Silvie Cinková: ***Words that matter. Towards a Swedish-Czech Colligational Lexicon of Basic verbs***

This thesis by Silvie Cinková (SC) is intended to lay the foundations for a detailed, machine-readable Swedish-Czech lexicon (SveVALLEX/PNL), focused on basic verbs and their constructions. Although there are good monolingual construction dictionaries of Swedish that are also useful for Czech learners, several morpho-syntactic phenomena need an in-depth description that is not yet available in existing dictionaries. The main part of the thesis is devoted to discussions of different theoretical approaches to the information categories that the SweVALLEX/PNL is intended to provide. Moreover, some chapters are devoted to contrastive, corpus-based studies of Swedish and Czech basic verbs. Also, SC accounts in some detail for the creating of a lemmatized, searchable Swedish corpus.

The thesis can thus be regarded as a contribution to several linguistic areas, namely (a) to theoretical linguistics, (b) to contrastive linguistics, especially concerning Swedish and Czech (c) to computational linguistics, and (d) to some extent to lexicography. Although there might be a risk that the thesis gives a somewhat heterogeneous impression, I would rather like to stress that it bears witness to the broad competence of the author. In the following, I will comment a little on the four issues, (a) to (d).

Chapters 2 through 8 deal with issues of theoretical linguistics. The concept of grammaticalization is discussed in chapter 3, a discussion that in the end leads to a definition of the notion of light verb constructions (LVCs, mainly verbs with a 'weakened' sense + (prepositions+) nouns). The definition is maybe a little vague, but it is, admittedly, difficult to give a more precise one. Different aspects of valency are dealt with, especially in connection with LVCs (SC provides, e.g., an interesting discussion on the question whether certain complements belong to the verb or the noun). Since the lexicon is to provide information on the collocates of the (predicate) nouns of the LVCs, the theory of lexical functions is highly relevant; it is (competently) dealt with in chapter 8. Chapter 5 may be regarded as a small digression (on the so-called Transitivity hypothesis), but SC's discussion is interesting, relating to a recent investigation on Swedish, Polish and Greek, and, as SC shows, the hypothesis has some relevance for the thesis.

In chapters 11 to 14, partly contrastive, corpus-based investigations of the behaviour of basic Swedish verbs are carried out. The contrastive aspects are perhaps most interesting, but it should be said that some of them lead SC to the discovery of new data on Swedish in its own right. Especially interesting is maybe the fact that, contrary to Czech, locative adverbs in Swedish are

sometimes only possible in pseudo-coordinations, e.g. *vi kan inte stå här och diskutera den saken hela dagen* ('we cannot stand here and discuss that matter all day long'), while in the sentence without *stå* ('stand'), the adverb *här* ('here') would be questionable.

Chapter 15, and to some extent chapter 16, belong to the area of computational linguistics. In order to make possible the use of the internationally well-renowned Word Sketch Engine – a program designed for computer-aided lexicography purposes which is provide much of the 'raw material' for the information in the SweVallex/PNP lexicon – , a reasonably big lemmatized corpus of Swedish had to be developed. (Until now, only a small one was available.) A Swedish lemmatizer, the so-called LEMPAS, was thus constructed, being based on Swedish morphological rules (but not very much on any existing machine-readable dictionary). In chapter 15, the most important rules are accounted for, as well as the prestanda of LEMPAS. Thanks to the initiative of SC, the Word Sketch program thus became available for analysis of Swedish corpora (in this case, the so-called PAROLE corpus). It is quite remarkable that this work, which will facilitate investigations by Swedish corpus linguists, has been carried out in the Czech Republic and not in Sweden. (For instance, a Gothenburg colleague of mine, studying lexical innovations in the colloquial register, made use of the LEMPAS lemmatizer and Word Sketch analyses of an Internet corpus, created by himself.)

This said, a couple of minor remarks may be made in connection with chapter 15. The reader might get the impression that until now, no lemmatizer of Swedish was available. Actually, already in the early 90s an excellent lemmatizer of Swedish (in fact, better than the Czech one) was developed in Finnland (the so-called SWETWOL lemmatizer, see Karlsson 1992). As far as I know, the SWETWOL lemmatizer has been available ever since. Since it is commercial, however, it was probably never an option to use it in this case. But the SWETWOL lemmatizer might well have been mentioned in chapter 15.

