Morphological Tagging of Ugaritic

Petr Zemánek

Charles University, Praha

1. Introduction

Ugaritic is, as many other ancient languages, also affected by the swift progress of the corpus linguistics today. Data for Ugaritic are now available in electronic form – the texts are interchanged among Ugaritologists, several institutions have created a corpus by themselves, usually based on the edition of KTU[2]. There has been effort to build databases on Ugaritic texts.

This also leads to the need of a linguistic exploitation of these texts. The available data, from the point of view of corpus linguistics, can be described as raw text, on which subsequent annotation can be construed. However, as Ugaritic is attested only in fragmentary form, several additional requirements have to be satisfied in order to meet the needs of a proper linguistic analysis.

In this paper, the problem of a morphological tagging will be treated, in the frame of the works on the Ugaritic Treebank, which is currently being prepared.

The morphological analysis has to face several problems, such as tokenization of the strings available on a tablet, as well as the fragmentary character of Ugaritic texts. Apart from these types of problems, which can be viewed as technical (or not purely linguistic in nature), there are also conceptual problems connected with the type of analysis that is applied to the language. Such a choice has its consequences for the shape of the tagset. The basic problem of choosing between a morpheme-oriented and function-based approach will be treated in the article, and finally, the solution chosen for the treebank will be discussed.

2. General frame: Ugaritic Treebank

We believe that Ugaritic can serve as a good example for creating a treebank of an ancient language. These reasons are both conceptual and quantitative, that come back to the size of the texts attested in Ugaritic.

The extent of the texts in Ugaritic is rather limited and is not comparable to the situation e.g. in Akkadian with hundreds of thousands of tablets. The number of texts discovered so far is about 1,400, and although new texts are still being found, it is not to be expected that the number of these texts will increase considerably in the future. The edition of Ugaritic texts (KTU[2]) contains approximately 50,000 of strings (out of which, about 12,300 are unique strings with damaged signs, and about 8,800 are without damaged signs), the literary texts containing about half of it (approx. 7,300 unique strings, 5,200 strings without damaged signs).

Such a limited amount raises questions about the utility of an electronic corpus for Ugaritic with complex linguistic annotation; however, when such a corpus is available, it can be expected that it will help in the development of Ugaritic studies. The use of corpus linguistics methods should, in our view, help also in the reconstruction of some of the Ugaritic passages and/or could help in a better understanding of some of those.

Ugaritic brings all the problems that can be met at a construction of a corpus of an extinct language, and even adds some more: its attestation is fragmentary and the usual ambiguity of a language is even increased by the script whose characteristics are very similar to those of Hebrew or Arabic, yet the situation is worsened by the fact that the information on vocalization comes only from secondary sources.[1] Ugaritic offers texts with manifold styles, ranging from narrative to poetry, treaties, personal letters, economic texts, etc.; there are fairly disparate views at reconstruction of some passages of the texts and some parts of its grammar. A wide variety of problems can be expected to appear, and it is probable that similar problems will have to be solved when constructing treebanks for other ancient languages.

As the reconstruction of some passages of the Ugaritic texts is of complex nature, a complex annotation scheme should be chosen in order to reflect the linguistic pecularities of the language. We believe that a treebank is the right choice for such a representation.

The construction of a treebank is also important for the study of Ugaritic itself. The treebank can be viewed as a complex linguistic analysis of the whole corpus of the attested texts on morphological and syntactic levels. In such a corpus, an approximation to the accepted and standard text must be made; however, variant readings ought to be preserved as well. The POS tags and annotation of syntactic structures will allow to approach the reconstruction of the language in a structuralist way and study and analyze both morphological and syntactic structures, which can and most probably will bring new insights and impulses for further reconstruction and better understanding of the language. It should e.g. enable the comparison of sentence structures, which is very helpful in the reconstruction of some damaged passages. However, it can be also expected that when such a tool is offered, the language will be made accessible to a broader group of scholars from neighbouring fields. More information on the treebank, esp. on the annotation on the syntactic level, can be found in Zemánek 2007.

3. Tokenization process

Tokenization can be viewed as a division of strings in the text into meaningful units. It has to be applied during the analysis of many languages, even such as

1 With the exception of the three *ʾaleph* signs, which signal the use of *a*, *i* and *u* in the neighbourhood of a glottal stop.

