Automatic identification of construction candidates for a Swedish constructicon

Linnéa Bäckström, Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Julia Prentice, and Emma Sköldberg

Dept. of Swedish, University of Gothenburg, Sweden

{linnea.backstrom, lars.borin, markus.forsberg,
benjamin.lyngfelt, julia.prentice, emma.skoldberg}@svenska.gu.se

Abstract

We present an experiment designed for extracting construction candidates for a Swedish constructicon from text corpora. We have explored the use of hybrid n-grams with the practical goal to discover previously undescribed partially schematic constructions. The experiment was successful, in that quite a few new constructions were discovered. The precision is low, but as a push-button tool for construction discovery, it has proven a valuable tool for the work on a Swedish constructicon.

Keywords: hybrid n-gram, Swedish, constructions, constructicon.

Linnéa Bäckström, Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Julia Prentice and Emma Sköldberg 2013. Automatic identification of construction candidates for a Swedish constructicon. *Proceedings of the workshop on lexical semantic resources for NLP at NODALIDA 2013.* NEALT Proceedings Series 19 / Linköping Electronic Conference Proceedings 88: 2–11.

1 Introduction

The research within the project *A Swedish constructicon* (see section 2) is targeted at Swedish constructions that are collected, analyzed, described, and published in a freely available resource.[1] Since no exhaustive construction description has ever existed for Swedish – or, to our knowledge, for any language – an important methodological question for the project is how to discover those constructions that have not been recognized as such before.

At least in the initial experiment presented here, where we have explored the use of hybrid n-grams (see section 3) as a tool for construction discovery, the search for construction candidates is restricted to partially schematic patterns, i.e. structures where at least one component is lexically fixed and at least one component is schematic, i.e., a morphosyntactic category. In this way, we target patterns with both lexical and grammatical properties, which neither purely lexical nor purely grammatical tools can capture.

2 Swedish constructicon

The Swedish constructicon (SweCxn; Lyngfelt et al., 2012) is a collection of variable multi-word units, based on principles of Construction Grammar and designed as an addition to the Swedish FrameNet (Borin et al., 2010).[2] It is still in its early stages, but the intention is for it to be developed into a large-scale, freely available resource for linguistics and language technology. At present, SweCxn consists of around 100 Swedish constructions, and is growing continually. A major concern of SweCxn is to account for linguistic patterns that are too specific to count as general rules of grammar and too general to be attributed to individual lexical units. Such constructions are peripheral both from a grammatical and a lexical perspective, and are therefore easily overlooked and neglected in grammars and lexical resources alike. Special attention is given to constructions deemed problematic for L2 acquisition. The project is a collaboration between grammarians, language technologists, lexicographers, phraseologists, semanticists, and L2 researchers.

One of the goals of SweCxn is to develop tools for automatic identification of constructions in authentic texts. This is a highly desirable research objective in itself, with potential uses in a number of NLP applications. In addition, the same methods provide the project with a heuristic tool. By automatically extracting various kinds of regularities in texts, we may discover patterns that might otherwise have been overlooked. This especially concerns seemingly insignificant constructions that do not stand out against the context the way spectacular idioms do. The resulting findings are treated as construction candidates, a subset of which may be considered actual constructions after manual evaluation.

3 Experiment setup

The general setting for our experiment is the resource infrastructure of *Språkbanken* (the Swedish Language Bank),[3] a modular set of resources and tools in the form of web services for accessing, browsing, editing and automatically annotating resources. The two facets of the infrastructure most relevant for the present purposes are the corpus infrastructure *Korp* (Borin et al., 2012b) and the lexicon infrastructure *Karp* (Borin et al., 2012a). Together, these provide a set of interoperable web services and downloadable resources which enable experiments like the one described here to be quickly set up and executed.

[1] The Swedish constructicon is accessible from here: <http://spraakbanken.gu.se/resurs/konstruktikon>

[2] See <http://spraakbanken.gu.se/swefn>.

