ANNA SÅGVALL HEIN

Lemmatising the Definitions of Svensk Ordbok by Morphological and Syntactic Analysis. A Pilot Study

Lemmatizing Dictionary Definitions

Abstract

In this paper we present the results of a study of the definition vocabulary of *Svensk ordbok.* It is part of our on-going work on the generation of a machine-tractable dictionary from this dictionary, in specific, of making its definitions exploitable to a parser. Aiming, in particular, at the automatic lemmatisation of the definition vocabulary, the study includes an automatic morphological analysis of a subset of it, and an examination of its results. Two major issues were addressed, i.e. the coverage of the dictionary in relation to the definition vocabulary, and the feasibility of homograph resolution by syntactic analysis.

1 Introduction

The primary lexical source of the Swedish parser developed in the project *A Lexicon-Oriented Parser for Swedish*, *LPS*, (Sågvall Hein 1987a), is *Svensk Ordbok* 'A Dictionary of Swedish' (1986). It comprises 58,536 head words, lemmas, representing 65,568 lexemes. The *lemma* is defined as a group of word forms belonging to the same word class and the same inflectional pattern (Allén 1970: Introduction). The distinct senses that can be expressed by a lemma are referred to as *lexemes* (Allén 1981: 382).

Svensk Ordbok is drawn from a lexical data base, developed in the *Lexical Data Base*, *LDB*, project (Allén 1976). Thus, the lexical material is not only machine-readable, but also systematically organized in a flexible file handling system (Sjögreen 1988). The database was, however, primarily compiled for human use, thus presupposing linguistic background knowledge to be fully exploited. In other words, it is not readily *machine-tractable* in the sense of the word used by e.g. Wilks (1988).

The basic semantic information in *Svensk Ordbok* is provided by the *definitions of the word senses*, expressed in a subset of Swedish. In order to make this information useful to the parser, we have to make the implied linguistic knowledge explicit, and to formalize it. A step towards this goal is to parse the definitions, morphologically and syntactically. Here we will discuss the problems involved in lemmatizing the (graphic) words of the definitions automatically. The procedure involves two basic steps, i.e. an automatic morphological analysis, followed by a homograph resolution process, optionally carried out by automatic syntactic analysis. In the carrying through of such a scheme, aspects of the implied linguistic background knowledge will be brought up and concretized. In specific, the following questions are addressed:

- To what extent are the words of the definitions explicitly defined?
- To what extent are the homographies of the definitions solvable within the contexts in which they occur?

The results that we present are based on an automatic morphological analysis of a subset of the definition vocabulary.[1] First, however, we give a short description of the definitions and the underlying principles behind them.

2 The Definitions of *Svensk Ordbok*

In the semantic model developed in the *LDB* project and applied in *Svensk Ordbok*, every lexeme is supposed to have a well identifiable and relatively pregnant *kernel* meaning, around which a number of subordinate, derived meanings are grouped (Järborg 1988). The kernel senses of the lexemes are described by means of a *definition*, and, optionally, a *definition complement*, whereas for the derived meanings their relations to the kernel meaning are stated. The definition "is subject to the restriction that it should be capable of replacing the lexeme in question (the definiendum) in all syntactic or morphological contexts (sometimes with the help of some natural transformation). Given this restriction, the definition should describe the kernel sense in an analytic way, i.e. with each semantic factor being recognized during the analysis being assigned a separate word or phrase in the definition. The words of the definition should, as far as possible, be more semantically central in the lexicon than the definiendum, as a first step towards establishing an inherent defining vocabulary."(Järborg ibid.: 144).

The corpus of the definitions (disregarding definition complements) comprises 360,144 tokens and 43,184 inflectional types (distinct graphic words). This is the *primary corpus* of our study. In the pilot study reported on here, a subset of the primary corpus has been analysed comprising the 2,500 first (in alphabetical order) types, from *abborre* 'perch' to *behovet* 'the need'.

[1] By definition vocabulary, as opposed to *defining vocabulary* we simply understand the inventory of words actually used in the formulation of the definitions.

