Automatic Detection of Lexicalised Phrases in Swedish

Janne Lindberg
Dept of Linguistics
Stockholm University
beb@ling.su.se

I will present a system under development, called LP-DETECT. The system detects and analyses Swedish lexicalised phrases (LPs) in order to enhance subsequent parsing. LPs are one of a number of stumbling blocks related to word sequences that must be dealt with when parsing unrestricted text. LPs include semantic idioms, syntactic idioms and morphological idioms and so called valency breaking LPs. The system reported on consists of an LP lexicon of some 8000 LPs with analyses, a detection program written in perl and rules for disambiguating between and discarding LP analyses. A small evaluation of the system is also presented.

1. Lexicalised phrases

LPs are expressions that belong to the lexicon and consist of more than one word. Included are **semantic idioms** the meaning of which is not built up compositionally from the individual meanings of the words in the idiom. (An English example is *kick the bucket* meaning 'die', a Swedish equivalent *ta ner skylten* (lit. take the sign down) also meaning 'die'.) They should be lexically listed because of semantic reasons. Another group consists of **syntactic idioms** containing "ungrammatical" or non-standard combinations of words, in syntactic terms *(inte så värst* (lit. not so worst) meaning 'not particularly'; English example: *by and large).* The syntactic idioms do not make up regular grammatical structures and must therefore be listed as wholes in a parsing system. Of course, since syntactic idioms are syntactically irregular, no general function of semantic interpretation can apply over them and therefore a syntactic idiom is also a semantic one.

A third group consists of phrases containing words that are unique to the phrases where they occur, often lexical relics, for instance the Swedish LP *med nöd och näppe* meaning 'with difficulty'. The word *näppe* is not used outside this phrase and should therefore not be individually listed in a word lexicon. Related to the latter group are also phrases containing foreign words *(anno domini)* or nonce forms *(hux flux).* These could be called **morphological idioms** and they of course qualify as semantic idioms too.

In addition there are a large number of LPs, partly overlapping with the LP types described above, which I call **valency breaking LPs**. They simplify the phrase structure of sentences when detected as LPs. They tend to end in function words. An example of these is multi-word prepositions, i.e. *in spite of* (Swedish example: *på grund av*, lit. on ground of, 'because of'). "Mechanically", the PP *in spite of the weather* consists of a P (in) and a complex NP (*spite of the weather*, itself containing a PP). *In spite of* here functions as a complex preposition, actually replaceable with the single preposition *despite*, making the syntactic structure of the PP simpler and more one-to-one with its semantics. Other examples of such LPs are phrases, often described using some kind of subcategorisation, are so called prepositional verbs *(tro på*, eng: believe in), V+NP+P

Proceedings of NODALIDA 1999, pages 103-114

constructions seen as "nouns with supportive verbs" by Dura (1997) *(få grepp om*, lit. get grip about; 'understand') and complex auxiliary verbs *(kommer att*, 'will')[1].

My intuitive opinions as to what a lexicalised phrase is was initially shaped largely by an article by Anward & Linell (1976). Those opinions have since been somewhat altered but their article is still very important for my thinking on the subject of lexicalised phrases.

1.1 Criteria

I have two obligatory criteria that have to hold for every potential candidate as an LP. The first (1:a) is that the candidate should be a standard way of expression in Swedish; it should be an institutionalised way of saying something. The second (1:b) is the simple formal criterion that they must consist of more than one word but less than a clause.

In addition to these, either of the following two characteristics has to hold: (2:a) The candidate has to be semantically non-compositional OR (2:b) a "better" phrase-structure analysis is obtained when grouping the word combination in question as an LP.

Non-compositionality can either arise from the candidate being partially non-compositional, that is, one or more but not all words are used in a non-standard way, e.g. *fatta eld* (eng. 'catch fire'), where the word *fatta* is used in a marked way[2], or the candidate as a whole is used to mean something quite different than what is expressed with a literal interpretation[3] *(sticka under stol med,* 'hide').

Below are some examples of the reasoning behind criterion (2:b).

(1) [på [grund [av [faran]]]] lit: on ground of danger-DEF
(2) [på grund av [faran]]

A "mechanical" division of the above PP produces constituents whose semantic status is clearly debatable. Both *grund av faran* (ground of danger-DEF) and *av faran* (of danger-DEF) in (1) are uninterpretable without context[4]. I regard *på grund av* ('because of') as a complex preposition taking an NP, here *faran.* (See (2).)

