Morphological Analysis for Less-Resourced Languages: Maximum Affix Overlap applied to Zulu

Uwe Quasthoff[1], Sonja Bosch[2], Dirk Goldhahn[1]
[1]Natural Language Processing Group, University of Leipzig, Germany
[2]Dept. of African Languages, University of South Africa

Less-resourced Languages

Lack of text resources:

- Dictionaries are often outdated
- No machine-readable dictionaries
- Limited amount of text in the Web

Some corpora exist (e.g. University of Pretoria, Language Resource Management Agency and Leipzig Corpora Collection) - limited in size, not annotated and often not accessible.

Lack of NLP data and tools:

- Training data for POS-tagger
- Training data for morphology

Complex Morphology of Zulu

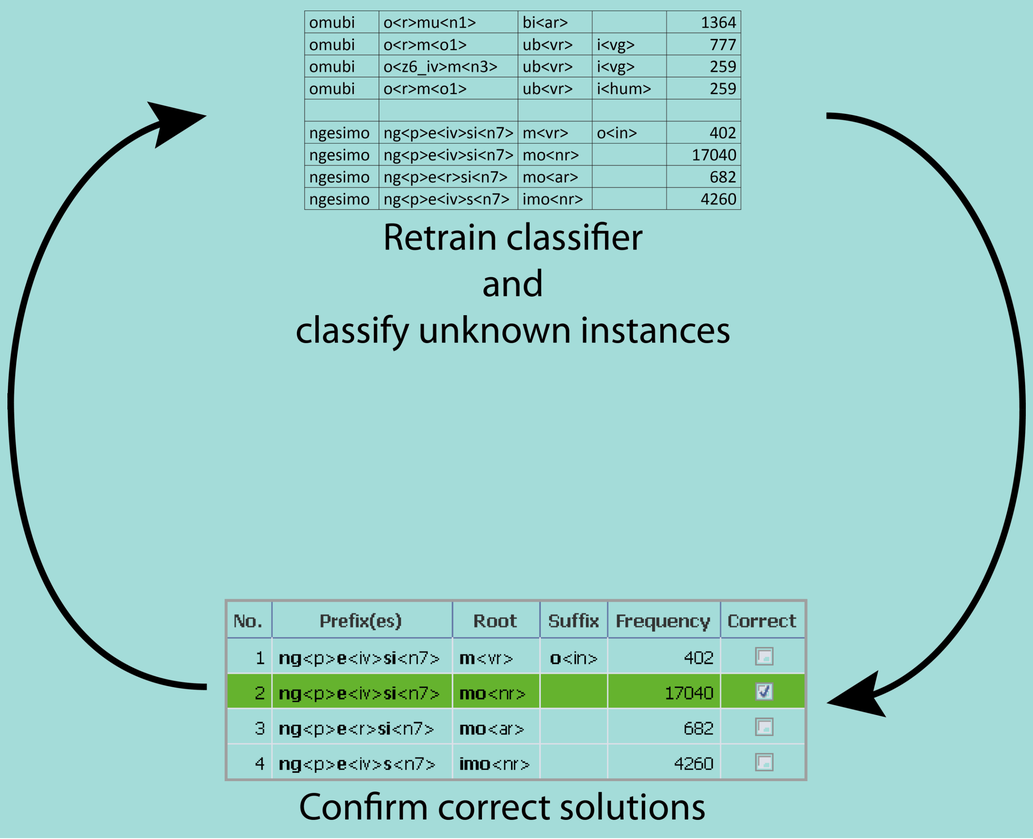
Zulu [ISO 639-3: zul]

Nominal classification system: nouns categorized by prefixal morphemes (with class numbers)