English, however, it usually touches only a small part of the system. The situation is dependent on the logic of the graphemic system and it is very common in ancient languages to find graphemic systems not marking the word and sentence borders, or doing it in a way different from the modern systems.

The Ugaritic scripture is in principle derived from the systems used by the Central Semitic languages (esp. Phoenician, Hebrew or Arabic), which means that as an inspiration for the solution, there are many approaches that can be adapted to Ugaritic. For our purposes, we have in principle adopted the system developed for the Prague Arabic Dependency Treebank (see Smrž and Hajič 2007).

The Ugaritic scribes used a "word divider", a sign that marks the borders of some strings; however, the applied logic is similar to the one used for Arabic or Hebrew. From the point of view of modern linguistic approaches to the analysis of textual flow, this type of segmentation is insufficient, as many issues are left unresolved. Content word is usually in the central position in such a cluster, and is surrounded by one or several function words.

The system does not address sentence division at all – no unit higher than "word" is being distinguished (with the exception of division of paragraphs on some texts).

On the "word" level, the Ugaritic system works in this way: the items from the list are connected with the following or preceding string:

- short prepositions and particles, esp. those consisting only of one grapheme, such as *w-* – "and", *b-* – "in", but also *km* – "like, likewise; while"; etc.
- suffixed pronouns are attached to the preceding string: *ʿnh* = *ʿn-h* – "his/her eyes"; *qštk* = *qšt-k* – "your bow"; etc.
- other "understandable" strings can be also joined together, esp. some genitive constructs, such as *mlkugrt* > *mlk ugrt* – "the king of Ugarit"; etc.

In the edition of Ugaritic texts used for the Ugaritic treebank (KTU[2]), the prepositions and particles, as well as the genitive constructs, are already disjoined from the following or preceding strings and can be taken over as an analysis of the strings attested on the Ugaritic clay tablets.[2] However, for the analysis in the treebank, the suffixed pronouns must also be marked as separate strings. Having in mind the extent of the Ugaritic texts, the easiest way to accomplish this task is a manual annotation of these two features, especially in a situation when most of the task has been already carried out by other scholars while preparing their edition, and the remaining task is to mark the suffixed pronouns.

The final style of the tokenization for the treebank is shown in the following example:

Tablet:	*ltbrknn . lṯr . ilaby*
KTU[2]:	*l-tbrknn . l ṯr . il-aby*
Treebank:	*l-tbrkn-n . l-ṯr . il-ab-y*

2 In some cases, it is possible that the annotation will take over a solution from other representative editions, such as Pardee 2000.

The basic principle in the choice of the tokens' borders has been their syntactic functions, i.e. the fact that they should be treated as a separate item on the level of surface syntax.
In the analysis of higher units, such as sentences or poetic cola, manual annotation will be used as well. As many passages of the attested texts are of poetic character, both sentence and cola borders must be used. The basic principle used in our treebank is based on the predicative function which should be present in any sentence – this function, if used in the main clause, marks the top of the sentence, and by definition there cannot be two such predicates in one main clause.

4. Morphological tagging

4.1 Conceptual possibilities of tagset construction

From the morphological point of view, Ugaritic is a Central Semitic language, which means that approaches used for the languages of the same group should be equally applicable to it as well.
Arabic and Hebrew, two prominent members of this group, have received most attention from the Semitic language group. As both these languages are in many respects similar to Ugaritic, approaches developed for them can serve as an inspiration for further analyses.
The number of studies and approaches that have been applied to those languages is quite high, cf. for Arabic e.g. Buckwalter 2004, Freeman 2002 or Khoja – Garside – Knowles 2001, for Hebrew one can refer to e.g. Adler – Elhadad 2006 or Bar-Haim – Sima'an – Winter 2006; there are even studies on Ugaritic itself – García-Serrano and Contreras 1998 and Cunchillos Ilarri and Cervignon Moreno 1998.
These studies can be divided into two major groups according to their approach to the Semitic word. The basic standard, which has governed the construction of morphological analyzers, viewed the word as a sequence of morphems, and the analysis concentrates on these morphems. As a result, the word in a text falls apart into an association of morphems, and even tokenization can be viewed as an analysis following the morphematic analysis. An important characteristics of this type of approach is a concept of a morpheme as a relatively independent member of a cluster, and the presence of a grammatical feature is bound to the presence of such a morpheme. The other approach sees the word as a complex of grammatical categories, with the POS tag as the highest category. Words can bear grammatical features that are not explicitly expressed by individual morphemes, the system is not incremental.
The differences between the two approaches will be treated in the following part of the article along with the analysis of their advantages and disadvantages for the analysis of Ugaritic.

