

Fred Karlsson

SWETWOL:

A COMPREHENSIVE MORPHOLOGICAL ANALYZER FOR SWEDISH

(Originally published in *Nordic Journal of Linguistics*, 15, 1992, 1-45. Unchanged text launched on the internet, 20 June 2007.)

Abstract

SWETWOL is implemented in the framework of Koskenniemi's (1983) **two-level model**. It contains a 48,000 item lexicon and a full inflectional description. Special attention was paid to the design of a computational analysis of productive Swedish compounds. Recall (coverage) and precision of SWETWOL meet high standards. SWETWOL has been extensively tested on various types of texts.

1. Introduction

SWETWOL is a computer program capable of morphologically analyzing Swedish text. It is implemented in the framework of Koskenniemi's (1983) **two-level model**, TWOL, which has become a near-standard of computational morphology.

SWETWOL incorporates a full description of Swedish inflectional morphology. It contains a fairly large lexicon, presently (January 1992) comprising some 40,000 basic core vocabulary items, excluding productive derivatives and compounds, and an additional 4,400 abbreviations and 3,500 proper names. The description spots all and only the proper forms of the lexical items. Special attention was paid to the design of a computational analysis of productive Swedish compounds. Recall (coverage, retrieval of correct analyses) is close to perfection when SWETWOL is applied to "ordinary" text. New methods (i.a. local disambiguation) are proposed for optimizing precision, especially by imposing a strict regime on the potential overgeneration of the compound analysis mechanism.

Here is the SWETWOL analysis of the word sequence en svensk tiger which is ambiguous between the readings 'a Swedish tiger' and 'a Swede keeps silent'. The word-forms occur at the far left and the readings are indented:

("<<en>>")

 ("en" ADV)

 ("en" NUM UTR INDEF SG NOM)

 ("en" N UTR INDEF SG NOM)

 ("en" PRON UTR INDEF SG NOM)

 ("en" ART UTR INDEF SG NOM))

("<<svensk>>")

 ("svensk" N UTR INDEF SG NOM)

(“svensk” A UTR INDEF SG NOM))

(“<<tiger>>”

(“tiger” N UTR INDEF SG NOM)

(“tiga” V ACT PRES))

All readings are retrieved that are feasible in relation to the lexicon and inflectional description. SWETWOL is capable of generating the base form of inflected forms, most clearly visible in the second reading of the word tiger.

SWETWOL is meant to be used as the basic morphological tool in systems for text analysis, information indexing, storage and retrieval, and machine translation. It is an integral part of a parsing system where the other major components are (i) a preprocessor converting text to the format presupposed by SWETWOL, and (ii) the **Constraint Grammar Parser** CGP (Karlsson 1989, 1990, 1992a,b, Karlsson, Voutilainen, Heikkilä, and Anttila 1991). CGP provides a language-independent formalism and a computer program for morphological disambiguation and surface-syntactic analysis. The sequential set-up of analysis in the present framework is thus:

- preprocessing,
- morphological analysis by SWETWOL,
- local disambiguation (on top of TWOL, or by CGP),
- context-sensitive disambiguation, clause-boundary determination, and surface-syntactic analysis by CGP.

Local disambiguation and preprocessing are discussed in sections 12, 13. Context-sensitive disambiguation of Swedish is treated in Karlsson (in preparation).

2. Previous computational approaches to Swedish morphology

The pioneering effort of Swedish computational lexicology and morphology was Allén’s Nusvensk frekvensordbok (NFO). Many aspects of computational morphology faced this project, such as text preprocessing, morpheme identity and segmentation, the distinction between homonymy and synonymy, homograph separation, and text type effects on vocabulary composition (Allén 1970).

One offspring of NFO was Hellberg’s (1978) algorithm for Swedish morphology, a precise description of word inflection and derivation, partly also of compounding. Hellberg set up a system of 235 inflectional paradigms, plus a basic vocabulary containing 8,609 lemmas.

This vocabulary had a coverage of 90% in the corpus of newspaper text that NFO was based on (Hellberg 1978:12). The algorithm was implemented by M. Eeg-Olofsson. Precise data on coverage in the realm of new texts are not included in Hellberg (1978). It is somewhat unclear how precise the morphotactic description of compounds was.

Hellberg’s system was later reimplemented by Rankin (1986). This dictionary contained some 7,200 entries (Rankin, ms., p. 18). Rankin (1986:166-) reports considerable overgeneration in the analysis of compounds. E.g., bil ‘car’ yielded spurious analyses such as “bi-l” (N - abbreviation), and ironiska ‘ironic (pl.)’ yielded a spurious interpretation “ironi-ska” (N - V). (In this paper, compound boundaries are indicated by a dash, “-“, and other morpheme boundaries by a plus sign, “+“.) The morphotactic description of compound formation was obviously not strict enough.

Sågvall Hein (1988, 1990) initiated the project “A Lexicon-Oriented Parser for Swedish” (LPS). The aim was to create a broad-coverage Swedish parser based on the machine-readable dictionaries compiled by Språkdata in Gothenburg, especially NFO and Svensk ordbok (SOB, 1986). The parser contains modules for morphological analysis and syntactico-semantic analysis, the latter based on the Uppsala Chart Processor (Sågvall Hein 1983, 1987, Sågvall Hein & Ahrenberg 1985).

The morphological analyzer of LPS incorporates a stem dictionary covering the lemmas of SOB (N=58,536) and an inflectional grammar. The analyzer rejects productive derivatives and compounds if they are not listed in SOB. Sågvall Hein (1990:345) reports that, in a test corpus of 2,500 word-form tokens, 452 compounds and 80 derivatives failed to get an analysis.

Brodda (1983) has developed a preliminary rule set for heuristically based analysis of inflectional morphology. Källgren (1984a,b, 1990a,b) continued working on heuristic parsing along similar lines. Her MorP (morphology-based parsing) system assigns part of speech labels on the basis of surface pattern matching. MorP contains a small lexicon of form words plus a set of inference rules. Källgren (1990b) reports a 91% accuracy level in the analysis of unrestricted Swedish text.

Ejerhed and Bromley (1986) designed a program called MORPH for morphological analysis of Swedish. It was specially geared for testing certain psycholinguistic hypotheses. Eeg-Olofsson (1988) has designed a bidirectional model for Swedish morphology in the framework of machine translation.

Blåberg (1984) was the first linguist to describe Swedish morphology in the framework of Koskeniemi’s two-level model. Blåberg’s description is fairly complete in regard to inflectional morphology. The lexicon contains some 4,000 entries. The morphophonological alternations of Swedish are described by 35 two-level rules. The description is abstract in the sense that 13 morphophonemes are in use. Compound formation and derivation are not included in this system. Borin (1986) offers a number of critical comments concerning Blåberg’s description of verbs and numerals, and, on a more general level, concerning the aptness of the two-level formalism for morpholexical description.

3. Central aims and properties of SWETWOL

Here are some important aims and properties of SWETWOL:

- a) All aspects of inflectional morphology are treated, down to the very last exception.
- b) Adequate morphological descriptions are assigned to (almost) all word-forms in running text. Of course, it is far from clear what constitutes an adequate morphological description. We take it to mean full specification for traditional inflectional parameters including part of speech, gender, case, number, definiteness, voice, tense, and mood.
- c) Syntactic and semantic properties such as valency, animateness, and abstractness are presently not taken into account.
- d) The most productive derivational phenomena are treated by way of affixation.
- e) A morphotactically precise mechanism for analyzing productive compounds is available.

This mechanism (i) does not generate an unmanageable multitude of spurious readings, (ii) eliminates the need for lexical listing of many transparent compounds, and (iii) enables proper analysis of accidental new compounds.

f) The lexicon contains the base forms of the simplex words belonging to the **core vocabulary** of Swedish. This contains all those words that are not text-type specific, dialectal, obsolete, or marked in some other way. Productively formed compounds and derivatives are not part of the core vocabulary. The rough magnitude of such a core vocabulary can be estimated to be 30,000-50,000 lexical entries. Note, in passing, that the intersection of the lexical entries of Longman's Dictionary of Contemporary English (ASCOT version) and Oxford Advanced Learner's Dictionary of English contains some 30,000 items (CELEX News 4/1988:8). Presently the number of entries in the SWETWOL lexicon is around 48,000. The number of entries in SOB is around 59,000 (Sjögreen 1988:330-331). All compounds and derivatives are included in this figure. Also cf. Sågvald Hein (1988) concerning the problem of compiling a Swedish core vocabulary.

g) Maximally successful use of SWETWOL presupposes the use of domain-specific lexicons in addition to the core lexicon. If SWETWOL is to be applied to a text containing words not in the core lexicon, a proper domain-specific lexicon should first be constructed. This is the safe way of treating words unknown to SWETWOL. Constructing a domain-specific lexicon (e.g., for chemistry or forestry) is not a very time-consuming procedure. Koskenniemi's TWOL program provides tools for determining which words in a text or dictionary that have no matches in the current lexicon. Alternatively, the CG parser offers as an option (applied during context-sensitive disambiguation) the possibility of using string-matching rules for heuristically predicting the morphological analysis of unknown words (Karlsson 1992b). Both procedures thus make it possible to morphologically analyze unrestricted Swedish text.

h) Basic-level morphological analysis involves no heuristics. On the other hand, some regimented heuristics is indispensable in the higher-level processes of disambiguation and syntactic parsing.

i) Apart from the specifics of the treatment of compounds, the SWETWOL morphological description has no pretense to linguistic originality. However, it may be used for uncovering interesting morphological phenomena in Swedish.

j) SWETWOL thus is a tool for morphological analysis. It is mainly a servant of higher-level linguistic analysis. It is also a tool useful for various more or less practical purposes.

k) SWETWOL accounts for written, not for spoken Swedish. All representations are graphemic and all segment alternations morphographemic.

4. Compilation of the SWETWOL lexicon

Work on the SWETWOL lexicon and morphological description started in June, 1988, and reached a quasi-final stage in January, 1992.

By courtesy of Språkdata, University of Gothenburg, our department had obtained in 1985 a

machine-readable version of the tenth edition of Svenska Akademiens ordlista över svenska språket (SAOL-10) which first appeared in 1973 (cf. Sigurd 1986 for a linguistic characterization of SAOL editions). The form of the file was fairly raw with hyphens and certain typographical codes retained:

:2abandon :1(-angång') :3-en :4s. :1ledighet,
:1otvungenhet; vårdslöshet
:2abbé :3-[e]n -er :4s. :4fransk :1präst
:2abbédiss/a :1(-iss'a) :3-an -or :4s. :1före-
:1ståndarinna för kloster :4el. stiftelse
:2abborr/e :3-en -ar :4s. :2-fiske -grund :4pl.
:3-grund :2-nate -pinne :4vard. :1liten ab-
:1borre

I wrote a program (in muLisp) generating a dehyphenated word list with all abbreviations spelt out and containing the requisite morphological information, especially parts of speech and ending sets. Thus the following kind of output was obtained:

abandon -en s.
abbé -[e]n -er s.
abbédiss/a -an -or s.
abborr/e -en -ar s.
abborr=fiske
abborr=grund pl. -grund
abborr=nate
abborr=pinne

This SAOL-10 word list contained 137,870 items. The number is of the same magnitude as the approximate number 140,000 quoted by Sigurd (1986:197) in his exposition on the history of SAOL.

The number of entries separated in SAOL from their main headword by a hyphen “-“ approximates the number of compound words. For SAOL-10, the figure I got was almost 104,200. This stresses the need for a proper treatment of compound morphology. However, this number also includes several thousand words that are derivatives (cf. -are, -het, -ig, -isk, -ning), and some that could be regarded as inflectional (e.g. -ad, -as).

Thus there seem to be some 33,600 simplex words in SAOL-10. This is in good harmony with the figures given above for a reasonable core vocabulary.

Using the part of speech codes and the endings provided for each simplex word I assembled the members of the inflectional types. This phase lasted for months.

