On the Coverage of a Morphological Analyser based on "Svensk Ordbok" [A Dictionary of Swedish]

Anna Sågvall Hein
Uppsala University

Introduction

In the project a *Lexicon-oriented Parser for Swedish* a stem dictionary (Sågvall Hein & Sjögreen 1991; Sjögreen, forthcom.) covering the 58,536 entry lemmas of *Svensk Ordbok* (1986) along with a complete inflectional grammar of Swedish (Sågvall Hein, forthcom.) was generated. This language description together with the *Uppsala Chart Processor*, UCP (Sågvall Hein 1987) constitute a morphological analyzer of Swedish, henceforth referred to as *SMU*, short for Swedish Morphology in the UCP framework.

So far, there are no word formation rules in the SMU grammar, and words outside the scope of Svensk Ordbok don't get an analysis[1]. Eventhough closed in its present version, the coverage of SMU is well-defined; prior to any processing we may consult Svensk Ordbok to find out for any word form whether it will get an analysis or not; the dictionary provides an intuitive, familiar format through which we may explore the (present) competence of the SMU analyser without any prior knowledge of its formalisms or operation. SMU is also well-defined in the sense, that for any of its lemmas, Svensk Ordbok provides links to the corresponding lexemes (basic senses), and for each lexeme a definition.

In our ongoing work on a machine-tractable dictionary for Swedish, we are approaching problems concerning the distinction between general and domain specific vocabulary, and the present coverage of SMU is our starting-point for delimiting a general Swedish vocabulary. For an evaluation of the generality of the dictionary, the analyser has been applied to different sets of Swedish text. For one of them, consisting of the 10,224 most frequent types of the 7,3 million word newspaper corpus of The Language Bank (Gellerstam 1989) the words outside he scope of the analyser have been examined at some detail. Here we will present the results achieved so far, and also discuss their impact on our continued work on the dictionary.

First, however, we will briefly characterize the SMU analyser with regard to morphological descriptions, and dictionary representation of inflection.

The morphological descriptions of the SMU analyser

The morphological descriptions generated by the analyser are expressed as attribute-value structures (Sågvall Hein & Ahrenberg 1985; cf. directed acyclic graphs, dags, for short, Shieber 1986). For a first illustration, we present the description of the noun *festernas* [*of the parties*] (see fig. 1).

Proceedings of NODALIDA 1991, pages 119-131

It comprises four general attributes, i.e. LEM for lemma, WORD.CAT for word category (part of speech), DIC.STEM for dictionary stem, and INFL for inflection type), and, four attributes specific to the nouns i.e. GENDER, NUMBer, FORM (species), and CASE. The general attributes are present in the descriptions of all the words, regardless of part of speech (noun, adjective, pronoun, verb, adverb, numeral, preposition, conjunction, interjection, article, and infinitive marker).

```
FESTERNAS   :
(* = (       LEM=FEST.NN
              WORD.CAT=NOUN
              INFL=PATTERN.FILM
              DIC.STEM=FEST
              GENDER=UTR
              NUMB=PLUR
              FORM=DEF
              CASE=GEN)
```

Figure 1. An analysis of the noun festernas [of the parties]

The value of the lemma attribute is identical to the basic form of the lemma with a (two letter) word class marker for the distinction between homograph lemmas, i.e. *springa1.nn* [chink; slot] and *springa2.vb* [run]. In addition, the homographs are numbered as they are in Svensk Ordbok (cf. **1spring/a** subst. [noun] and **2spring/a** verb), whereby immediate reference to this background material is facilitated. If there are two homograph lemmas of the same word category, the homograph number alone will keep them apart, e.g. *bok1.nn* [book] (plur. *böcker*) and *bok2.nn* [beech (tree)] (plur. *bokar*). The well-defined lemma marker supports the distinction between external and internal homography, a basis for subsequent lemmatization. Further, it provides a basis for the selection of domain tuned lemma dictionaries from a general dictionary; the lemmas specific to the domain are recognized by the morphological analysis of texts typical of that domain. The dictionary stem attribute may serve the same function in building a domain tuned stem dictionary from a general stem dictionary.

The value of the inflection attribute is a *pattern word* which is also the name of an *inflectional rule* defined in the grammar. The inclusion of this information in the morphological descriptions provides a basis for frequency studies of inflectional types in current text.

