of the method presented in the previous section, we designed an experiment where three linguists were presented with 1,200 top-ranked cxn candidates. The candidates were divided into 6 groups, 200 candidates in each group, where one half were schematic (no lexical constituent) and the other partially schematic (with at least one lexical and one schematic constituent). These were further divided into groups with different constituent lengths (2, 3, and 4).

© 2014. John Benjamins Publishing Company
All rights reserved

The candidates were derived from the Swedish PAROLE corpus, which is a balanced corpus consisting of novels, newspaper texts, periodicals, and web texts, from the years 1976–1997, with a size of 19 million words.[2]

The cxn candidates were presented with their five most frequent instances. As an example, (1) shows a partially schematic cxn candidate with the length of 4, where the candidate is **Pn *som* Adj N** 'Pn as Adj N', and the most frequent instances are listed together with their frequency in parentheses.[3] The numbers at the top after the pattern are the unique instance frequency followed by the candidate frequency.

(1) 117 **Pn *som*$_{KN}$ Adj N** 121 124
han som svenska kyrkans (2) 'him as the Swedish church's'
dem som andra parter (2) 'them that other parties'
vi som goda kommunister (1) 'we as good communists'
vi som enskild klubb (1) 'we as separate/individual club'

The task for the evaluators was to answer the questionnaire below for each candidate. The questions are formulated in such a way that if an evaluator answers "yes" on 1–4 and 6, and gives the highest score on 5, then this should entail that she considers the candidate to be a perfect fit for a Swedish constructicon.

1. *Is the construction candidate a pattern in Swedish?* [yes/no]
2. *Is the construction candidate a complete pattern?* [yes/no]
3. *Is the construction candidate without extraneous material?* [yes/no]
4. *Is the construction candidate homogeneous?* [yes/no]
5. *Is the construction candidate relevant for a Swedish constructicon?* [0–3]
6. *Is the relevance judgment based on the pattern rather than the examples?* [yes/no]
7. *Any additional comments?* [free text]

Question 1 is intended to filter out what an evaluator judges to be noise in the data. The rationale behind the question is to save time for the evaluator: If the answer is "no", no subsequent questions were answered. Questions 2–3 direct the attention of the evaluators to certain central formal characteristics of the candidate, in particular whether the posited pattern matches the proposed candidate n-gram in length. Question 4 concerns the matter of ambiguity; "no" means that the candidate represents more than one cxn pattern. Question 5, which is the most

2. The Swedish PAROLE corpus is available from Språkbanken (the Swedish Language Bank); see <http://spraakbanken.gu.se/eng/resource/parole>.

3. Note that the second example differs structurally from the remaining ones, reflected in the use of *that* instead of *as* in the English paraphrase. Hence it is not an instance of the same construction.

© 2014. John Benjamins Publishing Company
All rights reserved

important question, is an indicator of priority; irrelevant patterns receive a score of 0, whereas candidates awarded 1–3 are all considered relevant for a Swedish constructicon, although to different degrees. And, since it is not clear if the judgment in question 5 is primarily based on the actual candidate or its instances, question 6 asks for this distinction. Finally, question 7 adds the possibility for the evaluator to comment on the current candidate.

Before the actual evaluation, we had a long discussion about the questionnaire with the evaluators, after which they made a pilot evaluation of a small sample set of cxn candidates. The differences in judgment were presented to them, so that they could discuss the differences and try to reach a consensus. They reported having no problem with agreeing on the differences. The goal of this procedure was to reduce the confusion about what kind of judgment was expected for each question.

The main part of the discussion was concerned with question 1 and 2: How much linguistic material may be missing or extraneous in a candidate that is not a non-cxn? Another topic under discussion was the strong influence of the examples, where it was all too easy to forget about the actual candidate. Moreover, it became obvious that the different backgrounds of the evaluators had a strong impact on their judgment, such as their views on conventionalized expressions. Finally, for question 4, it was unclear how to judge the occurrence of both literal and metaphorical usages; it was decided to consider the pattern as homogeneous even though both usages occurred, since the distinction was too arbitrary in most cases.