Also, any lemmatizer will of course face the problem of homonymy. SC mentions this briefly on p. 181, giving a very simple rule for disambiguating nouns in connection with LEMPAS. She doesn't mention, however, homonymy where different parts of speech are involved, for instance verbs and nouns. Given the word form *cyklar* (*bicycles* or *ride(s) a bicycle*), how does the lemmatizer choose between the noun and verb interpretation? Probably by means of generalization of the rule given for nouns, but this is not completely clear. (Cf also p. 214, where SC discusses possible improvements of LEMPAS.)

However, this seems to be a relatively minor problem. There is no doubt that the 20-million words Swedish corpus lemmatized by means of LEMPAS (the

PAROLE corpus), is good enough to be processed by the Word sketch program. Intuitively, this is verified by the extracts on pp 208 and 209. And again, it should be underlined that this is the first time that the Word Sketch Engine has been used on Swedish for lexicographic purposes.

In chapter 16, the Swe-Vallex-PNL lexicon is presented. Although it is a Swedish-Czech one (with strong focus on Swedish, however, being rather a bilingualized than a bilingual dictionary), it is mainly intended for production of Swedish texts by advanced Czech learners. This is interesting, since bilingual dictionaries intended for text production are normally of the L1>L2 (in this case Czech-Swedish) type. But as SC rightly points out, L1>L2 dictionaries are often biased by the structures of the source language. They are thus not sufficient for advanced text production. For such purposes, monolingual or bilingualized (L2) dictionaries should be preferred.

As is clear from the tentative example articles (pp. 247–258) the dictionary will provide much information that is not to be found in conventional Swedish dictionaries. For instance, when does the noun of an LVC typically occur in indefinite form, with an adjectival modifier, and so on? The ambitions are obviously very high, and possibly it would be good to reduce them a little. In particular, in the valency frames, actants of the free modification type might sometimes be omitted (see, e.g., p. 2 in the PNL lexicon).

As far as (meta)lexicography is concerned, it should be mentioned that SC, in chapter 9, gives a useful survey of Czech, Swedish and English construction dictionaries. Possibly, SC over-estimates the BBI dictionary a little (it is indeed very far from being a combinatorial dictionary in the sense of Melchuk); in her survey, it might have been a good idea to replace it with the more modern, excellent Oxford Collocations Dictionary for Students of English.

It should be clear that the content of SCs thesis is very rich. The theoretical background of the dictionary is rather complicated but well motivated and accounted for; in fact, SC demonstrates good insights in quite a few linguistic theories. The work on the necessary lemmatized corpus, initiated by SC and performed by her together with computational linguists, is extremely useful. The dictionary prototype seems promising; with modifications, it will be useful to human users as well as computer programs.

It should finally be added that the thesis is well-written (as far as I can see, not being an L1-speaker of English). With two possible reservations, it is reasonably easy to read. The reservations concern the great amount of abbreviations (an appendix of abbreviations would have been welcome) and the fact that technical

terms are sometimes used long before they are defined (e.g. *FGD-based* on p. 17, *Oper2* on p. 28).

I strongly recommend that Silvie Cinková is granted permission to defend her thesis in order to achieve the doctoral degree.

Gothenburg, 14. July 2009

Sven-Göran Malmgren

Reference:

Karlsson 1992 = Fred Karlsson, SWETWOL: a comprehensive morphological parser for Swedish, in Nordic Journal of Linguistics 15, pp. 1–45.

Appendix: some minor errors (mostly detected by chance)

p. 31, note 10: *be hålla på* > be *hålla på*

p. 35, (11) and (16): *do to* > *to do*

p. 57, note 2 (and elsewhere). "Konkordanser" should be replaced by "Språkbanken", which is the official name.

p. 186; "The … sixth sentence" - but there are only five sentences

p. 207: it should be observed that "prep" also includes verbal particles (for instance, presumably in connection with *radio*, *kaffe* and *TV-program*)

p. 215 impossibly > not possibly

p. 216, (268): någon > något

p. 228, [18]: Lise > Ilse

p. 241, [153]: Karen > Karin

p. 4 in the SweVallex/PNL lexicon (verbal part) *sätta pilsnern i halsen* and *sätta fyr på bilen* do not belong to this group; *sätta i halsen* and *sätta fyr på ngt* are rather idioms

p. 2 in the PNL part: *slå rekord* can also mean 'sätta rekord' which is the case on line 7 (and thus, this a case of Oper1).