In addition to the general attributes, each part of speech (except for the prepositions and the infinitive marker) is characterised by its own set of attribute value pairs. In fig. 2 we present the morphological descriptions resulting from the analyses of an adjectival paradigm, the adjective *festlig [festive; grand]*. The descriptions illustrate, among other things, the representation of *internal homography*. The form *festliga* gets two descriptions, one corresponding to an analysis of it as a definite singular positive form of the adjective *festlig.nn*, the other one as a plural positive form of the same adjective. In other words, this is a case of internal homography. The forms differ with regard to number (singular versus plural) and form (definite in the singular case, unstated in plural). The description is *underspecified* in the sense that the form value is left out in plural with the effect that it will unify with definite as well as with indefinite contexts in a unification-based syntactic analysis; it allows for a definite or an indefinite reading (cf. Karlsson forthcom. representing underspecification by means of composite values, e.g. DEF/INDEF). A fully specified representation would have implied the establishement of two descriptions, one plural indefinite and one plural definite. There are many such cases in the Swedish inflectional system, consequently, underspecification has a substantial impact on representational economy. In Table 1 we summarize the attributes and values specific to the different parts of speech.

FESTLIG	:	*FESTLIGT*	:
(* = (	LEM=FESTLIG.AV WORD.CAT=ADJ INFL=PATTERN.BLEK DIC.STEM=FESTLIG GENDER=UTR NUMB=SING FORM=INDEF COMP=POS)	(* = (	LEM=FESTLIG.AV WORD.CAT=ADJ INFL=PATTERN.BLEK DIC.STEM=FESTLIG GENDER=NEUTR NUMB=SING FORM=INDEF COMP=POS)
FESTLIGA	:		
(* = (	LEM=FESTLIG.AV WORD.CAT=ADJ INFL=PATTERN.BLEK DIC.STEM=FESTLIG COMP=POS NUMB=SING FORM=DEF)	(* = (	LEM=FESTLIG.AV WORD.CAT=ADJ INFL=PATTERN.BLEK DIC.STEM=FESTLIG COMP=POS NUMB=PLUR)
FESTLIGE	:	*FESTLIGARE*	:
(* =	(LEM=FESTLIG.AV WORD.CAT=ADJ INFL=PATTERN.BLEK DIC.STEM=FESTLIG GENDER=UTR NUMB=SING FORM=DEF SEX=MASC COMP=POS)	(* = (	LEM=FESTLIG.AV WORD.CAT=ADJ INFL=PATTERN.BLEK DIC.STEM=FESTLIG COMP=COMP
FESTLIGAST	:	*FESTLIGASTE*	:
(* = (	LEM=FESTLIG.AV WORD.CAT=ADJ INFL=PATTERN.BLEK DIC.STEM=FESTLIG COMP=SUP FORM=INDEF))	(* = (	LEM=FESTLIG.AV WORD.CAT=ADJ INFL=PATTERN.BLEK DIC.STEM=FESTLIG COMP=SUP FORM=DEF))

Figure 2. Analyses of the adjective festlig [festive; grand]

Part of speech	Attribute	Values		
NN	GENDER	Neutr	Utr	
	NUMB	Sing	Plur	
	FORM	Indef	Def	
	CASE	Basic	Gen	
	PROPR	+		
	ABBREV	+		
AV	GENDER	Neutr	Utr	
	NUMB	Sing	Plur	
	FORM	Indef	Def	
	COMP	Pos	Comp	Sup
	FUNC	Attr	Pred	
	SEX	Masc		
	(+ CASE for the description of nominalized adjectives)			
VB	TENSE	Pres	Sup	Pret
	INFF	Inf	Pp	Ap
	VOICE	Pass	Act	Depon

	IMP	+		
	CONJ	+		
	(+ NUMB, GENDER, FORM, FUNC, SEX, CASE for the description of the participles)			
AL	GENDER	Neutr U	tr	
	NUMB	Sing	Plur	
	FORM	Indef	Def	
PN	PRON.TYPE	Pers	Poss	Rel
	ATTR.TYPE	Select	Quant	Comp
	DET.TYPE	Tot	Det	
	GENDER	Neutr	Utr	
	NUMB	Sing	Plur	
	FORM	Indef	Def	
	CASE	Basic	Gen	Obj
	SEX	Masc		
	PARTITIV	+		
	DUAL	+		
	NOUN.INDEF	+		
NL	ATTR.TYPE	Select	Quant	
	GENDER	Neutr Utr		
	NUMB	Sing	Plur	
	FORM	Indef	Def	
	CASE	Basic	Gen	
	SEX	Masc		
AB	COMP	Pos	Comp	Sup
KN	SUBJU	+		

Prepositions, interjections, and the infinitive marker have no attributes.

