Semi-automatic Building of Swedish Collocation Lexicon

Silvie Cinková, Pavel Pecina, Petr Podveský, and **Pavel Schlesinger**

Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic
Malostranské náměstí 25, Prague, Czech Republic
{cinkova,pecina,podvesky,schlesinger}@ufal.mff.cuni.cz

Abstract

This work focuses on semi-automatic extraction of verb-noun collocations from a corpus, performed to provide lexical evidence for the manual lexicographical processing of Support Verb Constructions (SVCs) in the Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns. Efficiency of pure manual extraction procedure is significantly improved by utilization of automatic statistical methods based lexical association measures.

1. Introduction

1.1. The Notion of Support Verb Construction, Support Verb, and Predicate Noun

Support Verb Constructions (SVCs) are combinations of a lexical verb and a noun or a nominal group containing a predication and denoting an event or a state (henceforth "predicate noun"), e.g. to *to take/make a decision, to undergo a change*. From the semantic point of view, the noun seems to be part of a complex predicate rather than the object of the verb, whatever the surface syntax may suggest. The meaning of an SVC is concentrated in the predicate noun, whereas the semantic content of the verb is reduced or generalized. The notion of SVC and related concepts has already been studied elswhere, see (Grefenstette and Teufel, 1995), (Tapanainen et al., 1998), (Lin, 1999), (McCarthy et al., 2003), (Bannard et al., 2003).

In general, SVCs are easily understood by foreign learners. Their meaning is concentrated in the predicate noun, and, cross-linguistically, the noun is the common denominator of an SVC in the foreign language and its equivalent in the learner's first language. The equivalent is typically also an SVC. Even though the equivalent occasionally may have another form or even be absent (Schroten, 2002), SVCs mostly remain understandable for the learner. On the other hand, SVCs pose substantial problems for foreign language production (Heid, 1998), (Målmgren, 2002) due to the unpredictability of the support verb. E.g. the predicate noun *question* in an SVC meaning *to ask* takes different support verbs in Czech and in Swedish: Czech uses the verb *položit* (i.e. *to put horizontally*) while Swedish uses the verb *ställa* (i.e. *to put vertically*). The translation equivalent of the support verb is unpredictable, though common semantic motivation can be traced back. The unpredictability of the support verb places SVCs into the lexicon, while the semantic generality of support verbs and their productivity move them to the very borders of grammar.

The initial attempts at identifying a fixed group of Swedish support verbs resulted in the insight that verbs, rather than *being* support verbs *become* support verbs by joining a predication-containing noun cf also (Baron and Herslund, 1998). Many verbs belonging to the basic vocabulary have a shifting potential for occurring together with predication-containing nouns. Some enter such constructions frequently and productively, while others only occur in one or few lexicalized cases such as *bjuda* (*to offer*) in *bjuda motstånd* (lit. *to offer resistance*). Only few lexical verbs occur almost exclusively in SVCs (e.g. *genomföra - to perform*). Most verbs with this ability have quite concrete meanings, e.g. *komma, ställa, stå* or *få* (*to come, to put vertically, to stand* and *to get*).

1.2. SVCs as a Means of Event-Structure Specification

Czech learners encounter difficulties when expressing aspect in Swedish (as well as in Germanic languages in general). Due to the morphological category of aspect present in most Czech verbs, Czech learners constantly miss aspect as a morphological category in Swedish. Therefore they may even be ignoring indications of event structure specifications rendered by various lexical means. Having this in mind, the issue of SVCs becomes interesting in connection with their impact on the event structure of the entire utterance.