(3) [Jag [tror [på [tomten]]]] lit: I believe on Santa-Claus-DEF
(4) [Jag [tror på [tomten]]]

In the VP *tror på tomten* (believes in Santa-Claus), the predicate meaning is described by the string *tror på* (believes in) and not by the verb (believes) alone. The PP *på tomten* in (3) is not meaningful without the verb. I see *tror på* as a complex transitive verb and the NP *tomten* here is the direct object. (See (4).)

(5) [Det [kommer [att regna]]] lit. It comes INF rain
(6) [Det [kommer att [regna]]]

Examples (5)-(6) treat the Swedish future tense expression. *Kommer att* is the string that expresses future tense in Swedish. *Kommer* alone does not, at least not in the written language. I see *kommer att* as a complex auxiliary verb.

Specific criteria are formulated for different structural types of lexicalised phrases where certain tests can be applied to candidates. Below are some examples:

Particle verbs

The general criterion saying that an LP should be semantically non-compositional should be applied to particle verbs in the following way:

Particle verbs that qualify as LPs should satisfy the following description: 1. The particle does not modify the verb meaning spatially in a way that can be predicted from the spatial meaning of the particle. *Slänga in* ('hurl in') could be an example of a non-qualifying candidate. However *lägga in* lit. 'lay in' qualifies because one of its meaning is non-compositional 'pickle'. 2. The particle verb is not an instance of a productive metaphor. Often spatial expressions can be used metaphorically as for instance: *slunga ut* 'hurl out' and *sväva ut* 'float away'. These do not qualify as LPs, because they are instances of productive spatial metaphors.

Reflexive verbs

The reflexive in a reflexive verb should not have argument status. The transitive reflexive verb *tvinga sig* 'force oneself' is thus not considered an LP.

Prepositional verbs

The noun phrase that follows the preposition should have argument status. It should also be replaceable with other noun phrases. If the whole of the verb+prep plus a following noun phrase forms an idiomatic phrase and the verb+prep combination can not be combined with other noun phrases, it is not a prepositional verb. The preposition plus the noun phrase should also not form an adverbial.

2. The use of "idioms" in tagging and parsing

The English Constraint Grammar, ENGCG (Karlsson et al 1995), uses a list of 700 syntactic "idioms" and 5000 complex nominals that are deterministically grouped and analysed as complex words at tokenisation. The two examples below are taken from Karlsson et al (1995).

in spite of --> in=spite=of/PP
run time --> run=time/NN NOM SG

The CLAWS 4 tagging system (Leech, Garside & Bryant 1994) used for annotating the British National Corpus (BNC) with parts-of-speech has a sub-component called IDIOMTAG, which makes use of multi-word units to correct prior erroneous taggings. The lexicon in IDIOMTAG has over 3000 entries representing "general idioms" (e.g. *as much as*), multi-word proper names (e.g. *Dodge City*) and foreign expressions (e.g. *annus horribilis*). IDIOMTAG introduces ambiguities where more than one "idiom" analysis is possible and later resolve these ambiguities. When Blackwell (1987) reported on an earlier version of CLAWS, IDIOMTAG made the idiom analyses deterministically, i.e. the

analyses could not be un-done at a later stage. From what I can make out of the documentation in Leech, Garside & Bryant (1994), the determinism is now abandoned, but I am not sure of that. The system can express several traits of variation and discontinuity of the "idioms".

Harald Berthelsen (Berthelsen 1997) used an earlier version of the LP lexicon that I will report on below and wrote a prolog program detecting LPs in texts. It was used to augment a Constraint Grammar (SWECG) analysis with LP membership on words. The analysed words had indexes pointing to the lexicon. Constraints could be written discarding or selecting LP analyses.

Other related implementations

Stephan Bopps (Bopp 1996) Phrase Manager is a framework for databases of multi-word units. The program makes it possible to specify e.g. transformation restrictions on idioms. Bopp does not report on an actual lexicon though. It is the database format itself that constitutes Phrase Manager.

André Schenk (Schenk 1994) analysed idioms and collocations and incorporated them into compositional M-grammar used for automatic translation in the Rosetta machine translation system. Schenk describes a formal apparatus for dealing with idioms and collocations but does not report on an actual lexicon either.