3 The Morphological Analysis

As a result of our previous work in the LPS project, we dispose of a morphological analyser of Swedish, comprising a stem dictionary covering the head words of *Svensk Ordbok*, along with a corresponding inflectional grammar (Sågvall Hein 1988; 1989). The parsing engine of the LPS parser is the *Uppsala Chart Processor*, *UCP* (Sågvall Hein 1987b).

```
AGAT :
        (* =  (WORD.CAT = VERB
              LEM = AGA.VB
              DIC.STEM = AGA
              INFL = PATTERN.ÄLSKA
              TENSE = SUP))
        (* =  (WORD.CAT = VERB
              LEM = AGA.VB
              DIC.STEM = AGA
              INFL = PATTERN.ÄLSKA
              PP = +
              GENDER = NEUTR
              ADJ.INF = STRONG
              NUMB = SING))
        (* =  (WORD.CAT = NOUN
              LEM = AGAT.NN
              DIC.STEM = AGAT
              INFL = PATTERN.FILM
              GENDER = UTR
              FORM = INDEF
              NUMB = SING
              CASE = BASIC))
```

Figure 1: An example of morphological analysis

The morphological descriptions generated by the LPS parser are represented as *Directed Acyclic Graphs*, *DAGs* (see e.g. Shieber 1986). Information on *word class*, *lemma*, *inflectional type*, and *dictionary stem*[2] is a compulsory part of each morphological description. For the rest, the information provided differs with the different word classes. For an illustration, we present an example of a morphological description generated by the morphological analyser (Fig. 1). The word *agat* (Fig. 1) has been recognized as a supine or past participle form of the verb *aga* 'flog', *internal homography*, or as the noun *agat* 'agate', *external homography*. However, with the automatic lemmatisation as the primary goal of the current study, internal homographies will, for the time being, be disregarded. In other words, *homography* in the following presentation should be understood as *external homography*. Likewise, when we present data on number of analyses below (Table 1), we disregard cases of internal homography.

As shown in Table 1, a single analysis was given in 73 percent of the cases, no analysis in 23 percent of the cases, and, finally, more than one analysis for a

[2] The information on dictionary stem is saved to be used as a tool in extracting a subdictionary for the syntactic analysis to follow; the information on inflectional type is included to provide a basis for subsequent statistical calculations.

Number of analyses	Absolute frequency	Relative frequency
0	572	23
1	1817	73
2	93	4
3	16	1
4	2	0
Total	**2,500**	**100%**

Table 1: Number of analyses resulting from the inflectional morphological analysis.

minority (5 percent) of the (graphic) words. Below we will examine the different cases, starting with analysis failure.

3.1 Lexical Gaps

The analysis failures are due to missing stems in the dictionary, excluding two cases of insufficient grammar coverage, and one case of an unforeseen use of parentheses (Table 2).

Type	Absolute frequency	Relative frequency
no stem: compound	452	79
no stem: derivative	80	14
no stem: proper noun	36	6
grammar coverage	2	< 1
orthographic convention	2	< 1
Total	**572**	**100%**

Table 2: Types of analysis failure.

Focusing for a moment on the missing stems, we have to state, that some 23% of the words in the definitions are not themselves head words in the dictionary, and, consequently, not explicitly defined. They are, basically, of two kinds, i.e. *proper nouns* and *derived words* (compounds and derivatives). The very fact that there are derived words missing in the dictionary should not surprise us (even if the high number of them did). A "complete" dictionary is theoretically impossible, due to, among other things, the rich potential for word formation, in specific, accidental compounding.

"Av samma skäl [utrymme] ges inte den triviala betydelsen hos en rad sammansättningar, avledningar och partikelverb. Denna betydelse täcks indirekt av ordboken genom beståndsdelarnas definitioner." 'For the same reason [space] the trivial meaning of a number of compounds, derivatives, and particle verbs is not given. This meaning is indirectly covered by the definitions of the constituent parts.' (*Svensk Ordbok* 1986: Preface). What can be objected to, however, is the use of such words in the definitions. "Vidare har en strävan varit att hålla antalet ord som används i definitionerna relativt litet och att välja så enkla ord som omständigheterna medger." 'Further the aim has been to keep the number of words used in the definitions relatively small and to choose as simple words as the circumstances permit.' (ibid.). Here we won't discuss this issue further, but rather examine the derived words used in the definitions to see if they keep up with the *transparency claim*, i.e. if their meaning is derivable from the definitions of their constituent parts by means of a set of general word formation rules. The rules on which those derivations are based, are part of the background knowledge implied by the lexicographers, and as such of primary interest to us. Thus, if they fullfil the transparency claim, they should be formalized and included in the word formation component of the LPS machine dictionary as its first piece.