[3] <http://spraakbanken.gu.se>

word	msd	lemma
Hur	HA	hur
är	VB. PRS. AKT	vara
det	PN. NEU. SIN. DEF. SUB+OBJ	den
då	AB	då
i	PP	i
Mellanöstern	PM. NOM	Mellanöstern
?	MAD	

Figure 1: SUC 2.0 annotations

The data source for the experiment is SUC 2.0 (Ejerhed and Källgren, 1997; Ejerhed et al., 1992), a balanced text corpus for Swedish consisting of 1.17M tokens that have been manually annotated with lemmas and MSDs (morphosyntactic description). A random example sentence from SUC 2.0 is given in Fig. 1: *Hur är det då i Mellanöstern?* 'What about the Middle East?'. The first part of the MSD is the part-of-speech, e.g., *VB* for *är* 'is'.

SUC was selected in order to avoid annotation errors confounding the experiment results, but the experiment can (and has been) run on any of the more than hundred corpora of Språkbanken that have been automatically annotated with the same information.

The experiment is based on the work on StringNet (Tsao and Wible, 2009; Wible and Tsao, 2010, 2011), where the notion of *hybrid n-gram* plays a central role. A hybrid n-gram is a generalization of an n-gram where not only the word forms are included in the process, but also the information from the annotation layers. If we limit ourselves to lemmas and part-of-speech, which is the case for this experiment, then the 2-gram *Hur är* 'How is' would generate four construction candidates: *hur vara* 'how be', *hur VB* 'how VB', *HA vara* 'HA be', and *HA VB*.

Since the aim is to capture partially schematic constructions, we discard all candidates that are fully schematic or fully lexical, i.e., consisting of only PoS tags (e.g., *HA VB*) or lemmas (e.g., *hur vara* 'how be'). Moreover, we remove all hybrid n-grams containing punctuation marks and/or words marked as foreign. They are not necessarily uninteresting, but since they did introduce a lot of noise in the candidate list, we decided to remove them. For SUC 2.0 with 2-, 3-, and 4-grams we ended up with 16M hybrid n-grams of which 8.8M were unique.

The next step is to rank all hybrid n-grams, which can be done with a wide range of association measures. We have followed StringNet in using point-wise mutual information (PMI). PMI has a known shortcoming in these kinds of experiments – it has a preference for the low-frequency items – which can be remedied by multiplying PMI with the absolute frequency. This does not solve another problem, however, which is boilerplate text, e.g., "For subscription enquiries e-mail:...". But with a small modification – instead of counting hybrid n-grams, we count UIF (unique instance frequency), which is the number of unique n-grams underlying the target hybrid n-gram – we can counteract that problem too. In sum, we end up with the following formula:

$$\text{PMI-UIF(H)} = UIF * log_2\left(\frac{P(H)}{\prod_{x \in H} P(x)}\right)$$

There is still one more problem that needs to be solved: since the bulk of the hybrid n-grams are subsets of other hybrid n-grams, we arrive at a ranking list with massive redundancy. This is solved, in the same spirit as StringNet's vertical/horizontal pruning (Tsao and Wible, 2009; Wible and

$vara_{VB}$ ute_{AB} och_{KN} VB	*är ute och letar (3)*	15	0.93	52.24
$vara_{VB}$ JJ $för_{PP}$ att_{IE}	*är viktiga för att (2)*	26	1.61	52.83
$stänga_{VB}$ av_{PL} NN	*stängt av motorn (1)*	11	0.68	52.25

