Ronald Ruzicka

10 KALAM: A Word Analyzer for Sabaic

Abstract: The word analyzer KALAM for Sabaic is a tool for the automatic detection of morphological attributes of a Sabaic word, like stem, conjugation, case and person. It has been developed as part of a Masters thesis in Arabic Studies. Connected to a computer-based dictionary, it also provides the translation, including prefixes and postfixes, like possessive pronouns and particles. New research work has connected the new system “KALAM reloaded” to online dictionaries like the Sabaic Online Dictionary, and is now extended to Minaic, Qatabanic and Ḥaḍramitic, too. The final aim is the automatic translation of sentences of Ancient South Arabian languages. The development of the project will be aided by using the newly digitized texts of the Glaser collection at the Austrian Academy of Sciences and building up annotated trees in a database in an iterative process, improving the algorithms.

Keywords: Sabaic, Ancient South Arabian, word analyzer, translation, squeezes

10.1 An Automatic Word Analyzer for Languages Epigraphically Attested

The word analyzer KALAM for Sabaic words has been developed within the framework of a Masters thesis (Ruzicka, 2016) in Arabic Studies at the University of Vienna's Institute for Oriental Studies with the main emphasis on South Arabia.

The objective was to find the roots of words or word-phrases, i.e. everything between two word dividers in Ancient South Arabian languages. Figure 10.1 shows a sample Sabaic text from Ethiopia.

Figure 10.1: Altar inscription, Wukro, Ethiopia (photo by R. Ruzicka)

Ronald Ruzicka, Simutech and Austrian Academy of Sciences, Vienna

© 2018 Ronald Ruzicka

This work is licensed under the Creative Commons Attribution 4.0 International license (CC BY 4.0)

The word analyzer supported the decipherment of all the theoretically possible meanings and readings of a word-phrase or a word-phrase with *lacunae*. The squeezes of the Glaser collection (Höfner, 1944) at the Austrian Academy of Sciences (Figure 10.2), which are being digitized by a project in which the author is also participating, contain a significant volume of unreadable text, meaning that sometimes only small portions of words are available. Within the digitizing project¹ words or word-phrases are tagged for computational usage. This means that names and ordinary words are separate and that, in due course, all words will be classified according to their grammatical features.

Figure 10.2: Squeeze sample collected by Eduard Glaser

¹ Glaser project's web-application [<http://glaser.acdh.oeaw.ac.at/>].

This tagging is intended to be carried out automatically by a computer. This is an important desideratum given that, in the DASI project,² the grammatical features of a word can at present be assigned only manually (Avanzini et al., 2015).

The idea was to develop a computer service, which takes a word-phrase in an Ancient South Arabian language as input and delivers the grammatical analysis as output: all possibilities for root, stem, person, number, gender, time, and origin, i.e. Sabaic, Minaic, Qatabanic, Ḥaḍramitic (see Avanzini, 2009 for the background of the classification of the languages). It should also separate prefixes, suffixes, conjunctions, and so on.

10.2 Requirements of the Word Analyzer for Sabaic

In this section, the operations that the word analyzer should be able to perform will be explained with examples taken from the Ancient South Arabian languages. Let us consider the word *stṣr*³. The root of this verb is:

nṣr
ST-stem, 3rd masculine singular, 3rd feminine plural
Middle and Late Sabaic
 to call somebody for support

Given that Ancient South Arabian displays weak radicals in some words, however, the root could conceivably be construed as (Multhoff, 2012):

**ṣwr, which is a hypothetical root*
ST-stem
**ṣyr*
ST-stem

The pattern of words where the roots cannot be found easily, or where the result is not unique, in a form like ABCD, could be:

stem ABC + suffix D ?
prefix A + stem BCD
prefix A + stem BCC + suffix D

The word analyzer should provide all possible variations of grammatical features that fit ABCD in a certain case.

² DASI project [<http://dasi.cnr.it/>].

³ In the whole project, we only deal with transcribed characters. There is an online keyboard available to enter diacritic characters.

The base of the KALAM service can be found in the most recent book on Sabaic grammar, the *Lehrbuch der sabäischen Sprache* by Peter Stein (Stein, 2013), which contains the whole paradigm for Sabaic – as far as it is known today.

The algorithm for the computer service is obtained through a synthetical approach. The program reproduces the scholar's reasoning. It first searches for the root, then synthesizes all possible epigraphically attested forms based on this root. Then it compares these forms to the input word-phrase.