4.1.1 Morphematic analysis

The analysis of a text in a Semitic language can be treated as the analysis of strings in such texts. This may appear convenient in case of a language where certain grammatical words in their graphemic representation are joined together with content words into a single string, as is the case of Ugaritic and other Semitic languages (see above on tokenization). In such an approach, strings are analyzed according to the rules of the language and their nature is identified. Subsequently, grammatical words can be separated from content ones; however, the content words are still a sequence (or association) of morphems, one of which is lexical (stem/lemma), while the others can render some grammatical meaning. This approach is based on the division of the consonantal root as a bearer of a semantic meaning, which together with vocalization forms a lemma, and affixes, which render only synsemantic notions, such as expression of some grammatical categories, etc. Such an approach has elaborated theoretical background, starting from McCarthy's pioneering study (1985) to theoretical systems adapted to a computational treatment of natural languages (e.g., Kiraz 2001) and to applications of the morpheme-based approach (such as Buckwalter's, e.g. 2004 or Beesley's, e.g. 2001) that prove the strength of this type of analysis. On the other hand, the analysis based on morphemes can be misleading at times, as it needs a proper planning of the application, so that some parts of the language system, such as some verbal forms (e.g., prefixed and suffixed conjugations) are not disjoined. Results of such a type of analysis of some strings can be complex and difficult to resolve into a utilizable result.

For the analysis of individual morphemes that can build up a string in Ugaritic, as well as in other Semitic languages, the following scheme shows an idealized form of a structure of verbal constructs in Ugaritic:[3]

-4	-3	-2	-1	0	+1	+2	+3
neg	subj	der	der	stem	subj	der	obj
				nzl			
	y	š	t	ḫwy			
l	y			ᶜms	n		n(<h)

Table 1: Tentative structure of verbal forms in Ugaritic

This analysis is based on the distinction of a stem (i.e. the root and vocalization) as the basic unit or a core of the verbal unit, and peripheral affixes. It is a cumulative analysis, where not all the positions need to be filled in at individual verb forms –

3 *nzl* = "go down"; *yštḫwy* = "he prostrates himself"; *yᶜms* = "he carries". Borders between -4 to -3 and +2 to +3 can be seen as limits of the content word. The instance *n* (<*h*) shows that additional rules for various assimilations need to be introduced; many more complicated examples could be given.

the analysis of a prefixed conjugation will fill in more than the suffixed one. It is an idealized structure, where no linguistic processes such as assimilation are taken into account – in other words, additional rules must be applied in order to meet the correct outputs.[4] Another problem is that some of the morphemes are not rendered by the Ugaritic script (e.g., plural morphemes in short verbal forms), which means that such information would be missing in the morpheme-based tag clusters, unless the analysis is done on texts with completely reconstructed vocalization. Moreover, this model does not represent strings that can be met in real, non-tokenized texts, where more function words, such as prepositions or particles, can be attached to a content word. However, for the purposes of our discussion here, such representation is sufficient, as it clearly shows the complexity of such a task.

The basic advantage of this approach is a good rendering of the word structure and the derivational information. Another important thing in favour of this type of analysis is the fact that the method has been successfully used both for Arabic and Hebrew (see Buckwalter 2004 and 2005; Adler and Elhadad 2006); even analysis of Ugaritic is available – see García-Serrano and Contreras 1998 and Cunchillos Ilarri and Cervignon Moreno. 1998. It is expectedly better in performance in the automatic analysis of real texts, where it works with real strings that appear in the language. It can also help in the tokenization process, as during this type of analysis, the prefix and suffix types are clearly identified.

On the other hand, it does not exactly correspond with the type of the language – Ugaritic is a flective language, and the analysis based on morphemes has some drawbacks, especially when combined with the analysis of surface syntax. As such, it is not directly applicable (or linkable) to higher levels of analysis and needs to be somehow translated into other types of tags, which may turn up to be a complex task. A complex library of affixes with concatenation rules would have to be created in order to convert morpheme clusters into a shape directly linkable to syntactic nodes.