3. LP-DETECT

My system for detection of LPs (LP-DETECT) takes text files from the SUC Corpus (Ejerhed et al 1992) as input, converted and simplified so that only <word>:<tag> pairs remain. Optionally, sentence enumeration can also be displayed in the output. <word> are word forms except for verbs where <word> are base forms. <tag> are the POS tags from SUC. All other information such as the morphological features are deleted[5]. The output of the LP detection is also <word>:<tag> pairs but more information is added. 1. The words in the LPs are given an additional tag for LP membership. 2. The LPs are given an analysis. 3. The LPs are enclosed within square brackets.

Input: ... word:tag word:tag , ... , word:tag word:tag ...
Output: ... word:tag [word:tag_LPtag , ... , word:tag_LPtag LPanalysis] word:tag ...

LP-DETECT consists of three components, an LP lexicon, a detection program and Dis rules (**Dis**ambiguation and **Dis**carding).

3.1 The LP lexicon

The LP lexicon comprises 8057 LPs distributed over 4775 LP Sets. LP sets reflect the fact that many LPs have variants[6] and those variants are grouped together in the lexicon and are given the same main enumeration index. Examples of LP Sets are *vara/ligga/stå i maskopi med* ('be/lie/stand in collusion with') and *så in i baljan/norden/helvete* ('as hell'). The lexicon contains morpho-syntactic and some phonetic/prosodic information about the LPs. The phonetic/prosodic information is left from an earlier phase where LPs were collected to enhance text-to-speech synthesis (Lindberg 1996).

My main source for the LP lexicon has been Johannisson & Ljunggrens "Handordbok" (1966). To overcome problems connected with the age of "Handordbok", I have also used "Wörterbuch Der Schwedischen Phraseologie In Sachgruppen" (Schottman & Petersson 1989). I have also taken "slang" expressions from Haldo Gibson's "Svensk slangordbok" (1969). Other written sources have been "Svenska Akademiens ordlista över svenska språket" (1979) and Svenskt uttalslexikon (Hedelin et al 1989). Phrases following certain POS patterns or containing certain words have been excerpted from the SUC corpus (Ejerhed et al 1992) and the PAROLE corpus[7]. Collocations have also been excerpted from the SUC corpus and the Dragon corpus of Dept of Speech, Music and Hearing at KTH, Stockholm. I have also found a great deal of the phrases by leading an ordinary life; listening to people talk, watching TV and reading books, newspapers etcetera, always with a pencil and a paper within reach (if that counts as an ordinary life).

Table 1. The fields in LP entries with two examples.

Field name	**Example 1**	**Example 2**
Index number	*2076:2*	*4030*
Index word	*finna*	*hållet*
LP tag	*VT?VC?VI*	*AB*
Reg exp	*finna:VB_<RP>*	*helt:AB_och:KN_hållet:..*
Word info	....	3:obs!
Prosodic markers	1:acc;2:ob	1:acc;2:ob;3:acc
S-form indication	0	....
Imperative form indication	1	....
Phrase info	....	....
Internal POS structure	1:VB;2:PN	1:AB;2:KN;3:Y

The LP entries consist of ten fields and are examplified in Table 1 above. *Index number* is an enumeration of the LP sets. The second figure is an enumeration of entries within LP sets. *Index word* is an approximation of the least frequent word of the phrase, using the longest word of the phrase. The *LP tag* field contains a syntactic analysis for the LP as a whole. The tags for the *LP tag* field use the SUC part-of-speech analyses except for the verbal LPs where subcategorisation information is encoded. The LP in example 1 is given three possible analyses; as a mono-transitive, copulative or intransitive verbal LP. Example 2 is analysed as an adverbial LP. In the *Reg exp* field resides a regular expression used to find the LP in the input sentences. The regular expressions allow for alternatives both for words and tags as well as variables that are expanded in the main detection program described below. *Word info* is for various information such as the note *obs!* indicating that the word could have more than one pronunciation and only one is correct here. It can also contain information on tense agreement in LPs with more than one verb. The prosodic marker field contains information stating the extent to which the words of the LP retain their word accent when pronounced together. The next two fields bears information on allowed inflexional forms of the verb in verbal LPs. In the *Phrase info* field there may be additional morpho-syntactic information such as morphological features restricting the context of the LP. The last field is an enumeration of the parts-of-

speech in the phrase, following the SUC tagging conventions as strictly as possible. However, some words in LPs are notoriously difficult to annotate, since many LPs are grammatically deviant. Therefore two additional tags are used. One tag, X, is used for words that are only found in LPs, often archaic forms (e.g. *i **sinom tid,*** 'in due time'). The other tag, Y, is used for words that are used in a non-standard way in the LP (e.g. ***till sist,*** 'at last'). Also, the word tag NN (noun) has the morphological tags i=indefinite, d=definite, s=singular and p=plural as an extension of the SUC POS tag. In the current implementation, only the first four fields are used.

