A Complete FS Model for Amharic Morphographemics

Saba Amsalu and Dafydd Gibbon

Universität Bielefeld, Germany

Our aim was to develop a complete morphographemic model for Amharic, the official language of Ethiopia, which urgently needs computational linguistic tools for information retrieval and natural language processing. Amharic is a Semitic language, with SOV word order and a complex morphology with consonantal roots and vowel intercalation, extensive agglutination, and both consonantal and vocalic stem modification. Previous computational models of Amharic lexemes are fragmentary, being restricted to affix stripping and radical extraction [2], [4], [3], [1]. The verb analysis by Fissaha and Haller [8] is the only previous FS based approach. FS and related approaches to other Semitic languages have also tended to concentrate on selected features of theoretical interest, such as the well–known analyses of Arabic intercalation [9], [5], [10].

In contrast, we have developed the first complete FS generator/analyser of Amharic morphology for all parts of speech (POS), including loan and native noun morphology, biradical, triradical and quadradical verb root generation, with vowel intercalation, conditioned internal vowel changes, agglutinative affixation of 13 affix classes, and full and partial reduplication. Phonological gemination is not represented in the Ethiopian Fidel orthography, and thus is not implemented.

Our development approach is linguistic rather than statistical, and includes novel features for modelling intercalation and reduplication. The analysis results are evaluated for precision and recall. The software used is XFST, with SERA (System for Ethiopian Representation in ASCII) romanisation. A port to Fidel Unicode is in progress.

Part of the system architecture is outlined in the activity diagram in Figure 1, which shows the FST verb cascade in generation direction, but is interpretable in both directions. Biradicals are generated from triradicals and quadradicals are independently generated; cf. [11], [6], [7], then vowels are intercalated, affixes are concatenated and phonological alternations processed.

Amharic has noun stem reduplication (with epenthetic vowel) (cf. Figure 2). A shell wrapper outside the FS system feeds XFST with a stream of words; the actual reduplication is then performed in the FS context using a novel bracketing 'diacritic' convention (not 'flag diacritic' [5]). Formally, this is a heuristic which treats the surface lexicon as the union of singleton sets of surface forms and applies the reduplication FST to the singleton sets individually.

For evaluation purposes we generate/analyse all POS separately. The FSTs for each POS are not unioned, because the individual FSTs are to be integrated into an FST chunk parser/tagger. Each POS is evaluated individually on a test corpus for standard recall and precision scores (ambiguity scores are currently implicit in the precision values). Recall/precision values for small finite

A. Yli-Jyrä, L. Karttunen, and J. Karhumäki (Eds.): FSMNLP 2005, LNAI 4002, pp. 283–284, 2006.

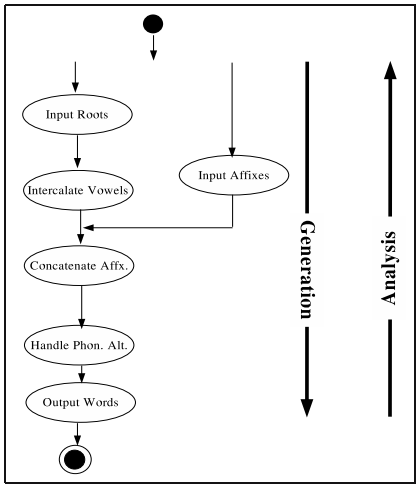

Fig. 1. FST cascade architecture

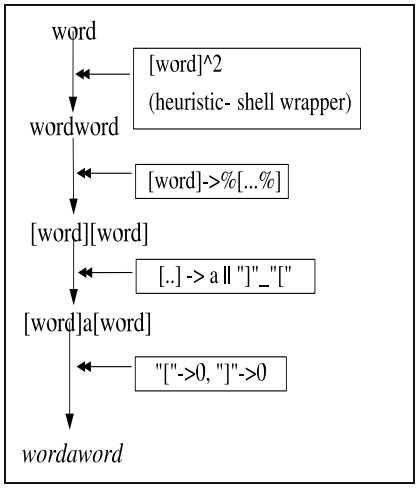

Fig. 2. Reduplication cascade

POS sets are, trivially, 1/1; verbs attain 0.94/0.54, nouns attain 0.85/0.94, and adjectives 0.88/0.81. The lower precision value for verbs is due to affix ambiguities (morphological syncretism).

References

1. N. Alemayehu and P. Willett. Stemming of Amharic words for information retrieval. *Literary and Linguistic computing*, 17(1):1–17, 2002.
2. L. A. Atalech Alemu and G. Eriksson. Building an Amharic lexicon from parallel texts. In *Proceedings of: First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, a workshop at LREC*, Lisbon, 2004.
3. A. Bayou. Developing automatic word parser for Amharic verbs and their derivation. Master's thesis, Addis Ababa University, Addis Ababa, 2000.
4. T. Bayu. Automatic morphological analyzer for Amharic: An experiment involving unsupervised learning and autosegmental analysis approaches. Master's thesis, Addis Ababa University, Addis Ababa, 2002.
5. K. Beesley and L. Karttunen. *Finite State Morphology*. CSLI, Stanford, 2003.
6. M. L. Bender, H. Fulas, and C. H. Dawkins. *Amharic Verb Morphology*. Michigan State University, African Studies Center, East Lansing, 1978.
7. C. H. Dawkins. *The Fundamentals of Amharic*. Sudan Interior Mission, Addis Ababa, Ethiopia, 1960.
8. S. Fissaha and J. Haller. Amharic verb lexicon in the context of machine translation. *TALN*, 2003.
9. M. Kay. Nonconcatenative finite–state morphology. In *EACL Proceedings*, pages 2–10, 1987.
10. S. Reinhard and D. Gibbon. Prosodic inheritance and morphological generalisations. In *Proceedings of EACL*, 1991.
11. B. Yimam. Root reductions and extensions in Amharic. *Ethiopian Journal of Languages and Literature*, 9:56–88, 1999.