Improbable morphological forms in a computational lexicon

Kristin Hagen and Lars Nygaard
The Text Laboratory, University of Oslo
http://www.hf.uio.no/tekstlab/
{kristin.hagen|lars.nygaard}@iln.uio.no

Abstract

In the construction of a computational lexicon, one of the problems is how to handle cases where words have a partial morphological paradigm. In this paper we will describe this problem and sketch how we implemented a system for capturing the degree to which forms should be considered improbable. Also, we will describe how our results can be used in language applications.

1 Introduction

For semantic and morphological reasons some words are considered only to have a partial morphological paradigm. This can be abstracts like *kjærlighet* (love) or uncountable nouns like *melk* (milk) that only occur in singular. Or it can be adjectives like *entusiastisk* (enthusiastic) not inflected for degree because the adjective has five syllables.

For Norwegian, words with a partial morphological paradigm include:

- Nouns only used in singular. Most nouns have plural forms:

  stein → steiner
  (stone → stones)
  sang → sanger
  (song → songs)

  But not all:

  snø → *snøer
  (snow → *snows)
  musikk → *musikker
  (music → *musics)

- Adjectives not inflected for degree. Many adjectives have morphological comparative and superlative forms:

  pen → penere → penest
  (pretty → prettier → prettiest)
  god → bedre → best
  (good → better → best)

  But not all:

  abnorm → *abnormere → *abnormest
  (abnormal → *abnormaler → *abnormalest)
  spesiell → *spesiellere → *spesiellest)
  (special → *specialer → *specialest)

- Verbs not used attributively. Many verbs have attributive forms:

  en skrevet bok
  (a written book)
  et spist eple
  (an eaten apple)

  But not all:

  * en gått tur
  (* a walked walk)
  * et abonnert tidsskrift
  (* a subscribed magazine)

2 Partial paradigms in Norwegian dictionaries

In Norwegian dictionaries there is no information about whether an adjective can be inflected for degree or not. Grammars normally list some morphological criteria claiming that an adjective can not be inflected for degree if the adjective is too long, normally estimated as an adjective with three or more syllables. Adjectives with suffixes like *-ende*, *-et(e)*, *-a*, *-sk* and probably *-s* and *-en* also have a partial paradigm according to the rules.

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 43–46
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

For verbs, there is no systematic information about attributive use in the dictionaries, but *Bokmålsordboka* (Wangensteen, 2004) lists examples for some verbs in the definition part of the dictionary. For nouns, *Bokmålsordboka* has classified some nouns as singular nouns, but the classification is not complete.

For the computational lexicon the present authors use, *Norsk ordbank*, all words were originally given full paradigms.

3 Improbable, not impossible

The problem with many of these «extra» inflected forms is that they are not totally impossible, only *improbable* to a varying degree. When searching for an abstract like *musikk* on Google, *musikkene* is actually found more than twenty times. *Gåtte* is also frequently used according to Google, and *spesiellere* is used once:

> Men selv om de to *musikkene* har fellestrekk, er mye ulikt.
>
> (Though the two musics do have similarities, there are many differences.)
>
> Fikk dere vekttall per antall *gåtte* fotturer?
>
> (Were you awarded points per walked walk?)
>
> ... men det som er enda *spesiellere* i Gawadar er sjøen.
>
> (... but what is even specialer in Gawadar is the lake.)

4 Including improbable forms is problematic

To handle examples like *musikkene* and *gåtte*, improbable forms have to be present in a computational lexicon. Including the forms is, however, problematic as well:

- From a linguistic perspective because the representation does not reflect the typical usage
- From the perspective of computational linguistics and language technology because the extra forms introduce unnecessary ambiguity:
  - In analysis, the forms are homographs with other forms. Example: *gjelder* (improbable plural form for «debts») is homonymous with *gjelder* (verb, present tense of «be valid for» or «applies to»)
  - In generation, the application will be presented with forms that are not idiomatic to use.

    *Han er prinsipiellere enn jeg trodde
    (* He is fundamentalier than I thought)
    Han er mer prinsipiell enn jeg trodde
    (He is more fundamental than I thought)

When Norsk ordbank was going to be used in the LOGON machine translation project (Oepen et al., 2004), a project which uses deep linguistic knowledge for both analysis and generation, the need to identify the lemmas with partial morphological paradigms became more urgent.

5 A heuristic score

The main task was identifying lemmas with improbable forms. Additionally, we needed to store and use this information in a way that would give minimal ambiguity, while retaining full coverage.