4.1.2 Morphosyntactic analysis

At an analysis based on the morphosyntactic categories, the annotated strings are closer to the units that serve as nodes in a sentence tree. The word is seen as a meaningful unit, which unites both semantic and grammatical information. The morphological features are associated with the word, and flective elements are introduced into the word as such.

As it has been pointed above, Ugaritic is a flective language and scholars working with Ugaritic will expect it to be treated in such a way – in other words, they will

4 E.g., Buckwalter sees the morpho-phonetic deviations as orthographic variations and adds them directly to the lexicon. In case of Ugaritic, where some of these processes can be formulated as simple rules, it would be advisable to take advantage of such rules for a more economical shape of the system.

expect a tag based on such characteristics, where the main part of the tag will point to a POS characteristics, such as Noun, Verb, etc. The connection of the morphological characteristics with the syntactic ones is also important for future interplay of the tags and syntactic functions assigned to the nodes in a treebank.
It should be, however, noted that this type of annotation can be applied only after the tokenization process has been completed, as it works with the strings resulting from this type of analysis. It can also be seen as a subsequent step to the analysis based on morphemes, esp. in cases when there are automatic tools developed for such an analysis.
In case of the Ugaritic treebank, the analysis based purely on morphemes can be considered obsolete, as there are no tools that could be used for automatic analysis and it is not to be expected that such tools will be available in the near future. This means that manual analysis needs to be made, which, having in mind the extent of the material, is possible. In such a case, there is no need for an intermediate type of solution in a project oriented on a syntactic analysis; however, derivational information, which is so nicely rendered by the morphematic analysis, can be important for some types of reconstruction, and it is thus advisable to include such information into the morphological tag.

4.2 Technical solution for the treebank

The tagging process itself is rather time- and work-consuming task. As stated before, our tagging of Ugaritic is a manual process, since the amount of the attested texts allows for such a solution. Therefore, the process itself is divided into several phases, in which different types of tags may be used. From the technical point of view, the first phase of annotation is to ensure that the tag used for this phase is readable and easily understandable by a human, so that the error rate is diminished. The following table offers an example of the state of the art of the tagset used for the first phase of annotation.

4.2.1 Overview of the current state of the art (NOUN)

The current tagset used for the first phase of annotation is shown in the following table. It is based on morphosyntactic approach, and it is human-readable.

Tag	Description of Word Category	Example	Translation
NCSgMN	sing., masc., nom., common noun	*mlk / malku*	the/a king
NCSgMG	sing., masc., gen., common noun	*mlk / malki*	the/a king
NCSgMA	sing., masc., acc., common noun	*mlk / malka*	the/a king
NCSgMV	sing., masc., voc., common noun	*mlk / malk- (?)*	(O) king!
NCSgFN	sing., fem., nom., common noun	*mlkt / malkatu*	the/a queen

NCSgFG	sing., fem., gen., common noun	*mlkt / malkati*	the/a queen
NCSgFA	sing., fem., acc., common noun	*mlkt / malkata*	the/a queen
NCSgFV	sing., fem., voc., common noun	*mlkt / malkat- (?)*	(O) queen!
NCDuMN	dual, masc., nom., common noun	*mlkm / malkāmi*	the/- kings
NCDuMG	dual, masc., gen., common noun	*mlkm / malkêma*	the/- kings
NCDuMA	dual, masc., acc., common noun	*mlkm / malkêma*	the/- kings
NCDuMV	dual, masc., voc., common noun	*mlkm / malk- (?)*	(O) kings!
NCDuFN	dual, fem., nom., common noun	*mlktm / malkatāmi*	the/- queens
NCDuFG	dual, fem., gen., common noun	*mlktm / malkatêma*	the/- queens
NCDuFA	dual, fem., acc., common noun	*mlktm / malkatêma*	the/- queens
NCDuFV	dual, fem., voc., common noun	*mlktm / malkat- (?)*	(O) queens!
NCPlMN	plural, masc., nom., common noun	*mlkm / mal(a)kūma*	the/- kings
.....	.....	.....	.....