Several thousand words are morphologically idiosyncratic in one way or another and require separate treatment. To mention just one example, the number of strong verbs in Germanic languages is normally taken to be around 150. This is the number of simplex strong verbs such as skriva ‘write’ - skrev - skrivit, bryta ‘break’ - bröt - brutit. But there are additional hundreds of lexicalized strong complex verbs such as avlida ‘die’, beskriva ‘describe’, and utkomma ‘appear’. One by one, these had to be described. A further complication is that many strong verbs have either defective paradigms or alternative weak forms. E.g., lyda ‘promise’ has the past tense forms lydde and löd but only the weak supine lytt.

Several thousand of the simplex entries of SAOL-10 were judged to be marginal and were therefore not included in SWETWOL.

Next, several thousand new lexical items were manually excerpted from SAOL-11 (1986) and SOB (1986). New words were also picked from Nyord i svenskan från 40-tal till 80-tal (1986) which contains some 7,500 neologisms. The 1989 edition of this book, Från rondell till gräddfil, was also checked. - Of course, all new entries in these sources were not included in SWETWOL.

Apart from SAOL-10 and SAOL-11, Thorell (1987) was used as the main authority on matters relating to morphological description. Thorell (1984) proved useful in the design of SWETWOL word formation properties.

All along, the evolving SWETWOL system was matched against machine-readable Swedish texts. Språkdata's corpus PRESS-76 containing Swedish newspaper text from 1976 was very useful. It was available to us as a subset with 700,000 words. The 23 Swedish books rendered machine-readable at our department by way of optical scanning were also useful. These books represented a broad range of text types and topics such as novels, administrative reports, computers, aeroplanes and flying, political science, psychology, social relations, car repair, chemistry, ecology, folk medicine, autobiography, and history. This "Helsinki Corpus of Swedish Texts" presently contains some 1,4 million word form tokens.

These texts were instrumental i.a. in providing a check on what proper names (presently some 3,500) and abbreviations to include in the SWETWOL lexicon. A systematic check was made of the more than 5,000 distinct abbreviations contained in Collinder and Svenblad (1987) most of which were included in SWETWOL.

A useful additional check on the SWETWOL treatment of compounds was provided by courtesy of John Giertz, Esselte, who put at my disposal a file containing some 12,500 new compounded word forms recently picked from newspaper texts.

Allén's 10000 i topp (1972) containing the 9,748 most frequent Swedish word forms (in newspaper texts) was rendered machine-readable by optical scanning. SWETWOL performance on the (Swedish) words in this corpus was carefully checked. These 9,748 words alone have a coverage of 83% of the 1 million words in the PRESS-65 corpus.

In March, 1991, SWETWOL was matched against two of Språkdata's large corpora, one containing almost 197,000 word-form types (those in a 25 million word text corpus that have their base-form in SAOL-11), the other containing the entries of SOB. Of course, these tests led to amendments and some revisions.

In December, 1991, SWETWOL was matched against more than 1,1 million word-form tokens (some 98,800 word-form types) of text from Finland's biggest Swedish newspaper Hufvudstadsbladet (several issues from August, 1991). This yielded more than 5,000 useful names and abbreviations typical of Finland and Finnish culture. These words were coded in the ordinary way but with the additional special feature <<Finn>>. They were included in a separate lexicon that may be added to SWETWOL proper when Finland-Swedish texts are to be analyzed.

A check of Swedish place names is presently (January, 1992) being made on the basis of the roughly 3,000 names in Svenska ortnamn - uttal och stavning (1991). The additions called for by this subcorpus have not yet been incorporated into the SWETWOL lexicon. Another check presently being performed concerns the roughly 3,500 person names of Modéer (1989) most of which, however, are historical in nature (e.g. Ormhild, Praetorius, Sighni).

5. Inflectional morphology

With few exceptions, SWETWOL is based on the morphological categories of “classical” Swedish grammar. In naming the parts of speech and other grammatical categories, fairly long codes were used. This increases readability and facilitates debugging and evaluation.

SWETWOL gives every item to be analyzed at least a base form (the lexical citation form), and a part of speech code. The **readings** of uninflected words contain nothing more. In the examples, the word-form occurs at the far left margin, enclosed in double quotes and angular brackets. The readings, one or more, that SWETWOL retrieves are below the word-form, one per line. The base form is enclosed within double quotes. The term **cohort** denotes a word-form and its associated SWETWOL readings. Here is a maximally simple cohort (TWOL has run in so-called Lisp mode, optimal as input to Constraint Grammar parsing, enclosing each cohort as well as all individual readings within parentheses):

```
(“<<och>>”  
  (“och” KONJ))
```

SWETWOL uses 12 basic parts of speech. Every lexical item is tagged for at least one of these. For each part of speech in the ensuing list, there is at the far right a number in parentheses indicating how many inflectional dimensions SWETWOL recognizes for it.

A	adjective (4)
ABBR	abbreviation (2)
AD-A	ad-adjective (i.e. adjectival intensifier) (-)
ADV	adverb (-)
INFMARK	infinitive mark (-)
INTERJ	interjection (-)
KONJ	conjunction (-)
N	noun (4)
NUM	numeral (4)
PREP	preposition (-)
PRON	pronoun (4)
V	verb (2; participles additionally 4)

The system of inflectional categories for nouns, adjectives, and pronouns includes gender, definiteness, number, and case. There are neutralizations where it is unclear how a category is realized. The comparison of adjectives is treated separately below. The subsequent schema has columns containing all codes possible in that morphotactic position. The left-right order is the same as in SWETWOL readings. The schema thus generates all combinations of inflectional properties in SWETWOL nominal readings (also overgenerating a few).

<u>Part-of-speech</u>	<u>Gender</u>	<u>Definiteness</u>	<u>Number</u>	<u>Case</u>
N	UTR	DEF	SG	NOM
A	NEU	INDEF	PL	GEN
PRON	UTR/NEU	DEF/INDEF	SG/PL	ACC
	UTR-MASC ?		?	?

Most codes are self-explanatory. UTR is non-neuter gender. UTR-MASC occurs only in definite e-forms of adjectives. ACC occurs in just a few pronouns:

(“<<gamle>>”
 (“gamma” A UTR-MASC DEF SG NOM))

(“<<henne>>”
 (“hon” PRON UTR DEF SG ACC))

There are three **underspecified** or **composite** codes: UTR/NEU, DEF/INDEF, SG/PL. Consider the comparative form vackrare ‘more beautiful’:

(“<<vackrare>>”
 (“vacker” <<Cmp>> A UTR/NEU DEF/INDEF SG/PL NOM))

It could be considered eight-ways ambiguous under the current grammatical features, cf. en vackrare blomma ‘a more beautiful flower’ (UTR INDEF SG), ett vackrare hus ‘a more beautiful house’ (NEU INDEF SG), vackrare hus ‘more beautiful houses’ (NEU INDEF PL), de vackrare husen ‘the more beautiful houses’ (NEU DEF PL), etc.

Underspecified codes like UTR/NEU are preferred when the following conditions are satisfied: (i) a form class (e.g., comparative) never formally realizes the opposition, (ii) the opposition is “unsyntactic”, and (iii) the ambiguous solution would be unintuitive. Comparatives never make a distinction between UTR and NEU, DEF and INDEF, or SG and PL forms.

The underspecified UTR/NEU code is used for present participles that have the code PCP1 (hoppande ‘jumping’ V PCP1 UTR/NEU DEF/INDEF SG/PL NOM), certain forms of past participles having the code PCP2 (givna ‘given’ V PCP2 UTR/NEU DEF SG NOM, and V PCP2 UTR/NEU DEF/INDEF PL NOM), certain adjective forms (enstaka ‘single’ A UTR/NEU DEF/INDEF SG/PL NOM, akut ‘acute’ A UTR/NEU INDEF SG NOM, vita ‘white’ A UTR/NEU DEF SG NOM, and A UTR/NEU DEF/INDEF PL NOM), comparatives, and a few superlative forms (största ‘biggest’ <<Sup>> A UTR/NEU DEF SG NOM, <<Sup>> A UTR/NEU DEF/INDEF PL NOM).

UTR/NEU is also used for a few noun types, especially amfetamin ‘amphetamine’, ar ‘are’, durra ‘durra’, fotogen ‘paraffin oil, kerosene’, and intervall ‘interval’ with fluctuating gender (def. sg. amfetaminen - amfetaminet, aren - aret, durran - durrat). Numerals are taken to be UTR/NEU, with the obvious exception of en - ett ‘one’.

The underspecified DEF/INDEF occurs in most of the verb forms mentioned under UTR/NEU above, in comparatives, superlatives, numerals and certain adjectives (enstaka ‘separate, individual’), and in nouns inflected like anmodan ‘request’, datum ‘date’, diarium ‘diary’, examen ‘examination, exam’, yen ‘yen’.

The underspecified SG/PL also occurs in the verbs and adjectives just mentioned, and in the noun types adjektiv ‘adjective’ (N NEU INDEF SG/PL NOM), akademiker ‘university graduate’, anagram ‘anagram’, anhängare ‘supporter’, ar ‘are’, datum ‘date’, hus ‘house’, intervall ‘interval’, garage ‘garage’, vatten ‘water’.

Because nouns normally make the number distinction, and presupposing principle (i) above, it could be argued that these nominal singulars and plurals should rather be treated as distinct readings:

(“<<hus>>”
 (“hus” N NEU INDEF SG NOM)
 (“hus” N NEU INDEF PL NOM))

However, this solution is not adopted. One reason is principle (ii): the number opposition is not a strongly syntactic one. Second, it is an overriding concern to minimize ambiguities, both paradigm-internal (as here) and external ones.

Another instance of the reductive inclination of SWETWOL is that no separate manner readings of adjectives are postulated. The NEU INDEF form is regarded as unambiguous and it is the task of syntax to decide the function of this form:

(“<<effektivt>>”
 (“effektiv” A NEU INDEF SG NOM))

The question mark “?” is occasionally used as an undetermined or unclear value of gender, definiteness, number or case, e.g. the case of fossilized words like allom, sinom (historical datives of respectively alla ‘all’, sin ‘time(s)’, occurring in a few idioms), or the definiteness of proper names. “?” and underspecification such as DEF/INDEF denote different phenomena. Underspecification is determinable in context.

A few minor features are also used. These occur in angular brackets to the left of the major features (<<Retain!>> and <<Coerce!>> are exception features, manually inserted in the lexicon, that relate to local disambiguation, cf. section 12):

<<Cmp>>	comparative (adjectives only)
<<Sup>>	superlative (adjectives only)
<<Number>>	e.g. 2, 34, 456
<<Ord>>	ordinal number
<<Prop>>	proper name
<<F-Name>>	first name
<<S-Name>>	surname
<<P-Name>>	place name
<<O-Name>>	other name than first name, surname, or place name
<<Retain!>>	blocks elimination by local disambiguation
<<Coerce!>>	indicates the correct reading in an ambiguous cohort that is otherwise not possible to disambiguate
<<Finn>>	word typical of Finland-Swedish usage

Examples

(“<<1991>>”
 (“1991” <<Number>> NUM ? ? ? ?))

(“<<femte>>”
 (“femte” <<Ord>> A UTR/NEU DEF/INDEF SG/PL NOM))

(“<<*carl-*gustaf>>”
 (“carl-gustaf” <<Prop>> <<F-Name>> N ? ? SG NOM))

(“<<effektivare>>”
 (“effektiv” <<Cmp>> A UTR/NEU DEF/INDEF SG/PL NOM))

(“<<finskor>>”
 (“finska” N UTR INDEF PL NOM)
 (“fin_sko” <<Retain!>> N UTR INDEF PL NOM))

(“<<fältskär>>”
 (“fält_skär” <<Coerce!>> N UTR INDEF SG NOM)
 (“fält_skära” V ACT IMP)
 (“fält_skär” A UTR INDEF SG NOM))

The positive form of adjectives is coded just as the absence of <<Cmp>> and <<Sup>>. More examples of nominal cohorts are provided in Appendix 1.