Table 1. An Overview of the attributes assigned by the SMU analyser

Inflection in Svensk Ordbok and in the SMU dictionary

In fig. 3 we present the inflectional variety of the *-ar-declination* for an illustration of the relation between the inflectional format of SOB and that of the SMU stem dictionary. In the simplest case, there is only one stem, in SOB and in SMU, and, further, the stem is identical to the basic (lemma) form of the word (see **1** and **2**). In such cases, the entry form of SOB represents at the same time the lemma and the stem, whereas in the SMU formal dictionary the two concepts have to be individually represented. The pattern rules of the SMU dictionary cover the inflectional information given in SOB in terms of word class and significant endings, and, also, the grammatical background information to which these morphological keys refer, i.e. morphotactic rules determining the inflectional behaviour of the nouns. Sometimes, SOB recognizes a stem identical to an initial substring of the lemma form and delimits it from the succeeding ending by a slash, as in **3 – 8**, and **11**. The stem concept thus adopted is the *technical stem* (Hellberg 1978). The SMU analyser treats these cases in a different way, i.e. by means of a general rewriting rule (**4 – 8**) or stem handling operations in the pattern rules (as in **3** and **11**). In the first case, the non-vowel stem alternant is regarded as the canonical form of the stem, and its vowel counterpart is reduced to this form by a secondary vowel deletion rule[2]. Thus there is only one stem in dictionary, and, what is more important, we don't have to accept as endings strings such as *eln, lar, nen, nart* etc. In the second case, SMU analyses the stem in two steps, i.e. as a *dictionary stem* followed by a stem building element, e.g. *pojk-e, dröm-m, läm-mel*. A word form such as *pojken* is analysed *pojk-e+n, lämmeln* as *läm-mel+n*, etc. where the corresponding pattern rules (pattern.gosse and pattern.lämmel) are

responsible for the recognition of the stem building segments and their distribution in the paradigm (see further Sågvall Hein, forthcom.).

In all, there are 132 (stem) pattern rules for the nouns, 40 for the adjectives, 89 for the verbs, 1 for the articles, 30 for the pronouns, 5 for the numerals, 9 for the adverbs, 2 for the conjunctions, and one each for the prepositions, the interjections, and the infinitive marker. In most cases the analysis is based on one stem[3], and in these cases, the stem with its pattern rule is a sufficient characterisation of the inflectional behaviour of the lemma as such. If, on the other hand, the lemma is represented by more than one stem in the dictionary (cf. **17** and **18** in fig. 3), the *set of stems* involved along with their *pattern words* determine the inflection of the lemma, the lemma inflection as opposed to the stem inflection. Frequency data on the inflectional types of the SMU words have been presented elsewhere (Sågvall Hein & Sjögreen 1991).

SOB	stem	SMU lemma	pattern
1 gård subst. *~en ~ar*	gård	gård.nn	.stol
2 sjö subst. *~n ~ar*	sjö	sjö.nn	.fru
3 pojk/e subst. *~en ~ar*	pojk-e	pojke.nn	.gosse
4 speg/el subst. *~eln ~lar*	speg(e)l	spegel.nn	.nyckel
5 sock/en subst. *~nen ~nar*	sock(e)n	socken.nn	.öken
6 bott/en subst. *~nen* el. =, ~nar	bott(e)n	botten.nn	.botten
7 myrt/en subst. *~en ~nar*	myrt(e)n	myrten.nn	.fröken
8 fing/er subst. *~ret* el. ~em *~rar*	fing(e)r	finger.nn	.finger
9 dröm subst. *~men ~mar*	dröm dröm-m	dröm.nn	.kam
10 läm/mel subst. *~meln ~lar*	läm-mel läm-l	lämmel.nn	.lämmel
11 sum/mer subst. *~mern ~rar*	sum-mer sum-r	summer.nn	.hummer
12 sommar subst. *~(e)n somrar*	som som-mar	sommar.nn	.sommar
13 hammare subst. *hammar(e)n*, = el. *hamrar*, best. plur.	ham-r ham-mar ham-mare	hammare.nn	.kammare
14 himmel subst. *himmel(e)n* el. *himlen, himlar*	him-mel him-l	himmel.nn	.himmel
15 afton subst. *aftonen aftnar*	aft-on aft-n	afton.nn	.morgon
16 djävul subst. *~en djävlar*	djäv-ul djäv-l	djävul.nn	.djävul
17 moder äv. [1]**mor** subst. *modern mödrar*	moder mödrar mor	moder.nn moder.nn moder.nn	.moder .mödrar .far
18 tok subst. *~en ~ar* el. **tok/er** *~ern ~ar*	tok toker	tok.nn tok.nn	.stol .toker
19 stadgar subst. plur.	stadg	stadgar.nn	.vägnar

Figure 3. Inflection in SOB and in SMU. An example: the -ar-declination.