The evaluators then proceeded to the full evaluation of the 1,200 cxn candidates. As a starting point for the analysis, we computed Krippendorff's α on their answers, to calculate the inter-annotator agreement. The inter-annotator agreement was surprisingly small, ranging from low to moderate, which suggests that the task is difficult, and that the evaluators' backgrounds had a strong effect on their answers, even though it was deceptively easy for them to agree on the differences in the evaluation of the smaller sample. The low agreement should not be considered a failure of the experiment, which will be elaborated upon in the next section. Moreover, a higher agreement among the three evaluators would mean nothing more profound than that they were able to interpret the questions in the same way.

The most interesting question is the one about relevance, where a score of 1 is typically attributed to a very general cxn, such as a common NP patterns or subordinate clause types. While these are, of course, relevant for a constructicon, they are not our first priority. A score of 2 or more indicates a candidate of high priority for a Swedish constructicon. The results will be presented in the next section.

© 2014. John Benjamins Publishing Company
All rights reserved

5. Evaluation

The evaluation of the candidates revealed two major results: First, the method provided a good number of interesting cxn candidates. Of the 1,200 candidates considered in the experiment, 314 were assigned a relevance score of 2–3 by at least one of the evaluators. Adjusting for duplicates, a fair estimate is that the outcome is about 200 potential constructicon entries of high priority (see Section 5.1 below).

Second, as mentioned in the previous section, the judgments vary considerably. Full consensus is rare, and for some candidates the relevance scores range from 0 to 3. It is quite clear, however, that the high amount of inter-evaluator variation is not due to different opinions among the evaluators. When comparing their results afterwards they agreed on almost every point. Rather, the explanation seems to be two other factors. On the one hand, the task is difficult and the criteria turned out to be hard to apply consistently. On the other hand, the evaluators simply spotted different things. Presented with an abstract n-gram and five examples, a creative mind can notice different patterns, and this is what the evaluators did. On the negative side, no reliable quantitative conclusions can be drawn from the experiment. On the positive side, however, more cxns were identified this way. The evaluations will be exemplified and discussed in Section 5.2.

5.1 High-ranking candidates

The most interesting cxn candidates are the ones with a relevance score of 2 or better. 314 candidates received this rank from at least one evaluator, 89 got it from at least two, and 20 were highly ranked by all three. Since the different assessments are due to different observations rather than different opinions, a high score from one evaluator is enough for a candidate to be given high priority.

However, 314 candidates do not equal 314 constructions. On the one hand, more than one cxn was identified for some of the candidates, either because of ambiguous n-grams or because some observations concerned particular examples rather than the candidate structures as such. Such instances were documented in the answers to question 4, the homogeneity criterion, whereas as many as 225 of the 314 were found to produce more than one cxn; to what extent the additional patterns are considered relevant remains to be investigated. On the other hand, some candidates correspond to essentially the same cxn. For example, consider the candidates in (2).

(2) a. ***både*** **PP** ***och*** **PP**
'both' 'and'
b. ***både*** **PP** **Conj** **PP**

© 2014. John Benjamins Publishing Company
All rights reserved

Both candidates in (2) show the same Swedish cxn, corresponding to 'both PP and PP' in English. In addition, the analogous candidates [***både* NP *och* NP**], [***både* AP *och* AP**] etc. also instantiate the same basic cxn and will not be treated as distinct subtypes in the Swedish constructicon. Such duplicates reduce the number of potential cxns considerably.

Based on a sample of 89 candidates, the ones with at least two high priority scores, an approximate estimation of the number of duplicates was made. After filtering out the duplicates, 58 out of 89 candidates remained, i.e. 65%. Assuming roughly the same proportions, the experiment has provided about 200 cxn candidates (65% of 314) for immediate consideration as constructicon entries, many of which represent more than one potential cxn entry. Some of them are already covered in the Swedish constructicon, but the majority are new cxns. As for their distribution over grammatical categories, the ones in the 58/89 sample are distributed as in Table 1.