Table 2: Overview of the initial form of POS tags for Ugaritic noun

The tagging process is divided into several phases, where different types of information are being added to the corpus. It is probably not necessary to treat the minute details of the process here, but rather discuss the final shape of the tag.

4.2.2. The requirements on the final tagset

The basic requirements for the final form of the tagset for Ugaritic are not trivial, as the fragmentary character of the language's attestation can have influence on future analyses. It should be an open system, which will allow its change as new approaches and views are developed. The requirements can be summarized in the following way:

The tagset should:

- discretely represent grammatical categories
- be expandable for future analyses
- allow operations on the tag
- allow "protocolling" in the tag – status of the analysis of the tagged string (fully available in the text, partially reconstructed, fully reconstructed, only some features reconstructible, etc.).

The tag can be viewed as a storage space for linguistic information on the morphological level. As such, it should allow maximal discreteness in the analysis of individual grammatical categories that can be applied to Ugaritic, as the analysis and/or reconstruction of them should be in many instances done separately.

Expandability is one of crucial properties of such a tagset, as many features in the text can be interpreted in several ways, and the use of the tagset should allow for an approximative analysis, based on the use of major properties (or their combination). Another advantage of such an approach is the fact that it also offers the reusability of the tags for further analyses of the texts. E.g., when the texts attested on tablets are reconstructed into a vocalized form, as is the case in recent editions of Ugaritic texts (cf. Pardee 2000), such a reconstruction may bring a finer analysis of the grammatical properties, some of which are not easily recognizable in the attested non-vocalized texts (e.g., the so-called 'long' form of a prefixed verb, corresponding to Arabic or Hebrew imperfect). Such features can also be easily implemented into the existing tagset at subsequent stages of annotation.

However, the resulting tagset is to contain not only morphological information, but also other possible characteristics of the respective string. The most important of those is the state of the string, as well as the state of the respective categories (this issue is further elaborated in section 4.3.1). Also, the reconstruction of some of the features contained in the tag can be subject to further development, which has to be noted in the tag (see more in 4.3.2).

The resulting type of tag chosen for the morphological annotation of Ugaritic is a positional tag with special slots for each category; moreover, the tagset includes characteristics of respective strings and the state of the art of individual linguistic information. As the positional tag can be viewed as an orthogonal structure, it is certainly open for easy adaptations according to new analyses in the future.

4.2.3 The shape of the positional tag

The tag developed for Ugaritic is both positional and hierarchical. The positional character is represented by a fixed position for each category and its description. The hierarchy of the tagset lies in the fact that in its construction there are some groups of linguistic information which are of hierarchical nature. The basic group describes the highest level of morphological information, i.e. the POS label ("Major Label"), such as Noun or Verb; the subsequent group contains information on the type of the "Major Label", such as Perfect or Imperfect for verbs or Common, Personal or Theophoric for nouns, Personal or Demonstrative for pronouns, etc. On the following positions, grammatical categories are grouped together. This part, which occupies most of the slots in the positional tag, is hierarchical, too – categories shared by several POS are made prominent, and specialized categories, such as "distal", which concern only special cases of some POS (pronouns), are on positions closer to the end of the tag.

Also the extent of reconstruction must also be marked. At the very end, information on the state of the respective string is provided (cf. also section 4.3.1). Information on the reconstruction of individual categories is also marked (see section 4.3.2).

The structure of the tag is shown in the following table:

Position	Type	Possible values	Meaning of values
1 (G1)	Major label	N, V, P, C, ...	Nominal, Verb, Pronoun, Conjunction, ...
2 (G2)	POS type	C, T, P, ...	Common, Theophoric, Personal, ...
3 (G2)	Derivation	A, P, S, ...	Active participle, passive participle, ...
4 (G3)	Gender	M, F, X, ...	Masculine, Feminine, Non-applicable, ...
5 (G3)	Number	S, D, P, ...	Singular, Dual, Plural, ...
.....	Sequence: from categories shared by several POS to categories applicable only to one POS.		
12 (G4)	State / protocol	A, F, P, ...	Available in the text, Fully reconstructed, Partially reconstructed, ...

Table 3: Structure of the positional POS tag for Ugaritic.