A subset of the implicitly defined words of the definitions are included in the dictionary as morphological examples. The very existence of the derived words is confirmed by the examples, but still, nothing is said about their use and meaning; they are found to be transparent. In all, there are 3,934 morphological examples. Once the word formation rules behind the derived words of the definitions have been formulated, the total of the morphological examples might be used as a test corpus for the resulting word formation component.

The missing *proper nouns* make up a much smaller number (36) than the derived words (532). Thirty-three of them name geographical units, and the remaining name persons. As long as the words that they define are included in the dictionary, we see no other solution than to enter them as head words and thus define them. A frequent type is made up of country names (e.g. *Algeriet* 'Algeria') used in the definitions of people from those countries (e.g. *algerier* 'Algerian': *person från Algeriet* 'person from Algeria').

Two words were not analysed due to limitations in the inflectional grammar. They represent a type that can be illustrated by the graphic word *ansikts-* 'face' in coordinated compounds such as *ansikts- och hårvård* 'face and hair care'. The phenomenon is regular and has to be taken care of partly by the morphological, partly by the syntactic grammar. In the first instance, the LPS inflectional grammar will be accordingly extended.

Short for *befordrats* '(has been) sent' or *befordras* 'is (being) sent' we find *befordra(t)s* among the definition words. This abbreviated form of expressing alternatives was not foreseen by the LPS inflectional grammar. Nor will it be, but the short forms of this kind will be spelled out.

3.1.1 Compounds and Derivatives

To keep up with the *transparency claim*, the word formation competence on which the understanding of (some parts of) the definitions is based, should be more stereotypical than what is generally the case. Further, we must require that the *constituent words are themselves defined*, and, that they are *not homographic* or *polysemous*. We begin our examination with the most dominant word

Type	**Example**	**Absolute frequency**	**Relative frequency**
N-N	*aktieägare* 'stockholder'	376	83
N-AP	*abborrliknande* 'resembling a perch'	22	5
N-PP	*arvsberättigad* 'entitled to an inheritance'	11	2
N-A	*alkoholfri* 'non-alkoholic'	18	4
Ab-N	*bakåtspark* 'kick backwards'	8	2
A-N	*allmängiltighet* 'universal applicability'	4	< 1
Ab-Ab	*akterifrån* 'from the stern'	4	< 1
Ab-AP	*bakomliggande* 'lying behind'	3	< 1
Ab-PP	*bakåtriktad* 'pointing backwards'	3	< 1
Ab-A	*antidemokratisk* 'antidemocratic'	2	< 1
A-A	*allmänkulturella* 'general cultural'	1	< 1
Total		**452**	**100%**

Table 3: Types of compounds.

formation type, i.e. the *compounds*. In Table 3 we present the distribution of the compounds by syntactic types.[3]

Among the 11 compound types that were found, the quite dominating one is the *N-N* type making up some 83 percent of the total number of compounds. The dominance of this type is in accordance with our intuition and with other studies of modern Swedish (see e.g. Blåberg 1988). Our figure is, however, considerably higher than that presented by Blåberg (68%). His figures are based on an investigation of a corpus of 3,971 compounds, identified in newspaper text from 1985. Our higher proportion of *N-N compounds* seems to support the transparency claim, indicating a restricted use of the variation offered by the word formation potential. Additional support is given by the parameter *number of different types*, whose value is much lower in our material than in that of Blåberg, i.e. nine[4] versus 22. Our types constitute a subset of the set identified by Blåberg, including, in addition: V-N, V-A, A-V, Ab-V, Numeral-V, Numeral-N, Numeral-A, Numeral-Numeral, N-Proprium, A-Proprium, Proprium-N, Proprium-V, Proprium-Proprium. It remains to be seen though, how stable our figures remain through out the material. We don't, however, expect to find such a rich variation of types involving proper nouns as does Blåberg, proper nouns in general, being outside the scope of *Svensk Ordbok*.

Among the 376 N-N compounds, there are in all 154 different first constituents (simplex or derived). The majority of them (146) are explicitly defined, seven of them implicitly, and one of them (*A4* in the compound *A4-format* not at all). Among those that are only implicitly defined, there is a dominating type, i.e. process nouns derived from verbs by means of the *-(n)ing* suffix, such as *avfyrnings-* 'firing' (cf. *avfyrningsmekanism* 'firing mechanism') of the verb

[3] AP is short for active present participle, and PP for passive past participle.