The Scope of Svensk Ordbok and the SMU analyser

For the recognition, and subsequent examination, of missing entries in the dictionary, we apply the SMU analyser to different sets of Swedish text. So far, we have analysed four text materials of substantial size, i.e. the 10,224 most frequent types of the 7,3 million word newspaper corpus of the Swedish Language Bank (Gellerstam 1989), referred to as *PressFreq*, the pharmacological text of the Swedish drug catalogue *FASS* (1985) (660,000 current words), the *Professional Prose* corpus of the *Skrivsyntax* project (Teleman 1974), referred to as *ProfProse* (78,0366 current words), and, finally, the corpus of the definitions of Svensk Ordbok, referred to as *DefVoc* (360,144 current words). ProfProse consists of four types of text of equal size (textbooks, newspapers, debate books, and brochures).

corpus	tokens	types	ty/to	covered	uncovered
Press-Freq	5,091,965	10,224	0,02	8,424 (82%)	1,790 (18%)
Prof-Prose	78,036	13,766	0,18	10,083 (73%)	3,683 (27%)
DefVoc	360,144	43,934	0,12	31,350 (71%)	12,584 (29%)
FASS	664,314	39,884	0,06	9,767 (25%)	30,117 (75%)

Table 2. Results of the application of the SOB-based SMU analyser to four sets of Swedish text

As might be expected, the analyser covers best in relation to the high frequent words of the newspaper text and worst with respect to the highly domain-specific pharmacological text. In between comes the more general text of ProfProse and that of the definition corpus, covered, basically, to the same extent. In tables 3 to 6 we present more detailed data on the results of the processing of each of the text materials. They include information on homography of three kinds, i.e. (lemma) internal, (lemma) external, and mixed (internal and external). (For each text material, we also have detailed data on the different subtypes of homographies that were found, e.g. int. AV, ext. AB/AB, ext. AB/KN, ext. AB/NN, and their frequencies.)

Number of parses	lexical coverage types	%	textual coverage tokens	%
0	1,790	17,5%	299,235	5,9%
1	6,142	60,0%	2,270,961	44,6%
2 (int.)	593	5,8%	152,334	3,0%
2 (ext.)	959	9,4%	1,391,623	27,3%
3 (int.)	158	1,5%	24,934	0,5%
3 (ext.)	191	1,9%	435,066	8,5%
3 (mix.)	206	2,0%	54,176	1,1%
4 (ext.)	33	0,3%	43,814	0,9%
4 (mix.)	69	0,7%	24,544	0,5%
5 (ext.)	14	0,1%	258,713	5,1%
5 (mix.)	54	0,5%	20,382	0,4%
6 (ext.)	6	0,1%	83,669	1,6%
6 (mix.)	7	0,1%	32,118	0,6%
7 (mix.)	1	0,0%	96	0,0%
8 (mix.)	1	0,0%	280	0,0%
Total:	10,224	100,0%	5,091,965	100,0%

Table 3. Results of the application of the SOB-based SMU analyser to PressFreq

The lemma can be unambiguously determined for 6,893 types (67,4%) and 2,447,956 tokens (48,1%). In **PressFreq words outside the scope of SOB** below, we give an account of the kinds of words that got no parses.

Number of parses	lexical coverage [4] types	%	textual coverage tokens	%
0	3,683	26,8%	6,806	8,7%
1	7,828	56,9%	38,015	48,7%
2 (int.)	800	5,8%	2,693	3,5%
2 (ext.)	829	6,0%	17,638	22,6%
3 (int.)	161	1,2%	282	0,4%
3 (ext.)	149	1,1%	6,068	7,8%
3 (mix.)	168	1,2%	758	1,0%
4 (ext.)	24	0,2%	455	0,6%
4 (mix.)	63	0,5%	384	0,5%
5 (ext.)	8	0,1%	3,282	4,2%
5 (mix.)	44	0,3%	266	0,3%
6 (ext.)	3	0,0%	1,201	1,5%
6 (mix.)	4	0,0%	186	0,2%
7 (mix.)	1	0,0%	1	0,0%
8 (mix.)	1	0,0%	1	0,0%
Total:	13,766	100,0%	78,036	100%

Table 4. Results of the application of the SOB-based SMU analyser to ProfProse

The lemma can be unambiguously determined for 8,789 types (63,8%) and 40,990 tokens (52,5%). Roughly 13% (479) of the words that got no analysis are numerical expressions.