Figure 2: Some example hybrid n-grams from SUC 2.0 ranked by PMI-UIF

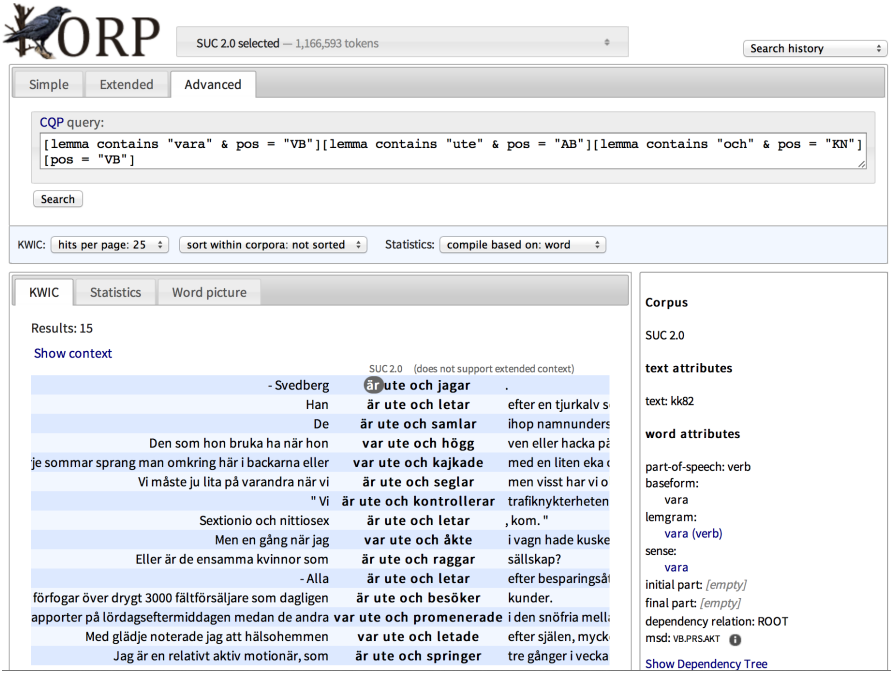

Figure 3: The instances of $vara_{VB}$ ute_{AB} och_{KN} VB

Tsao, 2010), by removing all hybrid n-grams that are subsets of other hybrid n-grams with a higher PMI-UIF. A hybrid n-gram is considered a subset of another if it occurs as a subsequence that are either equal or consisting of non-conflicting items sharing the same part-of-speech; e.g, $vara_{VB}$ is considered equal to VB.

Some sample candidates are given in Fig. 2. The hybrid n-grams are linked to the Korp interface to enable inspection of their instances (see Fig. 3). We also see the most frequent instance, followed by the absolute frequency, relative frequency, and the PMI-UIF. The full output of a top-2500 list is accessible from here: <http://spraakbanken.gu.se/eng/resource/konstruktikon/candidates> (may be subject to change). Here you will find other materials as well that have been annotated automatically using the Korp pipeline. More specifically for this experiment, we use the Swedish Hunpos tagger (Megyesi, 2009) for the part-of-speech tags, and the lexical analysis based on SALDO (Borin et al., 2008; Borin and Forsberg, 2009) for the lemmatization.

4 Data analysis

The construction candidate list makes it possible to go through a large amount of examples quickly, since every hybrid n-gram is directly linked to the instances in the corpus. However, it was a difficult task to draw the line between relevant and non-relevant constructions and this is still an ongoing matter of discussion in the project group. Of the 2500 items included in the list 50 constructions were decided to be relevant construction candidates according to our criteria, i.e., that they are partially schematic and productive multiword units that are "too general to be attributed to individual words but too specific to be considered general rules" (Lyngfelt et al., 2012).

The final list of 50 relevant constructions was extracted in several steps. First one project member went through the whole list extracting a list of 143 interesting candidates (approximately a day's work). This list was then, in consultation with the other members of the project group, gradually reduced and the final result of this process were, as mentioned above, 50 constructions that were found relevant for entries in the SweCxn. As the main goal was to discover constructions that are difficult to find with other methods the result of 50 is not the whole story – a construction candidate can also inspire descriptions of other similar constructions, which is a question of the researchers' capacity for creative thinking at a given moment in time.