[4] This is the figure we arrive at if we adapt to Blåberg, who includes the active and the passive participles in the verb group.

avfyra 'fire'. It accounts for five of the seven cases. One of the remaining cases is in itself an N-N compound, i.e. *bastuba* 'bass tuba' (cf. *bastubeinstrument* 'bass tuba instrument'). The other one is an A-N compound, i.e. *andraklass* 'second class' as in *andraklassutrymme* 'second class area'. When one of the nouns of an N-N compound is in itself only implicitly defined, the derivation of the definition word has to proceed in two steps. We think that such a situation should be avoided, and the two constituents of a compound used in a definition both be *explicitly* defined.

Another problem with the derivation of meaning from the N-N compounds of the definitions concerns homography and polysemy. Among the first constituents of the N-N compounds we found four cases of homography, i.e. *akter, arm, babords*, and *back*. *Akter, babord*, and *back* are all nouns or adverbs, and *arm* is a noun or an adjective. For instance, *akter* is a noun 'stern' or an adverb 'aft', and *back*, a noun 'back, reverse' or an adverb 'back'. The word *akter* occurs in *akterdäck*, *aktermast*, and *aktervägg*, and in deciding whether the noun or the adverb interpretation of *akter* was the intended one, we were guided by the morphological example *akterdäck* presented under the (noun) head word *akter*. By analogy, we chose the noun interpretation in the other two cases, too. As regards *back* it occurs in *backverkan* 'reverse effect' and in *backåkning* 'downhill going'. No morphological examples are given, but intuitively *back* in *backverkan* should be understood as an adverb instead of as a noun. This intuition is confirmed by its context *vända med utnyttjande av backverkan* 'turn using reverse effect'. The context of *backåkning*, i.e. *(typ av) kälke för backåkning* 'kind of sledge for downhill going' confirms the inutition that *back* here is an occurrence of *backe* 'hill' rather than of *back*. In addition to the four cases of homography mentioned above, there are 36 cases (9%) of polysemy concerning the first constituents of our N-N material. Without scrutinizing the individual cases further, we conlude that the derivation of meaning from the constituent parts of a compound in the definition has to be based on the identification of their definitions. When homography is involved, a choice has to be made. Only when a morphological example is there to facilitate the choice can we maintain the transparency claim. The compound words with an ambiguous first or second component will all have to be explicitly defined in our machine dictionary.

In addition to the 376 N-N compounds there are 76 compounds of different syntactic structure (Table 3). They will all be examined with regard to explicit definition and homography of constituent parts.

The total number of derivatives (not explicitly defined) that were identified is 80, which means that their share of the total number of implicitly defined words is only roughly 15 percent as compared to 85 percent for the compounds. In Table 4 we present the distribution of the derivatives by different types.

The total number of derivatives given in Table 4 (90), is higher than that presented in Table 3 (80), for the following reason. The figures in Table 3, are based on a 'top-level' classification of the words. In other words, a word such as *bakningsredskap* 'baking tool' is classified as a compound only, disregarding the fact that its first constituent is a derivative not explicitly defined. However, the figures presented in Table 4, include also such indirect derivation, and, in some

Type	Example	Absolute frequency	Relative frequency
N {*-(n)ing*}	*annonsering* 'advertising'	44	49
N {*-het*}	*aktsamhet* 'carefulness'	9	10
N {*-nde*}	*anskaffandet* 'acquirement'	7	8
A {*-bar*}	*avgränsbar* 'delimitable'	3	3
A {*-orisk*}	*artikulatorisk* 'articulatory'	1	1
Pref-V	*avstämpla* 'stamp'	6	7
V Particle	*hugga av* 'cut off'	20	22
Total		**90**	**100%**

Table 4: Types of derivatives incl. particle verbs.

cases, a word contributes to more than one of the derivation types. For instance a word such as *avfjällning* which has to be derived in two steps, i.e. *fjälla* 'peel' (explicitly defined), *fjälla av* 'peel off', and the process *avfjällning* is counted as an instance of both the V Particle and the N {*-(n)ing*} type. Cases of two-step derivation, though, are rare.