Number of parses	lexical coverage[4] types	%
0	12,584	28,6%
1	26,348	60,0%
2 (int.)	1,654	3,8%
2 (ext.)	2,205	5,0%
3 (int.)	251	0,6%
3 (ext.)	258	0,6%
3 (mix.)	372	0,8%
4 (int.)	1	0,0%
4 (ext.)	40	0,1%
4 (mix.)	137	0,3%
5 (ext.)	9	0,0%
5 (mix.)	63	0,1%
6 (ext.)	3	0,0%
6 (mix.)	7	0,0%
7 (mix.)	1	0,0%
8 (mix.)	1	0,0%
Total:	43,934	100,0%

Table 5. Results of the application of the SOB-based SMU analyser to DefVoc

The lemmas of 28,234 of the types (64%) can be determined unambiguously. Only 430 (~3,5%) of the 0-parses are numerical expressions.[5]

Number of parses	lexical coverage[4] types	%
0	30,117	75,5%
1	7,598	19,1%
2 (int.)	874	0,2%
2 (ext.)	766	0,2%
3 (int.)	147	0,0%
3 (ext.)	116	0,0%
3 (mix.)	142	0,0%
4 (ext.)	17	0,0%
4 (mix.)	52	0,0%
5 (ext.)	6	0,0%
5 (mix.)	42	0,0%
6 (ext.)	3	0,0%
6 (mix.)	3	0,0%
8 (mix.)	1	0,0%
Total:	39,884	100,0%

Table 6. Results of the application of the SOB-based SMU analyser to FASS

The lemmas of 8,619 of the types (22%) can be determined unambiguously. 11,148 (27%) of the 0-parses are numerical expressions or hybrids of numbers, special signs, and single letters, such as *75-80, 85%, 4$8, F100, E218, C++, 0,5-0,7, 20:e* etc. (A pharmacological stem dictionary covering the non-numerical words outside the scope of SOB has been built (see Sågvall Hein et al., forthcom.)) In fig. 4 we present a drug description from FASS to illustrate the special character of this text.

Abboticin®
Abbott

Dosgranulat 200 mg ℞

Antibiotikum Grupp 7B 3005

Deklaration. *1 dosgranulat innehåller:* Erythromycin. aethylsuccin. respond. erythromycin. 200 mg, mannitol. 1,5 g, constit. et aroma q. s.
Egenskaper. Dosgranulaten innehåller erytromycinetylsuccinat motsvarande 200 mg erytromycin. Granulatet löses i litet vatten (2–3 dessertskedar=20–30 ml). Beredningsformen är speciellt avsedd för barn och är även lämplig som jourförpackning. Beredningen är sockerfri och har körsbärssmak.
Erytromycinetylsuccinat är en ester av erytromycin och efter absorption sker hydrolys till fritt aktivt erytromycin. Se f ö ABBOTICIN tabletter.
Indikationer. Se ABBOTICIN tabletter.
Kontraindikationer. Se ABBOTICIN tabletter.
Försiktighet. Se ABBOTICIN tabletter.
Graviditet och amning. Se ABBOTICIN tabletter.
Biverkningar. Se ABBOTICIN tabletter.

Dosering. *Dosen för barn beräknas efter 30–50 mg per kg kroppsvikt och dygn fördelat på 2–4 doseringstillfällen.* 1 dosgranulat=200 mg erytromycin. För barn upp till 4 kg beräknas dosen i det enskilda fallet. Vid kroppsvikt överstigande 4 kg kan om dygnsdosen fördelas på två doseringstillfällen följande schema vanligen tillämpas:

Vikt kg	Dygns-dosering dosgranulat	mg/kg/dygn	Lämplig förpackning för 10 dagars behandling
4– 7	½×2	29–50	1×30
8–14	1×2	29–50	1×30
15–24	2×2	33–53	2×30
25–34	3×2	35–48	2×30

Inträffar gastrointestinala problem rekommenderas uppdelning av dygnsdosen på 3 eller 4 administreringstillfällen. För vuxna och barn över 35 kg ges 3 dosgranulat 3 gånger per dygn. Optimal absorption erhålles om dosen intages omedelbart före måltid.
Interaktion. Se ABBOTICIN tabletter.
Förpackningar och priser. *Dosgranulat 200 mg*
30 st 55:30