Table 1. A sample of 58 cxn candidates sorted by category

NP	AdvP	VP	Coord	PP	S	Other	Total
25	12	8	6	3	2	2	58

Except perhaps for the high number of NPs, the distribution of candidates over categories shows no apparent skew. While the scarcity of clausal cxns (S) is expected from the maximum n-gram length of 4, the low number of PPs is more surprising. In the set of 314 candidates, by contrast, there are 31 PPs. The coordination cxns stand out with regard to duplicates; in this small sample, 16 coordination duplicates were removed. As illustrated above, this depends partly on the high frequency of *och* 'and' among conjunctions, partly on the tendency of coordination cxns to apply equally to different phrase types. A more detailed investigation of the properties of particular cxns is beyond the scope of the present paper.

5.2 Examples and discussion

In the following the results of evaluation are discussed and illustrated with examples. First, the candidates were sorted according to the sum of the pairwise differences in judgment of the evaluators, where a "yes" was counted as 1 and "no" as 0. With three evaluators, this amounted to a difference range of 0 to 16. For approximately 50% of the candidates the evaluators's judgements differed by 4 at most, i.e. in half of the cases the difference in the participants' judgement was relatively small. For approximately 230 of the candidates the difference fell in the range 10–14.

© 2014. John Benjamins Publishing Company
All rights reserved

An example candidate with both a high relevance score and a high degree of agreement (difference: 2) is the pattern in (3):

(3) a. [***för***$_{\mathbf{P}}$ **NP** ***skull***$_{\mathbf{N}}$]
b. *för barn-en-s skull*
for children-DEF-GEN sake
'for the children's sake'
c. *för säkerhet-s skull*
for security-GEN sake
'just in case'

The candidate n-gram in (3a) received a relevance score of 3 from one evaluator and 2 from the other two. They also gave the same response to all the other questions except the fourth, where one of them indicated that more than one cxn is represented. Indeed, whereas the pattern is quite productive, there are also particular conventionalized instances of it, such as the example in (3c). While this specific instance is a lexical cxn, the general pattern is clearly a good candidate for the Swedish constructicon.

A typical case where the consensus is that the candidate in question does not represent a pattern in Swedish is illustrated in (4) below. The candidate generates a variety of examples without indicating any obvious patterns and is well suited to illustrate the function of question 1 as a filter.

(4) **[NP QP QP N]**
valör-er-na 1:70 och 2:30 mark-Ø
denomination-PL-DEF 1:70 and 2:30 mark-PL
'the denominations 1:70 and 2:30 marks'
k1 1 000 m-Ø
k1 1 000 m(eter-PL)
'k1 1,000 meters'

Questions 2 and 3 address to what extent the n-gram in its original form represents an actual pattern in Swedish, or whether it has too few or too many elements to be considered a complete pattern. Example (5) shows an n-gram that is clearly missing an element, i.e. an NP, to represent the cxn **såväl NP som NP** "NP as well as NP". As we can see, the Swedish word order deviates from the English one. However, **NP såväl som NP** is also a possible variant of the construction in question — although a less common one.

