Computational Analysis of Morphosyntactic Categories in Georgian

Sophiko Daraselia
University of Leeds, UK
s.daraselia@leeds.ac.uk

Part-of-speech tagging is an established procedure in corpus linguistics. There is a wide range of applications of part-of-speech tagging software and tagged texts and corpora. These include information retrieval, machine translation, sentiment analysis, production of corpus-based grammars and dictionaries etc.

This paper describes part-of-speech tagging of Georgian, a member of Kartvelian language family. Every language presents its own particular problems in developing part-of-speech tagging technology. Georgian is a morphologically complex agglutinative language with split ergativity, meaning that it presents number of interesting and possibly unique problems. For example, how to treat suffixaufnahme (double casing) case? How to tag argument agreement in verbs? How to treat numerous enclitic particles and postpositions?

In the presentation, I will discuss 1) the major challenges involved in the process of designing a hierarchical decomposable tagset for Georgian; 2) design principles of the Georgian tagset (KATAG), 3) the performance level of the probabilistic Markov model tagger using the KATAG tagset and outline 4) the main factors that affect the performance level of the tagger.

The TreeTagger program (Schmid 1994) has been used to perform tagging in Georgian. The TreeTagger (parameter files) was trained on the KaWaC corpus (Daraselia & Sharoff 2014, 2015). The annotated data used to train the TreeTagger program are as follows:

Table 1: Training data.

Fullform lexicon	8,488 words
Training set	90,872 words (7,425 sentences, 7,500 unique word forms)
Open class tags	133 tags
Auxiliary lexicon	84,683 words

Several variations of the Treetagger program have been tested applying different parameters, such as replacing zero frequencies by 0.1 before the tag probabilities. The influence of the pruning threshold on the accuracy of the trigram version and the quatrogram version of the TreeTagger was also tested. However, increasing the context did not result in any improvement.

Table 2: Comparison of accuracy.

Method	Context	Accuracy
TreeTagger	bigram	88.45%
TreeTagger (0.1)	bigram	88.45%
TreeTagger (auxiliary lexicon)	bigram	70.00%
TreeTagger (revised auxiliary lexicon)	bigram	92.41 %
TreeTagger (revised auxiliary lexicon)	trigram	92.41 %
TreeTagger (revised auxiliary lexicon)	quatrogram	92.41 %

The TreeTagger (as other Markov model taggers) can incorporate linguistic information to some extent. This can be done by manipulating the lexicon, the tagset and the initial tag probabilities (Voutilainen 1999). Thus, I have manipulated the training lexicon taking into consideration appropriate biases. Table 2 shows that the revised (manipulated/normalised) auxiliary lexicon improved the performance of the tagger by 20%.

In the presentation, I will also outline several factors that determine the performance level of the TreeTagger program. The size of the training corpus had some effects on the performance level of the TreeTagger. The other factor is the morphosyntactic complexity of the Georgian language. For example, one of the major problems in Georgian morphosyntax (for disambiguation) is *morphological syncretism*, when one wordform belongs to the same morphosyntactic category, but it is difficult to identify appropriate morphosyntactic features, such as tense and argument agreement in verbs. For example, the Georgian verb **გაწუხებთ** **[gac'uxebt]** can have at least two readings:

- Verb, 3rd person of Subject singular and 2nd person of object Plural ("S/he/it bothers you (PL))
- Verb, 1st person of Subject plural and 2nd person of object singular ("We bother you).

Thus, the paper discusses the process of designing a hierarchical decomposable tagset for Georgian, design principles of the Georgian tagset (KATAG), the accuracy of the TreeTagger program and the main factors that affect the performance level of the tagger.

References

Daraselia S. & Sharoff, S. (2014) Morphosyntactic specifications for KaWaC, a Web Corpus for Georgian. International Conference - *Humanities in the Information Society II*. Batumi, Georgia. 326-329.

Daraselia S. & Sharoff, S. (2015) The main steps of the Georgian Web-Corpus construction. Tbilisi. Arnold Chikobava Institute of Linguistics, *Journal of Linguistics*, Vol. XXXVIII, 52-62.

Schmid, H. (1994) Probabilistic part-of-speech tagging using decision trees. Proceedings of *International Conference on New Methods in Language Processing*, Manchester, UK.

Voutilainen, A. (1999) A short history of tagging. In van Halteren, H. (ed.) *Syntactic Wordclass Tagging*. Dordrecht: Kluwer Academic Publishers. 3-4.