

An Amharic Corpus for Machine Learning

Björn Gambäck **Fredrik Olsson**
Userware Laboratory
Swedish Institute of Computer Science AB
Box 1263, SE-164 29 Kista, Sweden
{gambäck, fredriko}@sics.se

Atelach Alemu Argaw **Lars Asker**
Department of Computer and Systems Sciences
Stockholm University
Forum 100, SE-164 40 Kista, Sweden
{atelach, asker}@dsv.su.se

The paper describes a tagged corpus of Amharic news texts and some machine learning-based tagging experiments that have been carried out on the corpus. Amharic is the second most spoken Semitic language in the World (after Arabic) and used for countrywide communication in Ethiopia. It is highly inflectional and quite dialectally diversified.

We have collected a total of 8715 Amharic news articles from the years 2001-2004. All the news texts in the corpus originate from the Walta Information Center, a private news and information service located in Addis Ababa, Ethiopia. At its website www.waltainfo.com, it provides Ethiopia-related news in English and Amharic on a daily basis. Part of the corpus, the 1065 news texts (210,000 words) from year 1994 in the Ethiopian calendar (parts of the Gregorian years 2001–2002), has been morphologically analysed [AA07] and manually part-of-speech tagged by staff at the Ethiopian Languages Research Center at Addis Ababa University [DG06]. The tagset consists of ten basic classes: Noun, Pronoun, Verb, Adjective, Preposition, Conjunction, Adverb, Numeral, Interjection, and Punctuation, plus one extra tag for problematic (unclassified) words. The ten basic classes were then further divided into a total of thirty subclasses.¹ The original corpus has been used as the basis for building efficient and accurate part-of-speech taggers semi-automatically using machine learning techniques and tools developed for other languages. The process has allowed us to spot several errors and inconsistencies in the corpus which has been refined by both automatic and manual measures.

The presentation will describe the extraction of the corpus from the web, the necessary clean-up measures that had to be undertaken after the manual tagging, and the application of several machine learning techniques to the corpus in order to create state-of-the-art part-of-speech taggers for Amharic.

[DG06] Girma A. Demeke and Mesfin Getachew. Manual annotation of Amharic news items with part-of-speech tags and its challenges. *ELRC Working Papers*, 2(1):1–17, March 2006.

[AA07] Atelach Alemu Argaw and Lars Asker. "An Amharic Stemmer : Reducing Words to their Citation Forms". In proceedings of Computational Approaches to Semitic Languages: Common Issues and Resources. A Workshop at ACL 2007, Prague, Czech Republic. June 2007.