The instances of the construction candidates display different properties regarding the form-function structure. The results represent patterns of lexical, idiomatic and syntactic character. A strong indication that the method identifies the correct items is that some of the qualifying constructions are already present in the SweCxn. One of these examples is:

(1) *RG NN* per_{PP} *NN*

The structure in (1) is realized in the corpus as, e.g., *en gång per dygn* 'once in 24 hours' and *500 kronor per månad* '500 Swedish Crowns per month'. This construction that can be regarded as a Swedish equivalent to a construction in the Berkeley English Cxn (Fillmore et al., 2012), the so-called Rate construction. Another construction already accounted for in the SweCxn is (2) below:

(2) den_{DT} *RO NN*

Instances of this structure found in the corpus are date expressions like *den 1 juli* 'the 1st of July' and *den tionde mars* 'the tenth of March'. Fillmore (2008) discusses this type of time expressions referring to dates or days of the week, which in English have a conventionalized structure with the preposition *on*, i.e., on_{PP} *NN RO* (*on June 17th*) and, hence, deviates from other time expressions like *in March*, *in the morning* and *at noon* (Fillmore, 2008). The Swedish date expression in (2) occurs without a preceding preposition, and differs in this respect from both its English counterpart and from the typical pattern of time expressions in Swedish as well as in many other languages, e.g., *i mars* 'in March', *på morgonen* 'in the morning', or *på eftermiddagen* 'in the afternoon'. This property makes the construction a challenge for language learners (Prentice, 2011).

A construction that is not previously included in the SweCxn is exemplified here:

(3) *RG* $år_{NN_{GEN}}$ $ålder_{NN}$

The genitive construction in (3) is realized in the corpus as, e.g., *vid sju års ålder* 'at the age of seven' and *från 17 års ålder* 'from 17 years of age'. The construction is not described in Swedish dictionaries in a sufficient way despite the fact that it can hardly be seen as completely transparent (cf. Köhler and Messelius, 2001). In, e.g., English and German the same content is expressed with a different kind of prepositional phrase (Eng. *at the age of...*; Germ. *im Alter von...*) (cf. Källström, 2012).

Other relevant candidates included in the list are the comparing constructions in (4)–(6) below:

(4) $varken_{KN}$ *NN* $eller_{KN}$ *NN*

(5) $vara_{VB}$ sig_{PN} *NN* $eller_{KN}$ *(NN)*

(6) $vara_{VB}$ sig_{PN} *PN VB* ($eller_{KN}$ $inte_{AB}$)

Examples of these structures from the corpus are (4) *varken uppehållstillstånd eller arbetstillstånd* 'neither residence permit nor work permit', (5) *vare sig fotboll eller ishockey* 'neither football or ice hockey', and (6) *vare sig vi vill eller inte* 'whether we want to or not'.[4] As we can see, (4) and (5) are used synonymously (in the sense of 'neither'), whereas (6) can be substituted by *oavsett (om)* 'regardless if'. The traditional normative rule is that *vare sig* in (5) requires explicit negation whereas negation is seen as part of the meaning of *varken*. The structure in (6) is obviously similar to (4) and (5) but here *vare sig* functions as a subjunction and does not require negation (Svenska språknämnden, 2005). In actual language use, however, both *vare sig* and *varken* are used with and without negation and, e.g., the usage notes in Svenska Akademien (2009) indicate that the candidates in (4)–(6) are subject of an ongoing language planning discussion. The potential for contamination – for both native and non-native speakers – is quite obvious, which is also reflected in the corpus. Considering the similarities and differences between the structures, as well as their discontinuity, the cluster can cause problems in relation to, e.g., language technology, lexicography, and language learning, which makes these structures excellent candidates for the SweCxn (cf. Lyngfelt et al., 2012).

