IVAN RANKIN,
DEPT. of COMPUTER and INFORMATION SCIENCE,
LINKÖPING UNIVERSITY,
SWEDEN.

SMORF - an implementation of Hellberg's morphology system

Abstract: *A brief account of Hellberg's morphology system [Hellberg, 78] is presented - its aims and structure, and how it deals with inflection, derivation and compounding. Then there follows a discussion of our experience when implementing and running the program and evaluating the success of the system. Discussion points are accompanied by run-time examples. Some improvements on the original system have been made and are illustrated in the text. A summary of our evaluation is given.*
***Key words**: inflection, derivation, compounding, explicitness, exhaustiveness, overproduction of analyses, filters, linguistic transparency, semantics.*

1. AIMS AND STRUCTURE

1.1. AIMS

Hellberg's intention is to provide a "detailed account of Swedish morphology" and he adds "A system of paradigms has been set up and a basic dictionary compiled, for the primary purpose of being used in algorithmic text analysis." [Hellberg, 78].
Hellberg makes two specific claims for his system - **explicitness** and **exhaustiveness**:

a) "As to **explicitness**, this means that the system accounts for all types of variations, limitations and extensions, including forms which may seem self-evident to speakers of the language."
eg. the unsettled use of the plural form of *meddelande* with both *meddelanden* or *meddelande* as acceptable forms.

b) "**Exhaustiveness** implies for one thing that attempts have been made to cover paradigms with just a few members, perhaps only three or four, ..."
eg. the paradigm for *prestanda* is an example of a paradigm with only a few members, as opposed to the one covering words like *flicka* with many members.

"... and for another that not only inflectional forms are taken into account, but also stem modifications and linking elements which may occur in derivations and compounding."
eg. the paradigm for a word such as *gata* will account for compounds with no linking element - *gatsopare*, for example, as well as compounds with the linking element - *gatubelysning*.

The borderline between what is an acceptable form of a word and what is not acceptable is not always clear in reality and Hellberg comments: "Where the limits

SMORF - an implementation of Hellberg's morphology system
Ivan Rankin
Proceedings of NODALIDA 1985, pages 161-171

should be drawn...cannot be decided once and for all, but gradually, on the basis of experience gained in the use of the system." op cit.

No claims are made to the effect that the system reproduces the way a human reader processes the text. Hellberg admits that the system fails to capture some of the morphological similarities between words or between groups of words, while at the same time this technical approach brings out similarities not adequately described in the traditional grammatical framework - such as "the distinction between strong and weak verb conjugation, where the system displays a multitude of transitional and mixed types between second, third and fourth conjugations". op cit.

1.2. STRUCTURE - THE DICTIONARY, THE PARADIGMS and THE SUBROUTINES.

1.2.1. The dictionary.

Words are stored in the dictionary in the form of a technical stem and a reference to a paradigm number (amongst other things). The technical stem is that part of the word which is common to all inflectional forms, eg.

Word	Technical stem	Paradigm nr.
flicka	*flick*	101
klaga	*klag*	715
nyckel	*nyck*	231
seger	*seg*	232
krypa	*kryp*	742
krupit	*krup*	744

We shall return to the consequences of the concept of the 'technical stem' later in the discussion of the implementation.

1.2.2. Paradigms and Subroutines.

To show how the paradigms and subroutines work, it may be clearer to follow an example. Imagine that the word *flickorna* is to be analyzed and assume for the time being its technical stem *flick-* has been located in the dictionary. The dictionary entry provides a reference to its related paradigm, 101.

Paradigm 101

nn utr

	lo # ---->92	*flicka(s)*
	---->*e*---->96	*flickan(s)*
>5		*flickor(s)*
		flickorna(s)
	sh ---->11	*flick(s)-*
	---->(lo)	*flicke-*

(Note that the possible forms of *flicka* have been added here for illustrative purposes; they are not part of the paradigm.)

a) First a check is made on the length of the rest of the word after the technical stem to determine whether it is so long that it must be a compound or derivative rather than an inflectional form. If so, it will be unnecessary to run through all the tests for inflectional endings as they will be bound to fail. The threshold value for this length is given in the paradigm (note >5 in the schema above). In the case of *flickorna* the length of the part of the word after the technical stem, *-orna* is not greater than 5 so the search is continued in the short (sh) branch. (Checking the length of the rest of the word and selecting a particular branch has no theoretical importance, it is a strategy designed purely to speed up the search.)