4.3 The extent of possible reconstruction

KTU 1:17: VI
[xxxxxxxxxx]x [xxxxx]
[xxxxxxxxx. l]ḥm[xxxxx]
[xxxxxxxx y]n . ay. ᶜ[d xxx]
[xxxxxx b ḥ]rb . mlḫ[t . q]ṣ
[mri . tšty . b ks . ksp] yn . b ks . ḫrṣ
[dm . ᶜṣm . ymlu]n . krpn . ᶜl . krpn
[xxxxxxxxx]qym . w tᶜl . trṯ
[xxxxxxxxx]n . yn . ᶜšy . l ḥbš
[w aqht . y]nḥtn . qn . yṣbt
[qšt . bnt . kṯ]r . b nši ᶜnh [.] w tphn
[xxxxxxxxxx]xl . kslh . k brq
[xxxxxxxxxx]k . yġḏ . thmt . brq
[xxxxxxx .] qnh . tṣb . qšt . bnt k
[ṯr . w ḫss . d qr]nh . km . bṯn . yqr

As stated several times, Ugaritic texts are attested in a rather fragmentary form. The tablets contain lots of lacunae, in many cases covering considerable part of a tablet, there are lots of places where the presence of a sign can only be expected (marked as "*x*" in the texts), and many signs are damaged. An example of the extent of damage can be found in the example from KTU 1:17 (the story of Aqhat), shown in the box (the parts with x-signs are in the damaged parts of the tablet, some parts are restored). Basically, the states of signs recognized in the editions are: 1) fully attested in the texts, 2) damaged, 3) unreadable and 4) empty. The second category is rather vague, as the damage could be expressed in a scale (the extent of damage in case of a single grapheme). However, this task should be completed during the edition of the texts.

On the other hand, such a situation calls for caution when working with the texts. There is a considerable degree of uncertainty that should be registered not only in the texts themselves, but also in the subsequent linguistic annotation. This uncertainty is then projected onto the syntactic level, however, some corrections of the uncertainty on higher levels are also possible due to the use of structurally conditioned information. Such a reconstruction is usually made based on the context or on other identical or similar passages attested in Ugaritic texts, in some cases also on external evidence from other Semitic languages (esp. Hebrew).

However, we believe that a certain part of the reconstruction can be reached based on linguistic information derived from both morphological and syntactic structures, although this reconstruction can be partial at times – so far, only reconstruction expressed in a presence or absence of a word/string in the text has been applied, however, we believe that it is equally important to start to fill the grid on the level lower than a word or a string, e.g. on the level of morphological categories, as there are cases where we can say that the missing string must have some characteristics, such as being plural or accusative. Even such piece of information can be very helpful in promoting the reconstruction of Ugaritic texts and can be used e.g. in the reconstruction of syntactic relations or other types of linguistic information.
However, reconstruction that works only with pieces of information, must be carried out in a very cautious manner. That is why we have introduced a type of protocol included in the tag, which reflects the extent of the reconstruction of a word/string.

4.3.1 Protocol in the tag:
There are several states that should be distinguished, which show the extent to which the individual word (string after tokenization) is available in the text. Currently, six states are explicitly marked, it is however easy to make the tagset more "fine-grained" or to join some of the states:

- the word is fully attested in the text;
- the word is partially attested in the text, but some or all graphemes are damaged to a certain degree, however, they are still fully or almost fully readable;
- the word is partially restored: some of the graphemes are completely missing, but can be restored;
- the word is fully restored: all the graphemes of the word in the text are missing, the whole word is restored according to some external information;
- the word is not available: the word is in lacuna, but due to some external arguments its presence in the lacuna can be expected, however, we cannot be sure of the type of such a word;
- the word is part of a cluster in a lacuna: there is a long lacuna in the text, where probably more than one word would fit, but it is probable that a certain type of a tag can be expected in the lacuna, without distinguishing its real position in it (initial, middle, final).

All of the above mentioned states concern the word/string as a whole and its position in the text.

