Morphisto - An Open Source Morphological Analyzer for German

Andrea Zielinski, Christian Simon

September 18, 2008

Morphisto - An Open Source Morphological Analyzer for German

- Motivation
  - Basic Idea
  - Short Introduction to SMOR
- Bootstrapping a Morphological Analyzer
  - Extracting morphological information from a MRD
  - Managing the Lexical Data
- Lexical User Interface of Morphisto
  - Cycle of Testing an Reengineering
- Test Results
- Integration into the TextGrid Workbench

Motivation

Motivation:

- Morphologcal analysis is the basis for a range of applications in TextGrid - an Infrastructure for ehumanities:
  - Lexical Lookup in Dictionaries
  - Annotation of Texts
  - Translation from/into other language stages of German
- No open-source morphological analyzer available for German at present
- This is particularly true for the lexicon component which is labour-intensive to build

Basic Idea

- Use GPL-licensed (except the lexicon) SMOR morphology for German as a starting point
- Define lexical entries for the most common 30,000 German words as defined by the German reference word list "DeReWo"
- Implement additional tools for the management of the lexical data
- Build easy-to-use web services and integrate them into the Eclipse Rich Client Platform

SMOR

SMOR is ...

- a computational morphology for German
- FST-based (SFST toolkit, cf. Schmid 2004)
- licensed under the GPL (except the lexicon)

What it can do

SMOR analyzes inflectional forms of German

- simple (or compound) lexemes

What it can do

SMOR analyzes inflectional forms of German

- simple (or compound) lexemes
- derivational constructions

What it can do

SMOR analyzes inflectional forms of German

- simple (or compound) lexemes
- derivational constructions
- complex word formations need not be stored in the lexicon but can be analyzed on-the-fly

What it can do

SMOR analyzes inflectional forms of German

- simple (or compound) lexemes
- derivational constructions
- complex word formations need not be stored in the lexicon but can be analyzed on-the-fly
- flat analyses:
  Bahn<NN>Hof<NN>Halle<+NN><Fem><Nom><Sg>

Lexical Aquisition for MORPHISTO

Resources used:

- DeReWo Lemma List
  `http://www.ids-mannheim.de/kl/derewo/`;
- Adelung(1793) - Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart
  `http://www.zeno.org/Adelung-1793`; lexicon published by the “Digitale Bibliothek"; early NHD dictionary; free for the public; covering more than 65,000 entries.
- Dictionary of Foreign Words (Deutsches Fremdwörterbuch)
- grammis `http://hypermedia.ids-mannheim.de/pls/public/gramwb.ansicht`

Example Entry from Adelung(1793)

An excerpt lemma entry from Adelung (1793) from which we aquired our data is given below:

1. Das Futter, des -s, plur. ut nom. sing. die Bekleidung eines Körpers von außen oder von innen; [..]
2. Das Futter, des -s, plur. ut nom. sing. 1) Alles, was Menschen und Thieren zur Nahrung dienet; ohne Plural [..]

Lexical Database Scheme

The acquired lexical data was joined with the DeReWo lemma list and stored in a lexical database.

Why is a lexical database helpful?

- developers want to to add, modify, remove lexical data
- finite-state based lexicon format is difficult to read and maintain

Lexical Database Scheme

What is needed?

- An exchange format that is independent of the specific finite-state platform
- Scripts that convert lexical data to the originial SMOR lexicon format
- Validating the lexicon against a schema
- Enhanced user interface for lexicographic work

Conversion of an XML-based Lexical Entry to SMOR

$< smor >$

$\quad < BaseStem >$

$\qquad < Lemma > Atlas < /Lemma >$

$\qquad < Stem > Atlanten < /Stem >$

$\qquad < Pos > NN < /Pos >$

$\qquad < Origin > nativ < /Origin >$

$\qquad < InfClass > NMasc/Pl < /InfClass >$

$\quad < /BaseStem >$

$< /smor >$

$=>< Basestems > Atlas : n <>: t <>: e <>: n$

$\quad < NN >< base >< nativ >< NMasc/Pl >$

Figure: Example Conversion for *Atlas (atlas)*

Cycle of Testing and Reengineering

- Checking the analysis of the DeReWo list (on the Adelung transducer)
- Adding missing base stems (for about 5,000 entries)
- Correcting false inflection classes or features (e.g., for *Geister* (ghosts))
- Adding word formation rules (e.g., for *StudentIn/Innen* (male and/or female student(s))
- Adding rules for the derivation of compound or derivation stems (e.g., *Peters-kirche* (Peter's church))

Intermediate Result

- Manually creating the simplex units together with their required features is time-consuming
- In the worst case, an intensive study of the documentation and software code is required
- Gold standard would be beneficial
- Fine-tuning for stems and affixes that were likely to produce ambiguities
  - Include complex words that produce segmentation errors (e.g., *Tee-nager* (tea rodent) instead of *Teen-ager*) into the lexicon
  - Assign the tag <NoDef> or <Initial> to short or antiquated words to restrict their productivity

Transducer Lexicon Statistics for Adelung and Morphisto

Lexicon	Adelung	Morphisto
Basestems	32152	17339
- Nouns	20605	7833
- Proper Nouns	2	1053
- Verb stems	7426	4300
- Adjectives	4061	3178
- Adverbs	2	781
- Closed Word Classes	28	190
Derivation Stems	63	67
Compound Stems	30	181
Prefix Stems	94	213
Suffix Stems (Derivation Rules)	404	410

Table: Frequency of morphological units in the transducer lexicons

Evaluation on the 'Ispell Test Corpus'

The wordform list provided by Ispell has been used for the evaluation of Morphisto. Approximately 225.833 words of Ispell are unknown to Morphisto. The tests on randomly selected subsets of 100 inflected German wordforms in different frequency ranges reveal the number of correct, missing and spurious readings, e.g. the precision and recall rates.

Evaluation on the 'Ispell Test Corpus'

Frequency Classes	Precision	Recall	F-Measure
$F_0 - F_4$	100.00	100.00	100.00
$F_5 - F_8$	99.63	99.63	99.63
$F_9 - F_{12}$	98.74	87.71	92.90
$F_{13} - F_{16}$	98.25	93.85	95.39
F_{17} - F_{20}	93.21	84.36	88.56
F_{21} - F_{25}	91.77	81.00	85.33
Average	96.93	91.09	93.63

Table: Test results on ispell for different frequency classes

Remaining Issues

- Dealing with unknown base stems

Remaining Issues

- Dealing with unknown base stems
- Dealing with ambiguous analysis due to homonyms, overgeneration or complex compounds

Using a Guesser for dealing with Unknown Words

- Implementation of 2 guessers for upper and lower case words
- Guesser lexicons include approx. 7.500 entries each
- The longest match analysis is preferred

Result: For about 50% of all unknown words the head (and paradigm) could be guessed
But: In many cases, the guess is wrong

Dealing with Ambiguous Analysis

- Implementation of a Markov Model based on wordclass probabilities
- Bigrams are trained on a small corpus of 100,000 words
- Probability that a word belongs to a certain category

Alternative Approach that does not automatically favour the analysis with the fewest constituents
Result: Disambiguation succeeds in many cases

Demo Applications

- Wordform Analysis with Morphisto
- Generation of Inflection Table with Morphisto

You are welcome to download our Morphisto Transducer Lexicon for German