Figure 4. An example of a drug description from FASS 1985

PressFreq words outside the scope of SOB

Proper nouns	1,433	(80,1%)
Abbreviations	137	(7,7%)
Compounds	127	(7,1%)
Numerical expressions	45	(2,5%)
Derivatives	20	(1,1%)
Foreign words	17	(0,9%)
Syntagmatic words	5	(0,3%)
Partial phrases	4	(0,2%)
Inflectional forms	2	(0,1%)
Total	1,790	(100%)

Table 7. Kinds of uncovered types in Pressfreq

Proper nouns. The dominating category is that of the proper nouns (including a small number of proper noun abbreviations, e.g. *ABF, AIK, DN, DDR, KFUM*). No normalization of spelling variation has been carried out, so far, so each appearance has been counted as an independent unit, e.g. *Anna, anna,* and *ANNA*; *Erik* and *Eric; Bernard* and *Bernhard, Bengtsson, Bengtson,* and *B-son, Lidingö* and *Lid-ö* etc. Roughly, half the number of the proper names refer to people (approximately, 440 first names and 280 second names.)

The high number of proper nouns, in specific, personal proper nouns, seems to be a characteristic feature of newspaper text. Most of the proper nouns that we found will be entered into the dictionary with a marking of their origin (PressFreq) as a clue to future work on domain.

Abbreviations. The abbreviation category comprises abbreviations (excl. those of the proper nouns) in various orthographic shapes, e.g. *bl.a* and *bl a, FEB, feb.* and *febr, kr* and *kr.* etc. They will all be included in the core of our dictionary (see Östling, this volume), and some of them be treated as functional core phrases (Sågvall Hein et al. 1990) and represented in the dictionary as such, e.g. *bl a* and *bl. a., d. v. s.* and *d v s* etc. The figures presented in table 9 are, however, based on individual text words, for instance *bl, bl., a, a.* etc.

Compounds. As is well-known, the Swedish compounds make up an open category, and in table 10 we present an overview of the different kinds of PressFreq compounds that were found to be outside the present scope of the SMU dictionary.

Type of compound	No of members	Examples
NN-NN → NN	97	arbetsuppgifter, hälso-
NL-NN → NN	10	andraplats, 30-talet
NN-AV → AV	6	medelstora, nordöstra
AV-NN → NN	4	nypris
NL-AV → AV	3	50-årig
NL-NN-NN → NN	2	50-årsåldern
AV-VB → VB	2	nybyggda
NN-NN → AB	1	förhoppningsvis
P-NN → AB	1	härområet
P-NN → NN	1	överåklagare
P-VB → VB	1	efteranmäld
NN-NN-VB → VB	1	text-TV-Textat
Total	129	

Table 8. PressFreq: Uncovered compounds

Following Blåberg (1988) we refer the participles to the verb category. Further, prepositions, and adverbs, are treated as members of one common category (in the compound context), denoted P.

Many of the productive compound types are indicative of domain (economy, politics, social security, sport, culture, weather, TV and broadcasting). The effect of compounding on domain is an important issue in our future work on a domain-sensitive extension of the dictionary. It is one of the criteria that should be taken into account when considering a rule-based (as opposed to a lexicalized) treatment of compounds.

Numerical expressions. The members of the numerical category are 39 expressions consisting of Arabian numerals, and hybrids of numberals, special signs, and single letters, such as *031/119840, 14.30, 25:E, and D0502.* (Some Roman numerals were also found, e.g. *VII*, but referred to the proper noun category as candidates for (parts of phrasal) proper nouns.)

Derivatives. Most of the derived words outside the scope of SMU (14 of 20) can be found in Svensk Ordboks as *morphological examples* (see table 11). This means, that their existence is confirmed, and that their meanings (definitions) should be derivable from the definitions of the words (lexemes) that they illustrate. When a lemma has more than one lexeme (definition), the morphological example tells us, on what definition the meaning of the derived word should be based, at least primarily (see *osäker*, in table 11). Five derivatives are not presented as morphological examples, but derived from one-lexeme lemmas, and so, the definition on which to base their derived meaning is uniquely determined. The remaining case, however, will cause overgeneration. *spelmässigt* is derivd from a noun with 9 definitions and an adjectival suffix with 2 definitions; the derivational power of the lexeme is in no way constrained, and we will have to consider them all equally well fit as bases of the derived words. (In all, there are 39,831 morphological examples in SOB.)