Verbs are normally inflected for voice and finiteness:

<u>Part-of-speech</u>	<u>Voice</u>	<u>Finiteness</u>
V	ACT	PRES
	PASS	PAST
	DEP	CNVJ
		IMP
		INF
		SUPINE
		PCP1 (+ nominal endings)
		PCP2 (+ nominal endings)

Finite verbs are those with tense-mood inflection (present, past, conjunctive, imperative). The infinitive, supine, and participles (PCP1 = present participle, PCP2 = past participle) are non-finite. Passives occur only with PRES, PAST, INF, and SUPINE. DEP refers to deponential verbs such as brottas ‘wrestle’. Participles have all nominal forms:

(“<<arbetar>>”
 (“arbeta” V ACT PRES))

(“<<arbete>>”
 (“arbete” V ACT IMP)
 (“arbete” V ACT INF))

(“<<arbetade>>”
 (“arbete” V ACT PAST)
 (“arbete” V ACT PCP2 UTR/NEU DEF/INDEF PL NOM)
 (“arbete” V ACT PCP2 UTR/NEU DEF SG NOM))

(“<<arbetats>>”
 (“arbete” V PASS SUPINE)
 (“arbete” V ACT PCP2 NEU INDEF SG GEN))

(“<<arbetande>>”
 (“arbeta” V ACT PCP1 UTR/NEU DEF/INDEF SG/PL NOM)
 (“arbeta” DER-nde N NEU INDEF SG NOM))

(“<<brottas>>”
 (“brottas” V DEP INF)
 (“brottas” V DEP PRES))

SWETWOL features are now morphological, in accordance with the surface orientation of Constraint Grammar (in the above example, DER-nde means “derivative with the ending -nde”). An optimal parsing system needs detailed information e.g. on valency and government properties of the lexical items, not to speak of their semantic properties. Designing and implementing such a lexicon for the 48,000 item SWETWOL vocabulary is a major undertaking (cf. Gellerstam (1988) and Järborg (1988)).

There is at least one deficiency in the present version of SWETWOL. Particles occurring as parts of particle verbs are not separately coded as particles, e.g. bryta ut ‘break out’, hugga av ‘cut off’, tycka om ‘like’. Neither are these properties coded on the verbs (also cf. Sigurd, ms.). Specifying which verbs go with which particles concerns syntactic government and valency. The particles are now analyzed as prepositions and/or adverbs:

(“<<ut>>”
 (“ut” ADV))

(“<<av>>”
 (“av” PREP)
 (“av” ADV))

Government features such as <<av>>, <<om>>, <<ut>> are presently (January, 1992) being coded on the respective verbs. This will yield cohorts such as the following one which is easy to identify as the verb body of a subsequent instance of om:

(“<<tycker>>”
 (“tycka” <<om>> V PRES))

6. Derivational morphology

Word-formation is easy to describe by postulating productive rules operating on derivational elements. However, the shorter such elements are segmentally, and the more they overlap with simplex word structure, the more they overgenerate and give rise to ambiguity problems. Such dangerous suffixes are e.g. -a, -ad, -al, -an, -ar, -d, -i. Furthermore, many formally segmentable derivatives are semantically opaque.

Therefore a conservative approach to derivational morphology was adopted. Most conventionalized derived words were kept in the lexicon even if they would have been possible to eliminate and have analyzed by derivational rules. This practice shows more clearly what there ‘really is’ in the lexicon, and it also makes easier a prospective future extension of SWETWOL to semantics. As soon as word meanings come to the fore, conventionalized lexical entries must

be available. Thus, e.g. the denominal adjectives ad+lig ‘noble’, akter+lig ‘pertaining to the stern’, kamrat+lig ‘friendly’ are in the lexicon.

First consider the derivational suffixes of which Thorell (1984:135-138) lists around 90. Some are productive, semantically transparent, and segmentally long, i.e. good candidates for productive treatment, e.g. -aktig, -artad, -mässig all forming adjectives meaning ‘-like’. As for the shorter nominalizing derivational suffixes, productive formation is postulated only for the denominal adjective types -ig, -isk, -lig, -sk. New words containing these will be properly analyzed, e.g. substandard porr+ig ‘pornographic’. This means that a word like kamratlig is assigned two readings:

(“<<kamratlig>>”

(“kamratlig” A UTR INDEF SG NOM)

(“kamratlig” DER-lig A UTR INDEF SG NOM))

Such cohorts will be reduced by local disambiguation in favor of the non-derived reading (section 12).

The suffix -het derives deadjectival nouns, e.g. billighet ‘cheapness’. The following deverbal derivational suffixes are recognized as productive: -ande/-ende (nomen actionis), -are (nomen agentis), -arinna (feminine agent), -bar ‘-able’, -else (noun), -erska (feminine agent), -ig (adj.), -lig (adj.), -Ning (i.a. nomen actionis). Some overgeneration will occur. For e.g. läsa ‘read’, the derived words läsande, läsare, läsarinna, läsbar, läslig, läsning do exist, but ?läselse, ?läserska, ?läsig do not.

Derivatives are indicated by **features** the names of which contain the string DER, a hyphen and the suffix: DER-are, DER-arinna, DER-bar, DER-else, DER-erska, DER-het, DER-ig, DER-isk, DER-lig, DER-sk.

Two suffixes require special treatment, -ande/-ende and -Ning. The specially marked features DER/-nde and DER/-ning are used. Many words are truly ambiguous between the derived interpretation and an inflectional or lexical one. The reductive scheme for local disambiguation must not apply to such cohorts:

(“<<läsande>>”

(“läsa” V ACT PCP1 UTR/NEU DEF/INDEF SG/PL NOM)

(“läsande” DER/-nde N NEU INDEF SG NOM))

(“<<försäkring>>”

(“försäkring” N UTR DEF SG NOM)

(“försäkra” DER/-Ning N UTR INDEF SG NOM))

Thorell (1984:133-135) lists some 50 derivational prefixes, e.g. ante, anti-, centi-, ex-, icke-, in-, jätte-, makro-. Here, a somewhat greater latitude of interpretation is possible. Many of the prefixes are segmentally fairly long. They are also often foreign. Therefore the risk of interference with simplex words is smaller. Some 30 prefixes are included as productive elements in SWETWOL, including most of those just mentioned except ex-, in-.

The negative prefix o- is very short and cannot be allowed to freely generate potential prefixed derivatives of all words in -o. All words with this prefix, many hundreds, are listed in the lexicon.

7. Compounds

Proper treatment of compounds, especially nominal ones, is indispensable for successful analysis of unrestricted text in Germanic languages. In her study of political language usage, based on corpora exceeding 500,000 words, K. Thelander (1986:158) observed the incidence of compounds to be around 8 %. Einarsson (1978:6,61) reports that the “Skrivsyntax” corpus (175,000 words) has 8,298 nominal compounds alone (4,7 %, containing N+N... combinations). A tractable solution must have good recall (coverage) without compromising precision. It is easy to reach perfect recall of compounds just by letting all roots and certain special stems and inflectional forms freely recurse back to the lexicon of roots. But this approach has devastating effects on precision. The output is flooded with spurious compound interpretations, noise.

Severe restrictions must be put on the combinatorial properties of individual lexical entries. Some words enter no compounds, at least not according to SAOL which SWETWOL uses as the central norm. Such entries must not recurse, especially not if they are segmentally short and likely to create spurious compound interpretations. However, one has to admit that using SAOL as the decisive criterion for when to allow compound recursion is too strict from the viewpoint of language use. SWETWOL does indeed, contrary to SAOL, allow recursion for many (long) words, especially if these occur in compounds attested in texts.

Other words are represented in compounds by special stems or alternations occurring at the compound boundary. These restrictions should be properly incorporated in the description.

A frequent nominal inflectional type is the utral one in -en -er, e.g. atom ‘atom’ - atomen - atomer (N=4,800 in SWETWOL). These words compound in four ways. Some have no attested compounds, e.g. abolition ‘abolition’, aborigin ‘aborigine’. Some enter compounds by plain concatenation, e.g. abiturient-fest ‘party for prospective freshmen’, ablativ-form ‘ablative form’. The third group has an obligatory s at the compound boundary, e.g. abdikations-akt ‘act of abdication’, abstraktions-förmåga ‘propensity for abstraction’. Fourth, s might be fully optional (atmosfär(s)-tryck ‘atmospheric pressure’), or optional in some compounds (akt(s)-förklaring ‘outlaw declaration’) but obligatorily absent in others (akt-stycke ‘document’).

A typical trait of several Swedish nouns is a compound stem ending in a vowel, e.g. kvinna ‘woman’ - kvinn-tjusare ‘lady-killer’, vara ‘article, product’ - varu-hus ‘department store’, träta ‘quarrel’ - träto-broder ‘adversary’, svan ‘swan’ - svane-sång ‘swan song’, svin ‘swine’ - svina-herde ‘swineherd’. Such words may in addition have obligatory or optional boundary-s in other compounds, or obligatory absence of this s.

This variability must be adequately coded. When the SAOL-10 file was converted, special track was kept of what compound stems each word had.

However, all compounds are not generated by productive analysis. There are several criteria for listing a compound in the SWETWOL lexicon, each criterion alone qualifying an item for inclusion:

(i) Some part of a compound does not occur as an individual word (free morpheme), e.g. bigga, fångare, ginnunga in arg-bigga ‘shrew’, fågel-fångare ‘bird catcher’, ginnunga-gap ‘bottomless gap’.

(ii) Some part is short and exempted from compound formation in order to minimize precision-lowering overgeneration, e.g. as, ö, ed, os in as-full ‘dead drunk’, ö-bo ‘island inhabitant’, domar-ed ‘judge’s oath’, stek-os ‘smell of frying’.

(iii) Some part occurs in a unique stem form, e.g. ape-katt ‘ape cat’ of apa ‘ape’.

(iv) Some part is exempted from productive compound formation because it would create disturbing ambiguity. This is true especially of the nouns bar ‘bar’ and hets ‘agitation’. If these

nouns freely could be later compound parts, deadjectival nouns formed by the nominalizer -het in the genitive singular would get two readings (e.g. rödhet+s DER-het N GEN SG ‘redness’ vs. the spurious röd-hets ‘red agitation’). In the same way derived adjectives in -bar like läs+bar ‘readable’ would get a parallel spurious compound reading läs-bar ‘reading bar’. Consequently a number of existing compounds like grill-bar ‘grill bar’, ras-hets ‘racial hatred’, krigs-hets ‘war agitation’ must be listed.

(v) A well-known morphographemic process of Swedish is the omission of one consonant at compound boundaries when three identical consonants meet (CC-C ->> C-C), e.g. brän-nässla ‘stinging nettle’, cigaret-tändare ‘lighter’, fjäl-luft ‘mountain air’, grup-process ‘group process’, rol-lista ‘cast’, äg-gula ‘yolk’, cf. bränn(a) ‘burn’, cigarett ‘cigarette’, fjäll ‘mountain’, roll ‘role’, ägg ‘egg’. Some 340 instances of such compounds were found in SAOL-10 but this list is not conclusive because compound formation is productive.

CCC-shortening could be accounted for by a two-level rule that lets simple surface consonants correspond to double ones in the lexicon when occurring in front of a compound boundary followed by an identical initial consonant. However, this would overgenerate. E.g., the adjective tal-lös ‘innumerable’ would get an additional reading “tall” ‘pine’ + “lös”. SWETWOL presently lists all (roughly 360) encountered compounds of this type. Of course, this is also suboptimal because new instances do occur in actual usage. Some recent ones spotted in Finland-Swedish are nol-lösning ‘zero solution’, nol-linje ‘zero policy’, both referring to current wage politics, cf. noll ‘zero’.

The optimal solution to this problem is still pending. Basically, it should be evaluated (by extensive dictionary and corpus study) to what extent the lexical listing solution lowers recall, as compared to how much the rule solution would cause precision to drop due to the emergence of additional spurious readings.