The dominating type (49%) is that of the verbal nouns, formed by means of the *-(n)ing* suffix. Its 44 members (inflectional forms) are derived from 42 different verbs. 32 of these verbs are explicitly defined, whereas 10 only implicitly so. Further, six of the 32 explicitly defined verbs are polysemous, and, consequenlty, the nouns derived from them not uniquely defined. The ten implicitly defined verbs, are either particle verbs formed by means of the particle *av*, or prefixed verbs formed by means of the prefix *av*. Intuitively, we identify them as derived from the particle verbs e.g. *klippa av*. However, formally they might as well be derived from the corresponding prefixed verbs e.g. *avklippa*. The same kind of ambiguity is a problem in the identification of the verbs of prefixed past participles, e.g. *avhuggen*. It might be derived from *avhugga* as well as from *hugga av*. As regards the semantic distinction between the prefixed verb and the particle verb, the meaning of the prefixed verb seems to be more abstract than that of the particle verb (see further Hellberg 1976; Ejerhed 1979). If we allow the use of implicitly defined particle verbs or prefixed verbs as a basis of verbal nouns of the *(n)ing* type, a systematic ambiguity will be created. It should be avoided, as should also the use of prefixed past participles in the definitions, inherently ambiguous as they are.

Concerning the remaining derivative types, their representatives will be examined for homographies and implicitly defined types in the same manner as the verbal nouns. In discussing the derivatives and their aptness for being handled by the word formation component rather than being registered as lexical units, the productivity of the affixes is an additional parameter to be considered. An example of a suffix sequence with low productivity is *-or-isk* with an absolute (lexical) frequency of one (*artikulatorisk* 'articulatory') in our material and of two in the NFO material (Allén et al. 1980).

Finally, derived words, which can be segmented in more than one way should be avoided. An example of such an instance in our material is *alliansfri-het* 'non-alignment' or *allians-frihet.*

3.2 Homography

In the automatic lemmatization process, the morphological analysis has to be followed by a homograph resolution procedure. The need for such a procedure in our material is evident from the figures on alternative analyses presented in Table 1. Roughly five percent of the definition words are homographic. Here we will give a presentation of the different kinds of homographies that we found, and discuss the possibility of solving them by means of syntactic analysis.

Type	Absolute frequency	Relative frequency
A/V	29	30
N/V	25	26
A/N	13	14
N1/N2	10	10
V1/V2	5	5
A/Ab	3	3
Ab/Pp	3	3
N/Ab	2	2
N/Pn	2	2
N/I	1	< 1
V/I	1	< 1
V/Pp	1	< 1
C/Im	1	< 1
Total	**96**	**100%**

Table 5: Types of homographies.

In Table 5 we present a word class based overview of the different kinds of homographies resulting from the morphological analysis. First we state, that in roughly 85 percent of the cases, the homograph components belong to different word classes, whereas 15 percent concern homography *within* one word class (the nouns or the verbs). The preconditions for solving the homographies by syntactic means, should, of course, be better within the first category than within the second one, where we may come close to purely lexical disambiguation. However, for an evaluation of this general assumption, we need to know more about the actual forms that coincide, and present such data for the major types in Tables 6 to 10.

The overwhelming number of cases of A/V homography (90%) are due to coincidence between the adjective and a particiipial form of the verb. The differentiation between them is notoriously difficult, to a great extent due to their partly overlapping distribution. However, one of the principles that was adhered

A/V

Type	**Example**	**Number of members**
1. A/V(pp)	*ansedd* 'respected/considered'	9
2. A(weak)/V(pp/sup)	*ansedda* 'respected/considered'	3
3. A(neutr)/V(pp/sup)	*avancerat* 'advanced'	6
4. A(weak)/V(pp/pret))	*allierade* 'allied'	6
5. A/V(ap)	*anslående* 'impressive/impressing'	2
6. A(weak)/V(inf)	*anrika* 'high-born/concentrate'	3
Total		**29**

Table 6: Types of A/V homographies.

to in the formulation of the definitions, i.e. the *replacement restriction* (see **2**), makes the task more realistic than would be the case with unrestricted text.

The nine members of the first type of Table 6 are in the basic form, and thus may occupy both an attributive and a predicative position. In all, they occur 73 times, and in 67 of these cases the adjectival interpretation should be preferred. Our choice was based on the following heuristics:

1. If the current word is in the attributive position, and not negated by an adverb, choose the adjective.
2. If the current word is the head of the definition of an adjective, and not negated by an adverb, choose the adjective.
3. If the current word is nominalized, choose the adjective.
4. If the current word occurs in the context *egenskapen att vara* ... the property of being '...', choose the adjective.
5. Else, choose the participle.