(5) [***såväl***$_{\mathbf{KN}}$ **NP** ***som***]
såväl kund-er som
as well customer-PL as
'customers as well as'

© 2014. John Benjamins Publishing Company
All rights reserved

såväl övernattning som
as well accomodation as
'accomodation as well as'

Example (6) illustrates some interesting results in relation to question 3, the without-extraneous-material criterion:

(6) a. **[NP *den*$_{\text{DT}}$ QP N]**
b. *brynäs-forward-en Anders Huss den 29 januari*
brynäs-forward-DEF Anders Huss the 29 January
'The Brynäs forward Anders Huss the 29th of January'
c. *fredag-en den 13:e*
Friday-DEF the 13th
'Friday the 13th'
d. *eu-val-et den 17 september*
EU-election-DEF the 17 September
'The EU election the 17th of September'

The cxn in (6a) contains an NP, followed by the determiner *den* 'the', followed by a QP and an NN denoting the number of the day in a particular month. According to the examples in (6) the pattern is used to describe a date or some activity happening at a certain date. Example (6b) on the other hand, contains some undesired elements, since the hockey player *Anders Huss* does not seem to have anything to do with the date cxn. The example illustrates the difference in judgement due to the fact that one can analyze the n-grams — or the examples they generate — in different ways, as mentioned in Section 5. While the pattern was considered to represent the candidate n-gram (i.e. question 3 was answered with "yes") by one evaluator, another considered the pattern without the NP, judging the candidate to contain extraneous material. In consequence, two different cxns were observed; one with an NP denoting the day of the week, as in (6c), or some activity, as in (6d), and one exclusively describing the date, represented as [***den*DT QP N**].

The following two examples (7–8) illustrate the function of question 4, the homogeneity criterion. In (7), the candidate was judged as homogeneous by all three evaluators:

(7) a. **[P NP *sedan*$_{\text{AB}}$]**
b. *för två år-Ø sedan*
for two year-PL ago
'two years ago'
c. *för några år-Ø sedan*
for some year-PL ago
'some years ago'

© 2014. John Benjamins Publishing Company
All rights reserved

The evaluators found this candidate to be consistent, as well as rather relevant (1,3,1) for the constructicon, even though it probably could be argued that the pattern follows the lexical properties of *sedan* 'ago', i.e. its valence. Example (8) shows a candidate which instead was judged as non-homogeneous:

(8) a. **[NP *efter*$_{p}$ NP]**
b. *år efter år*
year after year
'year after year'
c. *utrikesminister Lena Hjelm-Wallén efter möt-et*
Foreign.minister Lena Hjelm-Wallén after meeting-DEF
'Foreign Minister Lena Hjelm-Wallén after the meeting'

The pattern in (8a) is not homogeneous since it generates different cxns, which is illustrated by (8b–c).[4] The candidate was also judged as relevant for the constructicon, and especially the **[time after time]** variant in (8b), which means 'the time goes on without anything happening'. The cxn is difficult to handle lexically, since neither the preposition *efter* 'after' nor the lexical items that instantiate the open slots ('year', 'month', 'day', etc.) provide full transparency for the cxn.

The fifth criterion, question 5, concerns the candidate's relevance for the Swedish constructicon, ranging from "low relevance" (0) to "highly relevant" (3). Example (9) was judged as a highly relevant candidate by one evaluator, while the other two considered it to be of low relevance.

(9) a. **[PN VP-fin]**
b. *det som hände*
it which happen-PST
'what happened'

The candidate (9a) was generating examples such as (9b). One evaluator, however, observed the extended pattern in (10), hence regarding the candidate as an incomplete pattern, and judged this as a highly relevant cxn for the Swedish constructicon.

(10) a. **[VP VP]**
b. *det som hände hände (ju)*
it which happen-PST happen-PST (after.all)
'what happened happened (after all)'

4. (8c) could be an instance of a common way of introducing direct speech, especially in newspaper style.

© 2014. John Benjamins Publishing Company
All rights reserved

The double VP cxn represented in (10) is an unexpected cxn, since we do not expect a VP to be followed by another VP with identical lexical content, sometimes with the optional element *ju* as shown in (10b). This shows that the method allows for discovering phenomena not immediately apparent from the pattern, but rather triggered by an example (question 6), which means that one can find phenomena one did not know one was looking for in the first place.