8. The SWETWOL lexicon and morphology design

SWETWOL is presently morphology-oriented at the expense of semantic matters. Because the primary aim is to construct a morphologically efficient analyzer dealing properly with word-form ambiguities, **three simplifications must be kept in mind** that reduce the number of ambiguities generated by SWETWOL:

(i) Identical lemmata, i.e. semantically different words with precisely the same inflection (inflectionally identical external homographs), are not duplicated. E.g., there is only one noun bet/a -an -or in SWETWOL even if SAOL-11 lists three distinct ones (‘reminder’, ‘steep, mordant’, ‘beet’).

(ii) If one lemma formally is a proper subset of another (sharing everything except one or more endings), it is not included in SWETWOL. E.g., kräft/a -an ‘cancer’ is excluded since there is a word with a more extensive inflection, kräft/a -an -or ‘crayfish’.

(iii) If two or more inflectionally distinct lemmata have identical compound stems, only one is allowed to recurse. E.g., the words bas -en -er ‘base, basis, foundation; bass’ vs. bas -en -ar ‘boss, foreman’ form such a pair. Else, compounds such as bas-linje ‘base-line’, bas-röst ‘bass voice’ would receive (at least) two readings along compound paths.

The net effect of these simplifications was that several hundred potentially good lexical entries were excluded.

SWETWOL is fairly concrete. This means that the representations of lexical items adhere as closely as possible to ordinary spelling. Few abstract morphographemes and few two-level rules

are used. This makes the lexicon more perspicuous, and debugging easier. It also speeds up the compilation of run-time versions of the system because the rules do not have to be recompiled into finite-state automata frequently.

A consequence of this is that words with final morphographemic alternations are represented in the lexicon by the invariable part of their stems, often similar to the truncated stems of Hellberg (1978). The lexical entries of the word kvinna ‘woman’, hoppa ‘jump’, vaken ‘awake’ are thus kvinn, hopp, vak. Such brief lexical stems might be difficult to decode, especially when occurring dissociated from the morphotactic specifications.

SWETWOL could thus be characterized as an item-and-arrangement version of Swedish morphology. In TWOL it is left to the user to decide how the relations between lexical storage of items vs. rules should be described. No compelling requirements obtain a priori.

Almost all aspects of Swedish morphology are treated as listing of alternatives in the lexicon system. This concerns the root-internal morphographemic segment alternations (vaken - vak+na), the ending-class internal form variations (-en - -n), the morphotactic restrictions on combinations of stems and endings, and the combinatorial restrictions on compound formation.

The SWETWOL lexicon is a tree-structure consisting of some 320 major compartments or minilexicons (Koskenniemi 1983), each representing some important subclass of the phenomena just mentioned. Any compartment may bifurcate into any number of subcompartments.

Each item in a minilexicon must contain at least a pointer to another minilexicon where the analysis continues, or the designated symbol “#” if the end of a path has been reached. To the left of the minilexicon name, an item may (but need not) contain segments to be matched. To the right, it may (but need not) contain a string of features to be retrieved. The outlook options for an item in a minilexicon are thus four:

```
(segments) name-of-lexicon (“features-to-retrieve”)
      ENER1;
atom  ENER1;
en    NOMGEN      “ N UTR”;
      NOMGEN      “ N UTR”;
#     “ N UTR”;
```

The first line states that the analysis path continues in minilexicon ENER1; the second says that the path continues in ENER1 after the segments atom have been matched; the third says that the path continues in lexicon NOMGEN when the segments en have been matched, and then the features “N UTR” are retrieved; the fourth line requires no segments to be matched but the features “N UTR” are retrieved and the analysis goes on in minilexicon NOMGEN; the last line terminates a successful analysis path and retrieves the features “N UTR”.

The top lexicon is START. It allows an initial asterisk (for upper-case, introduced by the preprocessor), then continuing to lexicon START-2. A lower-case letter leads directly to START-2, the bulk of SWETWOL.

```
LEXICON START
* START-2 “ ”;
START-2 “ ”;
```

```
LEXICON START-2
MAJORS “ ”;
```

SINGLES “ ”;
NUMBERS “ ”;
ADVPREPS “ ”;
DIGITS “ ”;
NAMES “ ”;

START-2 splits up the vocabulary into six lexicons. MAJORS is the largest one. It contains most nouns, adjectives, and verbs, almost 40,000 entries:

LEXICON MAJORS

abakus ENER1;
abandon EN;
abbé EENER;
abbediss ANOR;
abborr ENAR4C;
abbot ENAR2;
abbotsdöme TN;
abbrevier V1;
...

Lexicon SINGLES contains some 600 short or otherwise problematic words exempted from occurring as later parts of compounds, e.g. aga, agna, alf, alv, am, amma, ami, ana:

LEXICON SINGLES

ag V1;
agn V1;
alk ANOR4BB;
alf ENER1;
alv ENER1;
am MAJ;
am V1M;

Lexicon SINGLES also contains abbreviations, interjections, conjunctions, and other items not partaking in normal compound formation. Lexicon NUMBERS accounts for numerals: en ‘1’, två ‘2’, ..., by recursion to itself also trehundra ‘379’,..., tre-fyra ‘3-4’, etc. Lexicon ADVPREPS contains adverbs, ad-adjectives (adjectival intensifiers), and prepositions not partaking in compound formation. Lexicon DIGITS accounts for words like 1, 2780, 12-13, 27,45, 27.45, 1300-talet ‘13th century’, 27-åring ‘person aged 27’. Lexicon NAMES contains some 3,000 proper names of persons (both first names and surnames) and places.

Productive compounding occurs via recursion from the appropriate inflectional minilexicons to lexicon Root (names recurse to NAMES):

LEXICON Root

MAJORS “ ”;
COMPOUNDCODAS “ ”;

Lexicon Root points to the beginning of MAJORS and to lexicon COMPOUNDCODAS which

contains items occurring only as post-elements in compounds (some of them, e.g. -aktig, are word-like derivational endings):

LEXICON COMPOUNDCODAS

aktig A;
arma AD;
axla AD;
axlig A;
barma AD;
...

At the level of individual inflectional types, the four ENER lexicons look as follows.

LEXICON ENER

NOMGEN “ N UTR INDEF SG”;
en NOMGEN “ N UTR DEF SG”;
er NAS “ N UTR”;
NDER “ ”;

LEXICON ENER1

NOMGEN “ N UTR INDEF SG”;
Z Root;
Root;
en NOMGEN “ N UTR DEF SG”;
er NAS “ N UTR”;
NDER “ ”;

LEXICON ENER2

NOMGEN “ N UTR INDEF SG”;
Z Root;
S Root;
en NOMGEN “ N UTR DEF SG”;
er NAS “ N UTR”;
NDER “ ”;

LEXICON ENER3

NOMGEN “ N UTR INDEF SG”;
Root;
Z Root;
S Root;
en NOMGEN “ N UTR DEF SG”;
er NAS “ N UTR”;
NDER “ ”;

ENER words not occurring as first parts of compounds are of type ENER, those forming compounds without š are of type ENER1 with recursion to LEXICON Root, those compounding with š are of type ENER2 with recursion to Root after an š has been spotted, and those forming

compounds with optional s are of type ENER3 with two paths to Root.

Z is a morphophoneme whose realization is determined by a two-level rule. Z is realized as s or 0 at compound boundaries other than the first, else as 0, cf. abonment-anslutning ‘customer connection’ vs. tre-abonnennt(s)-anslutning ‘connection for three customers’. The morphophoneme S is realized as s except after certain obstruents where it is manifested as zero.

The lexicon NDER stands for productive nominal derivation. The lexicon NOMGEN is simple (the last two lines cover coordinated truncated instances such as bil-, parkerings- in phrases like bil- och busstrafiken ‘car and bus traffic’, parkerings- och andra förseelser ‘parking and other offences’):

LEXICON NOMGEN

“NOM”;
S # “GEN”;
- # “N”;
s- # “N”;

There are thirteen subtypes of a/or-nouns like kvinna, depending upon compounding properties. E.g., the stem of begonia in compounds ends in -a or -e (note the convention that segments immediately following (without blanks) the first double quote signalling the features to be retrieved are to be substituted for the segments matched, e.g. a is to be substituted for or - this makes it possible to generate base forms, e.g. begonia from the matched string begonior):

LEXICON BEGONIA

a NOMGENMAX “a N UTR”;
e Root;
a Root;
or NAS “a N UTR”;
a- # “a N UTR INDEF SG NOM”;
e- # “a N UTR INDEF SG NOM”;

Alternative ending patterns are treated simply by listing, e.g. lexicon ENARER1 for check ‘cheque’ - check+en - pl. check+ar or check+er:

LEXICON ENARER1

NOMGEN “N UTR INDEF SG”;
en NOMGEN “N UTR DEF SG”;
ar NAS “N UTR”;
er NAS “N UTR”;
Root;
Z Root;
NDER “”;

More drastic stem truncation occurs in words like myller ‘crowd’ - myll+ret, lexically myll:

LEXICON ERRET1

er NOMGEN “= N NEU INDEF SG/PL”;
ret NOMGEN “er N NEU DEF SG”;

Root;
erZ Root;
ren NOMGEN “er N NEU DEF PL”;
rena NOMGEN “er N NEU DEF PL”;

Strong verbs require much listing. The entries and relevant minilexicons for e.g. omgiva ‘surround’ - omgav - omgivit are (“%” is explained in section 9):

omgiv /STRLID “omgiva”;
omgav PAST “omgiva”;

LEXICON /STRLID
STRBIND “ ”;
SUPINEIT “ ”;
PCPEN “ ”;

LEXICON STRBIND
a INF;
% # “a V ACT IMP”;
es% # “a V PASS PRES”;
s% # “a V PASS PRES”;
er% # “a V ACT PRES”;
ande% NOMGEN “a V ACT PCP1 UTR/NEU DEF/INDEF SG/PL”;
VDER “ ”;

LEXICON SUPINEIT
it SUPINE “a”;

LEXICON PCPEN
en NOMGEN “a V ACT PCP2 UTR INDEF SG”;
et NOMGEN “a V ACT PCP2 NEU INDEF SG”;
ne NOMGEN “a V ACT PCP2 UTR-MASC DEF SG”;
na NOMGEN “a V ACT PCP2 UTR/NEU DEF SG”;
na NOMGEN “a V ACT PCP2 UTR/NEU DEF/INDEF PL”;
enhet ENER2 “= DER-het”;

LEXICON PAST
% # “ V ACT PAST”;
s% # “ V PASS PAST”;

The 320 minilexicons give a rough indication of the morphological complexity of SWETWOL. The rough magnitude is the same as the 235 paradigms postulated by Hellberg (1978). Sågval Hein’s LPS inflectional grammar has 676 inflectional types for nouns, 162 for adjectives, and 547 for verbs, several with one member, reflecting the descriptive practice of SOB (Sågval Hein 1988:279; also cf. Sågval Hein and Sjögreen, in print). SWETWOL contains the same idiosyncrasies. In addition to the 320 patterns, SWETWOL contains more than 1,000 items listed as idiosyncrasies.

9. Two-level rules

Only eight two-level rules are used in SWETWOL. The rules were formulated by Kimmo Koskenniemi. The most important rules are presented here. Under Alphabet, an expression such as D:d states a lexical/surface character pair correspondence. The rule operator “ \Rightarrow ” states that the lexical/surface character pair left of the arrow must occur in the context specified right of the arrow. The operator “ \Leftarrow ” requires a correspondence pair with the same lexical segment as is left of the arrow, that furthermore occurs in the context specified right of the arrow, to correspond to the surface character specified to the left. The operator “ \Leftrightarrow ” requires a lexical segment to correspond to a certain surface segment and further prevents it from having other surface realizations, in the context expressed to the right.