SVCs are often referred to as one means of specifying event structure in non-aspectual languages. Support verbs add further semantic features to the event described by the given predicate noun, such as inchoativity, durativity, terminativity and causativity (called aspectual, diathetic and modal values by (Fontenelle, 1992), or simply aktionsart by others, e.g. Šmilauer (Šmilauer, 1972). The event structure of a given utterance can be modified by employing an SVC instead of the corresponding lexical verb (provided there is any), e.g. in *falla i sömn* (*to fall asleep*, lit. *to fall into sleep*) versus *sömna*. However, this opposition gives no direct correspondence to the Slavic category of aspect, which apparently is the product of several event structure features in combination. Some authors see aspect, or perfectivity, as a discourse-based phenomenon rather than a lexical feature of a given lexeme, often in comparison to telicity.[1] Therefore, we decided to observe the semantic as well as the morphosyntactic behavior of each SVC in context (i.e. in

[1](Hopper and Thompson, 1980): "Whereas telicity can be determined generally by a simple inspection of the predicate, perfectivity is a property that emerges only in discourse." and (Pustejovsky, 1991): "The lexical specification of a verb's event-type can be overridden as a result of syntactic and semantic compositionality of the verb with other elements in the sentence."

corpus concordances) in order to record the event structure modifications in the entire utterance.

Our attempt to make a link between the Swedish and the Czech ways of specifying event structure is based on (Lindvall, 1998). Lindvall has performed a comprehensive parallel-corpora based comparison of Greek, Polish and Swedish to look into verbal boundedness and object definiteness as two interacting components of Transitivity. Her point of departure was the Transitivity Hypothesis by (Hopper and Thompson, 1980). Transitivity is conceived as a semantic relation of an Agent affecting a Patient. The more the Patient is affected by the Agent, the more transitive the utterance is. This universal conception of Transitivity is not limited to the syntactic relation between a verb and its direct object and is gradual by its nature. According to the Transitivity Hypothesis, utterances with high Transitivity tend to have perfective verb forms and definite objects (whenever the morphology of the given language can indicate it), while utterances with low Transitivity tend to have imperfective verb forms and indefinite objects.

We are seeking to make use of Lindvall's observations by regarding the predicate nouns as objects of support verbs. Our considerations even include prepositional objects. To begin with, we gather the instances of the morphosyntactic behavior of a predicate noun linked to a given support verb, trying to decide which aspect the verb in the corresponding Czech utterances would get and whether the aspect would shift according to the shifting noun-definiteness in the original sentence. Regardless of what the outcome will be, we believe it valuable for the Czech learner to see how variable some predicate nouns can be in context in contrast to others that remain unchanged.

1.3. Outline of the Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns

The observations of Swedish SVCs have resulted in building a small machine-readable lexicon. As the predicate noun is the semantically heavier and more predictable part of a SVC, we decided to lemmatize the SVCs under their respective predicate nouns. SVCs are looked upon as collocations with the predicate noun as node and the verb as collocate. The verbal collocates are sorted by means of the Lexical Functions (Wanner, 1996). The entry structure has been described in more detail in (Cinková and Žabokrtský, 2005b) and (Cinková and Žabokrtský, 2005a). The lexicon seeks to itemize the commonest SVCs as well as to present their productive mechanisms in accordance with the Transitivity Hypothesis. It captures the morphosyntactic variability of predicate nouns in SVCs, i.e. number, article use and attribute insertion. It is yet to be noted that the actual lexicographical work is still in an early stage, and it is thus not meant to be the topic of this paper, which only describes the selection of entry candidates.

2. Tools and Data Sources

The collocations were extracted from the Swedish PAROLE-corpus of modern Swedish texts, which comprises more than 19 million running words. PAROLE belongs to Språkbanken, the set of corpora at Språkdata, University in Gothenburg, Sweden, and is available at `http://spraakbanken.gu.se/lb/parole/`. The PAROLE corpus was built within the EU project PAROLE (finished 1997), which aimed at creating a European network of language resources (corpora and lexicons).

PAROLE has automatic morphological annotation but no lemmatization. To be able to use our statistical collocation extraction method, we needed the corpus lemmatized. As we were not able to obtain any lemmatizer for Swedish from outside, we wrote a make-do lemmatizer ourselves (Cinková and Pomikálek, 2006).

The original frequency sorting was performed with the tool WinConcord (Martinek and Siegrist, 1995). The initial part of the manual extraction was carried out as a fully manual task without any hope of technical support through Språkbankens web interface in 2003. Concordances were copied directly from the web and pasted into a word-processor, which unfortunately limited the number of concordances recorded.