Translation programs for Semitic languages were already available, such as the *Stanford Log-linear Part-Of-Speech Tagger*⁴ and the *Arabic Language Analyzer with Lemma Extraction and Rich Tagset* (Aliwy, 2012), this last being a very complex but powerful translator for Arabic.

Nevertheless, all of the translation methods display disadvantages as word analyzers. Indeed, focusing on living languages, all of them are based on very large corpora of digital text available in the web, whereas the ancient languages of the Arabian Peninsula are only fragmentarily attested by inscriptions. They have further problems in cases where no vowels are available, or where rules changed over the course of time.⁵ Furthermore, some of them cannot deal with the assimilation of consonants or with weak radicals.

10.3 Functioning of the Word Analyzer

The aforementioned considerations prompted the creation of KALAM from scratch as a word analyzer for the Sabaic language.

KALAM uses three main categories with which to analyze the grammar:

- the so-called *Situation*, which consists of the *TimePeriod* and the *Locality*, for instance Middle and Late Sabaic;
- the *Rule*, together with the *rule-situation*, for instance verbs with suffix conjugation in a certain time frame, leading to
- a *Term*, which describes how, for example, the first common singular form is created.

Thus, looking at the 3 radicals (or 4 in some cases) we have possible prefixes, infixes and suffixes. Metathesis, gemination and weak radicals could occur as well. Moreover, all special cases of words, which do not fit in the ordinary schemes, must be taken into account.

Here are samples of the configuration list, for instance a verb in prefix conjugation, first person common singular, which carries the prefix ' (aleph) (symbolized by an “a”

⁴ [<https://nlp.stanford.edu/software/tagger.shtml>].

⁵ Sabaic and its brother languages Minaic, Qatabanic, Ḥaḍramitic have been written from the 9th century BCE up to the 6th century CE.

here), and some special rules for weak radicals. Some weaknesses are unavoidable (writing 0), some optional (writing 0 and 1).

v_PKO_1cs

*a null null null 0 0 n**=011 w**=011 *w*=101,111 *y*=101,111 *..=110 ydc=011*

The following one is another sample for a noun in status constructus and dual form:

n_SC_mdn

*null null null y 0 0 n**=011,111 w**=011,111 *w*=101,111 *y*=101,111*

In this way, about 670 rules have been created which build up the whole grammar of Sabaic, including special cases such as:

- cardinal and ordinal numbers,
- parts,
- personal/demonstrative pronouns,
- suffix/object pronouns,
- particles,
- conjunctions,
- *nomen loci*.

10.3.1 Using KALAM

Using KALAM is quite simple (Figure 10.3). Entering a word phrase, for instance “ybny”, and selecting the mode and the language (only “Sabaic” in this case), a possible root, the stem, a short and a long description of the grammar term are provided.

Kalam - 𐩧𐩣𐩪

Word Analysis for Ancient South Arabian Languages

☒ enter ASA ☐ English word
☒ use dictionary ☐ entire dictionary ☐ use SabaWeb ☐ 4 radicals, too
☒ sabaic ☐ minaic ☐ qatabanic ☐ hadramitic

ybny

Root	Stem	Term	Location	Time	Prefix	Suffix	Translation KALAM	Information
bny	O1	sabv_PKO_3ms	Saba'	middle+late-sabaic,old-sabaic			build, construct,building, edifice,(act of) building, construction	verb; prefix conjugation; 3; masculine singular
bny	O1	sabv_PKO_3md	Saba'	old-sabaic			build, construct,building, edifice,(act of) building, construction	verb; prefix conjugation; 3; masculine dual
bny	O2	sabv_O2PKO_3ms	Saba'	middle+late-sabaic,old-sabaic			build, construct,building, edifice,(act of) building, construction	verb; prefix conjugation; 3; masculine singular
bny	O2	sabv_O2PKO_3md	Saba'	old-sabaic			build, construct,building, edifice,(act of) building, construction	verb; prefix conjugation; 3; masculine dual
byn	O1	sabv_PKO_3md	Saba'	middle+late-sabaic			between; among; in both of (two things);intervene, separate (boundary);remove (punishment);surrounding are (of town); surrounding loculi (of tomb)	verb; prefix conjugation; 3; masculine dual

Figure 10.3: Analyzing *ybnny* using KALAM

If we enter the word “*stšr*” we will get any possible root – even the hypothetical ones listed above (see 10.2).