Examples:

1. [nn absess]: *begränsad ansamling var* 'limited amount of pus'
2. [av krystad]: *påfallande ansträngd och onaturlig* 'strikingly forced and unnatural'
3. [nn arv 1]: *övergång av egendom av (visst) värde från avliden till efterlevande* 'transition of property of (certain) value from deceased to surviving'
4. [nn begåvning/1]: *egenskapen att vara begåvd* 'the property of being talented'
5. [vb gälla 1/3]: *vara ansedd (som)* 'be considered (as)',
   [nn bricka/1]: *bärbar skiva begränsad av låg kant* 'portable plate surrounded by a low edge',

[av oavgjord/1]: *inte avgjord till någons fördel* 'not decided to anyone's advantage',
[nn ödesbygdsväg]: *väg genom ej bebodda trakter* 'road through not inhabited regions'.

The heuristic rule 2 is based on the replacement restriction; only if the definition belongs to the same syntactic category as the definiendum is it capable of substituting it. Further we require *same syntactic category* to mean *same word class*, (or same word class with respect to the head word of the definition). If, however, the homograph is preceded by a negating adverb, being a strong verb signal, the participle interpretation is chosen.

Concerning Type 2 (in Table 6) we have the same situation as with Type 1 with an additional supine form alternative. The supine form has a quite different distribution as compared to the adjective and the participle, and thus is easily distinguished in the syntactic analysis. (No supine forms, however, were found among the 31 occurrences of this type). In distinguishing between the adjectival and the participial forms, the heuristics listed above should be applied. It should also work out for Type 3, 4, and 5. The finite (past tense) form of Type 4 should be easily identified in the syntactic analysis. Finally, Type 6 should cause no problem.

A/N

Type	**Example**	**Number of members**
1. A/N	*alternativ* 'alternative'	8
2. A(neutr)/N	*basalt* 'basal/basalt'	1
3. A(weak)/N	*amerikanska* 'American (woman)'	4
Total		**13**

Table 7: Types of A/N homographies.

The dominating type of the adjective/noun homography is that between their basic forms (Type 1 in Table 7). Among the 70 occurrences of this type, 46 are adjectives, and, only 24 nouns. In the attributive position, there are hardly any problems in recognizing the adjectives on a purely syntactic basis. Making the homograph resolution though in the position after the finite verb requires additional knowledge. This is the case, for instance, for a word such as *bankrutt* 'bankrupcy' in the definition *göra bankrutt* 'become bankrupt' [vb bankruttera]. Further, Type 2 (Table 7) is an example of a causal coincidence between a neuter adjectival form and a noun. More systematic, however, is the homography between the weakly inflected adjective and the noun in Type 3. The rules for nominalizing the adjectives must be very restricted in the analysis grammar, if we are to solve these homographies syntactically, the reason being that the weakly inflected adjectives due to their inherent definitenesse constitute a productive basis for nominalization.

N/V

Type	Example	Number of members
1. N/V	*aga* 'flog(ging)'	6
2. N/V(imp)	*begär* 'desire/require'	3
3. N/V(ap)	*anseende* 'reputation/considering'	4
4. N/V(sup/Pp)	*agat* 'agate/flogged'	2
5. N/V(pret)	*bad* 'bath/asked'	3
6. N(+gen)/V(sup pass)	*ansats* 'approach(+gen)/(been) cultivated'	1
7. N(gen)/V(pres pass)	*begärs* 'desire(gen)/(is) required'	1
8. N(gen)/V(pret pass)	*bars* 'bar(gen)/(was) carried'	1
9. N(gen)/V(dep)	*andas* 'spirit(gen)/breathe'	1
10. N(pl)/V(pres)	*bakar* 'backs/bake(s)'	7
Total		**29**

Table 8: Types of N/V homographies.

In Table 8 we present the noun/verb homographies. All the different types seem to be solvable in their local contexts in combination with the syntactic prediction made by the word class marker of the definiendum, the substitution criterion.

N1/N2

Type	Example	Number of members
1. N1(utr)/N2(neutr)	*as* 'as/carcass'	5
2. N1(pl)/N2(pl)	*backar* 'backs/hills'	2
3. N1/N2	*bar* 'bar1/bar2'	1
4. N1(def sg)/N2(def sg)	*anden* 'the wild duck/the spirit'	1
5. N1/N2(pl)	*basar* 'bazaar/bass voices'	2
6. N1/N2(def)	*banan* 'banana/the path'	1
7. N1/N2(gen)	*askes* 'ascetisism/ash wood(gen)'	1
8. N1(indef gen)/N2(def gen)	*banans* 'banana(gen)/the path(gen)'	1
9. N1(gen)/N2(gen)	*bars* 'bar1(gen)/bar2(gen)'	1
Total		**15**

Table 9: Types of N1/N2 homographies.