Alphabet

a b c d e é f g h i j k l m n o p q r s t u v w x ü y z å ä ö
D:d N:n S:s Z:s : - _ * 0:* \$ @ . , ! ? ; ; ;

Diacritics

% ;

Sets

V = a e é i o u ü y å ä ö ;
Ve = a i o u ü y å ä ö ;
Cl = b c d f g h j k l m n p q r s t v w x z ;
Cr = b c d f g h j k l m n p q s t v w x z ;
Sb = s h z x c ;
Xs = a b d e é f g i j k l m n o p q r t u v w ü y å ä ö ;
X = a b c d e é f g h i j k l m n o p q r s t u v w x ü y z å ä ö ;

Rules

- (1) D:t \Leftrightarrow _ :t ;
- (2) N:0 \Leftrightarrow [e: r: | Cr: r: | Cl: l: | n:] _ ;
- (3) m:0 \Leftrightarrow :m _ N: ;
- (4) Z:s \Rightarrow #: :* _ #: :0* :X ;
- (5) S:s \Leftarrow :Sb _ ;
- (6) %: \Leftarrow %: :* _ ;

Rule (1) governs the d/t-alternation, e.g. vid ‘wide’ - neutr. vit+t, leda ‘(to) lead’ - supine let+t: D is realized as t before t, else as d. The lexical representations are respectively viD and leD.

Rule (2) accounts for the n/0-alternation in deverbal (n)ing-derivatives (lexically Ning), e.g. skola ‘educate’ - skol+ning, rena ‘clean’ - ren+ing.

Rule (3) requires m to be represented as zero after m and before N, e.g. simma ‘swim’ - sim+ning.

Rule (4) enables lexical Z to be realized as s when there is at least one compound boundary to the left, i.e. at later compound boundaries, but says nothing about its occurrence in other contexts.

Rule (5) discards s-genitives after certain obstruents, e.g. *hus+s, *Max+s.

Rule (6) constrains compound formation. The lexical segment % occurs in connection with

finite verb forms. (6) forbids more than one occurrence of % in the same word, i.e. a well-formed compound cannot consist of two or more finite verbs.

The rules are compiled into run-time finite-state automata by the TWOL rule compiler developed by Karttunen, Koskenniemi, and Kaplan (1987).

Consonant shortening over compound boundaries (presstöd ‘press subsidy’, cf. press ‘press’, stöd ‘subsidy’) is not taken to be rule-governed because this would lead to considerable overgeneration.

10. Performance

Kimmo Koskenniemi has designed efficient TWOL implementations. E.g., on a SONY News Workstation, a C version of TWOL runs SWETWOL in analyzing mode at 300-350 word form tokens per second, i.e. more than 1 million words per hour.

SWETWOL has been subjected to extensive corpus testing in the course of its development. In March, 1991, when a major check of the lexicon had been completed, the **recall** of SWETWOL was tested against the texts of two fresh, optically scanned books. By recall we mean the percentage of word form tokens (in relation to all word form tokens) that were assigned one or more readings by SWETWOL such that the appropriate reading(s) was/were among the readings received. Thus, deductions from optimal recall occur when SWETWOL assigns either no readings, or only one or more erroneous readings. ‘Fresh’ means that this was the very first confrontation of SWETWOL with these texts.

The first text was Ett barr är ett liv, written by Johan Ulfveng (Söderströms, Borgå 1989, 192 pp.), a book on ecology and the preservation of nature, containing 47,422 word form tokens (8,432 types). SWETWOL left 310 word form tokens unanalyzed (0,65%). Of these, 70 were misspellings like förekomsetn, skadefrekven, sten-yta, Ästerbotten pro förekomsten, skadefrekvens, stenyta, Österbotten. Another 50 were names or name-based nouns, often not Swedish, e.g. Amazonas, Amazonas-bäckenet, Brandenburg, Chengan, Joensuu, Karstula, Katowice, Martti, McCormick. Some 25 were English or Latin words. Some 10 were disputable Swedish word forms violating SAOL-11, e.g. ?bils-trafiken pro bil-trafiken ‘car traffic’, ?djurs-uppfödning pro djur-uppfödning ‘animal breeding’.

This leaves some 155 “interesting” word form tokens (0.3%) unrecognized. Most of these are text type specific items, often occurring in inflected forms, e.g. antropogen ‘antropogene’, avförsurning ‘acid making’, contorta-odling ‘contorta growing’, geoekologisk ‘geocological’, mineralogist ‘mineralogist’, mykorrhiza ‘mycorrhiza’, pah-halt ‘pah content’, pah-ämnen ‘pah substances’, ph4. Such words are not members of the core vocabulary.

A compound like solsken-kväveoxider-kolväten ‘sunshine-nitric oxide-hydrocarbons’ is not based on any formalizable pattern.

A handful of missing good words were also spotted, e.g. disk-hoar ‘washing-up sinks’, episodvis ‘episodically’, uppbevara ‘keep’. One instance was observed of a phenomenon known in advance not to be adequately covered, the word naturvetenskapligt-tekniska ‘scientific-technical’. Requisite descriptive data are presently lacking on what the conditions are for forming such internally inflected compounds.

No instances were found of cohorts containing only erroneous readings, as determined on the basis of a sample containing 10,000 cohorts.

The second text was a book on socialization and psychotherapy of children, Barn i grupper. Struktur, process och psykoterapi by Sten Lundqvist and Margareta Walch (Natur och Kultur,

Stockholm 1989, 223 pp.). This book contained 54,542 word form tokens (5,857 types). SWETWOL left 220 word form tokens unanalyzed (0,4 %). Observations highly similar to the preceding ones can be made: a handful of good words (e.g. oriktad ‘undirected’, ostrukturerad ‘unstructured’, rollek ‘role play’ from roll + lek, grupperspektiv ‘group perspective’ from grupp + perspektiv, a few somewhat strange compounds (antingenellerkaraktär ‘nature of being either-or’, en-till-en-relation ‘one-to-one-relation’, inåtvänd-utagerande ‘inwards turned - outwards acting’), some 20 text type specific words (e.g. co-terapeut ‘co-therapeut’, co-terapi ‘co-therapy’, individuation ‘individuation’, mutist ‘mutist’, narcissistisk ‘narcissistic’, oidipal ‘oedipal’), the rest being misspellings or names of various types. In a sample of 10,000 cohorts, no instances were found containing only erroneous readings.

SWETWOL recall thus seems to be well above 99,7 % in “ordinary” text, without invoking either the procedure of constructing a domain-specific lexicon for the new texts, or the option of heuristic morphological analysis provided by the Constraint Grammar Parser (cf. section 3). Both of these procedures are at hand if one wishes to obtain a perfect analysis. Of course, recall is lower than 99,7 % when the system is matched against highly specialized texts.

Even if the two 10,000 cohort samples did not contain cohorts with only erroneous readings, such examples were occasionally encountered when SWETWOL was compiled. Here are some early examples, spotted (and later corrected) when SWETWOL was applied to Nyord i svenskan från 40-tal till 80-tal (“_” henceforth indicates compound boundaries):

(“<<polypropen>>”

(“polyp_rop” N NEU DEF PL NOM))

(“<<reproteknik>>”

(“rep_ro_teknik” N UTR INDEF SG NOM))

(“<<makrobiotiken>>”

(“mak_ro_bio_tik” N UTR DEF SG NOM))

(“<<petrokemi>>”

(“pet_ro_kemi” N UTR INDEF SG NOM))

(“<<limbo>>”

(“lim_bo” V ACT INF)

(“lim_bo” V ACT IMP))

(“<<paranormal>>”

(“par_anormal” A UTR INDEF SG NOM))

Some word, often a specialized or otherwise marked one, is not in the dictionary but SWETWOL still analyzes it due to the productive compounding capacity and turns out only improper analyses. Second, some form of a lacking word may also intersect with the inflected forms of a word in the lexicon. E.g., when the male first name Kalle was not in the list of proper names, SWETWOL recognized that token as a form of the adjective kall:

(“<<*kalle>>”

(“kall” A UTR-MASC DEF SG NOM))

Consonant shortening over compound boundaries may also cause recall problems. The proper segmentations of bollek ‘ball play’ and borrigg ‘drilling offshore platform’ are “boll+lek” and “borr+rigg”. SWETWOL delivers the erroneous segmentations boll-ek, bor-rigg because the words ek ‘oak’, bor ‘boron’ happen to be in the lexicon. An interesting instance is the word glasskål which gets the proper reading glas-skål ‘glass bowl’, the spurious reading glass-kål ‘ice cream cabbage’, but not the other proper reading “glass-skål” ‘ice cream bowl’. The action needed is to judge whether the lexicon should be updated. Such errors are very rare in large amounts of varied text.

Precision refers to the incidence of ‘noise’ among the readings assigned. Detractions from optimal precision occur when a cohort contains spurious readings, either in addition to the correct one(s), or, very seldom, as the only ones. Prime examples of lacking precision are overgenerated compound readings such as:

```
(“<<kulturkrock>>
  (“kultur_krock” N UTR INDEF SG NOM)
* (“kult_ur_krock” N UTR INDEF SG NOM)
* (“kul_tur_rock” N UTR INDEF SG NOM)
* (“kul_tur_krock” N UTR INDEF SG NOM))
```

Here the readings prefixed by a star, “*”, are precision-lowering noise. A few powerful principles of local disambiguation suffice to pick the proper readings in such cohorts. Precision data and disambiguation principles are the topics of sections 11, 12.

11. Morphological ambiguity in Swedish

The notion morphological ambiguity is dependent on interpretation. Is e.g. the following (simplified) cohort eight-ways ambiguous?

```
(“<<arbetande>>”
  (“arbeta” V ACT PCP1 UTR/NEU DEF/INDEF SG/PL NOM))
```

In section 5 it was argued that it could be regarded as underspecified, rather than as eight-ways ambiguous. SWETWOL uses underspecification in the description of minor categories of participles and adjectives, and in number of nouns. A reading such as “hus N NEU DEF SG/PL NOM” is thus not ambiguous.

Operationally, we consider a word form ambiguous if SWETWOL generates more than one reading for it. Ambiguity as here defined concerns token and not lexical properties. Lexical ambiguity is normally called homonymy and is an identity relation between different lexical items. (Cf. the beginning of section 8 for some important simplifications of the treatment of lexical ambiguity in SWETWOL.)

When SWETWOL was applied on the token level to the “Helsinki Corpus of Swedish Texts” described in section 4, containing some 1,37 million words, the following ambiguity rate was obtained:

Table 1. Number of readings assigned by SWETWOL to the word form tokens of the “Helsinki Corpus of Swedish Texts”. N(r) = number of readings, N(w) = number of word form tokens, cum-% = cumulative percentage, %(-0) = percentage when words with 0 readings are omitted, cum-%(-0) = cumulative percentage when words with 0 readings are omitted.

N(r)	N(w)	%	cum-%	%(-0)	cum-%(-0)
0	35,508	2,6	2,6	-	-
1	570,049	41,6	44,2	42,7	42,7
2	517,890	37,8	82,0	38,8	81,5
3	112,861	8,2	90,2	8,5	90,0
4	45,513	3,3	93,5	3,4	93,4
5	53,611	3,9	97,4	4,0	97,4
6	8,098	0,6	98,0	0,6	98,0
7	16,714	1,2	99,2	1,2	99,2
8	9,448	0,7	99,9	0,7	99,9
9	493	0,04			
10	282	0,02			
11	80				
12	212				
13	28				
14	115				
15	32				
16	21				
18	8				
20	86				
21	1				
23	1				
24	6				
28	1				
32	1				
40	1				
42	1				
=====					
	1,371,061	100	100	100	100

The texts had been optically scanned and were not fully proofread and corrected at the time of the count. This explains the relatively high incidence of words lacking analysis (2,6%): the vast majority were misspellings. Disregarding these, the fifth column shows that 42,7% of the words are unambiguous, i.e. that 57,3% are (at least two-ways) ambiguous. This ambiguity rate is slightly lower than the 64,5% reported by Allén (1970:XXV) for the NFO corpus, due to differences in the interpretation and use of underspecification.