Forms beginning with the letter *s* are also displayed (Figure 10.4):

stšr

Root	Stem	Term	Location	Time	Prefix	Suffix	Translation KALAM	Information
nšr	O1	sabv_PK0_2ms	Saba'	middle+late-sabaic	sabpart_s		at,near to	verb; prefix conjugation; 2. masculin singular
nšr	O1	sabv_PK0_3fs	Saba'	middle+late-sabaic	sabpart_s		at,near to	verb; prefix conjugation; 3. feminin singular
nšr	O1	sabv_PK0_3fp	Saba'	middle+late-sabaic	sabpart_s		at,near to	verb; prefix conjugation; 3. feminin plural
nšr	ST	sabv_STSK_3ms	Saba'	middle+late-sabaic				verb; suffix conjugation; 3. masculin singular
nšr	ST	sabv_STSK_3fp	Saba'	middle+late-sabaic				verb; suffix conjugation; 3. feminin plural
nšr	T1	sabv_T1IMPk_2ms	Saba'	sabaic	sabpart_s		at,near to	verb; imperative; 2. masculin singular

Figure 10.4: Analyzing *stšr* using KALAM

If the hypothetical roots are not desired, an inbuilt word list can be used – a small dictionary that not only reduces the roots to already existing ones, but also provides a translation. If “use dictionary” is checked, only really existing words are displayed (Figure 10.5).

As one of the goals of the project is to support students as well as established scholars, the system allows a root, like *QTL*, to be entered, and the command “synthesize” will then display all grammatical forms the paradigm knows.

mhfd

Search Synthesize

enter ☒ ASA ☐ English word

☒ use dictionary ☐ entire dictionary ☐ use SabaWeb ☐ 4 radicals, too

☒ sabaic ☒ minaic ☐ qatabanic ☐ hadramitic

mhfd

Root	Stem	Term	Location	Time	Prefix	Suffix	Translation KALAM	Information
hfd	O1	minn_LOO1SI_cs	Macin	minaic			tower; projecting element (of wall)	nomen loci; status indeterminatus; communis singular
hfd	O2	minn_PZO2SI_cs	Macin	minaic			tower; projecting element (of wall)	passive/active participle; communis singular
hfd	O2	minn_LOO2SI_cs	Macin	minaic			tower; projecting element (of wall)	nomen loci; status indeterminatus; communis singular
hfd	O1	sabn_LOO1SC_ms	Saba'	sabaic			tower; projecting element (of wall)	nomen loci; status constructus; masculin singular
hfd	O2	sabn_PZO2SC_ms	Saba'	sabaic			tower; projecting element (of wall)	passive/active participle; masculin singular
hfd	O2	sabn_LOO2SC_ms	Saba'	sabaic			tower; projecting element (of wall)	nomen loci; status constructus; masculin singular

Figure 10.5: Using the dictionary. If one enters “*mhfd*” selecting “Sabaic”, “Minaic” and “use dictionary” one obtains “tower”. Similarly, entering “tower” and selecting “English word” produces “*mhfd*”

10.4 Future Perspectives

KALAM can be used freely online⁶ and presently supports all four Ancient South Arabian languages (Minaic, Qatabanic and Ḥaḍramitic in addition to Sabaic). The user can search for word-phrases with single missing characters, entering a question mark instead of a character. Connections to other online tools for the Ancient South Arabian languages, like the Sabaic Online Dictionary (University of Jena) or DASI (University of Pisa) are provided.

For DASI, some pre-work has been done. For the Sabaic Online Dictionary (hereafter, SabaWeb), which provides all known occurrences and state-of-the-art translations of the words in Sabaic, we have a full integration reached (Figures 10.6, 10.7).

The screenshot shows the KALAM search interface. At the top, there is a search bar with the text 'stṣr' and buttons for 'Search' and 'Synthesize'. Below the search bar, there are checkboxes for 'enter' (ASA, English word), 'use dictionary', 'entire dictionary', 'use SabaWeb', '4 radicals, too', 'sabaic', 'minaic', 'qatabanic', and 'hadramitic'. The 'use dictionary' and 'use SabaWeb' options are checked. Below the search bar, there is a table with the following columns: Root, Stem, Term, Location, Time, Prefix, Suffix, Translation KALAM, Translation SabaWeb, and Information. The table contains one row with the following data:

