

“Beja: How well does the current parser handle input texts?”

[Beja data bases, texts, and parser]

Paper to be presented at "Aethiopisches Forschungs-Colloquium", Berlin, 12-13 June 2009

Charlotte and Klaus Wedekind, Rottweil / Addis Abeba

Abstract: This presentation will have four parts, based on the heading of the paper: (1) "Beja" (2) "How Well", (3) "The Current Parser", and (4) "Input Texts".

Thus, (1) is a brief note about the relation of "Beja" to the Ethio-Semitic languages of this Colloquium, (2) "How Well", is about the status of the analysis, (3) "The Current Parser", is about the lexical data bases which are accessed by the software, and (4), "Input Texts", is about all texts which presently are available to linguists - and therefore have served as input to the parser.

(1) "Beja"

While Beja is habitually considered part of the Cushitic branch of the Afro-Asiatic family, its exact relation in this family has been the topic of various discussions, mainly because of its largely non-Cushitic lexicon. How "Semitic" is this Cushitic language?

In this presentation, the large number of Semitic loans - both from Tigre and Tigrigna, but mainly from Arabic - will be obvious when the lexical data bases will be characterized. In addition, the fact that nearly 50% of its verbs are prefix verbs - which is distinctly more than the percentage in Lowland East-Cushitic languages - makes it a semi-Semitic language.

In this presentation, the Beja lexicon as well as the Beja texts are taken from all areas where Beja is spoken. When gathering these data, we constantly were reminded of the fact that the Beja area is surrounded by Semitic speaking peoples.

(2) "How Well"

To characterize the status and the quality of the morphological parser, two very simple statistical measures can be employed: percentage of failures and percentage of ambiguities.

Currently, the parser still is largely limited to morphological parsing of words, and the goal is, of course, to have the lowest possible number of "failures", i.e. words which fail to be parsed - and the lowest possible number of "ambiguities", i.e. words which are parsed in more than one way.

Many "failures" simply indicate that the lexical entry is missing; other "failures" are due to different orthographic conventions. The interesting "failures", however, are those which show that the analysis is based on false assumptions.

Currently, the percentage of "failures" is about 1% or less, i.e. in a text of 2000 words, about 20 are failures). But the percentage of "ambiguities" is about 25%, i.e., in a text of 2000 words, about 500 parse in more than one way! Many words are genuinely ambiguous morphologically - e.g., *taku* means "man-is" and "man-my". Such words can only be disambiguated by their syntax (which is being implemented). Since, however, the present purposes of the parser are being satisfied by morphological parsing, the implementation of syntactic parsing is not pressing, and the rate of "ambiguities" is not too disappointing.

The present purposes include, among others, (a) to maintain consistency in orthographic conventions - even across dialects, and (b) to make the lexicon as inclusive as possible, given the diverse kinds of sources and dialects.

(3) "The Current Parser"

Two kinds of lexical "data bases" are being accessed by the software. The focus shall be on these data bases, not on the software: this presentation is about "Beja", not about software.

Only this much shall be said about the software: Two software packages are being used, "Toolbox" and "Carla". They have the advantage of (a) being freely available, and (b) being carefully maintained and updated by the persons who designed them: the Buseman family and the Black family.

The two kinds of data bases contain (a) Beja roots and names, and (b) Beja prefixes and suffixes. In the present parser, there are about 6000 roots and 400 names (loans included), but only 100 prefixes (24 demonstratives and articles included) and 120 suffixes (6 possessive pronouns included).

Opinions strongly differ - both among non-Beja and Beja linguists - concerning the status of articles, possessive pronouns, demonstratives, and certain adverbial particles: Should they be considered to be roots or affixes?

- Among Beja speakers, the tendency is to view everything as a root (there is a Beja linguist who even writes the articles as separate morphemes). There are borderline cases such as *oon ti a* "this the time", which most will write as one word *oont'a* "now", possibly patterned on Ar. "the time".

- Among western linguists, however, the tendency is to lump morphemes together - (starting with Almkvist and Reinisch, e.g. the demonstratives were attached). The view of linguists is based on the fact that morpho-phonemic changes only happen "inside word boundaries". The present parser allows for both options: The articles, for instance, are included both in the "affix" lexicon and in the "root" lexicon.

- The "root" dictionary recognizes 8 classes, the largest being "N" (5000 nouns) and "V" (2400) verbs, of which 1100 are prefix verbs ("Semitic"), and 1300 are suffix verbs ("Cushitic").

- The "affix" dictionaries recognize two kinds of affixes, (a) those which do not change classes, and (b) those which do:

(a) include "N/N" such as articles, or "V/V" such as tense/aspect/person affixes.

(b) the only "N/V" affix is the copula "to be", while the only "V/N" affixes are the participles and the nominalizer "thing" ("action nouns" and "agent nouns" are not included here.)