38,8% are two-ways ambiguous, then a rapid decline is observable. The frequency fluctuations at N(r)=7,8 are explained by the two words för ‘for’, så ‘so’, both with a cohort of readings belonging to almost all central parts of speech:

(“<<för>>”
 (“för” PREP)

("för" AD-A)
("för" ADV)
("för" N UTR INDEF SG NOM)
("för" KONJ)
("föra" V ACT PRES)
("föra" V ACT IMP))

("<<så>>")
("så" AD-A)
("så" ADV)
("så" V ACT INF)
("så" V ACT IMP)
("så" N UTR INDEF SG NOM)
("så" INTERJ)
("så" KONJ)
("så" PRON UTR INDEF SG NOM))

The current token precision of the compound mechanism of SWETWOL, arrived at after lengthy testing and optimization, is satisfactory in view of the fact that not more than 0,04% of the tokens get more than 10 readings (N=594). Only 12 get more than 20, 3 more than 40, and 1 more than 40 readings.

The word with 42 readings, barbitursyrehaltiga, is interesting. Because there are no elements barbitur- or barbitursyra in the lexicon, the wealth of readings arise by various combinations of the words bar 'bar', bi 'bee', bit 'part', tur 'luck', ur 'watch', syra 'acid', hal 'slippery', halt 'content' (with the derived adjective haltig), and tiga 'keep silent'. Of course, more work could -- and should -- be done on making the compound mechanism even more precise. The conspicuous jumps in Table 1 at N(r)=14,20,24 are due to accidental individual words. E.g., all instances of the word form psykologiska get 14 readings, and all forms of biologiska 20 readings.

In Table 2, the frequency distribution of readings over word form types is presented. The 1,37 million tokens of Table 1 reduce to 95,327 types.

Table 2. Number of readings assigned by SWETWOL to word form types. Cf. table 1.

N(r)	N(w)	%	cum-%	%(-0)	cum-%(-0)
0	11,193	11,7	11,7	-	-
1	46,081	48,3	60,0	54,8	54,8
2	22,574	23,7	83,7	26,8	81,6
3	8,427	8,8	92,5	10,0	91,6
4	3,872	4,1	96,7	4,6	96,2
5	879	0,9	97,6	1,0	97,2
6	1,330	1,4	99,0	1,6	98,8
7	229	0,2	99,2	0,3	99,1
8	347	0,4	99,6	0,4	99,5
9	129	0,1	99,7	0,2	99,7
10	97	0,1	99,8	0,1	99,8
11	24				
12	83				
13	6				
14	19				
15	7				
16	15				
18	5				
20	4				
21	1				
23	1				
24	1				
28	1				
32	1				
40	1				
42	1				
		95,327	100	100	100

Disregarding the unanalyzed words, almost 55% of the word form types are rendered unambiguous by SWETWOL.

12. Local disambiguation

In Karlsson (1989, 1990) it was shown that there are two types of morphological disambiguation. Local disambiguation is performed solely within cohorts, by examination of the readings at hand, using no cohort-external information. Context-sensitive disambiguation relies on examination of the contents of neighbouring cohorts, perhaps unboundedly far away. For a detailed exposition of context-sensitive disambiguation of Swedish, cf. Karlsson (in preparation).

Consider the following interesting cohorts analyzed by SWETWOL (some of which contain decidedly comical readings):

(“<<bytesbil>>”

(“bytes_bil” N UTR INDEF SG NOM)

(“by_tes_bil” N UTR INDEF SG NOM))

(“<<syndrom>>”

(“syndrom” N NEU INDEF SG/PL NOM)

(“synd_rom” N UTR INDEF SG NOM))

(“<<tonkontroll>>”

(“ton_kontroll” N UTR INDEF SG NOM)

(“ton_kon_troll” N NEU INDEF SG/PL NOM))

(“<<bankomat>>”

(“bankomat” N UTR INDEF SG NOM)

(“ban_koma” N NEU DEF SG NOM)

(“ban_ko_mat” N UTR INDEF SG NOM))

(“<<konkurrenssamhälle>>”

(“konkurrens_samhälle” N NEU INDEF SG NOM)

(“kon_kur_rens_samhälle” N NEU INDEF SG NOM))

(“<<konsultuppdrag>>”

(“konsult_uppdrag” N NEU INDEF SG/PL NOM)

(“konsult_uppdra” V ACT IMP)

(“kon_sul_tupp_drag” N NEU INDEF SG/PL NOM)

(“kon_sul_tupp_dra” V ACT IMP))

(“<<antirobotrobot>>”

(“anti_robot_robot” N UTR INDEF SG NOM)

(“anti_robot_ro_bot” N UTR INDEF SG NOM)

(“anti_robot_ro_bo” N NEU DEF SG NOM)

(“anti_rob_otro_bot” N UTR INDEF SG NOM)

(“anti_rob_otro_bo” N NEU DEF SG NOM)

(“anti_ro_bot_robot” N UTR INDEF SG NOM)

(“anti_ro_bot_ro_bot” N UTR INDEF SG NOM)

(“anti_ro_bot_ro_bo” N NEU DEF SG NOM)

(“anti_ro_bo_tro_bot” N UTR INDEF SG NOM)

(“anti_ro_bo_tro_bo” N NEU DEF SG NOM))

(“<<publikunderlag>>”

(“publik_underlag” N NEU INDEF SG/PL NOM)

(“publik_under_lag” N NEU INDEF SG/PL NOM)

(“publik_under_lag” N UTR INDEF SG NOM)

(“publik_under_lag” N UTR INDEF SG NOM)

(“publik_underlag” N NEU INDEF SG/PL NOM)

(“publik_under_lag” N NEU INDEF SG/PL NOM)

(“publik_under_lag” N UTR INDEF SG NOM)

(“publik_under_lag” N UTR INDEF SG NOM)
 (“pub_lik_underlag” N NEU INDEF SG/PL NOM)
 (“pub_lik_under_lag” N NEU INDEF SG/PL NOM)
 (“pub_lik_under_lag” N UTR INDEF SG NOM)
 (“pub_lik_under_lag” N UTR INDEF SG NOM)
 (“pub_lik_underlag” N NEU INDEF SG/PL NOM)
 (“pub_lik_under_lag” N NEU INDEF SG/PL NOM)
 (“pub_lik_under_lag” N UTR INDEF SG NOM)
 (“pub_lik_under_lag” N UTR INDEF SG NOM)
 (“pub_lik_underlag” N NEU INDEF SG/PL NOM)
 (“pub_lik_under_lag” N NEU INDEF SG/PL NOM)
 (“pub_lik_under_lag” N UTR INDEF SG NOM)
 (“pub_lik_under_lag” N UTR INDEF SG NOM))

(“<<marinbiolog>>”

(“marin_bilog” N UTR INDEF SG NOM)
 (“marin_bio_le” V ACT PAST)
 (“marin_bio_log” N UTR INDEF SG NOM)
 (“marin_bio_le” V ACT PAST)
 (“marin_bio_log” N UTR INDEF SG NOM)
 (“marin_bilog” N UTR INDEF SG NOM)
 (“marin_bio_le” V ACT PAST)
 (“marin_bio_log” N UTR INDEF SG NOM)
 (“marin_bio_le” V ACT PAST)
 (“marin_bio_log” N UTR INDEF SG NOM))

It is easy to see that many compound readings may be discarded as improper if the respective cohort contains some reading with fewer compound boundaries “_”. This is the Compound Elimination Principle:

Compound Elimination Principle

If a cohort C contains readings with n and m compound boundaries, discard all readings with m compound boundaries if m >> n.

This leaves marin-biolog, publik-underlag, kultur-krock, anti-robot-robot, bankomat (with 0 compound boundaries), etc. as the unique correct ones. Unambiguous cohorts, full disambiguation, optimal precision, were reached by the Compound Elimination Principle alone. Of course, more than one reading might be left pending if they are at the same level of compound complexity, cf. the two readings for konsult-uppdrag.

There are extremely few instances where the Compound Elimination Principle overgenerates and discards a reading that should have been retained. I have analyzed the whole material in Berg (1978), which contains the 5,619 externally homographic components in the NFO corpus (Allén 1970). Homographs are external when they concern distinct lemmas, homographs within the same lemma are internal. Furthermore, tens of thousands of cohorts have been inspected in the course of testing SWETWOL on texts.

Only a handful of words have so far been found which have two attested readings with variable number of compound boundaries (one reading always being non-compounded:

brunsten (brunst+en ‘the rut’ vs. brun-sten ‘manganese ore’)
finskor (finsko+r ‘Finnish women’ vs. fin-sko+r ‘fine shoes’)
kastrullen (kastrull+en ‘the pan’ vs. kast-rulle+n ‘the casting reel’)
snöras (snöra+s ‘lace (pass.)’ vs. snö-ras ‘snow slide’)
toppar (topp+ar ‘crests’ vs. top-par (“topp”-par) ‘top pair’)
vägras (vägra+s ‘refuse (pass.)’ vs. väg-ras ‘road slide’)

This is an interesting fact about Swedish word structure. An analysis of its causal basis shall not be attempted here, we just note that the same is true of Finnish as well (less than five documented exceptions), indicating the existence of a deep-rooted constraint. Technically, the erroneous elimination of the compounds brun-sten etc. is prohibited by including these compounds in the lexicon and supplying them with the minor feature <<Retain!>>.

There is also the Derivative Elimination Principle:

Derivative Elimination Principle

If a cohort C contains readings some of which are derived and others simplex, discard the derived ones.

Derivational status is technically indicated by the presence of a feature the name of which contains the initial segments “DER-“ (cf. section 6). Accordingly, in the following cohort the second reading is spuriously derived from bill ‘share’ and must therefore be discarded, the proper reading being the upper one meaning ‘cheap’:

(“<<billig>>”

 (“billig” A UTR INDEF SG NOM)

 (“billig” DER-ig A UTR INDEF SG NOM))

Of course, both our elimination principles have a common basis: **eliminate morphologically more complex readings in favour of simpler ones**. There are more aspects of local disambiguation that cannot be expounded here (cf. Karlsson 1989, 1992a).

Local disambiguation has been implemented in two ways, by myself as a Lisp program, and by Kimmo Koskenniemi as an AWK program running under Unix. The latter provides the option of being directly applied as a post-filter to TWOL output, thus delivering immediate locally disambiguated output. When the corpora presented in Tables 1, 2 are subjected to local disambiguation, clear -- almost dramatic -- effects are discernible:

Table 3. Number of readings for word form tokens after local disambiguation. Cf. Table 1.

N(r)	N(w)	%	cum-%	%(-0)	cum-%(-0)
0	35,508	2,6	2,6		
1	642,138	46,9	49,5	48,1	48,1
2	497,717	36,3	85,8	37,3	85,4
3	102,262	7,5	93,3	7,7	93,1
4	27,372	2,0	95,3	2,0	95,1
5	42,191	3,1	98,4	3,2	98,3
6	459	0,03	98,4	0,03	98,3
7	15,159	1,1	99,5	1,1	99,4
8	8,254	0,5	100,0	0,6	100,0
9	1				
=====					
	1,371,061	100	100	100	100

The share of unambiguous tokens has now risen by more than 4% from 42,7% to 48,1%. The most pronounced effect, however, concerns the reduction of cohorts with lots of readings. Maximal cohort size is now N(r)=9 and even of this there is just one instance. The jumps at N(r)=7,8 are due to the words för, så. Their impact is clearly visible in Table 4.

Table 4. Number of readings for word form types after local disambiguation.

N(r)	N(w)	%	cum-%	%(-0)	cum-%(-0)
0	11,193	11,7	11,7	-	
1	63,498	66,6	78,3	75,5	75,5
2	14,905	15,6	93,9	17,7	93,2
3	4,722	5,0	98,9	5,6	98,8
4	757	0,8	99,7	0,9	99,7
5	198	0,2	99,9	0,2	99,9
6	45				
7	6				
8	2				
9	1				
=====					
	95,327	100	100	100	100

The one word remaining with nine readings, renskrubgade, has a certain interest. These readings arise since ren ‘clean’, rens ‘offal’, and rensk ‘relating to Rhine’ are suggested as first parts of compounds, and ambiguous past participles of the verbs skrubba ‘scrub’, krubba ‘eat’ (substandard), and rubba ‘upset’ as second parts. Only two word form types have eight readings, så and meningsfyllda.

