Petr Zemánek
Charles University, Prague
Institute of Comparative Linguistics

Abstract

The paper presents an outline of a treebank of Ugaritic, an extinct Semitic language. It describes the basic structure of the treebank, and possibility of re-using approaches applied to other Semitic languages. It also discusses problems of analyzing a language attested in a fragmentary form and possible usage of a treebank based approaches for further reconstruction of text passages.

1 Introduction

Extinct languages offer new themes that can help in development of new ways of analysis also for the languages spoken today. The different concepts of some categories or syntactic notions can help in formulating more robust theories of languages. Building electronic corpora of these languages will also enable the access to these languages by a broader group of researchers.

The current progress in the computational treatment of the Semitic languages has mostly covered the living languages (esp. Arabic and Hebrew), while the ancient ones have received far less attention, focused on the formal description of a text on a medium, e.g. a cuneiform tablet (cf. esp. the Cuneiform Digital Library Initiative, `http://cdli.ucla.edu` or, e.g. Koslova and Damerow 2003). As more of these texts become available in electronic form, the linguistic exploitation of these data is to be discussed.

This contribution concentrates on the basic description of a treebank of Ugaritic, an extinct Semitic language attested from cca 1500-1200 B.C.E. in northwestern Syria. The relatively limited extent of the attested Ugaritic texts makes it ideal for testing the procedures of dealing with extinct languages while constructing a treebank.

A decision to create a treebank has one more reason – as a treebank is a complex tool demanding complex analysis, it is suitable for Ugaritic, where the reconstruction needs a combination of analyses at several layers, and an interplay of these analyses can offer a better insight to the knowledge of Ugaritic.

* This research has been supported by the project MSM 0021620823 "Český národní korpus a korpusy dalších jazyků" [The Czech National Corpus and Corpora of Other Languages] of the Ministry of Education of the Czech Republic.

2 A Treebank of Ugaritic – a Description

The project of the Ugaritic treebank started this year, and currently we have set up the basic rules for its creation. At present, a pilot part of the project, the complex treebank annotation of the legend of Aqhat, has been finished.

In case of Ugaritic, it has been noted earlier that the texts are preserved often in a very fragmentary form. Together with the lack of vowels in the texts this has contributed to the fact that some textual passages are still discussed by Ugaritologists.

The attested texts form only a part of the reconstruction of the language. In recent editions of Ugaritic texts, such as Pardee 2000, the reconstruction gets close to a standard (natural) language, the texts are vocalized and capable of formal treatment, although a certain level of variation has to be allowed.

The treebank has to meet all the requirements defined by the CDLI, but for a linguistic analysis, it has to add linguistic information, where all the levels of interpretation must be discrete.

2.1 Morphotactical and Morphological Analysis

The concept of a word as a syntactic unit is not very well rendered in the Ugaritic texts. A word-divider is used there, although not consistently – the situation is similar to that of Arabic or Hebrew: prepositions, some pronouns and even genitive construct may be connected and tokenization is necessary. A solution based on the one used for the Treebank for Arabic (PADT, Smrž and Hajič 2007) can be applied.[1]

An important part is also the estimate of the length of the lacunae in the text. Apart from words, they can also contain various borders, such as those between words, syntactic units and sentences and some of this information can be restored (either fully, partially or no reconstruction is possible).

Morphological annotation on one hand does not represent a major problem. In grammar, Ugaritic is a typical West-Semitic language, and thus a tagset can be derived relatively easily from similar ones prepared for Arabic or Hebrew.

However, it is important that the tagset meets several requirements. The most important among them are the following: functional perspective, expandability (and reductability) of the tagset, variability and total discreteness of individual categories included in the tags.

In order to keep the functional perspective, we have chosen to avoid the approaches based on morphematic analysis, such as Buckwalter's (e.g., Buckwalter 2004), although such an approach has been used for Ugaritic, e.g. by García-Serrano and Contreras (1998). Instead, we have derived the system

[1] With some minor adaptations, e.g., phantom words that result from the division of strings across the line boundaries have to be formally marked.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

for Ugaritic based on Khoja et al. 2001 for Arabic, which works with basic categories (*Noun, Pronoun, Quantifier, Verb, Conjunction, Preposition, Particle*), that can be understood as "major labels", each of them is expandable, and to each of them a finer analysis, covering grammatical categories, can be added. We have also included the derivational information (such as *Participle, Suffixed conjugation, Verbal noun*, etc.), as it can play role in the syntax of a Semitic language. Several other tags have been added.[2]

For further reconstruction, the expandability is one of crucial properties of such a tagset, and the use of the tagset should ensure that an approximative analysis, based on the use of major properties (or their combination) is possible. Such a situation is typical for the process of reconstruction of Ugaritic – a number of features can appear due to reconstruction of further layers, and need to be included in the former analysis. On the other hand, reductability is important for situations when some of the tags in the finer analysis are not considered safe or have, according to Ugaritologists, several solutions.