Type	Entry in SOB	Morph. ex.
avveckling avvecklingen	**avveckla** verb	+
mobbning	**mobba** verb	+
utvisning utvisningar	**utvisa** verb	+
sänkning	**sänka** verb	+ (1st lexeme)
pensionering pensioneringen	**pensionera** verb	- (1 lexeme)
enighet	**enig** adj.	+
osäkerhet	**osäker** adj.	+ (3rd lexeme)
skicklighet	**skicklig** adj.	+
trovärdighet	**trovärdig** adj.	- (1 lexeme)
öppenhet	**öppen** adj.	+ (4th lexemes)
författarinnan	**författare** subst.	+
socialdemokratisk socialdemokratiska	**socialdemokrat** subst.	+
mittfältare mittfältaren	**mittfält** subst.	- [6](1 lexeme)
spelmässigt	**spel** subst.	- (9 lexemes)
	-mässig	- (2 lexemes)

Table 9. Uncovered derivatives in PressFreq

Foreign words. To the foreign word catagory we have referred foreign words, not immediatly recognized as proper nouns. It includes function words (*der, des, die, the, til, to, with, you*) as well as content words (*chiffres, glasnost, labour, lettres, new, outs, science, télévisé, week*). We assume, that the function words, and most of the content words, are parts of phrasal proper nouns or other

phraseological expressions, and so they won't be entered as individual entries into the dictionary. *glasnost*, alone, will be entered into the SMU dictionary, and marked with respect to origin (PressFreq).

Syntagmatic words. Four missing types are examples of varying writing conventions, i.e. *iväg* (cf. *i väg*), *godnatt* (cf. *god natt*), *långtifrån* (cf. *långt ifrån*), *framförallt* (cf. *framför allt*). To the same category we refer the colloquial *gomiddag* (cf. *god middag*.) The one word variants will all be included in the SMU dictionary (the last one marked as colloquial).

Partial phrases. Four missing types are old inflectional forms appearing in phraseological expressions only, i.e. *godo* (till godo [to someone's credit etc.]; i godo [amicably etc.]), *sjöss* (till sjöss [at see]), *vintras* (i vintras [last winter]), and *somras* (i somras [last summer]).(*till godo* and *till sjöss* can be found among the examples of [1]**god** and [1]**sjö**.) The four expressions will be entered into the SMU dictionary as phrases.

Inflectional forms. Two underived inflectional forms were found to be unaccounted for, i.e. *måst* (supine of the verb *måste* [must]) and *törs* (present tense of the verb *tör/as* or *tord/as* [dare]), eventhough *törs* is part of an example of the verb. Frequent as they are found to be in newspaper text, both forms will be included in the SMU dictionary.

Conclusions

The SMU analyser, operating on Swedish text, works well as a tool for distinguishing between general vocabulary, as defined by the lemma entries of Svensk Ordbok (i.e. its explicitly defined vocabulary), and words outside that scope. As a result of the morphological analysis, members of the general vocabulary are identified and described in terms of lemma, part of speech, and form, and homographies are recognized in accordance with the lemma distinctions made in SOB.

The processing of four different Swedish materials has shown, that the SOB lexical coverage (in terms of types) ranges from 82% to 25%. The highest figures are valid for highfrequency words of newspaper text, PressFreq, and the lowest ones for highly specialized pharmacological text. In between we find some good 70% relating to general LSP (Language for Special Purpose) text.

The words outside the scope of the analyser indicate domain and type of the analysed text. The big amount of numerical expressions (and hybrids of numerals, special signs and single letters), i.e. 27% of the unanalyzed words, stand out in the pharmacological text as does that of proper nouns (close to 80% of the unanalyzed words) in PressFreq.

The PressFreq zero-parses have been examined in some detail, and categorized into: proper nouns, abbreviations, compounds, numerical expressions, derivatives, foreign words, syntagmatic words, partial phrases, and simple inflectional forms. Abbreviations, syntagmatic words, partial phrases, and inflectional forms form a, basically, closed set of a general character (in all, less than 150 items). They will all be included in the general part of the SMU dictionary (as one-word units or as phrases).

Most of the foreign words (except for *glasnost*) seem to be part of phraseological expressions (proper nouns), and so far, they will be disregarded, but *glasnost* be entered in the dictionary, marked by origin (PressFreq), as a first clue to domain. The proper nouns, forming a big, but, basically, closed and domain-related category, will be handled in the same manner. The numerical expressions, forming an open category, will be handled by means of rules, defined in the SMU grammar (see Sågvall Hein 1987).