The next example also concerns the question if the relevance judgment is based on the pattern rather than the examples:

(11) a. **[*det*$_{\text{PN}}$ *här*$_{\text{AB}}$ PP]**
b. *det här med uteslutning*
this here with exclusion
'this thing with exclusion'
c. *det här under ert eget tak*
this here under your own roof
'this under your own roof'

Regarding the pattern in (11) all three evaluators answered question 6 with "no". Thus the candidate was judged as relevant for the constructicon (2,2,2) based on the example (11b), rather than on the pattern. These could be described as [**det här med X**] "concerning X", where X could be a thing or an event.

Some of the differences in relevance scores are also due to special interests. For instance, one of the evaluators is a specialist on second language acquisition and therefore awarded a higher score to cases like (12):

(12) a. **[ADVP-wh NP V VP-att]**
b. *hur det kommer att gå*
how it come-PRS to go-INF
'how it is going to go'
c. *när regering-en kommer att fatta beslut*
when government-DEF come-PRS to make-INF decision
'when the government is going to make a decision'

The instantiations of the candidate in (12) are potentially problematic for learners of Swedish because they are easily confused with the more specific pattern in (13):

(13) a. **[ADVP-wh NP *komma* VP-att]**
b. *Hur det kom att prägla hennes liv*
How it come-PST to mark-INF her life
'How it came to mark her life'

The specific meaning associated with the verb form of *komma* 'come' in (13) is a certain unpredictability and lack of intention of the action described by the second

© 2014. John Benjamins Publishing Company
All rights reserved

VP. A learner who has not encountered the patterns in (12) and (13) before has no way of knowing these differences but even after he or she has analyzed the difference between the structures, a certain potential for confusion on a semantic-pragmatic basis remains. This is the reason why the candidate in (12) was regarded highly relevant by the evaluator with a special interest in L2 acquisition, while the other two evaluators considered it mostly reflecting the general valence relation between the fixed elements *komma att* 'come to' (which is used to form a periphrastic future, among other things) and the variable V.

6. Conclusions

We have presented an automatic, corpus-based method for suggesting potential constructions to include in the Swedish Constructicon. Our method first counts occurrences of linguistic patterns in a large corpus, and then ranks these patterns using a ranking measure based on two intuitions: strong interdependence between the parts of the pattern, and linguistic productivity. Using an automatic method to suggest constructions has advantages not only for efficiency but also from a methodological point of view: it forces the analyst to look more objectively at the constructions actually occurring in corpora, as opposed to focusing on 'interesting' constructions only.

Future efforts in this area will have to address the limitations pointed out by our evaluators. It was frequently observed that the potential construction patterns are often very similar, with only minute variation. Another issue is that the patterns are sometimes too restricted since our method considers short sequences only. To address these problems, we need to generate more flexible generalizations that allow more variation and contexts large enough to capture complete cxns. One possible solution could be to consider graph-theoretic methods instead of our current sequence-based pattern detection method. For instance, there are algorithms for finding frequent subgraphs in large graphs that could be used for this purpose (Jiang et al. 2013). Furthermore, while in this work we focus entirely on the form side of constructions, it is possible that their content side could be modeled using distributional semantics, as employed in methods for detecting non-compositional phrases (Biemann & Giesbrecht 2011).

As a heuristic for identifying potential constructions, the method has proved its worth. The outcome of the experiment is about 200 highly relevant construction candidates, to be considered as constructicon entries in the near future. Although each pattern of course has to be investigated further before inclusion in the database, the sheer number of high priority findings is a considerable contri-

© 2014. John Benjamins Publishing Company
All rights reserved

bution. Furthermore, since many of the patterns most likely will point to related constructions, we expect this number to increase.

To what extent and in which ways the candidates generated this way are different from constructions identified by other means is a question for further analysis. Looking closer at the properties of these constructions should also reveal something about what was *not* located and therefore should be taken into account for future experiments. In addition, relating the intuitive relevance assessments of the candidates to structural properties on the one hand, and known construction criteria on the other, may provide insights about what is perceived as a construction and thus contribute to our understanding of the concept.