b) Within the long/short branches the tests are run in a top-to- bottom order; this order is based on text frequency-counts.

c) The search continues through the paradigm; the path chosen depends on the next character to be matched. The branches of the paradigms lead into subroutines. A paradigm is entered only once for each technical stem. All numbers in a paradigm refer to subroutines.

d) While traversing the paradigms and subroutines the morphological features are picked up; in this example **nn utr** (**noun non-neuter**) are picked up on entering the paradigm.

e) In this example the short branch of the paradigm refers first to Subroutine 11 (and later to the long form of the same paradigm) and the first letter to be matched is the *o* of *-orna*.

Subroutine 11

---->*a* ---->91 **obe sin**
---->*n* ---->91 **bes sin**
---->*#* *o* ---->10

The ---->*a* branch is not followed up, but the ---->*o* branch is, (although there are no features to be picked up). The # sign is of no theoretical importance. It simply marks the end of the linguistic stem. The search is directed to Subroutine 10.

Subroutine 10

---->*r* ---->91 **obe plu**
---->*n* ---->*a* ---->91 **bes plu**

Here the *r* in *-rna* is matched and then the second of the two options (either absence or presence of *na* to be matched) is chosen. The features **bes plu** (**definite plural**) are picked up and the search is referred to Subroutine 91.

Subroutine 91

If position 0 = *s* then
if position 1 = blank then accept and label **gru/gen**
else reject
Else if position 1 = blank then accept and label **gru**
else if position 1 = *s* then
if position 2 = blank then accept and label **gen**
else reject
else reject

Position 0 is the position of the last character matched. Position 1 is the position of the next character to be matched. At this stage in the example it is blank as the ending *-orna* has been successfully matched already and no further characters remain. The feature **gru (base form)** is picked up.
The analysis returns: ***flickorna: flicka*** **nn utr bes plu gru**
(noun non-neuter definite plural base-form)

Note that the paradigms reflect the morphological differences between words - if two words are similar but display a morphological difference, they will belong to different paradigms. The subroutines reflect similarities; several groups of words may be referred to a single subroutine, take the *-rna* plural ending of Subroutine 10, for example.

In passing it may be observed that there are 235 paradigms in all.

1.3. INFLECTION, DERIVATION and COMPOUNDING.

1.3.1 Inflection.

The way in which inflection is dealt with was exemplified in case of *flickorna* above. The paradigms and subroutines contain all the possible inflectional forms, and the features of each word are picked up as the paradigms and subroutines are traversed.

1.3.2 Derivation.

Hellberg was faced with a choice: either a) include derivational endings in the paradigms and subroutines, or b) include them in the dictionary where they are treated as entries in their own right alongside *flick, kryp* etc., and then view them as compound elements as far as the paradigms and subroutines are concerned.

For a number of reasons he chose the latter, mainly because choice a) would cause an "enormous expansion" of the paradigms and also restrictions in the use of productive suffixes are semantic rather than morphological in nature (and therefore a morphology system need not account for such restrictions). This means that a word like *storhet* will be treated as a "compound" of *stor* + *het* The same is true of prefixes

such as *o, sam* or *bi*: *omöjlig* will be treated as the first part of a compound such as *o* + *möjlig*.

1.3.3 Compounds.

The fact that the paradigm system is adequately equipped to handle compounds is a theoretical necessity considering the seemingly unlimited possibilities of compounding in Swedish, and a practical necessity in that it drastically reduces the number of words stored in the dictionary. If *hand* and *duk* are entered in the dictionary, there is no need to add *handduk*.

There are some exceptions to this rule. An example will suffice to clarify the issue. The word *blick* is in the dictionary. So are the prefixes *in-* and *över-*. This means that *inblick* and *överblick* do not need to be entered separately. However, although *ögon* is also in the dictionary, *ögonblick* must be entered in its own right as it is not derivable from its component parts - it is of a different gender from the above examples. This point is discussed further in the second section.