4.3.2 Uncertainty zone at individual features of the tag
The uncertainty can also be found at individual categories in the positional tag. We can speak of a reconstruction (analysis) of e.g. a string of nominal or verbal character of the respective string, or maybe grammatical gender or number in the string

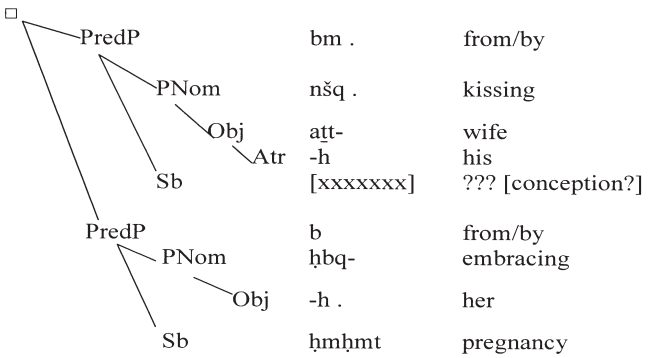
according to some reasons, which can be based on linguistic, structural or other type of information available to us from indirect sources or pointers. It can thus happen that only some of categories suitable for a string under analysis can be filled in, while others can remain "empty" – this emptiness can be, though, understood as a scale, where the "emptiness" can change and a feature can be reconstructed. That is why an uncertainty zone is defined also for individual slots in the positional tag, to allow for a description of the status of the position/feature in the tag in a more precise way. Currently, the following states are distinguished:

- no analysis has been applied yet
- no analysis is possible at the current state of knowledge
- the feature is not applicable

As some features can be reconstructed individually, the tag would exhibit a certain disbalance. The above mentioned information should diminish such a disbalance, as well as provide documentation of the process of reconstruction. It is also clear that with some types of POS, the usage of this scale is more frequent, while with others with less flection, the reconstruction is easier.

4.3.3 Example of an application of the tagset on reconstructed parts

The following example shows a possible approach to the reconstruction and application of some of the features of the tagset developed for Ugaritic:



The structure shows a phrase from a legend of Aqhat (KTU 1.17 I 39-40); the part expressed in [xxxxxxx] means that this part of the tablet can contain 7 Ugaritic signs; it can be analysed as a subject of a nominal sentence. The passage is structured – it is an example of a *parallelismus membrorum,* where in the two parts the structures are very similar, which allows further reconstruction based on linguistic and structural arguments. Within the string, some of the signs will represent a subject of the sentence, possibly expected to be expressed by a noun, as we have a parallel in the second structure (*ḫmḫmt* - "pregnancy"), and it will also be

parallel in meaning. The proposed insertion of a word meaning "conception" would thus mean that the string contains more than one word (the word "conception" – in Ugaritic *hr*, has only 2 signs).[5] Our description of the string in the lacuna at the present state of our knowledge is following: It is part of a lacuna, whose position within the lacuna we cannot exactly determine, however, due to the information from the parallel structure it is Nominal, most probably a common noun, and as its counterpart from the parallel structure is Subject, we can expect it to be in nominative case. The morphological tag for [xxxxxxx] then is the following: [NC----1----Z].

5. Conclusion

In the paper, the problems and requirements of the morphological tagging of Ugaritic have been discussed, mainly the conceptual issues of a morphological tagset.
We believe that Ugaritic should use a tagset based on the morphosyntactic approach, which reflects the shape of the language. Beside the match between the type of the language (flective) and the approach to the tagset, there are more reasons, among them there is the unavailability of a tagger (or even a morphological analyzer) for Ugaritic, and also the type of the corpus that is being discussed here, namely a treebank of Ugaritic, i.e. a corpus which apart from morphological annotation includes a syntactic description.
For us, the positional tag meets most of the needs of a tagset for Ugaritic. It enables a discrete treatment of individual grammatical categories and their independent reconstruction. As an orthogonal structure, it is easy to maintain or change it.
As Ugaritic is attested in a fragmentary form, a high degree of uncertainty is met. This fact needs to be manifested in the tag as well. For the representation of the degree of uncertainty, two forms have been developed: one related to the reconstruction of strings in the Ugaritic texts, the other in order to allow the description of the process of the reconstruction of individual features that are contained in the tag. The distinction between the uncertainty on the word/string level and on the level of individual grammatical categories allows to approach the reconstruction of Ugaritic as a step-by-step operation that can start from small pieces of knowledge.

References

Adler, Meni and Michael Elhadad, 2006: An Unsupervised Morpheme-Based HMM for Hebrew Morphological Disambiguation. In: *ACL 2006. Proceedings of the 21st*

5 This certainly has implications for syntactic analysis; however, in the theory we have chosen for the description of the Ugaritic syntax (dependency approach), the expansion of a node is possible, e.g. into a noun phrase; this would allow to change the proposed word, while its morphological and syntactic properties will remain the same regardless of its actual meaning.