In Table 9 and Table 10 we present the inherently most difficult homography cases, i.e. those in which the homograph components belong to the same word class, the noun class in Table 9 and the verb class in Table 10. Among the nouns there are four types (2, 3, 4, and 9) with no formal criteria to distinguish between them. Consequently, in these cases the homograph resolution amounts to purely lexical disambiguation, and the homography will be indifferent to the syntactic analysis. In Type 1 the homograph components differ with regard to gender only.

Its five members have, in all, 35 occurrences, but only in 9 of these cases is gender decisive for the choice. In Type 5, number is the distinguishing feature. It has two members with 10 occurrences. Three of these cases can be readily solved by syntactic means. Three of them can be solved, if number agreement is considered to be a precondition for coordinating nouns. For instance, *hög överbyggnad i för eller akter på medeltida fartyg* [nn kastell/2] 'high superstructure at the prow or at the stern of medieval vessels'. Two cases, finally, can be solved if phraseology is taken into account, e.g. *akter ut* 'astern'. In Type 6 and Type 8 definiteness is the distinguishing feature. It readily solves one of the three cases, i.e. *väg (till ngt) som ej följer den naturliga banan* 'road (to something) which doesn't follow the natural path' [nn bakväg]. As regards definiteness in the head word of prepositional phrases (the remaining two cases), the individual prepositions make their own demands, and the situation is more complicated. Finally, in Type 7, the homograph components differ with regard to case. It solves the homography, if the grammar doesn't admit elliptic NP heads.

V1/V2

Type	Example	Number of members
1. V1/V2	*avsluta* 'finish/conclude'	2
2. V1(+pres)/V2(+pres)	*avslutas* 'finish(pass)/conclude(pass)'	2
3. V1(pp/pret)/V2(pret)	*avlade* 'begetted/layed aside'	1
Total		**5**

Table 10: Types of V1/V2 homographies.

In Table 10 we present three types of verb homographies. The first two have no distinguishing formal feature, and thus cannot be solved by syntactic means. In the third type, however, there is a context in which they differ, i.e. in past participle constructions. In our material, this distinction doesn't emerge, and the homography cannot be solved syntactically.

Type	Example	Number of members
N1/N2/V	*backar* 'backs/hills/back(s)'	5
A/N/V(ap)	*avgörande* 'decisive/decision/deciding'	4
A/Ab/Pp	*bakom* 'stupid/at the back/behind'	1
A/Ab/C	*bara* 'bare/only/(as) long as'	1
N1/N2/Ab	*akter* 'stern/acts/aft'	1
N/Ab/Pn	*allt* '(the) universe/gradually/everything'	1
A/N/Ab	*akut* 'acute/casualty department/urgently'	1
V/Nl/Pn	*andra* 'state/second/other'	1
N1/N2/N3	*bas* 'bass voice/base/thrashing'	1
Total		**16**

Table 11: Three homograph components

Type	Example	Number of members
N1/N2/Ab/Pp	*bak* 'behind/baking/at the back/behind'	1
N1/N2/A/V	*bar* 'bar/bar/bare/carried'	1
Total		**2**

Table 12: Four homograph components.

In those cases where three or four analyses were generated (cf. Table 1) combinations of the homography types presented above (cf. Table 5) were involved. The actual combinations are presented in Table 11, and Table 12.

4 Conclusions

In this paper we have presented the results that we achieved in a pilot study of the definition vocabulary of *Svensk ordbok*. It is part of our on-going work on the generation of a machine-tractable dictionary from the lexical database from which the dictionary was drawn, in specific, of making its definitions exploitable to the LPS parser. Parsing them is a step towards this goal. The present study, aiming in particular at the automatic lemmatisation of the definition vocabulary, has included an automatic morphological (inflectional) analysis of a subset (2,500 graphic words) of it, and a, largely manual, analysis of the results. Two major issues were treated, i.e. the coverage of the dictionary in relation to the definition voaculary, and the feasibility of homograph resolution by syntactic analysis in the local (definition) contexts.