Note that the paradigms reflect the morphological differences between words - if two words are similar but display a morphological difference, they will belong to different paradigms. The subroutines reflect similarities; several groups of words may be referred to a single subroutine, take the *-rna* plural ending of Subroutine 10, for example.

In passing it may be observed that there are 235 paradigms in all.

2. ON IMPLEMENTING THE SYSTEM

The system has been implemented at Linköping Tekniska Högskola in two versions - in Interlisp-10 on DECSYSTEM-20 and in Interlisp-D on a Xerox 1108/1109. The basic dictionary was kindly supplied by Språkdata, Gothenburg. The system has been partially running and tested since summer 1984 and finally developed and tested since summer 1985.

2.1 INFLECTION, DERIVATION AND COMPOUNDING. THE OVERPRODUCTION OF ANALYSES.

The analysis of inflected forms has proved very accurate and complete. Examples such as *flickorna* and *drack* provide clear, unambiguous analyses.

```
SMORF Window
54-SMORF: FLICKORNA
(FLICKORNA (("FLICKA" (noun non-neut def plur base))))
55-SMORF: DRACK
(DRACK ((("DRICKA")
         (verb past))))
56-SMORF:
```

However, even seemingly uncomplicated word forms can display unexpected side effects, eg. *bil*.

```
57-SMORF: BIL
(BIL (("BIL" (noun non-neut indef sing base))
      ("BI" "L" (abbrev base))
      ("BI" "L" (abbrev))))
58-SMORF:
```

This problem stems from the fact that, in allowing for the almost unlimited ways of forming compounds - and in the formation of derivatives in this system - a profusion, if not to say confusion, of interpretations is possible. In the analyses of *bil* the correct interpretation comes first; the second and third interpretations have analyzed *bil* as a compound of *bi* and the abbreviation *l*. *Frukosten* illustrates the difficulties even more clearly.

```
SMORF Window
60-SMORF: FRUKOSTEN
(FRUKOSTEN (("FRUKOST" (noun non-neut def sing base)
                       )
            ("FRU" "KO" "STEN" (prop-name base))
            ("FRU" "KO" "STEN"
                   (noun non-neut indef sing base))
            ("FRU" "KO" "TE" "N" (abbrev base))
            ("FRU" "KO" "TE" "N" (abbrev))
            ("FRU" "KOSTA" "EN"
                   (art indef sing non-neut base))
            ("FRU" "KOSTA" "EN" (non-Swed))
            ("FRU" "KOSTA" "EN" (adv))
            ("FRU" "KOSTA" "EN" (pron nominal-fn
                                      base))
            ("FRUKOST" "EN"
                       (art indef sing non-neut
                            base))
            ("FRUKOST" "EN" (non-Swed))
            ("FRUKOST" "EN" (adv))
            ("FRUKOST" "EN" (pron nominal-fn base)))
           )
```

Filtering out the more unlikely analyses characterized a new stage of development. By preventing foreign words and abbreviations from forming compound elements it was possible to rule out the ubiquitous unlikely analysis of words such as *flicka* being interpreted as a compound of *flicka* and the English article *a* or **non-neuter definite singular** forms such as *bilen* being analyzed as a compound of the Swedish *bil* and the French *en* on the one hand, and *flicka* being analyzed as a compound of *flicka* + *a*, the latter part of the abbreviation of *bl a* (*bland annat*) on the other. These two filters alone reduced the number of "wrong" interpretations by about 30%. The following example shows the stepwise layering of filters on *flicka* the first version with no filters implemented, the last version with three filters implemented:

```
27-SMORF: FLICKA
(FLICKA (("FLICKA" (noun non-neut indef sing base))
         ("FLICKA" "A" (non-Swed))
         ("FLICKA" "A" (abbrev base))
         ("FLICKA" "A" (abbrev))))
28-SMORF:
```

SMORF Window

```
48-SMORF: FLICKA
(FLICKA (("FLICKA" (noun non-neut indef sing base))
         ("FLICKA" "A" (non-Swed))
         ("FLICKA" "A" (abbrev))))
49-SMORF:
```

SMORF Window

```
51-SMORF: FLICKA
(FLICKA (("FLICKA" (noun non-neut indef sing base))
         ("FLICKA" "A" (non-Swed))))
52-SMORF:
```

SMORF Window

```
54-SMORF: FLICKA
(FLICKA (("FLICKA" (noun non-neut indef sing base))))
55-SMORF:
```

So far filters have been put on in an ad hoc fashion. This is mainly because it has not yet been decided what use the system will be put to.