Another interesting example occurring in the list is:

(7) $vara_{VB}$ ute_{AB} och_{KN} *VB*

The corpus samples linked to (7) mostly contains instances of the construction with the literal meaning 'being out doing something one typically does outside', e.g., *vara ute och jaga* 'being out hunting'. These instances are not particularly relevant as candidates for the SweCxn since they can be referred to as a general syntactical pattern. However, a search for this general structure in a wider range of corpora provides metaphorical instances of the pattern, implying a certain 'disorientation', 'confusion', or 'lack of knowledge' on the part of the agent. One of the most conventionalized examples is *vara ute och cykla* (lit. 'being out biking'), meaning 'being mistaken' or 'not knowing what one is talking about' (cf. *talk through one's hat*). Other realizations are *vara ute och segla* (lit. 'being out sailing'), and *vara ute och snurra* (lit. 'being out spinning'), which both are used synonymously with *vara ute och cykla* in a metaphorical sense. Here, the form-function structure is by no means obvious, which also makes the construction relevant from an L2-perspective. The word combination *vara ute och cykla* is included in printed dictionaries. However, the productivity of the construction is far from evident; an information that can be straightforwardly described in the SweCxn format.

As mentioned before, a majority of the items generated by the method are not relevant candidates for entries in the SweCxn, or at least of no priority in the current state of the project. One of those structures is exemplified in (8), where the candidate is a noun phrase followed by a finite verb, thus a general pattern that can be described according to syntactic rules:

(8) den_{DT} *JJ NN VB*

[4] It is doubtful whether the first component of *vare sig* should be synchronically analyzed as a form of *vara* 'be' as it has been in examples (5) and (6). Rather, the sequence *vare sig* should probably be treated as an unanalyzed whole.

Examples from the corpus are *de senaste månaderna har* 'the last months have' and *de nordiska länderna är* 'the Nordic countries are'. Another example is the sequence in (9) below:

(9) *SN PN VB* en_{DT}

Instances of this sequence are, e.g., *att det var en* 'that it was a' (in *Så fort han hörde att det var en kvinnoröst...* 'as soon as he heard that it was a women's voice...') and *Om du är en* 'if you are a' (in *Om du är en mördare...* 'If you are a killer...').

The sequence in (9) exemplifies another problem with the method, namely that it generates construction fragments – the sequence in (9) is not a recognized linguistic unit of any kind – due to the fact that the method is based on 2-, 3-, and 4-grams. In fact, the sequence in (9) is similar to the *lexical bundles* of Biber and Conrad (1999), a term that they use to refer to high-frequency (word) n-grams. However, distinct to the work reported here, the only criterion used for recognizing lexical bundles is their frequency. No collocation co-ocurrence measures or other means of ranking or filtering the results are used. Instead, fixed-length text word n-grams are sorted according to frequency and the resulting lists manually inspected for interesting results.[5] Lexical bundles are said to differ from other kinds of multi-word units in three major aspects: first, they are extremely frequent, second, they have no idiomatic meaning and last, they are not perceptually striking in themselves. Another characteristic of lexical bundles is that they often transcend structural boundaries.[6]

Biber and Barbieri (2007) ascribe lexical bundles a pre-fabricated or formulaic status, solely on the basis of their high frequency. However, this view has not escaped criticism. Nekrasova (2009) maintains that high-frequent sequences are of different strengths: A bundle should be described in terms of its place on a continuum from more holistic to more compositional units. In the NLP literature it has been observed that although frequency certainly is a strong indicator of MWE-hood (termhood, collocational strength), much can actually be done – and has been done – to improve on frequency alone (Wermter and Hahn, 2006; Pecina, 2010).