Root	Stem	Term	Location	Time	Prefix	Suffix	Translation KALAM	Translation SabaWeb	Information
nṣr	O1	sabv_PK0_2ms	Saba'	middle+late-sabaic	sabpart_s		at,near to summon to one's support,aid, support, assistance, help,aid; provide support,gain the support of (s.o.)	um Hilfe nachsuchen , um Hilfe bitten	verb, prefix conjugation, 2, masculin singular

Figure 10.6: Connecting KALAM and SabaWeb. We take again “stṣr”, “use dictionary”, “Sabaic only” and select “SabaWeb”. Clicking on the SabaWeb link, all additional information is shown

The screenshot shows the SabaWeb search results page. At the top, there is a header with the text 'SABÄISCHES WÖRTERBUCH' and the logo of the Friedrich-Schiller-Universität Jena. Below the header, there is a navigation bar with links for 'Projekt', 'Suche', 'Corpus', 'Statistik', 'Bibliographie', 'Kontakt', 'Andere Projekte', and 'Hilfe'. The 'Suche' link is selected. Below the navigation bar, there is a search bar with the text 'stṣr' and a button for 'Suche'. Below the search bar, there is a table with the following columns: Zitiertform, Suchergebnis für nṣr ST, Übersetzung, and Ältere Übersetzung. The table contains one row with the following data:

Zitiertform	Suchergebnis für nṣr ST	Übersetzung	Ältere Übersetzung
um Hilfe nachsuchen, um Hilfe bitten	3 Belege in 3 Texten	um Hilfe nachsuchen, um Hilfe bitten	[Bitte um] Unterstützung richten Nebes 2005, 350 appeler au secours Calvet/Robin 1997, 144

Figure 10.7: Result page from SabaWeb

⁶ KALAM [<http://kalam.ruzicka.net>].

SabaWeb is linked to KALAM for two purposes:

- it detects already existing roots;
- it delivers all the translation features described above.

A new project at the Austrian Academy of Sciences starting in 2018 will try new methods of scanning the squeezes of the Glaser collection (Figure 10.2) and will provide not only digitized, but fully tagged texts. KALAM will be “reloaded”, too, and get new features. The aim is to advance from word-phrases to parts of sentences by including the automatic building of syntax trees. The currently manually tagged texts of the Glaser collection will then provide a small corpus with which the algorithms for automatic tagging in KALAM can be improved and parametrized.

The research question of this project is: is it possible to fill in missing parts of text automatically? Using word analysis, automatic tagging and “sentence formulars”, as they appear for instance in dedication and building inscriptions, it should be possible. The comparison of syntax trees will be used within the corpus. At least the question is: to which degree can we support semi-automatic filling of *lacunae* in the text?

Bibliography

- Aliwy, A. (2012). Arabic Language Analyzer with Lemma Extraction and Rich Tagset. In H. Isahara & K. Kanzaki (Eds.), *Advances in Natural Language Processing: Proceedings of the 8th International Conference on NLP, JapTAL 2012, Kanazawa, Japan, October 22-24, 2012* (pp. 168–179). Springer-Verlag. doi: 10.1007/978-3-642-33983-7_17
- Avanzini, A. (2009). Origin and classification of the Ancient South Arabian languages. *Journal of Semitic Studies*, 54, 205–220.
- Avanzini, A., De Santis, A., Gallo, M., Marotta, D., & Rossi I. (2015). Computational Lexicography and Digital Epigraphy: Building digital lexica of fragmentary attested languages in the Project DASI. In G. Guidi, R. Scopigno, J.C. Torres, & H. Graf (Eds.), *2015 Digital Heritage International Congress* (pp. 405–408). New York: IEEE. doi: 10.1109/DigitalHeritage.2015.7419535
- Höfner, M. (1944). *Die Sammlung Eduard Glaser* (Akademie der Wissenschaften in Wien, Philosophisch-historische Klasse, Sitzungsberichte 222.5). Brunn-München-Wien: Rudolf M. Rohrer.
- Multhoff, A. (2012). *Die Verbalstammbildung im Sabäischen* (PhD thesis). Jena: Friedrich-Schiller-Universität Jena.
- Ruzicka, R. (2016). *KALAM – Wortanalyse für Sabäisch* (MA thesis). Wien: Universität Wien.
- Stein, P. (2013). *Lehrbuch der sabäischen Sprache. Teil 1: Grammatik*. Wiesbaden: Harrasowitz Verlag.