Another important point is the annotation of the lacunae of the text. In some cases, the missing text can be fully reconstructed (e.g. based on other passages in the texts), in some cases not, sometimes only the reconstruction of the underlying grammatical form is possible, e.g. due to the so-called *paralelismus membrorum*, a construction that is very frequent in the poetic speech of the ancient Semites (cf. Čech 2005 and an example below). This means that in some instances, the annotation will be only partial – e.g., there will be instances when only a minor category, such as "genitive" or "plural" can be projected into the otherwise empty morphological tag.

As the major part of the annotation is being done manually, currently the tagset is in a human readable form (see Čech 2005), in order to decrease mistakes. For the final form, we plan to use a positional tagset, which allows to work discretely with the grammatical information – e.g. it allows to work with such combinations as Noun and Plural, etc.

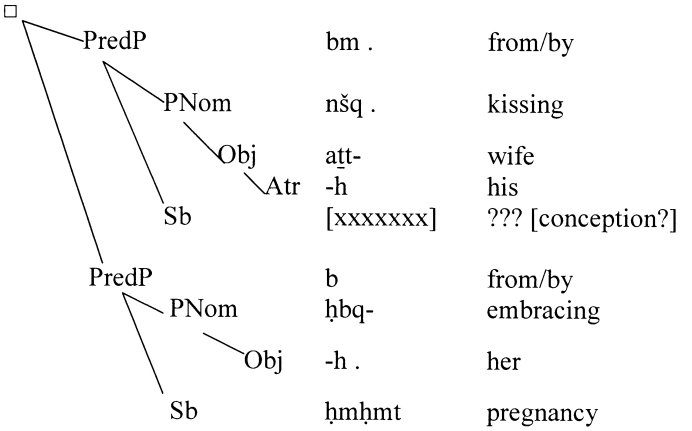
2.2 Syntactic Annotation

The syntactic annotation covers the annotation of the sentence boundaries and the analysis of the parts of the sentence (parsing). The main factor that heavily influences the syntactic analysis is the fragmentary character of Ugaritic texts. Many syntactical structures in the texts will be incomplete, without a clear beginning or end. A robust theory must be chosen. An important point is the

[2] Such as theophoric, topographical and personal names, information on the underlying root of individual words (to reduce reduce homography: e.g., a string *št* can be analysed as belonging to roots *št, štt, šty* with different meanings), reconstructed vocalization, etc. A detailed description of the preliminary stage of the analysis of Ugaritic morphology representation can be found in Čech 2005.

ability of the theory to expand or collapse various branches of the syntactic tree. This can be very useful when working with text fragments, where some parts of the text are missing, and yet it is possible to reconstruct the core of such a missing part and assign it a syntactic role, although at the same time one has to admit that this core and its function can be further expanded (see the example below). There is a need for a theory with a hierarchical approach (a clear representation of the top node of the sentence and the relations of other nodes in that sentence) and a functional perspective, to ensure harmony with a similar approach applied at morphological tagging.

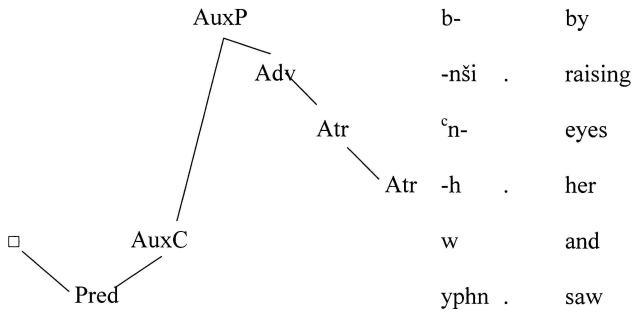
In our view, a theory that meets these requirements, is the dependency approach which has been used for the construction of treebanks of Czech and Arabic (e.g., Hajič et al. 2001, Smrž and Hajič 2007). The adaptation of the concept for Arabic at the analytical level (surface syntax) can be taken over only with some minor adaptations, as the basic syntactic concepts are shared by both languages. The application of the theory, minor changes and its use for reconstruction is shown in the following examples.