In Table 2 below, the phrase tags with subcategorisation codes for the verbal LPs are shown.

Table 2. Phrase tags with subcategorisation codes for the verbal LPs.

VI	Does not take a complement	VS	Takes a subordinate clause complement
VT	Takes an object (NP or S)	VTnexus	Takes an object (NP) LP-internally (in the nexus)
V2T	Takes two objects (NP or S)	VTnexusA	Takes an object (NP or S) LP-internally (in the nexus) and a post-posed *att*-clause as complements
Vadv	Takes an adverbial complement (of any form)	VTnexusS	Takes an object (NP or S) LP-internally (in the nexus) and a post-posed subordinate clause as complements
VH	Takes an infinitive VP	VTnexusH	Takes an object (NP or S) LP internally and a post-posed infinitive VP LP as complements
VC	Takes a predicative complement	V2Tnexus	Takes an object (NP, indirect object) LP-internally (in the nexus) and a post-posed direct object.
VA	Takes an *att*-clause (that-clause)	VHpass	Takes an infinitive VP with the verb in passive form

For certain verbal LPs, alternatives can be given:

VI?VC?Vadv *hålla_<RP>* Eng. lit: hold_REFL

Han var nödig men han höll sig (VI) 'He needed to go to the toilet but he restrained himself'
Hon höll sig underrättad (VC) 'She kept herself informed'
Han höll sig på kontoret hela dagen (Vadv) 'He stayed at the office all day'

The tags for the non-verbal LP analyses *(LP type)* are taken from the SUC part-of-speech tag set (where they of course are used to analyse words and not phrases). The one exception to that rule is the tag PD (multi-word pre-determiner). The LP analyses in question are shown in Table 3.

Table 3. Phrase tags for the non-verbal LPs with examples

AB	adverbial LP	*så länge*
DT	multi-word determiner	*en och annan*
NN	multi-word nominal	*mannen på gatan*
JJ	adjectival LP	*liten i maten*
PP	multi-word preposition	*tack vare*
SN	multi-word subordinating conjunction	*så länge*
KN	multi-word co-ordinating conjunction	*för att inte tala om*
IE	multi-word infinitive marker	*i akt och mening att*
PN	multi-word pronoun	*en och annan*
PD	pre-determiner	*en del av*
HP	multi-word interrogatory pronoun	*vem i all värden*
HA	multi-word interrogatory adverb	*hur i all världen*

3.2 The detection program

In the main detection program, written in perl, the regular expressions for the LPs are enhanced to allow for intervening material between certain words and specification of variables in LPs. Lookup speed is increased by only considering those LPs whose longest word is present in the sentence to be analysed (length of word approximating word frequency in an inverse relation). This is much more efficient than the usual method of using the first word as a trigger. When two or more words are of the same length, the last one is chosen for most structural types of LPs. The detection also allows for embedded LP analyses where one LP is embedded within another LP.

3.3 The Dis rules

The next step is the Dis rules. "Dis" stands for disambiguation and discarding of LPs. *Disambiguation*: The disambiguation rules pick one of several analyses for a given LP or discards an erroneous one, possibly leaving more than one analysis. *Discarding*: All LP analyses for a word sequence that is not an LP but happens to contain the words of an LP in the right order (juxtaposition) are deleted. Usually, word forms or tags in the immediate context are used for the Dis rules. 36 disambiguation rules and 75 discarding rules are implemented presently.

The algorithm for the detection of LPs with the Dis rules is shown below.

