Margareta Sjöberg
Center for Computational Linguistics
Uppsala University

ON THE IDENTIFICATION OF STEMS IN FASS

FASS is the Swedish drug catalogue. An information retrieval system FASSIS has been designed and implemented by LINFO AB and Uppsala University Computing Center (UDAC) with this as a basis. In this system every word occurrence is treated as a key word.

Work has been under way since spring 1985 at the Center for Computational Linguistics to produce a key word index of basic forms for FASSIS by semi-automatic means. One step towards this goal is to build a stem dictionary covering the material (see further Sågvall Hein 1985a). Two parallel lines have been followed side by side. On the one hand, the full text of FASS has been investigated and prepared for treatment by the Center's program package for linguistic text processing, TEXTPACK, which produces word lists, concordances, etc (see Rosén 1986 and Rosén & Sjöberg 1985 for further details). On the other hand, the present key word file in FASSIS has been used as a test file in attempting to identify stems in the material. This presentation will concentrate on these attempts.

For the sake of illustration an extremly short but in other respects representative drug description from FASS is presented in fig. 1 below.

On the identification of stems in FASS
Margareta Sjöberg
Proceedings of NODALIDA 1985, pages 185-193

Capsolin
Parke-Davis
Salva
Smärtstillande salva Grupp 12F 0510
Deklaration: 1 g innehåller: Oleoresin capsic. 12 mg, camphor. 52,5 mg, aetherol. tereb. 97,5 mg, eucalypti aetherol. 25 mg, cera flava, vaselin. et odor q.s.
Upplysningar. Parke-Davis, tel. 08- 82 03 50.
Egenskaper. Ökar hudgenomblödningen och ger en värmekänsla i det behandlade området.
Indikatorer. Lokal behandling vid tillfällig huvudvärk.
Försiktighet. Överkänslighet mot ingående beståndsdelar, speciellt terpentinolja, kan förekomma.
Dosering. Appliceras tunt och ingnides lätt några gånger dagligen. I barnpraxis spädes helst med 3-4 delar vaselin.
Observera. Capsolin skall ej komma i kontakt med ögon, slemhinnor eller skadad hud. Händerna tvättas väl.
Förpackningar och priser. Salva
35 g 17:40

- fig. 1. -

1. Strategy.

Our aim has been to identify automatically as many stems as possible starting from the present key word index. One problem with this file is that all words longer than fifteen letters have been chopped off at the fifteenth letter. Another problem is that the distinction between capital and common letters is not maintained in the file. When TEXTPACK produces a correct and complete file of the vocabulary of the material, with this difference kept and with deleted letters beyond the fifteenth restored, this file will replace the present test file.

A study of the material shows that there are a lot of exceptions from standard Swedish morphology. Many word forms are of foreign origin or are numerical expressions or hybrids as illustated below.

... ibland av karaktären bull's eye ...
... aktivt ulcus ventriculi et duodeni ...
... mixtur 40 mg/ml ...
... 24 st vnr 411496 ...
... cirka 1-2 timmar ...
... uppges till 80-85% ...
... 12:-/st ...

In all there are ca. 40,000 graphical words in the key word file among which ca. 10,000 are purely numerical expressions. These have been removed from the target set for the stem identification programme.

As a first step in the stem identification process I wanted to mark, and thus, for the time being, remove from further analysis, words of foreign origin, together with abbreviations and proper names. And FASS contains a lot of them! There are, for example, a large number of names of drugs and companies which produce them, of Latin names of chemical substances and abbreviations of these. There are, furthermore, a lot of Latin names of organs, illnesses, bacterial species and so forth, together with English expressions and quotations, with references. Quite a number of these "special words" can be identified by capitalising on the internal structure of the drug descriptions. Thus, names of drugs (1,372) and companies which produce them (150) occur in posts which are specially marked in the FASS source file. As for the greater part of the abbreviated Latin names of chemical substances, they occur under the subheading of 'DFK' - for "declarations" (see fig. 1). Ca. 1,200 different words ending with the full stop of abbreviation have so far been identified in these sections. These names and abbreviations are marked in the source file.

There is a concentration of Latin names of chemical substances in the 'DFK'-sections, and, as the result of the envisaged TEXTPACK treatment of the source file will allow us to treat different parts of the material as a corpus of its own, it will also for example be possible to gather a large number of the Latin words by picking out word forms in 'DFK' sections, which neither contain the full stop of abbreviation nor occur in the remainder of the text. With the distinction of words with capital and common initial letters maintained we will also be able to heuristically identify the remaining proper names in the text assuming that words which only occur with an initial capital letter are proper names.

