Creating Tools for Morphological Analysis of Sumerian

Valentin Tablan, Wim Peters, Diana Maynard, Hamish Cunningham

Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street
S1 4DP, Sheffield, UK
{valyt, wim, diana, hamish}@dcs.shef.ac.uk

Abstract

Sumerian is a long-extinct language documented throughout the ancient Middle East, arguably the first language for which we have written evidence, and is a language isolate (i.e. no related languages have so far been identified). The Electronic Text Corpus of Sumerian Literature (ETCSL), based at the University of Oxford, aims to make accessible on the web over 350 literary works composed during the late third and early second millennia BCE. The transliterations and translations can be searched, browsed and read online using the tools of the website. In this paper we describe the creation of linguistic analysis and corpus search tools for Sumerian, as part of the development of the ETCSL. This is designed to enable Sumerian scholars, students and interested laymen to analyse the texts online and electronically, and to further knowledge about the language.

1. Introduction

The Sumerian language of ancient Sumer is a long-extinct language documented throughout the ancient Middle East, in particular in the south of modern Iraq, from at least the 4th millenium BC. It is arguably the first language for which we have written evidence, the rival candidate being ancient Egyptian. Sumerian was replaced by Akkadian as a spoken language around 2000 BC, but continued to be used as a sacred, ceremonial and scientific language in Mesopotamia until about 1 AD (Wikipedia, 2006).

The Electronic Text Corpus of Sumerian Literature (ETCSL), based at the University of Oxford, aims to make accessible on the web over 350 literary works composed during the late third and early second millennia BCE. The corpus comprises Sumerian texts in transliteration, English prose translations and bibliographical information for each composition. The transliterations and the translations can be searched, browsed and read online using the tools of the website.

In this paper we describe the creation of linguistic analysis and corpus search tools for Sumerian, as part of the development of the ETCSL. This is designed to enable Sumerian specialists to analyse the texts online and electronically and to further knowledge about the language.

The benefit of having linguistic information as part of cultural heritage digital libraries has been demonstrated by the success of the Perseus Digital Library [1], where this information is used not only by scholars, but also by students and interdisciplinary researchers who want to study the electronic collections but are not proficient in the language (Crane, 1996). In a similar fashion, we expect that the creation of tools for Sumerian is likely to enhance the accessibility of this collection to scholars, students and the general public.

2. Characterization of Sumerian

The Sumerian language is generally regarded as a language isolate in linguistics. No languages related to it have so far been convincingly identified, although many of its grammatical features are attested in other living languages outside of the Indo-European family to which English belongs.

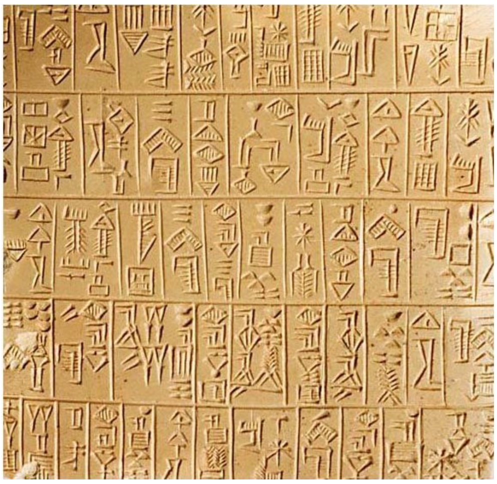


Figure 1: Example of Sumerian cuneiform

In its orthographic form, Sumerian is encoded in cuneiform script, as depicted in Figure 1, which shows an example of 120 compartments of cuneiform script written by an expert scribe. Note that this example is much clearer and more beautiful than standard scripts because it describes the gifts from highly placed persons to a priestess.

Originally, cuneiform was logographic in nature, and a sign represented a content word (a thing or an action). It gradually developed into a combined system, where the same set of signs could be used to represent logograms and phonograms or syllabograms. In texts of the period we concentrate on, i.e. late third and early second millennium B.C., logograms were used to write content words and the base (root) of a word, while phonograms were used to write bound morphemes and loan words. In transliterated form, i.e. signs represented in the Roman alphabet with a few additions, these logograms and syllabograms are separated

[1] http://www.perseus.tufts.edu/

by a dash, as in nam-lugal (kingship), where the base root lugal (king) is combined with a derivational affix that changes the word into an abstract entity.