Acknowledgements

We are grateful to two anonymous reviewers for their helpful comments on the first version of this article. The research presented here was supported by the Swedish Research Council (the *Swedish FrameNet++ project*; grant agreement 2010–6013), by the Bank of Sweden Tercentenary Foundation (the *Swedish Constructicon* project; grant agreement P12–0076:1), and by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken.

References

Biemann, C., & Giesbrecht, G. (2011). Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the workshop on distributional semantics and compositionality* (pp. 21–28). Portland: ACL.

Boas, H. C. (2014). Zur Architektur einer konstruktionsbasierten Grammatik des Deutschen. In A. Lasch & A. Ziem (Eds.), *Grammatik als Netzwerk von Konstruktionen. Sprachwissen im Fokus der Konstruktionsgrammatik* (pp. 37–63). Berlin: de Gruyter.

Borin, L., Dannélls, D., Forsberg, M., Gronostaj, M.T., & Kokkinakis, D. (2010). The past meets the present in Swedish FrameNet++. In *14th EURALEX international congress* (pp. 269–281). Leeuwarden: EURALEX.

Bybee, J. (2013). Usage-based theory and exemplar representations. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 49–69). Oxford & New York: Oxford University Press. DOI: 10.1093/oxfordhb/9780195396683.001.0001

Bäckström, L., Borin, L., Forsberg, M., Lyngfelt, B., Prentice, J., & Sköldberg, E. (2013) Automatic identification of construction candidates for a Swedish constructicon. In *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013* (pp. 2–12). NEALT Proceedings Series 19.

Bäckström, L., Lyngfelt, B., & Sköldberg, E. (this issue). Towards interlingual constructicography. On correspondence between constructicon resources for English and Swedish.

© 2014. John Benjamins Publishing Company
All rights reserved

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.

Croft, W. (2003). Lexical rules vs. constructions: A false dichotomy. In H. Cuyckens, T. Berg, R. Dirven & K.-U. Panther (Eds.), *Motivation in language: Studies in honour of Günter Radden* (pp. 49–68). Amsterdam: John Benjamins. DOI: 10.1075/cilt.243.07cro

Culicover, P. W., & Jackendoff, R. (2005). *Simpler syntax*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199271092.001.0001

Evert, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Fano, R. M. (1961). *Transmission of information: A statistical theory of communications*. New York: MIT Press.

Fillmore, C. J. (2008). Border conflicts: FrameNet meets construction grammar. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX international congress* (pp. 49–68). Barcelona: Universitat Pompeu Fabra.

Fillmore, C. J., Lee-Goldman, R., & Rhomieux, R. (2012). The FrameNet constructicon. In H. Boas & I. Sag (Eds.), *Sign-based construction grammar* (pp. 309–372). Stanford: CSLI.

Friberg Heppin, K., & Toporowska Gronostaj, M. (this issue). Exploiting FrameNet for Swedish: Mismatch?

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

Goldberg, A. E. (2013). Constructionist approaches. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 15–31). Oxford & New York: Oxford University Press. DOI: 10.1093/oxfordhb/9780195396683.001.0001

Gustafson-Čapková, S., & Hartmann, B. (2006). Manual of the Stockholm Umeå corpus version 2.0. Stockholm University.

Halácsy, P., Kornai, A., & Oravecz, C. (2007). HunPos – an open source trigram tagger. In *Proceedings of the 45th annual meeting of the Association for Computational Linguistics companion volume: Proceedings of the demo and poster sessions* (pp. 209–212). Prague: ACL.

Hilpert, M. (2013). *Constructional change in English. developments in allomorphy, word formation, and syntax*. Cambridge & New York: Cambridge University Press. DOI: 10.1017/CBO9781139004206

Hilpert, M. (2014). *Construction grammar and its application to English*. Edinburgh: Edinburgh University Press.

