Porting morphological analysis and disambiguation to new languages

Trond Trosterud
University of Tromsø

1 Introduction

This poster presents morphological parsers and disambiguators for North and Lule Sámi. Utaric languages spoken in the Northern parts of Norway, Sweden and Finland (the project's home page is http://giellatekno.uit.no). The parsers use Xerox tools (www.fondbook.com) for morphological analysis, and constraint grammar for disambiguation (sourceforge.net/projects/condyn/). It also report from the experiences with porting the system for Sámi to other Utaric languages.

2 The Tromsø disambiguation project

2.1 Morphological analysis

The morphological analyses of the project are based upon a two-level analysis with finite state automata, cf. [3]. We use Xerox software, btw for lexical analysis and segmental morphology, etc for morphophonological processes, perf for preprocessing and axt for case conversion and integrating the parts into a whole, e.g. [1] and http://www.fondbook.com. The lexical analysis and the segmental morphology operate on two levels, one surface level for roots and affixes, and one underlying level for lexemes and grammatical properties. The surface stage becomes the underlying level for the morphological rule set, taking a root- and suffix string as input, enriched with morphophonological information, and transforms it to the wordform we know from the written language, cf. [23].

2.2 Suffixes

The lexicon contains all the roots of the language. The roots are classified according to part of speech (POS) and stem class, directing words taking the same suffixes and undergoing the same morphophonological processes to the same continuation lexicon.

2.3 Morphomorphology

Morphophonological processes are taken care of in a different component. The rules involved in generating the forms shown in the float diagram are shown below. In the figure the lexicon MUORRA contains words undergoing consonant gradation, and NOADE contains words that do not. The case forms that show weak grade are kept in a separate lexicon, and the MUORRA nouns get a weak grade mark q1 before they enter this lexicon. This means that the root consonants (here: st and v) undergo gradation, change to sl and v, their accusative forms are buatit, buayit, and not *buatitit, buayeyit, although we have *garthit, gurgit. They undergo the following rules (using the bow ruformalism):

3 Disambiguating Sámi

As an example, let us take the sentence Mii ev baa jitan mansital a "We haven't told it", with the verbs baa ‘to be’ and mansital ‘to tell’. The sentence is given the following analysis, prior to disambiguation:

4 Extending the work on Sámi to other languages

4.1 Sharing infrastructure

The North Sámi transducer has been ported to several Utaric languages. In the beginning of the construction of fully languages independent files like lists of punctuation marks and letters for case conversion. Lexicographical resources like person names and place names may be reused (with different continuation lexica for different languages), the same goes for productive loan-words. Thus, certain classes of Norwegian verbs and nouns can potentially be borrowed into the different Sámi languages, with slightly different adjustment strategies. The same goes for Russian loan-verbs in Udmurt, Mari, etc. A way of re-using such resources is to create a common pool of potential loan-words, with language-specific adjustment strategies. This strategy also makes it easy to remove words from certain applications, e.g. partial spell-checkers.

A major advantage was found in replicating the basic file structure for each new language. The initial development phase was shortened, and work on multilingual projects was easier with a uniform file structure.

4.2 Localisation

North Sámi has 7 letters outside ascii (for example â). The origin of the project dates back to 1994, so it is localised in Latin 1, using digraphs for the 6 non-Latin1-letters: â, ê, ð, ñ, ò, ù, etc. The current versions of the Xerox tools support UTF-8. We have made UTF-8-based parser prototypes (for Komi and Hindi). The source code was written in Perl and C (and C++ if available) and not really an option for text editor. Unicode support on Linux is still not the de-facto standard (cf. http://www.cl.cam.ac.uk/mkp/utf8/manual.html), but once the technical problems are solved, there are several benefits of writing the source code in UTF-8: dictionary files may be imported directly, they are easier to proof-read, text may be analysed without conversion, and the morphophonological (actually morphomorphemical) rules may generalise directly over the actual letter sequences.

The following is an extract from the Komi analysis:

```
LEXICON NOUNSUF
+N+Sg: NOUNSUF-2 ;
+N+Pl:%>яс NOUNSUF-2 ;
LEXICON NOUNSUF-2
 PREPX ;
 PRECX ;
 ACCLEX ;
 LOCLEX ...
+Car+:%>а POSTPX ;
LEXICON ELALEX
+Ela:%>ысь #;              
+Ela:%>сь POSTPX ;
```

This is an extract from the Hindi analysis.

```
the
```

4.3 Working on similar languages in parallel

Porting grammatical analysis to new languages is especially useful for language continua, or groups of closely related languages, where the linguistic analyses may be reused. Examples include Turkkic, Bantu, Dravidian, Indo-Aryan, Slavic, Romance and Scandinavian.

4.4 Grammatical approaches in language technology

Grammar-based disambiguation has been known to provide good results, compared to stochastically-based approaches [5].

Looking at minority languages, the arguments in favour of grammar-based approaches are even stronger. In the cases of the Sámi languages or the Utaric languages of Russia, there is not a choice between using the multimillion electronically available corpus or not. There is no such corpus. Rather, what is available is a grammar, and in most cases a rescoped grammar dictionary. With these two tools (especially if the dictionary is electronically available, it is possible to build good transducers and disambiguators within a couple of years, or, after a while, within even shorter time. For inflectial languages with hundreds of inflected forms for each lexeme (and sometimes more), transducers based on stem classes and inflectional paradigms are the only way of ensuring good coverage of the morphological information. Another option than the manual writing of transducers is to apply to a combined version of human elicitation and machine learning, as argued by [4]. The approach should be more suited to families of very similar languages, like the Turkkic or Bantu languages. Whether these semiautomatic transducers are as easy to update as hand-made ones, or whether they will look more like a "black box", remains to see.

Most minority languages do not have many and rich enough speakers to attract commercial language technology projects. Linguists still write reference grammars for these languages. For grammar-based language technology, it makes perfectly sense to be integrated in this descriptive work. Making a morphological analyser is the way of checking the coverage of any language description. Practical applications like spell-checkers should then come as a side effect of this type of basic descriptive work.

5 Summary

The present poster has given an overview of work with morphological transducers and disambiguators for some related Utaric languages. The work conducted so far shows that the building of transducers and disambiguators will benefit from sharing code written in an as language-independent way as possible.

References


Porting morphological analysis and disambiguation to new languages

Department of Linguistics
N-9037 University of Tromsø, Norway
trond.trosterud@hum.uio.no
http://giellatekno.uio.no/