Among the zero-parse derivatives, six types were found, i.e. verb-to-noun by means of the suffix *-ing* (the process; 8 cases), adj-to-noun by means of the suffix *-het* (presence of the property; 5 cases), noun-to-adj by means of *-isk* (the property; 2 cases), noun-to-noun by means of the suffix *-inna* (feminine; 1 case), noun-to-noun by means of *-are* (agens; 1 case), and, noun-to-adj by means

of *-mässig* (according to the noun etc.; 1 case). The first four types will be handled by means of word formation rules in the grammar, whereas the remaining two cases will be entered into the dictionary, marked by origin. This treatment is supported by SOB, presenting the first four types as morphological examples.

The most difficult category to handle is that of the compounds, being an open, productive category, with a complex semantics, dominating the zero-parses of general LSP text (see Sågvall Hein 1990). Further, compounding has a bearing on domain. In our continued work on the dictionary we will approach the problems of the compounds from the point of view of the effect of compounding on domain. The material presented by the application of the SMU analyser to text of different type and domain is a valuable source for such studies.

Notes

1 CF. SWETWOL by Karlsson (forthcom.) performing rule-based structural analysis of compounds and derivatives.

2 The secondary vowel deletion rule in part of the inflectional grammar and invoked by the dictionary search process.

3 in the sense of dictionary stem

4 Textual frequency data were not at hand when the analysis was carried out, so only lexical coverage can be accounted for here.

5 In a pilot study of a fragment of (2,500) types) of DefVoc the (572) types outside the scope of SOB were examined (see Sågvall Hein 1990).

6 Eventhough *mittfältare* doesn't appear as a morphological example, the relative *mittfältsspelare* does.

References

Blåberg, O. 1988. A study of Swedish compounds. Umeå University. Department of General Linguistics. Report No 29.

FASS. Farmaceutiska specialiteter i Sverige. 1985. [Pharmaceutical Specialties in Sweden.] LINFO.

Gellerstam, M. 1989. The Language Bank. The Department of

Computational Linguistics. University of Gothenburg.

Hellberg, S. 1978. *The morphology of present-day Swedish.* Stockholm. Karlsson, F. SWETWOL: A comprehensive morphological analyzer for Swedish. Forthcoming.

Östling, A. A Swedish Core Vocabulary for Machine Translation. This volume.

Shieber, S. 1986. An introduction to unification-based approaches to grammar. CSLI. Lecture Notes Number 4. Sjögreen, C. 1988. Creating a dictionary from a lexical database. In: *Studies in computer-aided lexicology.* Stockholm. Pp. 299-338.

Sågvall Hein, A. 1987. Parsing by means of Uppsala Chart Processor, (UCP). In: L. Bolc (ed.) *Natural language parsing systems.* Berlin & Heidelberg. Pp. 203-266.

Sågvall Hein, A. 1988. Towards a comprehensive Swedish parsing dictionary. In: *Studies in computer-aided lexicology.* Stockholm. Pp. 268-298.

Sågvall Hein, A. 1990. Lemmatising the definitions of Svensk Ordbok by morphological and syntactic analysis. A pilot study. In: J. Pind & Rögnvaldsson, E. (eds.) *Papers from the seventh Scandinavian conference of computational linguistics.* Reykjavík. Pp. 342-357.

Sågvall Hein, A. The SMU inflectional grammar. Uppsala University. Department of Linguistics. Forthcoming.

Sågvall Hein, A. & Ahrenberg, L. 1985. A parser for Swedish. Status Report for Sve.Ucp. June 1985. Uppsala University. Center for Computational Linguistics. UCDL-R-85-2.

Sågvall Hein, A. & Sjögreen, C. 1991. Ett svenskt stamlexikon för datamaskinell morfologisk analys. En översikt. [A Swedish stem dictionary for computational morphological analysis. An overview.] In: M. Thelander et al. (eds.) *Svenskans beskrivning 18.* Lund. Pp. 348-360.

Sågvall Hein, A., Östling, A. & Wikholm, E. 1990. Phrases in the Core Vocabulary. Uppsala University. Center for Computational Linguistics.

Sågvall Hein, A., Starbäck, P. & Wikholm, E. A pharmacological stem dictionary based on FASS, Pharmacological Specialties in Sweden 1985. Uppsala University. Department of Linguistics. Forthcoming.

Svensk Ordbok. 1986. *[A Dictionary of Swedish.]* Stockholm.

Teleman, U. 1974. *Manual för beskrivning av talad och skriven svenska.* Lund.

Anna Sågvall Hein
Uppsala University
Department of Linguistics
Computational Linguistics
Box 513
S-751 20 Uppsala
E-mail: uduas@seudac21.bitnet