The morphological analysis was carried out by means of the LPS morphological analyser, disposing of a stem dictionary covering the head words of the dictionary. As a result of this analysis, we found that, roughly, 23 percent of the definition words were not explicitly defined. The lexical gaps were mainly due to the use of proper nouns (6%), and derived words (93%). As regards the missing proper nouns, they will be registered in our machine dictionary. The dominating type of the derived words are the compounds (79%), while the derivatives (incl. particle verbs and prefixed verbs) account for 14 percent of the gaps. The transparency claim with regard to the implicitly defined words cannot be maintained without certain modifications of the definition vocabulary. Two measures will be taken in the generation of the machine dictionary. First, derived words including homographic or polysemous constituents will be exchanged by explicitly defined words, or by derived words of unambiguous elements. Secondly, the use of deverbal nouns based on implicitly defined particle verbs or prefixed verbs, implying a systematic ambiguity, will be avoided, as will the use of prefixed past participles, inherently ambiguous as they are. In general, derivations of more than one step will be avoided. In other words, we require that the toplevel constituents of a compound or a derivative are themselves explicitly defined. These modifications, when applied to whole definition vocabulary, will bring it one step closer to a *defining* vocabulary.

Homography turned out to be a smaller problem than the lexical gaps. In all, external homographs constitute less than five percent of the lexical material that was examined. Data on the distribution of different types of homographies are presented. With a manual study of their contexts (definitions) as a basis, we conclude, that they can, to a large extent, be resolved by syntactic means, provided that the syntactic prediction made by the word class marker of the definiendum is taken into account. This information is of vital importance. A heuristics for distinguishing between adjectives and participles was proposed. It will be further evaluated in the course of the project.

References

Allén, S. 1970. *Nusvensk frekvensordbok baserad på tidningstext. 1. Graford. Homografkomponenter.* [Frequency dictionary of present-day Swedish based on newspaper material. 1. Graphic words. Homograph components.] Stockholm.

Allén, S. 1981. The Lemma-Lexeme Model of the Swedish Lexical Data Base. B. Rieger [Ed.]. *Empirical Semantics*:376–387. Bochum.

Allén, S., S. Berg, J. Järborg, J. Löfström, B. Ralph, & C. Sjögreen. 1980. *Nusvensk frekvensordbok baserad på tidningstext. 4. Ordled. Betydelser.* [Frequency dictionary of present-day Swedish based on newspaper material. 4. Morphemes. Meanings.] Stockholm.

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

Ejerhed, E. 1989. Verb-partikelkonstruktionen i svenska: syntaktiska och semantiska problem. [The verb-particle construction in Swedish: syntactic and semantic problems.] O. Josephson, H. Strand, & M. Westman [Eds.]. *Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning 11*:49–64. Dept. of Nordic Languages. University of Stockholm.

Hellberg, S. 1976. *Av* som partikel och preposition. ['Av' as particle and as preposition.] Dept. of Computational Linguistics. University of Göteborg.

Järborg, J. 1988. Towards a formalized lexicon of Swedish. *Studies in computer-aided lexicology*:140–158. Almqvist & Wiksell International. Stockholm. (Data linguistica 18).

Sågvall Hein, A. 1987a. Forskningsprogram för projektet En lexikonorienterad parser för svenska. [Research program for the project A Lexicon-Oriented Parser for Swedish.] Dept. of Computational Linguistics. University of Göteborg.

Sågvall Hein, A. 1987b. Parsing by means of Uppsala Chart Processor (UCP). L. Bolc [Ed.]. *Natural language parsing systems*:202–266. Springer Verlag. Berlin & Heidelberg.

Sågvall Hein, A. 1988. Towards a comprehensive Swedish parsing dictionary. *Studies in computer-aided lexicology*:268–294. Almqvist & Wiksell International. Stockholm. (Data linguistica 18).

Sågvall Hein, A. 1989. The LPS inflectional grammar. A listing of the rules. Dept. of Computational Linguistics. University of Göteborg.

Sjögreen, C. 1988. Creating a dictionary from a lexical database. *Studies in computer-aided lexicology*:299–338. Almqvist & Wiksell International. Stockholm. (Data linguistica 18).

Svensk ordbok. [A dictionary of Swedish.) 1986. Produced at Språkdata ([Dept. of Computational Linguistics. University of Göteborg.] Stockholm.

Wilks, Y., D. Fass, C. Guo, T. McDonald, & M. Slator. 1988. Machine tractable dictionaries as tools and resources for natural language processing. D. Vargha [Ed.]. *Proceedings of the 12th International Conference on Computational Linguistics (COLING)*:750–755. Budapest.

Center for Computational Linguistics
Uppsala University
Box 513
S-751 20 Uppsala
uduas@seudas21.bitnet