International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, 17-21 July 2006, pp. 665-672. http://acl.ldc.upenn.edu/P/P06/P06-1084.pdf (16.1.2008).

Bar-Haim, Roy – Khalil Sima'an – Yoad Winter, 2007: Part-Of-Speech Tagging of Modern Hebrew Text. *Natural Language Engineering* 1, 2007: 223-251. http://www.cs.technion.ac.il/~winter/papers/HebPOStagging.pdf (16.1.2008).

Beesley, Kenneth R.. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research. Status and Plans in 2001. In: ACL/EACL-2001 Workshop on Arabic Language Processing: Status and Prospects. Toulouse, France, pp. 1-8. http://www.elsnet.org/arabic2001/beesley.pdf (16.1.2008).

Buckwalter, Tim. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. LDC Catalog No. LDC2004L02, Linguistic Data Consortium, http://www.ldc.upenn.edu/Catalog (16.1.2008).

Buckwalter, Tim. 2005. Arabic Morphology. http://www.qamus.org/morphology.htm (16.1.2008).

Čech, Pavel. 2005. Ugaritic Narrative: Annotating Very Fragmentary Corpora. Poster at a Conference *Framing Plots: the Grammar of Ancient Near Eastern Narratives*, London. http://usj.ff.cuni.cz/veda/publikace/cech/UCLPoster05.pdf (16.1.2008).

Cunchillos Ilarri, Jesús-Luis and Raquel Cervignon Moreno. 1998. Analizador Morfologico Ugaritico (AMU). *First International Conference on Language Resources and Evaluation (LREC)*, poster. Granada 1998.

The Cuneiform Digital Library Initiative. http://cdli.ucla.edu (16.1.2008).

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

Freeman, Andrew. 2001. Brill's POS tagger and a morphology parser for Arabic. In *ACL/EACL-2001 Workshop on Arabic Language Processing: Status and Prospects.* 7 p. Toulouse, France. http://www.elsnet.org/arabic2001/freeman.pdf (16.1.2008).

García-Serrano, Ana and Jesús Contreras. 1998. A Computational Platform for Ugaritic Morphological Analysis. *First International Conference on Language Resources and Evaluation (LREC)*, Granada 1998, 6p. http://www.isys.dia.fi.upm.es/~agarcia/UMA-LREC.ps (16.1.2008).

Khoja, Shereen – Roger Garside and Gerry Knowles. 2001. A Tagset for the Morphosyntactic Tagging of Arabic. In: *Proceedings of the Corpus Linguistics.* Lancaster University (UK), Volume 13 - Special issue, 341, pp. 59-72. http://zeus.cs.pacificu.edu/shereen/CL2001.pdf (16.1.2008).

Kiraz, Georg Anton. 2001. *Computational Nonlinear Morphology. With Emphasis on Semitic Languages.* Cambridge: Cambridge University Press.

McCarthy, John. 1985. *Formal Problems in Semitic Phonology and Morphology*. New York: Garland Publishing.

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

Smrž, Otakar. 2007. *Functional Arabic Morphology. Formal System and Implementation.* Doctoral Thesis, Charles University in Prague, July 2007. `http://ufal.mff.cuni.cz/~smrz/ElixirThesis/elixir-thesis.pdf` (16.1.2008).

Smrž, Otakar and Jan Hajič. 2007. The Other Arabic Treebank: Prague Dependencies and Functions. In: Ali Farghali (ed.): *Arabic Computational Linguistics: Current Implementations.* Stanford: CSLI Publications (to appear). `http://ufal.mff.cuni.cz/~smrz/CSLI2006/csli-prague.pdf` (16.1.2008).

Tropper, Josef. 2000. *Ugaritische Grammatik*. Münster: Ugarit Verlag.

Zemánek, Petr. 2007. A Treebank of Ugaritic. Annotating Fragmentary Attested Languages. In: Koenraad De Smedt, Jan Hajič, Sandra Kübler (eds.): *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories.* December 7-8, 2007, Bergen, Norway. NEALT 1, pp. 213-218.