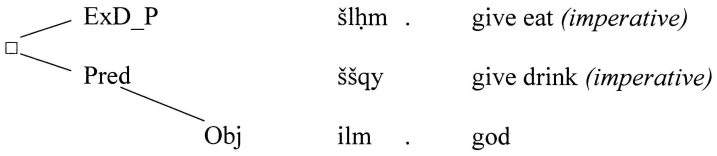
Example 1: A passage from KTU 1.17 I 39-40, showing a structure of a *parale-lismus membrorum* with a gap in the text (damaged part, marked as [xxxxxxx], can contain 7 Ugaritic signs). The structure is simplified (word-dividers are ignored in the analysis). The text in the damaged part consists most probably of more than one word (Ugaritic words are usually shorter), possibly a genitive construct. The mean-ing of the noun in the damaged part must have a semantic relation to that in the non-damaged part – it has to be usually synonymous or antonymous, thus in our exam-ple, a certain semantic evolution of the concept (from "conception" to "pregnancy") is allowed. We cannot be sure of the particular meaning of the inserted string, however, the major morphological category of the node / subbranch is Noun, and the major syntactic category is Subject. It is quite probable that there are more units in this subbranch, as the Ugaritic word for "conception" consists of only two signs

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

(hr), thus a space for another five should be explained at some point. The possible subbranch could be further expanded in case a finer reconstruction of the passage is reached.



Example 2: A passage from KTU 1.17 V 9-10. The example shows a function of a coordination particle as subordinate, connecting between an adverbial part ("by raising her eyes") and the main (verbal) part of the phrase.



Example 3: A passage from KTU 1.17 V 19-20. The example shows a sequence of two verbal phrases without formal coordination, but with only one object. The predicative function of the first verb is indicated as a suffix attached to the analytical (syntactical) function.

The syntactic analysis is important also for further reconstruction of Ugaritic. The comparison of syntactic structures, even their projection onto "empty space", is helpful in many ways. It can bring new insights in the emendations of missing passages, it can also bring a better understanding of existing sequences where an agreement on the meaning has not been reached, and help its gradual reconstruction.

Summary

The construction of a treebank for Ugaritic is a complex task involving a combination of several layers of information, from the description of extra-linguistic characteristics of the object and the text written on it, to linguistic annotation. In

most of the examples the methods and concepts developed for the Central Semitic languages (such as Hebrew or Arabic) can be applied. We have chosen the dependency approach, especially on the level of surface syntax.

In case of Ugaritic, a high degree of uncertainty is met, which has to be taken into account while analyzing it and constructing the treebank. For further reconstruction, several types of analysis can be used – such as a reductionistic one, a discrete analysis of various categories, collapsing or expanding a branch of a syntactic tree can be helpful and allows work even with incomplete data. A dependency approach that works also with syntactic roles is very useful for the analysis of Ugaritic.

We believe that our work will provide a better access to the Ugaritic language from a wider circle of the scholarly community and that methods developed and applied during the construction of the corpus of Ugaritic will be usable also for corpora of other extinct languages.

References

Tim Buckwalter. 2004. *Buckwalter Arabic Morphological Analyzer Version 2.0*. LDC Catalog No. LDC2004L02, Linguistic Data Consortium, http://www.ldc.upenn.edu/Catalog.

Pavel Čech. 2005. Ugaritic Narrative: Annotating Very Fragmentary Corpora. Poster at *Framing Plots: the Grammar of Ancient Near Eastern Narratives*, London. 2005.

The Cuneiform Digital Library Initiative. http://cdli.ucla.edu.

Manfried Dietrich, Oswald Loretz and Joaquín Sanmartín. 1995. *Cuneiform Alphabetic Texts from Ugarit, Ras Ibn Hani and Other Places.* Münster: Ugarit Verlag 1995.

Ana García-Serrano and Jesús Contreras. 1998. A Computational Platform for Ugaritic Morphological Analysis. *First International Conference on Language Resources and Evaluation (LREC)*, Granada 1998.

Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidová-Hladká. 2001. *Prague Dependency Treebank 1.0.* LDC catalog number LDC2001T10, ISBN 1-58563-212-0.

Shereen Khoja, Roger Garside and Gerry Knowles. 2001. A Tagset for the Morpho-syntactic Tagging of Arabic. *Proceedings of the Corpus Linguistics*. *Lancaster University (UK), Volume 13 - Special issue, 341*. http://zeus.cs.pacificu.edu/shereen/CL2001.pdf

Natalia Koslova and Peter Damerow. 2003. From Cuneiform Archives to Digital Libraries: The Hermitage Museum Joins the Cuneiform Digital Library Initiative. In: *Proceedings of the 5th Russian Conference on Digital Libraries RCDL 2003*, St. Petersburg, Russia 2003.

Dennis Pardee. 2000. *Les Textes Rituels.* Fasc. I, II., Ras Shamra – Ougarit XII. Éditions Recherche sur les Civilisations, Paris.

Otakar Smrž and Jan Hajič. 2007. The Other Arabic Treebank: Prague Dependencies and Functions. In: *Arabic Computational Linguistics: Current Implementations.* Edited by Ali Farghaly, to appear March 20, 2007. http://ufal.mff.cuni.cz/~smrz/CSLI2006/csli-prague.pdf

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace/handle/10062/4476