The 8,254 tokens with eight readings (Table 3) reduce to two types almost all of which are instances of the high-frequency word så. The 15,159 instances with seven readings reduce to six types almost all of which represent för, etc.

The figures given in Tables 3, 4 are based on local disambiguation using just the two disambiguation principles stated above. There are additional principles of local disambiguation,

plus the whole system of context-sensitive disambiguation in terms of CGP (Karlsson, in preparation). A precision rate well exceeding 95% is envisaged, 85% being the rate achieved with the current rudimentary system (January, 1992). Note, in comparison, that CGP disambiguation of English has reached a precision rate of 94-97%, depending upon text type (Voutilainen 1991). For both Swedish and English, recall exceeds 99,5%, i.e. less than 1 cohort of 200 does not contain the proper reading among the ones suggested by SWETWOL or ENGTWOL, respectively.

13. Preprocessing

Robust automatic morphological analysis requires preprocessing of the data fed to the analyzer. Relevant practical problems that need treatment are i.a. the conversion to ASCII characters, the distinction between upper and lower case, the dissociation of periods, commas, etc. from the words they are appended to, and the separation of those abbreviations terminated by a period that are sentence-internal from those that are sentence-final. The last problem is important from the viewpoint of context-sensitive disambiguation. This should be performed on chunks bounded by dependable sentence delimiters. E.g., initials of first names in clauses like N. A. Chomsky is a linguist. should be spotted as sentence-medial, as should abbreviations that are not sentence-final.

Such functions are performed on Swedish text by a preprocessing program (written in Benny Brodda's programming language Beta). We illustrate how Swedish preprocessing works by providing an example of a simplified laboratory sentence before (i) and after (ii) preprocessing:

(i) A. Norén var bl. a. lingvist.

(ii) *a. *nor`n var bl. a. lingvist \$.

Upper-case letters are converted to sequences of an asterisk followed by the corresponding lower-case letter. Sentence-final delimiters are made into quasi-words of their own headed by the character "\$", \$. Abbreviations like bl. and a. are in the lexicon with and without a final period.

In the SWETWOL lexicon, upper case is also represented as a sequence of an asterisk followed by the corresponding lower-case letter. Thus, the names Carl, Bildt, NATO, WordPerfect are rendered as *carl, *bildt, *n*a*t*o, *word*perfect. Some abbreviations occur in upper as well as lower case, both variants are included in SWETWOL (e.g. *a*b - ab, *a*i*d*s - aids).

Prior to this linguistic preprocessing, hyphenated words should have been dehyphenated, and paragraph boundaries should have been marked by the designated character "@". The linguistic preprocessor operates on paragraphs, i.e. chunks bounded by @.

Fixed phrases and idioms pose a major problems for high-quality morphological analysis and parsing. E.g., it makes little sense to deliver an analysis of the expression bit för bit 'piece by piece, piecemeal' which for bit gives the correct N reading but also the contextually spurious V IMP reading (cf. the verb bita 'bite'), and which furthermore gives all eight readings of för. Rather, bit för bit should get one reading only, ADV. In addition, such an immediate unambiguous correct analysis would minimize the work load for the disambiguation and parsing modules standing on top of SWETWOL.

Determining which expressions count as fixed phrases and which should be analysed by ordinary rule-based morphology obviously is a difficult and time-consuming task. For

SWETWOL purposes, some 2,500 such expressions have been identified. Of course, this list is from conclusive, and it has not yet been included in SWETWOL. The plan is (i) to have such expressions identified as quasiwords by the preprocessor, e.g. bit=för=bit, and (ii) to add entries with this outlook as lexical items to the SWETWOL lexicon, supplied with the appropriate morphological descriptions. This would yield SWETWOL output cohorts like:

(“<<bit=för=bit>>”
 (“bit=för=bit” ADV))

14. Applications

SWETWOL is intended to be used as a basic morphological module in parsing systems for Swedish. A comprehensive parser for Swedish, SWECG, is being designed at the Department of General Linguistics, University of Helsinki, using the Constraint Grammar formalism and program (Karlsson, in preparation; also cf. Karlsson & al. 1992).

SWETWOL is also useful for tagging texts and is presently used as a tool in the Stockholm-Umeå Corpus project started in 1990 that aims at compiling a 1 million word corpus of analyzed Swedish. SWETWOL is also going to be used in a Finland-Swedish corpus project started in 1991 at the Department of Scandinavian Languages and Literature, University of Helsinki.

SWETWOL is being used in a project trying to create an automatic analyzer for phonetic transcription of Swedish text with a prospective use in text-to-speech environments (Magnuson, Granström, Carlson, and Karlsson 1990).

Because of its morphological accuracy, especially in regard to compound formation, SWETWOL, or rather a reduced version stripped i.a. of the grammatical features, could be used as a spelling corrector.

Due to its morphological capacity and the property of retrieving base-forms, SWETWOL also has obvious applications in the domain of information storage and retrieval, including text indexing and abstracting.

15. Becoming a user of SWETWOL

A public version of SWETWOL, including preprocessing and local disambiguation, is available in the University of Helsinki Language Corpus Server (UHLCS) for direct network access. UHLCS is a network of Unix workstations containing corpora and programs for linguistic research, run by the Department of General Linguistics and the Research Unit for Computational Linguistics.

This gives the user an opportunity to analyze his/her own Swedish texts. UHLCS can be used from remote sites, given access to TELNET networking facilities. Text files are sent over the network to UHLCS, the remote user then applies e.g. SWETWOL to the texts, and finally dumps the result files to her/his local machine.

Contact the author for getting access to these facilities. Addresses are given at the end of the paper.

References

Allén, S. 1970. Nusvensk frekvensordbok baserad på tidningstext. 1. Graford. Homografkomponenter. Data Linguistica 1, Almqvist & Wiksell International, Stockholm.

-- 1972. 10000 i topp. Ordfrekvenser i tidningstext. Almqvist & Wiksell Förlag AB, Stockholm.

Allén, S., Eeg-Olofsson, M., Gavare, R. and Sjögreen, C. 1981. Svensk baklängesordbok. Esselte studium, Nacka.

Berg, S. 1978. Olika lika ord. Svenskt homograflexikon. Data Linguistica 12, Almqvist & Wiksell International, Stockholm.

Blåberg, O. 1984. Svensk böjningsmorfologi: en tvånivåbeskrivning. Master's Thesis, Department of General Linguistics, University of Helsinki.

-- 1988. A Study of Swedish Compounds. Department of General Linguistics, University of Umeå, Report Nr. 29.

Borin, L. 1986. "Swedish Two-level Morphology: Some Remarks". UCDL-R-86-1, Uppsala University, Center for Computational Linguistics.

Brodda, B. 1983. "An Experiment with Heuristic Parsing of Swedish". In Proceedings of the First Conference of the European Chapter of the ACL, Pisa, pp. 66-73.

Collinder, B. and Svenblad, R. 1987. Förkortningsordbok. Andra utökade upplagan. Liber, Kristianstad.

Eeg-Olofsson, M. 1988. "A Morphological Prolog System for Swedish Based on Analogies". In Papers from the First Nordic Conference on Text Comprehension in Man and Machine, ed. Ö. Dahl and K. Fraurud, Institute of Linguistics, University of Stockholm, Stockholm, pp. 49-62.

Einarsson, J. 1978. Talad och skriven svenska. Lundastudier i nordisk språkvetenskap, Serie C Nr 9. Walter Ekstrand Bokförlag, Lund.

Ejerhed, E. 1988. "Finding Clauses in Unrestricted Text by Finitary and Stochastic Methods". In Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas, ACL, pp. 219-227.

-- 1990. "A Swedish Clause Grammar". In Papers from the Seventh Scandinavian Conference on Computational Linguistics, ed. J. Pind and E. Rögnvaldsson, Institute of Lexicography and Institute of Linguistics, University of Iceland, pp. 14-25.

Ejerhed, E., and Bromley, H. 1986. "A Self-extending Lexicon: Description of a Word Learning Program". In Papers from the Fifth Scandinavian Conference on Computational Linguistics, Department of General Linguistics, University of Helsinki, Publications No. 15, ed. F. Karlsson, pp. 59-72.

Ejerhed, E. and Wennstedt, O. 1989. "Preliminärt förslag till märkning av ord i Stockhom - Umeå corpus". Manuscript, Department of Linguistics, University of Umeå.

Från rondell till gräddfil. Nya ord i svenskan från 40-tal till 80-tal. Svenska språknämnden. Esselte studium, Stockholm 1989.

Gellerstam, M. 1987. "Svenska Akademiens ordlista - elfte upplagan". Språkvård 2/1987:6-13.

-- 1988. "Verb Syntax in a Dictionary for Second-Language Learning". In Studies in Computer-Aided Lexicology, Almqvist & Wiksell International, Stockholm, pp. 103-122.

Hellberg, S. 1978. The Morphology of Present-Day Swedish. Word-Inflection, Word-Formation, Basic Dictionary. Data Linguistica 13, Almqvist & Wiksell International, Stockholm.

Järborg, J. 1988. "Towards a Formalized Lexicon of Swedish". In Studies in Computer-Aided Lexicology, Almqvist & Wiksell International, Stockholm, pp. 140-158.

Källgren, Gunnel 1984a. "HP: A Heuristic Parser for Swedish". In De nordiska datalingvistikdagarna 1983, ed. A. Sågvall Hein, Uppsala universitet, pp. 155-162.

-- 1984b. Automatisk excerpering av substantiv ur löpande text. Ett möjligt hjälpmedel vid datoriserad indexering? IRI-rapport 1984:1, Institutet för Rättsinformatik, Stockholms universitet.

-- 1988. "What good is Syntactic Information in the Lexicon of a Syntactic Parser?" Nordiske Datalingvistikdage og Symposium for datamatstøttet leksikografi of terminologi, Proceedings, LAMBDA Nr. 7, Institut for Datalingvistik, Handelshøjskolen i København, pp. 1-12.

-- 1990a. "Parsing without Lexicon: the MorP System". Manuscript, Department of Computational Linguistics, University of Stockholm.

-- 1990b. "Making Maximal Use of Surface Criteria in Large-scale Parsing: the MorP Parser". Manuscript, Department of Computational Linguistics, University of Stockholm.

Karlsson F. 1989. "Parsing and Constraint Grammar". Unpublished paper, Department of General Linguistics, University of Helsinki.

-- 1990. "Constraint Grammar as a Framework for Parsing Running Text". In Papers presented to the 13th International Conference on Computational Linguistics, Helsinki, ed. H. Karlgren, Vol. 3, pp. 168-173.

-- 1992a. "Considerations for the Design of a General-Purpose Parser". In Karlsson, Voutilainen, Heikkilä, and Anttila (1992, ed.).

-- 1992b. "User's Manual for the Constraint Grammar Parser CGP". In Karlsson, Voutilainen, Heikkilä, and Anttila (1992, ed.).

-- (in preparation) "Constraint Grammar for Parsing Swedish text". Department of General

Linguistics, University of Helsinki.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. 1991. "Constraint Grammar: A Language-independent System for Parsing Unrestricted Text, with an Application to English". In Workshop Notes from the Ninth National Conference on Artificial Intelligence (AAAI-91), Natural Language Text Retrieval, Anaheim, California.

-- 1992, ed. Constraint Grammar: A Language-Independent Parsing Formalism. In print.

Karttunen, L., Koskenniemi, K., and Kaplan, R. M. 1987. "A Compiler for Two-level Phonological Rules". In Tools for Morphological Analysis, ed. Dalrymple, M., Kaplan, R. M., Karttunen, L., Koskenniemi, K., Shaio, S., and Wescoat, M., Report No. CSLI-87-108, Center for the Study of Language and Information, pp. 1-61.

Koskenniemi, K. 1983 Two-level Morphology: A General Computational Model for Word-form Recognition and Production. Publications of the Department of General Linguistics, University of Helsinki, No. 11.