Another characteristic feature of Sumerian is the large number of homophones (words with the same sound structure but different meanings) - or perhaps pseudo-homophones, since there might have been differences in pronunciation (such as tone) that we do not know about. The different homophones (or, more precisely, the different cuneiform signs that denote them) are marked with different numbers by convention. For example: du = ”to go”, du3 = ”to build”.

In terms of language typology, Sumerian is agglutinative. Word roots have grammatical elements glued on before or after them to build up complex grammatical forms. Many words (msotly verbs) consist of a root form (possibly reduplicated) and a chain of more or less clearly distinguishable and separable affixes or clitics. Nouns may have affix chains before as well as after the root. Overall, slightly less than 100 clitics have been postulated. Many of these clitics have allomorphs, depending on properties of the morphological context, such as progressive or regressive assimilation phenomena for vocals and consonants. On top of this, the null morpheme can occur as an allomorphic variant of a number of clitics. Overall, if we take all morphological rules into account, this leads to a huge number of possible interpretations.

Sumerian distinguishes the grammatical genders animate and inanimate, as do Polish, Russian, and some Native American languages, such as Navajo. There are also a large number of cases - nominative, ergative, genitive, dative, locative, comitative, equative (”as, like”), terminative (”to”), and ablative (”from”). The case markers for nouns convey information about case and animacy, e.g.:

- -ak (genitive)
- -da (comitative)
- -e (animate ergative)
- -ra (animate dative/directive)
- -a (inanimate locative/directive)
- -e (inanimate dative)
- -ta (inanimate ablative)

Many of these clitics have allomorphic representations. For example, the genitive clitic ’ak’ can be written by means of, amongst others, the following variants:

- a in final position: dumu lugala the son (dumu) of the king (lugal-ak)
- Ca-kV In this case, ak bridges two syllabograms, where ’C’ stands for consonant and ’V’ for vocal: an sky an-na-ke4 of the sky (an-ak-e4)

In order to approach the linguistic complexity of Sumerian in a constructive way, we have created an incrementally complex automatic analysis module, starting with a slot based lookup and adding increasingly complex combinatorial constraints.

3. Morphological Analysis Tools for Sumerian

The main aim of our work is to create a set of tools for performing automatic morphological analysis of Sumerian. This essentially entails identifying the part of speech for each word in the corpus (technically, this only involves nouns and verbs which are the only categories that are inflected), separating the lemma part from the clitics and assigning a morphological function to each of the clitics. In order to do this, we used the model of Sumerian morphology defined by a team of Sumerologists, which we then represented in a way that can be used for automated language processing.

The morphological model we used consists of noun and verb templates comprising a lemma plus a number of morphological slots that could be filled. The nouns have a lemma and up to six suffix slots while the verbs have up to twelve prefix slots, a lemma and two suffix slots. For each slot there is a known list of morphemes that can fill it and a set of restrictions encoding dependencies between the slots, such as agreement in gender. The lists of candidate slot fillers have non-null intersections – the same morpheme can appear in several lists, though usually with different functions.

There are two main phases involved in our morphological analysis of transliterated Sumerian – a normalisation stage which deals with various surface phenomena which affect the way Sumerian words are written, such as reduplication or assimilation, and the actual morphological analysis which identifies parts of speech and assigns functions to the various morphemes.

The implementation for the morphological analysis tools has been done using the GATE language engineering platform (Cunningham et al., 2002a) and includes the following high-level tasks:

Tokenisation: splits the input text into syllables while identifying special text components such as determiners and markers for damaged regions in the original clay tablet.

Input normalisation: makes explicit the ambiguity caused by some phenomena in written Sumerian by generating all possible normalised interpretations for each particular text fragment.

Slot fillers look-up: identifies syllables in the input that are candidate fillers for morphological slots.

Non-inflected words lookup: identifies words that are not inflected by looking them up in a predefined list.

Morphological analysis: identifies nouns and verbs and generates structure information by labelling the lemma and all the other constituents.

The tokenisation step is performed by a customised variant of the GATE English Tokeniser. Its role is to identify syllables – defined as contiguous sequences of letters, dashes, determinatives – marked by square brackets in the transliteration schema, and entities – marked by round brackets.

The output of the tokeniser consists of a set of annotations of various types (syllable, dash, entity, determinative and whitespace) which cover the entire input text.