Algorithm:
Reads in and stores the LP lexicon
Regular expressions for LPs are augmented
For each SUC sentence:
    Converts the SUC file format
    Stores relevant lexicon entries
    For each stored lexicon entry matching the sentence:
        Marks matching words and gives the LP an analysis
    Runs dis rules if that is chosen and if at least one LP is found

4. Output

As indicated above, not all possible analyses residing in the lexicon are present in the output, partially because of the simplified input format. The examples (7)-(9) contain lexicon entries, output and an english traslation. Examples (8)-(9) also contain dis rules. The example entries only contain the information used in the implementation. In example (7), two LPs have been detected, the adverbial *i alla fall* ('at least') and a verbal LP taking a complement adverbial *ha det* (lit. have it).

(7)
*4236 fall AB **i**:PP_**alla**:DT_**fall**:NN*
*295:1 det Vadv (**få**:VB|**ha**:VB)_**det**:PN*

Men:KN ösatyskarna:NN kan:VB nu:AB alltså:AB skryta:VB med:PP att:SN de:PN [**i**:PP_LP-4236 **alla**:DT_LP-4236 **fall**:NN_LP-4236 @AB-4236] [**har**:VB_LP-295:1 **det**:PN_LP-295:1 @Vadv-295:1] bättre:AB i:PP sängen:NN än:KN västtyskarna:NN .:MAD

Eng translation: But the east Germans can brag about the fact that they at least are better off in bed than the west Germans.

In example (8), a discarding rule has worked, deleting a particle verb reading (here VI, 'takes no complements') of the string *gå bort* (the idiomatic reading meaning 'die', here retaining the literal meaning 'walk away') using prepositions in the immediate right context.

(8)
Dis rule:
gå_PL^{8} *--> NIL / _ (mot|till|från)*

Gumman:NN [reste:VB_LP-2174 sig:PN_LP-2174 @VI-2174] och:KN [**gick**:VB_LP-2693 långsamt:AB **bort**:AB_LP-2693 @VI-2693] mot:PP sitt:PS bylte:NN .:MAD
>>>>>
Gumman:NN [reste:VB_LP-2174 sig:PN_LP-2174 @VI-2174] och:KN gick:VB långsamt:AB bort:AB mot:PP sitt:PS bylte:NN .:MAD

Eng translation: The old woman rose and walked away slowly towards her bundle.

In example (9), a disambiguation rule has chosen the analysis V2T (takes two object complements) using the right context (the LP *böna och be* means 'beg and plead'). Another disambiguation rule has chosen the analysis AB (adverbial LP) over SN (complex subordinating conjunction) for the LP *varje gång* 'each time', also using the right context.

54 böna V2T?VA?VI ***böna:VB_och:KN_be:VB***
*4771 gång AB?SN (****varje****:DT|****varenda****:DT)_****gång****:NN*

V2T?(VT?VA)?VI *--> V2T / __ (NN|PN) NPbörjan*
AB?SN *--> **AB** / __ (KN|PP|MID|MAD|VB)*

Hon:PN hade:VB [**bönat**:VB_LP-54 **och**:KN_LP-54 **bett**:VB_LP-54 @V2T?VA?VI-54] honom:PN att:IE vara:VB tystare:JJ ,:MID för:KN tänk:VB om:SN flickan:NN vaknade:VB och:KN [**varje**:DT_LP-4771 **gång**:NN_LP-4771 @AB?SN-4771] lovade:VB han:PN .:MAD
>>>>>
Hon:PN hade:VB [**bönat**:VB_LP-54 **och**:KN_LP-54 **bett**:VB_LP-54 @V2T-54] honom:PN att:IE vara:VB tystare:JJ ,:MID för:KN tänk:VB om:SN flickan:NN vaknade:VB och:KN [**varje**:DT_LP-4771 **gång**:NN_LP-4771 @AB-4771] lovade:VB han:PN .:MAD

Eng translation: She had begged him to be more quiet, because what if the girl would wake up, and each time he promised.

5. Small evaluation

I trained the system on 50.000 words of Swedish text from the SUC corpus. The genres represented newspaper text, legal text and novels. As a simple base-line setting test I tested the precision on one of the texts from novels from the training material. 144 sentences were scanned, 90 LPs were found. After the Dis rules 80 LPs remained. Below are the results. Figures in the evaluation are rounded because of the small amounts of text used.

90/144 = 0.6 LPs were detected per sentence as a mean
86/90 = 96 % of LPs were unambiguous initially, 100 % finally
76/90 = 84 % of LPs were correctly detected initially
75/80 = 94 % of LPs remaining after the dis rules were correctly detected

The precision figure rose from 84 % to 94 % using the dis rules. All ambiguities were eliminated.