Another troublespot, also caused by the inclusion of derivative endings in the dictionary as items in their own right, was the common collocation *ska*.It was found that all adjectives of the *-sk* type in plural and singular definite attributive forms were interpreted not only as adjectives (ie. correctly) but also as verbs, eg. *ironi + ska*. To make matters worse *-ska* is included in the dictionary as a noun ending, eg. *ilska*. Thus *ironiska* also became a **noun**. As there is only a small number of nouns ending in *-ska* in Swedish, it would perhaps be a better choice to list them separately in the dictionary.

Before leaving this brief glimpse at the problem of the overproduction of analyses, it is in place to emphasize that the "correct" interpretation is always provided; the difficulty has been in reducing the number of erroneous alternatives.

2.2 WHAT THE SYSTEM LACKS

2.2.1. Re-insertion of deleted consonants.

One feature that the system lacks is the ability to deal with deleted consonants in compounds as in *ful(l)-lärd* or *glas(s)-skål*. What happens is that *full* is recognized first and then *ärd* is checked (and found not to exist in the dictionary), then *ful* is recognized and *lärd* is then checked (and found) which explains the two interpretations shown below. A similar explanation goes for *glasskål*.

```
SMORF Window
79-SMORF: FULLÄRD
(FULLÄRD (("FUL" "LÄRA"
                 (verb past-prt indef sing
                       non-neut base))
          ("FUL" "LÄRD"
                 (adj indef sing non-neut base)
                 )))
80-SMORF: GLASSKÅL
(GLASSKÅL (("GLAS" "SKÅL"
                   (noun non-neut indef sing
                         base))
           ("GLASS" "KÅL"
                    (noun non-neut indef sing
                          base))))
81-SMORF:
```

2.2.2 Dictionary search.

When using Hellberg's description of the system as a basis for our implementation, we found that no algorithm or even general criteria for organizing the dictionary search were provided. The solution we decided best fulfilled the requirements of the system was to conduct an exhaustive backward search on the string being analyzed such that for *bildrulle*, for example, the following technical stems would be sought in the dictionary:

bildrulle, bildrull, bildrul ... bild, bil, ... b

In this way all the possible technical stems of the initial string are found and the rest of the string is subsequently analyzed in the same way.

2.2.3 Linguistic transparency

One of Hellberg's claims was to provide "a detailed account of Swedish morphology". The system is not immediately comprehensible to alinguist on account of the large number of paradigms (235), many with only marginal differences (see [Brandt, 1985]). Only brief descriptions of each paradigm have been provided (which has not made it easy to add new words to the dictionary along with the correct paradigm

reference). Hellberg mentions in his introduction that some linguistic generalizations have given way to technical considerations. I would agree with Brandt in his criticism of the fact that the regular phenomenon of e-syncope is spread out over several paradigms, but not in his preference for another model with fewer paradigms that can capture broader linguistic generalizations. This point of view ignores the use the system was designed for - which was not just a theoretical model but the design for a practical implementation which is required to handle the individual idiosyncrasies of a word or group of words in authentic texts.
Also clouding the linguistic transparency is the fact that while most of the derivative endings have been included in the dictionary, a small subset have been tucked away in a subroutine of their own (93) as they differ from the others in that they demand the lack of an unstable vowel in the stem. The consequence of this is that derivative endings are to be found both in the dictionary and in the subroutines - a deviation from Hellberg's original aim not to build out theparadigms and subroutines in this way.

During an analysis the grammatical features are picked up when the paradigms and subroutines are being traversed. However, some features are attached to words in the dictionary. These are words which are simply accepted in their dictionary form (eg. *att* the infinitive marker) or others which are referred to a cross-reference dictionary (eg. *trivas* which is labelled as a verb). Features are, therefore, to be found in the paradigms, subroutines and, for certain words only, in the dictionary.