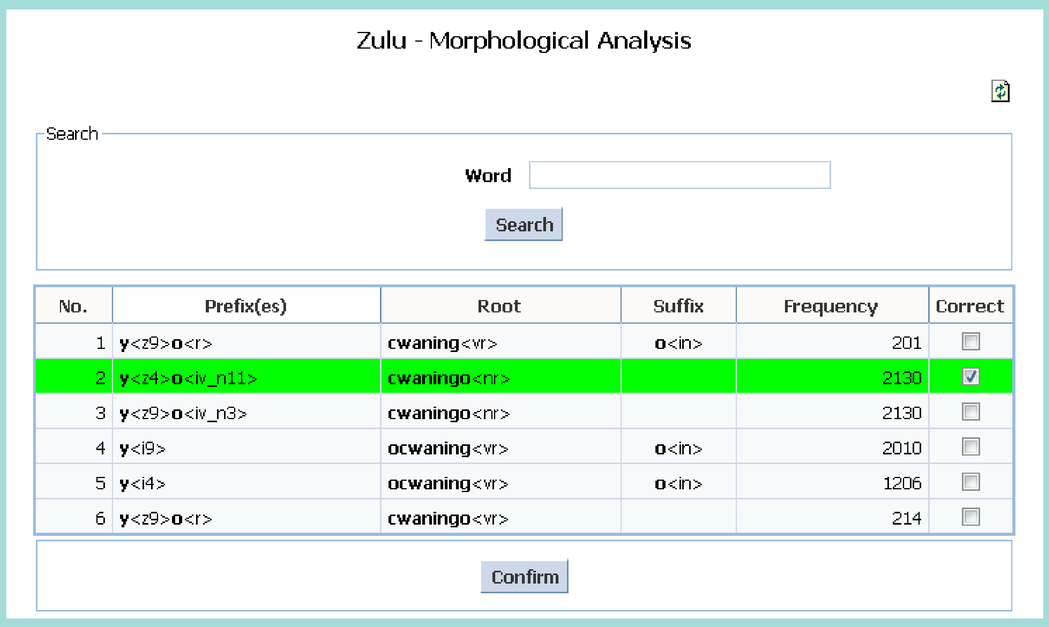
Concordial agreement system: noun class prefixes use concordial agreement for linking nouns to verbs, adjectives, pronouns, possessives etc.

abantu	**abaningi**	**bangayichitha**	**imali**	**yabo**
aba-ntu	**aba**-ningi	**ba**-nga-**yi**-chitha	**i**-mali	**ya**-**bo**
people	many	they-may-it-waste	money	of-them
[Many people may waste their money.]				

The Collaboration Loop



Training Data Web Editor



Machine Learning: Maximum Affix Overlap Algorithm

For all segmentations of the word w into **three segments** w1 (prefixes), w2 (root) and w3 (suffixes), where w1 and w3 might be of zero length:

- For each word x in the training set having exactly the **prefix sequence** w1 we collect the pair (morphological analysis of w1, w2 with the tag of the root of x).
- For each word x in the training set having exactly the **suffix sequence** w3 we collect the pair (w2 with the tag of the root of x, morphological analysis of w3).

From this, form triple by joining on **identical root tags**: (morphological analysis of w1, w2 with the tag of the root of x, morphological analysis of w3).

Interesting features: **length** of w2 and **frequency** of identical triples above.

Procedure above allows considering affixes next to stem as part of stem, therefore **shorter stems** should be preferred.

In case of **multiple decompositions** with same stem (or different stems of same length): decompositions ranked according to frequency.

In general, we set a **frequency threshold** of 2 for decompositions to be considered.

Applications for the Collaboration Loop

- Inflection type / baseform reduction, morphological decomposition, compound decomposition
  - Special problem: Identifying new roots for dictionaries
  - Special problem: Identifying writing variants
- Classification tasks for subject areas or relations (as in WordNet)
- Bilingual translation equivalents

Contact: *{quasthoff, dgoldhahn}@informatik.uni-leipzig.de, boschse@unisa.ac.za*

Gedruckt im Universitätsrechenzentrum Leipzig