To test recall, a text from a novel from the training material was manually scanned. It contained 144 sentences. 76 correct LPs had already been found by the detection program, nine more were found in the manual scan. Thus, the total number of LPs in the text was 85.

Recall: 76/85 = 89 %

Five of the misses were due to the LPs not being in the lexicon.

Lexicon recall: 80/85 = 94 %.

Since the lexicon lacked five LPs found in the manual scan, only 81 LPs would have been possible to find with the present lexicon, even with a perfect detection algorithm. 76 LPs were found. Therefore, the recall for the detection was 94 %.

Detection recall: 76/81 = 94 %

As for the shortcomings of the detection program, one flaw was due to the way the regular expressions work, scanning the texts from left to right. When a situation of LP overlap occurs, only the leftmost LP is detected. The other flaw was due to the very simple formulation on the allowed tags for intervening material in LPs as an unordered set of tags[9]. That forces prohibition of some tags, e.g. prepositions, conjunctions and verbs in order to avoid overgeneration. In fact, all tags can appear breaking up a verbal LP but not in any order. Other sources of misses were from faulty tagging in SUC.

I did a somewhat larger test in texts not from the training material. The proportions were 40 % newspaper text, 40 % text from novels and 20 % political text. 782 sentences were used. 415 LPs were found initially. After the Dis rules, 380 LPs remained.

415/782 = 0.53 LPs were detected per sentence as a mean
386/415 = 93 % unambiguous initially
369/380 = 97 % unambiguous finally
340/415 = 82 % correctly detected initially
339/380 = 89 % correctly detected finally

Here, the precision figure rose from 82 % to 89 % using the dis rules. That means that about a third of the initial false positives were eliminated by the dis rules. The proportion of unambiguous analyses rose from 93 % to 97 %. Thus more than half of the ambiguities were eliminated by the dis rules.

6. Discussion

My mini-tests have several flaws but I have chosen to present it here anyway just to get a rough idea of the magnitude of the problem of over- and undergeneration of the system. One flaw is the mere sparseness of the testing material. The different files varied considerably as to the number of false positives both before and after the discarding rules. Perhaps not surprisingly, there was a clear tendency for texts from novels to produce more false positives than those from the other two genres. The amount of remaining ambiguity in LP analyses was smaller for these texts though. There was also a tendency for newspaper texts to have more LPs in them but fewer different LPs, the use of LPs were more stereotyped in the newspaper texts.

The precision result on the small portion of the training material that was investigated shows that I have not been able to eliminate the overgeneration of LP analyses. Those cases that I could not write discarding rules for were either "parsning-complete" (i.e. a

parser not only for constituents but for grammatical functions would have been needed) or even "AI-complete" (i.e. only full text understanding would have sufficed to eliminate these false positives).

There were clear tendencies for the false positives. These instances were very often two-word verbal LPs ending in either a preposition (false prepositional verb), a particle (false particle verb) or a reflexive pronoun (false reflexive verb), where both the verb and the second word were high frequency words. Examples of false LPs are the prepositional verbs *vara till* 'have the function of', *bli till* 'become' and *gå för*[10]. The latter has later been reconsidered as not being a qualifying candidate as a prepositional verb by the author.

As for the recall figures, the method I used was not optimal. The author (also the compiler of the lexicon) scanned the text for LPs not found by the program and marked candidates that satisfied my criteria. A better way would have been to have another person look through the text and mark every LP that person could find using my criteria and after that run the program on the same material and compare the results.

Clearly, LP-DETECT deserves a better evaluation.

To sum up, the fact that about half of the sentences contained an LP as an average suggests that LPs are common enough to spend time on detecting in the first place. However, in addition to helping out a system of syntactic and/or semantic parsing, such a system could also be hindered by false positives, some of which are very difficult to avoid.

---

[1] Of course even idioms can be said to be valency breaking since syntactic arguments in idioms are often not part of the argument structure of the sentences where the idioms occur.

[2] Such phrases are often termed collocations (see e.g. Schenk 1994).

[3] A description most often attributed to idioms.

[4] Note that the Swedish genitive is not productively signalled by a PP with *av* (of) as in English. "The cause of the danger" is expressed *farans grund* (danger-DEF-GEN ground) in Swdish and not *grunden av faran* (ground-DEF of danger-DEF). The preposition *till* (to) can be used instead, however.