The algorithm also generates sequences like the one in (10):

(10) $till_{AB}$ och_{KN} med_{AB} *VB*

An example from the corpus is *till och med börja* 'even start' (in the context *skulle jag till med och börja dricka igen* 'would I even start drinking again'). As in example (9), the phrase has been cut off in an inadequate way. In addition, the structure in (10) contains the fixed phrase *till och med*, and can be compared with other examples from the list:

(11) i_{PP} all_{DT} $fall_{NN}$ *VB*

(12) $över_{PP}$ $huvud_{NN}$ $tagen_{PC}$ *VB*

The main parts of the items in (11)–(12) constitute lexically filled fixed phrases, *i alla fall* and *över huvud taget*, which makes them more suitable as candidates for a dictionary, and indeed, they are well covered in Swedish dictionaries of today.

In some cases the results represent a structure which at first sight does not seem very interesting or relevant according to our criteria. In some of those cases, however, the link to the corpus sample leads to interesting examples of subtypes of a general structure. (13) is one of those cases:

[5]This is a bit like attempting to discover the words in un-word segmented text by looking at the frequency of, e.g., four-character sequences, which seems to be an exercise of doubtful value.

[6]However, it is not very obvious what can be concluded about the language system or the mental lexicon of the language user from the attested high text frequency of a sequence like the example *in the case of the* cited by Biber and Conrad (1999).

(13) $komma_{VB}$ *IE VB*

The structure in (13) ('come IE VB') reflects 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 *VB*. These types of relations are generally well described in dictionaries and therefore not a main priority for SweCxn. However, a search in the corpus for the structure in example (13) highlights a more specific form-function structure:

(14) $komma_{VB_{PAST,SUP}}$ *IE VB*

The pattern in example (14) indicates that the action described by the verb was accidentally initiated, as in *Det var så jag kom att lösa det urgamla filosofiska problemet* 'that was how I came to solve the ancient philosophical problem'. The specific meaning associated with the verb forms is difficult to describe in a classical dictionary format, and the partially schematic structure make this construction – which is also quite productive – a relevant candidate for SweCxn. Looking at the construction in (14) from an L2-perspective, one can assume that it can cause problems, e.g., in relation to the more general structure in (13). How is the learner, who has never met the structure in (14), to know that *jag kom att lösa det urgamla filosofiska problemet* is not simply the past tense of *jag kommer att lösa det urgamla filosofiska problemet* 'I will solve the ancient philosophical problem'? And even once the learner has analyzed the difference between structures in (13) and (14), a certain potential for contamination on a semantic-pragmatic level can be expected (cf. Prentice and Sköldberg, 2011).

5 Conclusions and future work

From a methodological standpoint the experiment has been a success, in that we have been able to discover quite a few previously undescribed partially schematic constructions. The precision is low, but as a push-button tool for construction discovery, it has proven a valuable tool for the work on a Swedish constructicon.

The main issues with the construction candidates are that they often end up being too syntactic (e.g., a candidate may correspond to a regular NP pattern), too lexical (e.g., because of internal inflection of a multi-word unit), or fragmented (due to the nature of n-grams). Planned future work includes the exploration of whether a combination of existing Swedish lexical resources together with the syntactic analysis from the MALT parser (Nivre et al., 2007), accompanied by a more flexible notion of candidates than pure hybrid n-grams, can be used to counteract these issues.

An alternative approach we intend to explore is using MDL (Minimum Description Length) for the construction extraction, an approach that has been previously explored by Lagus et al. (2009).

Acknowledgments

The research presented here was supported by the Swedish Research Council (grant agreement 2010-6013), by the Bank of Sweden Tercentenary Foundation (grant agreement P12-0076:1), by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken, and by Swedish Academy Fellowships for Benjamin Lyngfelt and Emma Sköldberg, sponsored by the Knut and Alice Wallenberg Foundation.

References

Biber, D. and Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26:263–286.

Biber, D. and Conrad, S. (1999). Lexical bundles in conversation and academic prose. In Hasselgard, H. and Oksefjell, S., editors, *Out of corpora: Studies in honor of Stig Johansson*, pages 77–85. Rodopi, Amsterdam.