While waiting for the TEXTPACK preliminaries to be completed, Latin, English and French words and expressions as well as proper names (other than the ones already marked) have been entered manually into special files as they crop up on inspection of preliminary results of rough test marking in the key word file. At present the English "dictionary" contains 110 words and expressions, the latin "dictionary" 380 and the list of names 145, the latter comprising mainly names of persons, journals and drugs. After this marking there remain ca. 27,000 unmarked word forms for further analysis.

The next step towards building a stem dictionary was to identify automatically as many stems as possible among these remaining word forms. For this purpose I have chosen to work with heuristic rules for the recognition of word endings which I have expressed in Brodda's BFTA system (for a description of the system see for example Brodda & Karlsson 1980). Because of the size of the material and the definite need for manual inspection of the result produced by the heuristic rules I found it necessary to concentrate on smaller parts of the material at a time, gradually trying to correct mistakes in the analysis more or less manually by adding "exception rules". I therefore divided the set of rules into groups that tentatively mark approximately 2,000-5,000 word forms each, the markings then being carefully checked and new rules added until the set of words is correctly analysed and marked.

A stem dictionary has been created with the help of the markings introduced into the key words, and word forms containing these stem are removed from the source file. So far, ca. 16,500 different stems have been identified in FASS, and with word forms containing these stems removed from the source file there remain ca. 7,000 unmarked forms. Of these, ca. 2,800 are words which have been chopped off at the fifteenth letter, the majority of which will be ascribed a correct analysis by our rules when restored to their

full length. The rest remain to be treated.

2. Comments on the BETA rules.

The rules I have devised analyse word endings only. They mark stems in word forms depending primarily on whether they contain a characteristic suffix or a derivational component. The rules are divided into five mutually independent groups, two of which concentrate on identifying noun suffixes, one on verbs and one on adjectives. In the fifth group I employ a list of final derivational strings which occur frequently in the text.

The first set of rules includes the most distinctive suffixes '-arna(s)', '-erna(s)', '-orna(s)', '-ar(s)','-er(s)' and '-or(s)', the central rules being*

```
'ARNA' --> '=ARNA/s'
'ERNA' --> '=ERNA/s'
'ORNA' --> '=ORNA/s'

 'AR'  -->  '=AR/s'          after 'D','G','L','M','P','S'
 'AR'  -->  'A=R/v'          otherwise

 'ER'  -->  '=ER/v'          after 'G','J'
 'ER'  -->  'ER=/s'          after 'C','K'
 'ER'  -->  '=ER/s'          otherwise

 'OR'  -->  'OR=/s'          after 'U'
 'OR'  -->  '=OR/s'          otherwise
```

There is also a set of rules by which incorrect analyses generated by these heuristic rules can be avoided. We call them "exception rules" simply. Let the following examples suffice as an illustration.

```
  'KAR'   -->  'K=AR*/s'          after 'C','J','N','S'
 'KVAR '  -->  'KVAR/adv'
'DELAR'   -->  'DELAR/hom'
'MOLAR'   -->  'MOLAR/adj'
'SPELAR'  -->  'SPELA=R/v'
```

---

* '/s', '/v' , etc. are form class tags denoting respectively nouns, verbs, etc.

In all there are 230 rules in this set, and they mark a total of 2,813 stems (455 are verbs, 50 adjectives, 22 homographs and the rest nouns. Examples of homographs are 'delar', 'klumpar', 'pumpar', 'isomer'.). The number of "exception rules" is 210.

The main rules in the next set of rules which deal with possible noun suffixes are

```
 'AN'  -->  '=AN/s'
'ANS'  -->  '=ANS/s'
'ATS'  -->  '=ATS/s'

 'EN'  -->  '=EN/s'
'ENS'  -->  '=ENS/s'
 'ET'  -->  '=ET/s'
'ETS'  -->  '=ETS/s'
```

Even more rules for exceptional cases must be added here in order that mistakes in the analysis should not multiply inordinately. There are altogether 287 rules (278 "exceptional" ones) here which together mark 3,100 stems (32 adjectives, 38 adverbs, 110 verbs, 20 homographs and the rest nouns).