[5] Actually, the surface forms of the verbs are stored and retrieved later to be present in the output.

[6] See Sköldberg (1999) for an interesting discussion on types of idiom variation.

[7] The PAROLE corpus is part of "Språkbanken" at Dept of Swedish, Göteborgs University.

[8] **"gå_PL"** means "an LP consisting of some form of the verb *go* followed by a particle (PL)". Actually this is just a way of representing the Dis rules in a somewhat more readable form. In real life, the Dis rules are formulated as regular expressions.

[9] That is, <word>:<tag> pairs. The actual form of the regular expression for the intervening material is "([^:]+:(AB|PN|NN|DT|JJ|PC|PM|PS|RG|RO))*?".

[10] In the expression *det går för PERSON* ('PERSON is having a sexual climax') or in the expression *visa (PERSON1) vad PERSON2 går för* ('show PERSON1 what PERSON2 can do').

References

Anward, Jan & Linell, Per. 1976. Om lexikaliserade fraser i svenskan. *Nysvenska studier* 55-56, 77-119.

Berthelsen, Harald. 1997. The Constraint Grammar Idea applied to two Problems in Text-to-Speech: Detecting Multi-word Units and Prosodic Boundaries. D-level essay. Stockholm: Computational Linguistics, Stockholm University.

Blackwell, Susan. 1987. Syntax versus Orthography: Problems in the Automatic Parsing of Idioms (Ch 9). In: *The Computational Analysis of English. A Corpus-Based Approach.* (Eds. Roger Garside, Geoffrey Leech & Geoffrey Sampson) London & New York: Longman, 110-119.

Bopp, Stephan. 1996. *Phrase Manager: a System for the Construction and the Use of Multi-word Unit Databases.* EURALEX'96 Proceedings, 55-64.

Dura, Ela. 1997. *Substantiv med stödverb.* Meddelanden från Institutionen för Svenska Språket (MISS) 18. Göteborg: Göteborgs universitet..

Ejerhed, Eva; Källgren, Gunnel; Wennstedt, Ola & Åström, Magnus. 1992. *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project, Version 4.31.* Umeå: Publications from the Department of General Linguistics, University of Umeå, no 32.

Gibson, Haldo. 1969. *Svensk slangordbok.* Stockholm: Bonniers.

Hedelin, Per; Jonsson, Anders & Lindblad, Per. 1989. *Svenskt uttalslexikon del 1-2.* Teknisk rapport nr 4. Inst för informationsteori. Göteborg: Chalmers University of Technology.

Johannisson, Ture & Ljunggren, K.G. 1966. *Svensk Handordbok. Konstruktioner och fraseologi.* Nacka: Svenska språknämnden och Esselte studium AB.

Karlsson, Fred; Voutilainen, Atro; Heikkilä, Juha och Anttila, Arto. (eds.) *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text.* Berlin - New York: Mouton de Gruyter.

Leech, Geoffrey; Garside, Roger & Bryant, Michael. 1994. *CLAWS4: The Tagging of the British National Corpus.* Kyoto: Proceedings of COLING Kyoto.

Lindberg, Janne. Detektering av lexikaliserade fraser för text-till-talkonvertering. 1996. *The Nordic Languages and Modern Linguistics.* Proceedings from the Ninth International Conference of Nordic and General Linguistics, (eds: Kjartan G. Ottósson, Ruth V. Fjeld and Arne Torp). Oslo: Novus, 191-203.

Målande uttryck. 1990. *En liten bok med svenska idiom.* Uppsala: Esselte ordbok, 1990.

Schenk, André Y. 1994. Idioms and Collocations in Compositional Grammar. Utrecht: OTS Dissertation Series.

Schottmann, Hans & Petersson, Ricke. 1989 *Wörterbuch Der Schwedischen Phraseologie In Sachgruppen.* Münstersche Beiträge Zur Deuchen Und Nordischen Philologie 6 1. Aufl. Münster: Kleinheinrich Verlag Für Kunst, Literatur Und Wissenschaft.

Sköldberg, Emma. 1999. *Varianter av idiom.* Svenskans beskrivning 23. (Eds. Lars-Gunnar Andersson, Aina Lundqvist, Kertin Norén & Lena Rogström). Lund: Lund University Press, 384-392.

Svenska Akademiens ordlista över svenska språket. 1979. Stockholm: P.A. Nordstedt & söners förlag.