Borin, L., Danélls, D., Forsberg, M., Kokkinakis, D., and Toporowska Gronostaj, M. (2010). The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281, Leeuwarden. EURALEX.

Borin, L. and Forsberg, M. (2009). All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, Odense. NEALT.

Borin, L., Forsberg, M., and Lönngren, L. (2008). The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In Nivre, J., Dahllöf, M., and Megyesi, B., editors, *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, number 7 in Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia, pages 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala.

Borin, L., Forsberg, M., Olsson, L.-J., and Uppström, J. (2012a). The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 3598–3602, Istanbul. ELRA.

Borin, L., Forsberg, M., and Roxendal, J. (2012b). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Ejerhed, E. and Källgren, G. (1997). Stockholm Umeå corpus 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. ISBN 91-7191-348-3.

Ejerhed, E., Källgren, G., Wennstedt, O., and Åström, M. (1992). The linguistic annotation system of the Stockholm-Umeå corpus project - description and guidelines. Technical report, Department of Linguistics, Umeå University.

Fillmore, C., Lee-Goldman, R., and Rhomieux, R. (2012). The framenet constructicon. In Boas, H. and Sag, I., editors, *Sign-Based Construction Grammar*, pages 309–372. CSLI, Stanford.

Fillmore, C. J. (2008). Border conflicts: FrameNet meets Construction Grammar. In Bernal, E. and DeCesaris, J., editors, *Proceedings of the XIII EURALEX International Congress*, pages 49–68, Barcelona. Universitat Pompeu Fabra, Universitat Pompeu Fabra.

Köhler, P. O. and Messelius, U. (2001). *Natur och Kulturs svenska ordbok*. Bokförlaget Natur och Kultur, Stockholm.

Källström, R. (2012). *Svenska i kontrast. Tvärspråkliga perspektiv på svensk grammatik.* Studentlitteratur, Lund.

Lagus, K., Kohonen, O., and Virpioja, S. (2009). Towards unsupervised learning of constructions from text. In Sahlgren, M. and Knutsson, O., editors, *Proceedings of the Workshop on Extracting and Using Constructions in NLP of 17th Nordic Conference on Computational Linguistics, NODALIDA*. SICS Technical Report T2009:10.

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

Megyesi, B. (2009). The open source tagger HunPoS for Swedish. In Jokinen, K. and Bick, E., editors, *Proceedings of the Nordic Conference on Computational Linguistics (Nodalida)*, volume 4 of *NEALT Proceedings Series*, pages 239–241, Odense, Denmark.

Nekrasova, T. M. (2009). English l1 and l2 speakers' knowledge of lexical bundles. *Language Learning*, 59(3):647–686.

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

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–158.

Prentice, J. (2011). "jag är född på andra november" konventionaliserade tidsuttryck som konstruktioner – ur ett andraspråksperspektiv. Technical report, Institutionen för svenska språket, Göteborgs universitet.

Prentice, J. and Sköldberg, E. (2011). Figurative word combinations in texts written by adolescents in multilingual school environments. In Källström, R. and Lindberg, I., editors, *Young urban Swedish. Variation and change in multilingual settings*. University of Gothenburg.

Svenska Akademien (2009). *Svensk ordbok*. Norstedts, Stockholm.

Svenska språknämnden (2005). *Språkriktighetsboken*. Norstedts Akademiska Förlag, Stockholm.

Tsao, N.-L. and Wible, D. (2009). A method for unsupervised broad-coverage lexical error detection and correction. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 51–54, Boulder. ACL.

Wermter, J. and Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge) – A qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of COLING-ACL 2006*, pages 785–792, Sydney. ACL.

Wible, D. and 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*, pages 25–31, Los Angeles. ACL.

Wible, D. and Tsao, N.-L. (2011). The StringNet lexico-grammatical knowledgebase and its applications. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 128–130, Portland. ACL.