We implemented a system for creating a heuristic score, attempting to capture the degree to which forms should be considered improbable. The score was based on frequencies in the Oslo Corpus of Tagged Norwegian texts (Johannessen et al., 2000). The Oslo Corpus is tagged with the Oslo-Bergen tagger (Hagen et al., 2000), a constraint grammar tagger where ambiguity is left if none of the constraints can disambiguate between two or more readings. In the Oslo Corpus this results in both ambiguous occurrences of word forms and unambiguous word forms. The formulas for nouns look like this:

For nouns with one or more occurrences in plural form:

$$\frac{P+Q}{Fk}$$

For nouns with zero occurrences in plural form:

$$(0-F)m$$

P is the total number of occurrences in plural form (both ambiguous and unambiguous), Q

is the number of unambiguous occurrences in plural form. F is the total frequency, and m and k are weighting constants (we used $m = 2$ and $k = 100$).

A positive score indicates that a word has a full paradigm. A negative score indicates the opposite. The score also says something about the probability: A high negative score says that it is more unlikely that the noun can be used in plural than if the negative score is low.

Results are given in tables 1 and 2. The results are based on a medium-size corpus, where infrequent forms where penalized, since the score was likely to be less reliable: If there are two occurrences in singular and none in plural that is not necessarily an indication that the word only has a singular form.

We also found some problems with homonymy: *land* (a rare word for «urine from domestic animal») is clearly not a plural word, but since it is ambiguous with *land* (country), a frequent homonym in the corpus, it got a score indicating plural.

6 Improbable forms in Norsk ordbank

Instead of removing improbable forms from the lexical database Norsk ordbank, we will choose to flag them as improbable using the heuristic improbable-score. In this way, application developers can select what forms will be used. For example, a language generation application can choose not to include the improbable forms. For a tagger like the Oslo-Bergen-tagger, the forms can be included in the initial analysis, but removed later unless they are unambiguous.

In the Oslo-Bergen-tagger we mark the improbable words as <sjelden> (<rare>), and choose them as the correct reading only if the context is unambiguous.

In the following example *disse* in the meaning *huske* or *gynge* (swing) is marked as <rare>, but will still be disambiguated in a sentence like:

> Parken var en liten grønn plett med ei disse og ei sandkasse.
>
> (The park was a small green patch, with a swing and a sand pit.)

In some contexts it is hard for a tagger without semantic rules to disambiguate between the noun *disse* and the pronoun *disse*. When the noun *disse* is marked as <rare>, this reading can be deleted after the ordinary linguistic rules are applied:

> Ved behov kan disse allikevel kontaktes ved første anledning.
>
> (If need be, they can be contacted at the first opportunity.)

7 Further work

Although the initial results look promising, a full scale evaluation of the method remain. We plan to evaluate

- against a gold-standard set of hand annotated lemmas
- for application-specific tasks, including analysis and generation

References

Kristin Hagen, Janne Bondi Johannessen, and Anders Nøklestad. 2000. A constraint-based tagger for norwegian. In *Proceedings of the 17th Scandinavian Conference of Linguistics*.

Janne Bondi Johannessen, Anders Nøklestad, and Kristin Hagen. 2000. A web-based advanced and user friendly system: The oslo corpus of tagged norwegian texts. In *Proceedings of the Second International Conference on Language Resources and Evaluation*.

Stephan Oepen, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, and Victoria Rosén. 2004. Som å kapp-ete med trollet? Towards MRS-based Norwegian-English Machine Translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*.

Boye Wangensteen. 2004. *Bokmålsordboka*. Universitetsforlaget.

lemma	*F*	*P*	*Q*	*score*
norsk (norwegian)	4598	0	0	-9196
død (death)	3360	0	0	-6720
musikk (music)	2934	0	0	-5868
politikk (politics)	2825	0	0	-5650
forskning (research)	2483	0	0	-4966
undervisning (teaching)	1979	0	0	-3958
kaffe (coffee)	1871	0	0	-3742
litteratur (literature)	1418	0	0	-2836
folketrygd (social security)	1316	0	0	-2632
bistand (aid)	1234	0	0	-2468
snø (snow)	1216	0	0	-2432
tillit (trust)	1143	0	0	-2286

Figure 1: Nouns least likely to have plural forms.

lemma	*F*	*P*	*Q*	*score*
år (year)	55928	44563	34462	3951179
krone (crown)	14396	13671	13671	1367005
prosent (per cent)	12801	12408	12025	1221554
barn (children)	20488	15114	7913	1151293
folk (people)	13445	11576	10342	1095818
menneske (human being)	14254	10322	10322	1032127
forhold (relation)	19407	14323	5014	966800
million (million)	10033	8981	8981	898010
øye (eye)	10509	8644	8644	864317
kvinne (woman)	14739	8283	8283	828243
mann (man)	20910	8462	6977	771913
dag (day)	37388	7100	7100	709981

Figure 2: Nouns most likely to have plural forms.