The output from the tokeniser is then normalised by adding new syllables in parallel with the existing ones where the surface form can be interpreted in several different ways. For example, in the word `sizkur2-ra` the second syllable can either be seen as `ra`, or `a` in the case when the `r` was a reduplication of the final consonant in the previous syllable, or `ak` where `r` is a reduplication and the final `k` was dropped. The normalisation phase will add all possible alternative interpretations for all syllables where such phenomena occur. This will generate noise (as only one of all the alternate interpretations is the right one); the intention is to let the later analysis phase choose the right one based on combinatorial restrictions encoded by the morphology model.

The next phase in the processing identifies the candidate slot fillers. This is essentially a list look-up operation based on the pre-compiled lists of possible fillers for each morphology slot. This was implemented using the GATE Gazetteer which needed to be modified in order to deal with the alternative syllable interpretations generated by the tokenisation phase. All sequences of one or more syllables that could play the role of a morphological slot are annotated with their type (i.e. part of a noun or a verb), slot position in the noun or verb template and morphological function.

The following step is that of identifying words that are non-inflected – another list look-up operation, based on a list of known words which, although multi-syllabic, do not show inflection. This was necessary because our analysis method is based on looking at the internal structure of each word and trying to identify morphemes that could play the role of clitics. Multi-syllabic non-inflected words can trick the system when some of the syllables in their lemmas could also be interpreted as morphological markers.

The final phase of the analysis process is that of identifying nouns and verbs and performing morphological analysis determining which syllables form the lemma and which are clitics with various functions. This was implemented as a cascade of JAPE transducers (Cunningham et al., 2002b). The first one identifies exceptional spellings and normalises them – for instance the `nuc` clitic is sometimes written as `ni-ic`. After this final normalisation, grammars that identify nouns and verbs are run. These essentially try to find words that satisfy regular expression-like patterns based on slot candidate and syllable annotations created by the previous modules. The word patterns encode both the correct position of the various slots in relation to the lemma and the other restrictions to do with various type of agreement (for example one cannot have both an `animate` and `inanimate` marker as part of the same word). The assignment of priorities to the various rules is done in such a way that versions that assign more syllables to morphological functions are preferred to the ones that assign more syllables to the lemma. This assures that when a particular syllable can be successfully assigned a morphological function, then this will be done in preference to considering it part of the lemma.

Although the application was designed to address both nouns and verbs at the same time, we have concentrated our efforts first on the noun morphology, which is the simpler case because of the fewer slots, aiming to direct our attention to the more complex case of verbs after we get a good understanding of the phenomena we need to address, and when we are confident that the architecture of our application is well suited for Sumerian morphology. Work is currently in progress on improving the analysis of the verbs.

4. Evaluation of Results

To evaluate the results, we obtained a copy of the corpus automatically annotated with morphological information using a tool developed at the University of Pennsylvania[2]. Although that annotation is not perfect (the tool does make some mistakes and also the model of Sumerian morphology used differs slightly from the one defined by the Oxford group) it does give us a good indication of where problems might occur. In the current development state of the application, the results as evaluated over a document containing some 2300 nouns and 1400 verbs are as follows:

Type	P	R	F
noun recognition	59%	84%	69%
verb recognition	65%	67%	66%
morphological analysis	52%	73%	61%