Magnuson, T., Granström, B., Carlson, R., and Karlsson, F. 1990. "Phonetic Transcription of a Swedish Morphological Analyzer". PHONUM, Reports from the Department of Phonetics, University of Umeå, 1, pp. 58-61.

Malmgren, S.-G. 1988. "On Regular Polysemy in Swedish". In Studies in Computer-Aided Lexicology, Almqvist & Wiksell International, Stockholm, pp. 179-200.

Modéer, Ivar 1989. Svenska personnamn. Anthroponymica Suecana 5. Studentlitteratur, Lund.

Nyord i svenskan från 40-tal till 80-tal. Svenska språknämnden. Esselte studium, Stockholm 1986.

Rankin, I. 1986. "SMORF - An Implementation of Hellberg's Morphology System". In Papers from the Fifth Scandinavian Conference on Computational Linguistics, Department of General Linguistics, University of Helsinki, Publications No. 15, ed. F. Karlsson, pp. 161-172.

--, ms. "SMORF User's Guide" (mimeo), Department of Computer and Information Science, Linköping University.

SAOL. See Svenska Akademiens Ordlista över svenska språket.

Sigurd, B. 1986. "Ordboken, ordlistan och några andra av Svenska Akademiens språkliga insatser under 1900-talet". In Svenska Akademien och svenska språket, ed. S. Allén, B. Loman, and B. Sigurd, Norstedts, Stockholm, pp.145-231.

--, ms. "Particle verbs and their treatment in the MT-system SWETRA". Department of Linguistics, Lund University.

Sjögreen, C. 1988. "Creating a Dictionary from a Lexical Database". In Studies in Computer-Aided Lexicology, Almqvist & Wiksell International, Stockholm, pp. 299-338.

SOB. See Svensk ordbok.

Svensk ordbok. Esselte Studium, Stockholm 1986.

Svenska Akademiens ordlista över svenska språket. 11. upplagan. Norstedts förlag, Stockholm 1986.

Svenska ortnamn - uttal och stavning. Utgiven av Lantmäteriverket och Svenska språknämnden. Norstedts, Stockholm 1991.

Sågvall Hein, A. 1983. A Parser for Swedish. Status Report for Sve.Ucp. February 1983. UC DL-R-83-2. Center for Computational Linguistics, Uppsala University.

-- 1987. "Parsing by means of Uppsala Chart Processor (UCP)". In Natural Language Parsing Systems, ed. L. Bolc, Berlin & Heidelberg.

-- 1988. "Towards a Comprehensive Swedish Parsing Dictionary". In Studies in Computer-Aided Lexicology, Almqvist & Wiksell International, Stockholm, pp. 268-298.

-- 1990. "Lemmatizing the Definitions of Svensk Ordbok by Morphological and Syntactic Analysis. A Pilot Study". In Papers from the Seventh Scandinavian Conference on Computational Linguistics, ed. J. Pind and E. Rögnvaldsson, Institute of Lexicography and Institute of Linguistics, University of Iceland, pp. 342-357.

Sågvall Hein, A., and Ahrenberg, L. 1985. A Parser for Swedish. Status Report for SVE.UCP. June 1985. UC DL-R-85-2. Center for Computational Linguistics, Uppsala University.

Sågvall Hein, A. and Sjögreen, C. (in print) "Ett svenskt stamlexikon för datamaskinell analys. En översikt.". In Svenskans beskrivning, 18.

Thelander, K. 1986. Politikerspråk i könsperspektiv. Ord och stil. Språkvårdssamfundets skrifter 17. Liber, Stockholm.

Thorell, O. 1984. Att bilda ord. Skriptor, Stockholm.

-- 1987. Svensk grammatik. Andra upplagan. Norstedts, Stockholm.

Voutilainen, A. 1991. "Lexical Disambiguation". In Natural Language Processing for Information Retrieval Purposes, ed. F. Karlsson & al., SIMPR Document No. SIMPR-RUCL-1990-13.4e. Department of General Linguistics, University of Helsinki. (in print)

Appendix 1. The analysis of a sentence by SWETWOL

(Readings that will be discarded by local disambiguation are marked by a prefixed star, “*”.)

(“<<*i>>”

(“*i” PREP)

(“*i” ADV))

(“<<jämförelse>>”

(“jämförelse” N UTR INDEF SG NOM)

* (“jämförelse” DER-else N UTR INDEF SG NOM)

* (“jäm_förelse” DER-else N UTR INDEF SG NOM))

(“<<med>>”

(“med” PREP)

(“med” ADV))

(“<<de>>”

(“de” PRON UTR/NEU DEF/INDEF PL NOM)

(“de” ART UTR/NEU DEF PL NOM))

(“<<starka>>”

(“stark” A UTR/NEU DEF/INDEF PL NOM)

(“stark” A UTR/NEU DEF SG NOM))

(“<<motiv>>”

(“motiv” N NEU INDEF SG/PL NOM))

(“<<som>>”

(“som” KONJ)

(“som” PRON UTR/NEU DEF/INDEF SG/PL NOM))

(“<<anför>>”

(“anför” V PASS PRES))

(“<<för>>”

(“för” PREP)

(“för” AD-A)

(“för” ADV)

(“för” N UTR INDEF SG NOM)

(“för” KONJ)

(“föra” V ACT PRES)

(“föra” V ACT IMP))

(“<<ett>>”

(“ett” NUM NEU INDEF SG NOM)

(“ett” ART NEU INDEF SG NOM))

(“<<omfattande>>”

* (“om_fatta” V ACT PCP1 UTR/NEU DEF/INDEF SG/PL NOM)

* (“om_fatta” DER/-nde N NEU INDEF SG NOM)

(“omfatta” V ACT PCP1 UTR/NEU DEF/INDEF SG/PL NOM)

(“omfatta” DER/-nde N NEU INDEF SG NOM))

(“<<nordiskt>>”

(“nordisk” A NEU INDEF SG NOM)

* (“nordisk” DER-isk A NEU INDEF SG NOM))

(“<<biståndssamarbete>>”

(“bistands_samarbete” N NEU INDEF SG NOM)

* (“bistands_sam_arbete” N NEU INDEF SG NOM)

* (“bi_stand_samarbete” N NEU INDEF SG NOM)

* (“bi_stand_sam_arbete” N NEU INDEF SG NOM))

(“<<såväl>>”

(“såväl” KONJ)

* (“så_väl” N NEU SG NOM))

(“<<i>>”

(“i” PREP)

(“i” ADV))

(“<<det>>”

(“det” PRON NEU DEF/INDEF SG NOM)

(“det” ART NEU DEF SG NOM))

(“<<nya>>”

(“ny” A UTR/NEU DEF/INDEF PL NOM)

(“ny” A UTR/NEU DEF SG NOM))

(“<<nordiska>>”

(“nordisk” A UTR/NEU DEF/INDEF PL NOM)

(“nordisk” A UTR/NEU DEF SG NOM)

* (“nordisk” DER-isk A UTR/NEU DEF/INDEF PL NOM)

* (“nordisk” DER-isk A UTR/NEU DEF SG NOM))

(“<<samarbetsprogrammet>>”

(“samarbets_program” N NEU DEF SG NOM)

* (“sam_arbets_program” N NEU DEF SG NOM))

(“<<som>>”

(“som” KONJ)

(“som” PRON UTR/NEU DEF/INDEF SG/PL NOM))

(“<<i>>”
 (“i” PREP)
 (“i” ADV))

(“<<andra>>”
 (“andra” <<Ord>> A UTR/NEU DEF/INDEF SG/PL NOM)
 (“annan” PRON UTR/NEU DEF/INDEF SG/PL NOM)
 (“andra” V ACT INF)
 (“andra” V ACT IMP))

(“<<nordiska>>”
 (“nordisk” A UTR/NEU DEF/INDEF PL NOM)
 (“nordisk” A UTR/NEU DEF SG NOM)
 * (“nordisk” DER-isk A UTR/NEU DEF/INDEF PL NOM)
 * (“nordisk” DER-isk A UTR/NEU DEF SG NOM))

(“<<avtal>>”
 (“avtal” N NEU INDEF SG/PL NOM))

(“<<och>>”
 (“och” KONJ))

(“<<överenskommelser>>”
 (“överenskommelse” N UTR INDEF PL NOM)
 * (“överenskommelse” DER-else N UTR INDEF PL NOM))

(“<<finner>>”
 (“finna” V ACT PRES))

(“<<kommittén>>”
 (“kommitté” N UTR DEF SG NOM))

(“<<inte>>”
 (“inte” ADV))

(“<<att>>”
 (“att” INFMARK)
 (“att” KONJ))

(“<<programmet>>”
 (“program” N NEU DEF SG NOM))

(“<<har>>”
 (“har” N UTR/NEU INDEF SG/PL NOM)
 (“ha” V ACT PRES))

(“<<fått>>”

(“få” V ACT SUPINE))

(“<<en>>”

(“en” ADV)

(“en” NUM UTR INDEF SG NOM)

(“en” N UTR INDEF SG NOM)

(“en” PRON UTR INDEF SG NOM)

(“en” ART UTR INDEF SG NOM))

(“<<utformning>>”

(“utforma” DER/-ning N UTR INDEF SG NOM))

(“<<som>>”

(“som” KONJ)

(“som” PRON UTR/NEU DEF/INDEF SG/PL NOM))

(“<<innebär>>”

* (“inne_bära” V ACT PRES)

* (“inne_bära” V ACT IMP)

* (“inne_bär” N NEU INDEF SG/PL NOM)

(“innebära” V ACT PRES)

(“innebära” V ACT IMP))

(“<<någon>>”

(“någon” PRON UTR INDEF SG NOM))

(“<<stark>>”

(“stark” A UTR INDEF SG NOM))

(“<<satsning>>”

(“satsa” DER/-ning N UTR INDEF SG NOM))

(“<<på>>”

(“på” PREP)

(“på” ADV))

(“<<en>>”

(“en” ADV)

(“en” NUM UTR INDEF SG NOM)

(“en” N UTR INDEF SG NOM)

(“en” PRON UTR INDEF SG NOM)

(“en” ART UTR INDEF SG NOM))

(“<<i>>”

(“i” PREP)

(“i” ADV))

(“<<egentlig>>”
 (“egentlig” A UTR INDEF SG NOM))

(“<<mening>>”
 (“mena” DER/-ning N UTR INDEF SG NOM)
 (“mening” N UTR INDEF SG NOM))

(“<<gemensam>>”
 * (“gem_ensam” A UTR INDEF SG NOM)
 (“gemensam” A UTR INDEF SG NOM))
 * (“gemen_simma” V ACT PAST)

(“<<nordisk>>”
 (“nordisk” A UTR INDEF SG NOM)
 * (“nordisk” DER-isk A UTR INDEF SG NOM))

(“<<biståndsverksamhet>>”
 (“bistånds_verksamhet” N UTR INDEF SG NOM)
 * (“bistånds_verksamhet” DER-het N UTR INDEF SG NOM)
 * (“bi_stånds_verksamhet” N UTR INDEF SG NOM)
 * (“bi_stånds_verksamhet” DER-het N UTR INDEF SG NOM))

Acknowledgements

The helpful advice of Kimmo Koskenniemi on various aspects of two-level morphology is gratefully acknowledged. SWETWOL would not have been possible to design without recourse to a machine-readable version of Svenska Akademiens Ordlista, SAOL. I am most grateful to Språkdata, University of Gothenburg, especially to Christian Sjögreen and Martin Gellerstam, for making the tenth edition of SAOL available to me. Helpful comments have been made also by Sture Berg, Benny Brodda, Gunnar Eriksson, Gunnel Källgren, as well as two anonymous NJL referees. John Giertz, Esselte, provided me with good material for testing compound analysis performance. Timo Järvinen and Kari Pitkänen have given me valuable assistance in the compilation of the SWETWOL lexicon and preprocessor and during the time-consuming testing phase. Taina Rantala and Outi Sihvola excellently compiled and scanned several test corpora.

Address of the author

Department of General Linguistics
PB 9, FI-00014 University of Helsinki
www.ling.helsinki.fi/~fkarlsso