3. Manual Collocation Extraction

The initial extraction procedure was inspired by Heid (1998), Dura (1997), Ekberg (1987) and Malmgren (2002). It comprises three steps:

1. extraction of word expressions whose morphosyntactic behavior suggests that they could be SVCs
2. subsequent manual elimination of non-collocations
3. sorting of collocations into three groups.

Step 1 comprised formulating several corpus queries and obtaining their results. The queries basically varied the distance between the verb and the noun. Some queries introduced article, number and adjective insertion restrictions. To ensure that the noun be the object of the verb, the verbs had to follow a modal or an auxiliary verb.

To carry out the steps 2 and 3, the collocations were ordered according to their frequency in the corpus. Each collocation interval (i.e. the distance between the noun and the verb) was processed in a separate file. Equally frequent collocations were sorted alphabetically according to their verbs. This facilitated the manual processing, as some very frequent verbs could be instantly recognized as "never-support-verbs", and ignored in blocks, i.e. *köpa (to buy)* or *säga (to say)*. Step 3 included a fine-grained semantic classification. Three groups were set at the beginning: "SVCs", "Quasimodals" and "Phrasemes". The group "SVCs" included collocations with nouns denoting an event (also a state) or containing a predication, e.g. *få hjälp (to get help)* and *få betydelse* (lit. *to get significance - to become significant*). In the group "SVCs" it is the event described by the predicate noun that actually "takes place". In "Quasimodals", on the other hand, the verb and the predicate noun form one semantic unit that resembles a modal verb (e.g. *to get the chance to V = to start to be able to V* etc.) (Cinková and Kolářová, 2004) and must be completed by the event in question (here marked as V). "Phrasemes" include frequent collocations in which the noun is not a predicate noun and the meaning of the entire unit is idiomatic (e.g. *ta hand om X* ? lit. *to take hand about X - to take care of X*). Naturally, this sorting was strongly based on intuition. Basically, the phraseme and quasimodal

groups also allow for nouns which do not contain any predication (e.g. hand), while the "pure SVCs" are supposed to be denoting events and states. In this respect, we were not able to find a consistent solution for constructions like *begå en dummhet* (lit. *to commit a stupidity*), which underspecify the given event.

Extraction procedure yielded 10 235 SVC candidates out of which 9 442 were classified as negative examples, not collocations of interest. 689 collocations were classified as SVC, 27 were labeled as quasimodal, 77 were labeled as phrasemes. Careful manual classification of word pairs took about 3 days and was done by one person.

4. Automatic Collocation Extraction

The approach described in the previous section is very time-consuming and thus expensive. Query results in the first step of the extraction procedure contained only 8% of word expressions to be included in the lexicon. This implies that approximately 92% of the time in the second step was spent on elimination of useless material (assuming uniform data distribution).

We improve the first step of the procedure by applying methods for automatic collocation extraction. These methods employ lexical association measures to determine the degree of association between words in order to obtain a list of candidates further processed in the second and third step of the manual procedure. Items in this list are ranked according their association scores: the higher the score, the higher the chance of the candidate being a collocation.

4.1. Methods

A number of methods for automatic collocation extraction were proposed in the last decades. An overview of the most widely used ones is given e.g. in (Pearce, 2002) or in (Evert, 2004). They include Mutual information, Student's t-test, Pearson's χ^2 test, Log likelihood, and others. Pecina (2005) presents a comprehensive overview of known association measures applicable for collocation extraction together with empirical results showing that combining these measures (by logistic regression) leads to a significant performance improvement. His experiments were performed on (manually) morphologically and syntactically annotated Czech data from the Prague Dependency Treebank and his notion of collocations was much wider than ours. In our work we tried to duplicate his experiments with these two differences: a) our data came from the morphologically tagged PAROLE corpus and b) we focused only on verb--object collocations.