The structure of entries - especially verb entries - is best exemplified by inspecting some entries and their "fields". The main fields are "\gl Gloss", "\rt Root name", "\al Allomorphs", "\cl Class", and "\mp Morpheme Property" as well as "\mc Morpheme Co-occurrence restrictions".

(4) "Input Texts"

In the past, Beja text collections have been published by various scholars - both Beja and non-Beja. They have been adapted to the present parser by means of minor orthographic changes (this is especially true for texts written before the days of "phonology").

Adaptations or "harmonizations" were made to accommodate the published texts to the "accepted" Eritrean orthography. This uses only the following letters and digraphs: the Vowels *a, ee, i, oo, u* (lengthened: *aa, ii, uu* - early publications erroneously also have short *e* and *o*), and the Consonants *' b, d, dh, f, g, gh, h, j, k, kh, l, m, n, r, s, sh, t, th, w, y* (lengthened: *bb dd dhdh ff* etc.) To distinguish digraphs like *dh* from sequences like *d+h*, a "silent *e*" is now being employed by all who publish Beja texts in the Latin script. Thus, "*dh*" is the "retroflex *d*", but "*deh*" has the two phonemes "*d*" and "*h*"; i.e. *dhibaa* "fall" has a retroflex *d*, but *dehay* "people" has "*d*" and "*h*".

Texts

Texts from the following publications serve as input to the parser. They are of very different genres, as shown by the remarks at the end of every reference :

- [Hudson1976.txt] Hudson, Richard A., 1976, Beja, pp. 97-132 in M.L.Bender, ed., *The Non-Semitic Languages of Ethiopia*, Carbondale [Sample Sentences]
- [Mahmud2004.txt] Mohammed Mahmud (ed.), 2004, "Baakwidhayt Alaama", Asmara: IRC [Songs and Sagas]
- [Ministry2005.txt] Ministry of Education, 2005, "Bidhaawyeeti bhali 2" [Beja Textbook for Grade 2], Asmara: MoE [Educational Texts]
- [Morin1995.txt] Morin, Didier, 1995, "Des paroles douces comme la soie", Paris: Peeters [Fables]
- [Ohaj1971.txt] Ohaj Muhammad, 1971, "Min turat al-Bega al-sa'bi", Khartum [Traditional Texts]
- [Reinisch1895.txt] Reinisch, Leo, 1893. "Die Bedauye Sprache in Nordost Afrika", Wien: Hoelder [Anecdotes]
- [Roper1928.txt] Roper, E. M., 1928, "Tu Bedawie: An Elementary Handbook for the Use of Sudan Government Officials", Hertford: Stephen Austin [Fables]
- [WedekindAbuzeinab2008Koepppe.txt] Wedekind, Klaus and Charlotte, and Abuzeinab Musa, 2008, "A Learner's Grammar Beja", Cologne: Koepppe [Text Collection]
- [WedekindAbuzeinab2008Web.txt] Wedekind, Klaus and Charlotte, and Abuzeinab Musa, 2008, "Beja Pedagogical Grammar", Cologne University, Institut fuer Afrikanistik, www.afrikanistik-online.de/archiv/2008/1283 [Conversations]
- [WedekindAli2003.txt] Wedekind, Klaus, and Mohammed Ali, 2003, "Report on the 2nd Survey of Social Services and Needs in NDA-held Areas of North Eastern Sudan", Asmara: International Rescue Committee (RIC) [Social Questionnaire]
- [WedekindMahmud2008.txt] Wedekind, Klaus, and Mahmud Mohammed, 2008, "A Beja saga in four dialects: lexical and other differences", pp. 366-378 in: Gabor Takacs (ed.), *Semito-Hamitic Festschrift for A. B. Dolgopolsky and H. Jungraithmayr*, Berlin: Reimer [Bishaari Song]
- "Harmonization" of orthographies is a term used by one or two Beja linguists. "Harmonization meets the desire of several influential Beja leaders both in Eritrea and in Sudan: They fear that a diversity of writing systems and a neglect of writing (for publications, esp. for schools) may hasten the decline of their language and culture. Not only is there a desire to "harmonize" the writing system, but also the selection of lexical items - with a focus on the ones most widely used. This is a concern that has been expressed repeatedly, in Eritrea, in Kassala, and recently in Port Sudan. In 1999 a first international meeting addressing these concerns was held in Cairo 1999 - more meetings are being planned.

Input

Note that for each text in this collection the following "fields" are being used:

- \id "Identity" of the text
- \txt "Text" to be parsed (with or without orthographic "harmonization")
- \trs "Free translation" as provided by the author of the publication.

The parser then generates the following additional "fields":

- \wrd "words" (with spacing of "words" adjusted to the interlinear output)
- \dcm "decomposition" (splitting words into morphemes)
- \ana "analysis" (providing the "gloss" of the morphemes)

To appreciate the various kinds of texts and their various properties, some of them shall now be parsed "in real time".

For references, see the list above.