The rules in the third set are mainly concerned with the identification of adjective/participle endings. The emphasis lies on suffixes like '-da', '-dd', '-ld', '-rd', '-ad', '-at', '-ade', '-ande', and '-ende'. Word forms ending with '-as' and '-es' are also marked here. There is a total of 310 rules in the group, marking 3,275 stems (ca. 2,300 verbs, including participles, 80 adjectives and the rest nouns). The number of "exception rules" is 276.

With the rules in the fourth group 2,484 stems have been identified, the majority of which (1,925) are adjective stems and the remainder nouns and verbs. The word endings recognised here are '-igt', '-iga', '-iskt', '-iska', '-fritt', '-fria', '-bart', '-bara', '-lt', '-la', '-mt', '-mma', '-nt' '-na', '-vt', '-va', '-ta', '-ärt', '-ära', and '-are', '-ast',

'-aste'. The group consists of 230 rules in all, 180 "exceptional".

A large number of words in the text contain no explicit inflexional ending and many of them therefore remain unmarked by the above rules. But many of these contain a derivational component indicating that the word as whole belongs to a given category, the stem being identical to the word itself. Examples of derivational strings of this kind are '-id' (hexicid, jodid, ureid, karbamid; gravid, fungicid, cyticid, vermicid, baktericid, tyfoid, myceloid, ...), '-fri' (valfri, alkoholfri, symptomfri, kaliumfri, ...), '-isk' (biologisk, urologisk, allergisk, kirurgisk, alkalisk, ...), '-är' (bacillär, lågosmolär, bipolär, muskulär, högmolekylär, ...), '-ig' (mjölkig, flockig, lindrig, ...), '-ing' (odling, mässling, pensling, rubbning, välling,..), '-tion' and '-sion'. A list of a number of final components in compound words frequently occurring in the text such as 'terapi' (67), 'medel' (50), 'dos' (70), 'virus', 'status' and 'enzym' is also used here. Altogether the rules in the group number 445 (ca. 210 "exceptional") and correctly mark around 7,000 words.

3. Future plans and concluding remarks.

Not all the stems in FASS have yet been identified, and work on the remaining stems continues. Certain suffixes, for example '-a', are so insignificant that it has not been practically feasible to consider them as yet, but now they might be of some help. And purely manual methods will presumably have to be resorted to in the final stages.

It also remains to deal manually with the ca. 300 words and stems marked as homographs (beroende, buffrar, format, ... and allergen=, hosta=, ...) and to check for pure adjectives in the words marked as participles.

The BETA system has proved to be practical to work with. It is a simple

matter to express rules within the system for the sort of treatment we wanted to give to our word form file. That the number of exception rules is large is not, of course, something for which the BETA system can be blamed. Certainly it is possible that we could have managed with fewer; however, the aim has not been to build up as compact a system of rules as possible, but rather to identify the stems in the text as efficiently as possible. The number of rules which are of a purely lexical character is also so large that it would seem to be very difficult to reduce them dramatically. A trial run, albeit preliminary, with a version of Brodda's SWEMORF pointed in the same direction.

In parallel with the work described in this paper preparartions are under way for the next step in our project, i.e. the integration of the stem dictonary in an automatic morphologic analyser. One morphological model for the inflectional analysis has been tested for some of the identified stems (Sägvall Hein 1985b). The specification of the particular word class to which the identified stems belong was made in order to facilitate this phase.

BIBLIOGRAPHY

Brodda Benny, and Karlsson Fred, 1980. "An experiment with Automatic morphological analysis of Finnish". Papers from the Institute of Linguistics, University of Stockholm, Publication 40. (Reprinted 1981: Department of General Linguistics, University of Helsinki, Publication, No.7.)

Rosén Valentina, 1986. "Förberedande undersökning av FASS för automatisk nyckelordsindexering." Center for Computational Linguistics, University of Uppsala. (forthcoming)

Rosén Valentina, and Sjöberg Margareta, 1985. "TEXTPACK, programpaket för språkvetenskaplig textbearbetning". Center for Computational Linguistics, University of Uppsala.

Sågvall Hein Anna, 1985a. "Automatic Key-Word indexing for FASSIS. Proposals for a project". Center for Computational Linguistics, University of Uppsala.

---- 1985b. "Parsing by means of Uppsala Chart Processor". In Bolc, L (ed.) Natural Language Parsing Systems. Springer-Verlag. (forthcoming)