Our primary goal is to develop a method combining multiple association measures and employ it as an alternative to the first step of the extraction procedure. The secondary goal is to estimate quality of this procedure by *precision* (the fraction of collocation predictions correct) and *recall* (the fraction of collocations correctly predicted) curves, and compare this automatic approach with the simple manual procedure mentioned above. Both training of the automatic extraction procedure and its evaluation require some manually annotated data – word expressions assigned to two categories (collocations and non-collocations). For this purpose we utilized the results of the manual extraction.

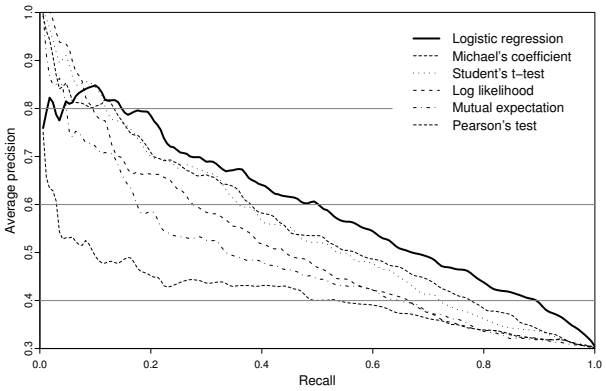

Figure 1: Performance of collocation extraction methods. Precision-recall curves obtained by logistic regression combining all association measures (thick line) compared with curves of selected individual association measures (thin lines). The closer to the top and right the better the method is.

4.2. Data and Experiments

From the PAROLE corpus containing 22 883 361 runing words in 2 639 283 sentences, we extracted 898 324 verb--object pairs appearing within a four-word collocation window sliding over all sentences in the corpus. 35 668 of these expressions occurring in the corpus more than five times were selected as collocation candidates in our experiments For all the selected 35 668 verb-object pairs we extracted their joint and marginal frequencies and distributions of content words occurring in their immediate and empirical contexts, and computed all 82 association scores, as described in (Pecina, 2005). a sample of 2 858 collocation candidates appearing also in the manually extracted data an was used for evaluation: 851 (29.77%) of them were collocations (of any kind), 2 007 (70.22%) were non-collocations. We followed the experiments described in (Pecina, 2005) and split this data into five stratified folds, obtained averaged precision-recall curves for individual association measures and a curve for logistic regression combining all 82 association measure in one model, and visualized the results in Figure 1.

4.3. Results

Based on our evaluation data, we estimated that the first step of the manual extraction procedure would operate with constant precision 29.77% within the entire interval of recall (assuming uniform distribution of the manually extracted data). The best performing automatic method is based on *Michael's coefficient* and achieved 70.83 % precision at 20% recall, 57.04% precision at 50% recall, and 37.98% precision at 80% recall. Similar results were obtained also by *Student's t test* (which surprisingly did not perform well in Pecina's experiments with Czech collocations). The combination of multiple association measures improved the performance even more: logistic regression on all 82 association measures achieved 80.95% precision at 20% recall, 62.5% precision at 50% recall, and 44.29% precision at 80% recall.

We applied this method on all 35 668 collocation candidates extracted from the PAROLE corpus and used the resulting ranked list as an input for the second step of the ex-

traction procedure. We estimated that in order to extract 20% of all collocations from the data, the amount of non-collocations eliminated from the manually processed data is only 19.04%. In order to extract 50% and 80% of collocations we have to eliminate 37.05% and 55.7% of useless material, respectively which is a substantial improvement over fully manual extraction.

5. Conclusion

We studied extraction semi-automatic procedures of Swedish collocations. Standard manual lexicographic approach was enhanced by statistical data preprocessing based on combination of multiple lexical association measures. Manually extracted data was used both for training parameters of applied statistical models and for estimation of possible efficiency increase of lexicographic work.

6. Acknowledgment

This work has been supported by the MSMT CR Project LC536, MSM0021620838, GAUK 489/2004, and GACR 405-06-0589. We would like to thank our colleagues from Sprøakbanken for providing us with the PAROLE corpus.

7. References

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A Statistical Approach to the Semantics of Verb-Particles. In Diana McCarthy Francis Bond, Anna Korhonen and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72.