For anyone else intending to implement the system it may be mentioned that there is a misprint (?) in paradigms 825-827 in [Hellberg, 78] where the threshold level given as 0 should in fact be 1 (to allow for the genitive *s*). Also the list of abbreviations on page 130 is incomplete; the following are missing : *psp - present participle, pas - passive, kom - comparative, suv - superlative, opt - optative.*

2.3 SEMANTICS

In the first section of this paper one of Hellberg's reasons for not including the derivative endings in the paradigms and subroutines was given: restrictions governing possible derivative endings are of a semantic rather than morphological nature. The system accepts, therefore, **bilig*, **biling*, **bilarinna*, **obil*, etc.

```
SMORF Window
60-SMORF: BILIG
(BILIG (("BI" "LIG" (adj indef sing non-neut base))
        ("BIL" "IG" (adj indef sing non-neut base))))
61-SMORF:
NIL
61-SMORF: BILING
(BILING (("BI" "LING" (noun non-neut indef sing base))
         ("BIL" "ING" (noun non-neut indef sing base))))
62-SMORF:
```

Later on, though, Hellberg goes on to draw some semantic distinctions; consider the example given earlier: *Ögonblick* should be included in the dictionary as a separate item "since it is not semantically derivable" from its component parts. He continues "...the principle should be that compounds and derivatives whose meaning cannot be derived from their components must be entry words of their own in the dictionary". This means that *avlasta*, *avsluta* and *avskilja* are not listed in the dictionary, whereas *avlida*, *avse* and *avrätta* are as the latter only have a transferred meaning. Thus the following differences appear:

```
SMORF Window
96-SMORF: AVLASTA
(AVLASTA (("AV" "LASTA" (verb inf))
          ("AV" "LASTA" (verb imp))))
97-SMORF: AVSE
(AVSE (("AVSE" (verb inf))
       ("AVSE" (verb imp))
       ("AV" "SE" (verb inf))
       ("AV" "SE" (verb imp))))
98-SMORF:
```

The attention paid to semantics, then, is somewhat inconsistent and if the system is to be used for purposes other than word analysis, a more systematic approach may be required. To justify the need for including semantic considerations, try the following example, *smörgås*.

```
SMORF Window
57-SMORF: SMÖRGÅS
(SMÖRGÅS (("SMÖRGÅS" (noun non-neut indef sing base/gen))
          ("SMÖR" "GÅS" (noun non-neut indef sing base/gen))
          ("SMÖR" "GÅ" (verb inf pas))
          ("SMÖR" "GÅ" (verb pres pas))
          ("SMÖR" "GÅ" (verb imp pas))))
58-SMORF:
```

3. CONCLUSIONS (+ advantage, - disadvantage)

+ The system is a powerful analyzer of Swedish word forms.

+ It meets its claims of explicitness and exhaustiveness.

- The over-production of unlikely analyses will have to be further restricted for most purposes.

- It is not as linguistically transparent as it could be.

- It has no morphotactic rules built in; the anaylysis is purely on a character matching level.

+ It has potential as a generator where, given a word and a set of features, it would generate appropriate word forms. This would have to be limited to inflected forms (and possibly derivations), as it is difficult to predict the type of linking element (if any) otherwise required.

REFERENCES:

Brandt, Søren. *The Influence of Computers on Linguistics and Language*, Eighth Scandinavian Conference of Linguistics, ed. Tugely. 1985.

Hellberg, Staffan. *The Morphology of Present-Day Swedish,* Almqvist & Wiksell International. Stockholm. 1978.

ABBREVIATIONS USED IN THE TEXT

Sprakdata's original abreviations retained in the paradigms and subroutines:

bes	definite (Sw. *bestämd*)	obe	indefinite (Sw. *obestämd*)
gen	genitive	plu	plural
gru	base form (Sw. *grundform*)	sin	singular
nn	noun	utr	non-neuter (Sw. *utrum*)

Abbreviations used as output of the system as in the examples shown:

abbrev	abbreviation	inf	infinitive
adj	adjective	nominal-fn	nominal function
adv	adverb	non-Swed	non-Swedish
art	article	pas	passive
base	base-form	past-prt	past participle
def	definite	plur	plural
imp	imperative	pron	pronoun
indef	indefinite	sing	singular