Hutchinson, L. G. (1974). Grammar as theory. In D. Cohen (Ed.), *Explaining linguistic phenomena* (pp. 43–73). New York, etc.: Wiley.

Jackendoff, R. (2002). *Foundations of language. Brain, meaning, grammar, evolution*. Oxford & New York: Oxford University Press.

Jiang, C., Coenen, F., & Zito, M. (2013). A survey of frequent subgraph mining algorithms. *The Knowledge Engineering Review* 28, 75–105. DOI: 10.1017/S0269888912000331

Kay, P. (2013). The limits of (construction) grammar. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 32–48). Oxford & New York: Oxford University Press. DOI: 10.1093/oxfordhb/9780195396683.001.0001

Kilgarriff, A., & Tugwell, D. (2002). Sketching words. In M.-H. Corréard (Ed.), *Lexicography and natural language processing: A Festschrift in honour of B. T. S. Atkins* (pp. 125–137). EURALEX.

© 2014. John Benjamins Publishing Company
All rights reserved

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., Marinov, S., & Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2), 95–135.

Nivre, J., Megyesi, B., Gustafson-Čapková, S., Salomonsson, F., & Dahlqvist, B. (2008). Cultivating a Swedish treebank. In J. Nivre, M. Dahllöf & B. Megyesi (Eds.), *Resourceful language technology: Festschrift in honor of Anna Sågvall Hein* (pp. 111–120). Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia 7.

Ohara, K. (2013). Toward constructicon building for Japanese in Japanese FrameNet. *Veredas*, 17(1), 11–27.

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44, 137–158. DOI: 10.1007/s10579-009-9101-4

Svanlund, J. (2002). Lexikalisering [Lexicalization]. *Språk och stil*, 12, 7–45.

Torrent, T. T., Lage, L. M., Sampaio, T. F., Tavares, T., & Matos, E. (this issue). Revisiting border conflicts between FrameNet and construction grammar: Annotation policies for the Brazilian Portuguese Constructicon.

Van de Cruys, T. (2011). Two multivariate generalizations of pointwise mutual information. In *Proceedings of the workshop on distributional semantics and compositionality* (pp. 16–20). Portland: ACL.

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: ACL.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Ziem, A., Boas, H. C., & Ruppenhofer, J. (2014). Grammatische Konstruktionen und semantische Frames für die Textanalyse. In J. Hagemann & S. Staffeldt (Eds.), *Syntaxtheorien. Analysen im Vergleich* (pp. 297–333). Tübingen: Stauffenberg.

Authors' addresses

Markus Forsberg
Språkbanken, Dept. of Swedish
University of Gothenburg
Box 200
SE-405 30 Gothenburg
Sweden

markus.forsberg@svenska.gu.se

Richard Johansson
Språkbanken, Dept. of Swedish
University of Gothenburg
Box 200
SE-405 30 Gothenburg
Sweden

richard.johansson@svenska.gu.se

Linnéa Bäckström
Department of Swedish
University of Gothenburg
P.O. Box 200
SE-40530 Gothenburg
Sweden

linnea.backstrom@svenska.gu.se

Lars Borin
Språkbanken, Dept. of Swedish
University of Gothenburg
Box 200
SE-405 30 Gothenburg
Sweden

lars.borin@svenska.gu.se

© 2014. John Benjamins Publishing Company
All rights reserved

Benjamin Lyngfelt
Department of Swedish
University of Gothenburg
P.O. Box 200
SE-40530 Gothenburg
Sweden

benjamin.lyngfelt@svenska.gu.se

Julia Prentice
Department of Swedish
University of Gothenburg
P.O. Box 200
SE-40530 Gothenburg
Sweden

julia.prentice@svenska.gu.se

Joel Olofsson
Department of Swedish
University of Gothenburg
P.O. Box 200
SE-40530 Gothenburg
Sweden

joel.olofsson@svenska.gu.se

© 2014. John Benjamins Publishing Company
All rights reserved