Irène Baron and Michael Herslund. 1998. Support verb constructions as predicate formation. In H. Olbertz, K. Hengeveld, and J.S. García, editors, *The Structure of the Lexicon in Functional Grammar*. John Benjamins, Amsterdam/Philadelphia.

Silvie Cinková and Veronika Kolářová. 2004. Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. In *Korpusy a korpusová lingvistika v zahraničí a na Slovensku*.

Silvie Cinková and Jan Pomikálek. 2006. A Makedo Lemmatizer for the Swedish PAROLE Corpus. In progress.

Silvie Cinková and Zdeněk Žabokrtský. 2005a. Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns: Describing Event Structure in Support Verb Constructions. In G. Kiss F. Kiefer and J. Pajzs, editors, *Papers in Computational Lexicography: Complex 2005*, page 50?59. Hungarian Academy of Sciences.

Silvie Cinková and Zdeněk Žabokrtský. 2005b. Treating Support Verb constructions in a Lexicon: Swedish-Czech Combinatorial Valency Lexicon of Predicate Nouns. In Sabine Schulte Katrin Erk, Alissa Melinger, editor, *Interdsciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbrücken.

Ela Dura. 1997. *Substantiv och stödverb*, volume 18 of *Meddelanden från Institutionen för Svenska Språket*. Göteborgs universitet.

Lena Ekberg. 1987. *Gå till anfall och falla i sömn. En strukturell och funktionell beskrivning av abstrakta övergangsfraser*, volume A 43 of *Lundastudier i nordisk sprakvetenskap*. Lund University Press, Lund.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

Thierry Fontenelle. 1992. Co-occurence Knowledge, Support verbs and Machine Readable Dictionaries. In *Papers in Computational Lexicography, COMPLEX'92, Budapest*.

Gregory Grefenstette and Simone Teufel. 1995. A Corpus-based Method for Automatic Identification of Support Verbs for Nominalisations. In *Proceedings of the EACL*, Dublin, Ireland.

Ulrich Heid. 1998. Towards a corpus-based dictionary of German noun-verb Collocations. volume 1, pages 301–312, Université de Liège, Départements d'anglais et de néerlandai.

Paul J. Hopper and Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language*, 56(2).

Dekang Lin. 1999. Automatic Identification of Non-Compositional Phrases. In *Proc. of the 37th Annual Meeting of the ACL*, pages 317–324, College Park, USA.

Ann Lindvall. 1998. *Transitivity in Discourse. A Comparison of Greek, Polish and Swedish*. Ph.D. thesis, Faculty of Humanities, Lund University.

Sven-Göran Målmgren. 2002. *Begøa eller ta självmord? Om svenska kollokationer och deras förändringsbenägenhet 1800-2000*. Rapporter från ORDAT. Göteborgs universitet. Institutionen för svenska språket, Göteborg.

Zdenek Martinek and Leslie Siegrist. 1995. Winconcord. http://www.ifs.tu-darmstadt.de/sprachlit/wconcord.htm.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In Diana McCarthy Francis Bond, Anna Korhonen and Aline Villavicencio, editors, *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80.

Darren Pearce. 2002. A Comparative Evaluation of Collocation Extraction Techniques. In *Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Pavel Pecina. 2005. An Extensive Empirical Study of Collocation Extraction Methods. In *Proceedings of the ACL 2005 Student Research Workshop*, Ann Arbor, USA.

James Pustejovsky. 1991. The syntax of event structure. *Cognition*, 41:47–81.

Jan Schroten. 2002. Light verb Constructions in Bilingual Dictionaries. In *From Lexicology to Lexicography*, pages 83–94. University Utrecht. Utrecht Institute of Linguistics OTS., Utrecht.

Pasi Tapanainen, Jussi Piitulainen, and Timo Jarvinen. 1998. Idiomatic Object Usage and Support Verbs. In *COLING/ACL*, volume 2, pages 1289–1293, Montreal.

Vladimír Šmilauer. 1972. *Nauka o českém jazyku*. Praha.

Leo Wanner, editor. 1996. *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 of *Studies in Language Companion Series (SLCS)*, Amsterdam-Philadelphia. John Benjamins.