Table 1: Evaluation of POS recognition and morphological analysis

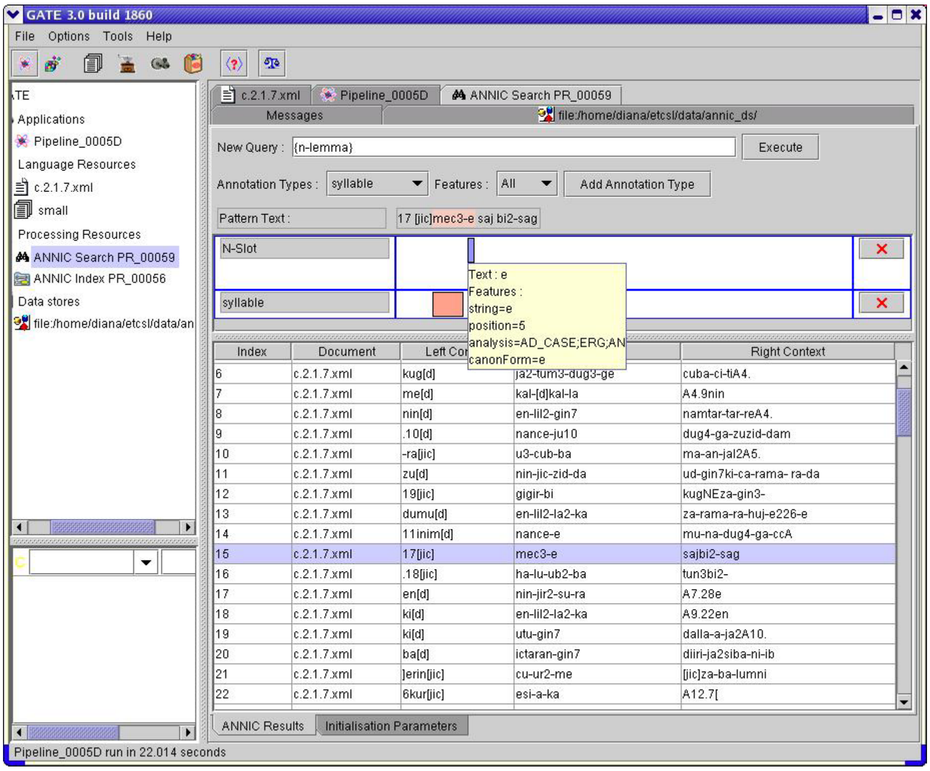
The only other system for automatically analysing Sumerian that we know of is the work at Pennsylvania which we are using as the gold standard. So we can only compare our work with this. A manually created gold standard is forthcoming from ETCSL.

Considering the difficulty of the task and the stage of the work, these results are very promising. We have not yet measured the morphological analysis of verbs but this is forthcoming. Note also that since there is much ambiguity between nouns and verbs, errors in the identification of nouns will generally also have an impact on identification of verbs, and vice versa, because missing nouns will often be falsely identified as verbs and so on.

5. Corpus Search Tools

The linguistic analysis tools described above are complemented by the development of a tool for advanced search and visualisation of linguistic information, ANNIC (ANNotations In Context) (Aswani et al., 2005). This provides an alternative method of searching the textual data in the corpus, by identifying patterns in the corpus that are defined both in terms of the textual information (i.e. the actual content) and of metadata (i.e. linguistic annotation and XML/TEI markup). ANNIC is similar to a KWIC (Key-Words In Context) index, but where a KWIC index provides simply text in context in response to a search for specific

[2] http://psd.museum.upenn.edu/epsd/index.html



New Query : {n-lemma}

Annotation Types : syllable

Features : All

Pattern Text : 17 [jic]mec3-e saj bi2-sag

N-Slot

syllable

Text : e
Features :
string=e
position=5
analysis=AD_CASE;ERG;AN
canonForm=e

Index	Document	Left Co		Right Context
6	c.2.1.7.xml	kug[d]	ja2-tum3-dug3-ge	cuba-ci-tiA4.
7	c.2.1.7.xml	me[d]	kal-[d]kal-la	A4.9nin
8	c.2.1.7.xml	nin[d]	en-lil2-gin7	namtar-tar-reA4.
9	c.2.1.7.xml	.10[d]	nance-ju10	dug4-ga-zuzid-dam
10	c.2.1.7.xml	-ra[jic]	u3-cub-ba	ma-an-jal2A5.
11	c.2.1.7.xml	zu[d]	nin-jic-zid-da	ud-gin7ki-ca-rama- ra-da
12	c.2.1.7.xml	19[jic]	gigir-bi	kugNEza-gin3-
13	c.2.1.7.xml	dumu[d]	en-lil2-la2-ka	za-rama-ra-huj-e226-e
14	c.2.1.7.xml	11inim[d]	nance-e	mu-na-dug4-ga-ccA
15	c.2.1.7.xml	17[jic]	mec3-e	sajbi2-sag
16	c.2.1.7.xml	.18[jic]	ha-lu-ub2-ba	tun3bi2-
17	c.2.1.7.xml	en[d]	nin-jir2-su-ra	A7.28e
18	c.2.1.7.xml	ki[d]	en-lil2-la2-ka	A9.22en
19	c.2.1.7.xml	ki[d]	utu-gin7	dalla-a-ja2A10.
20	c.2.1.7.xml	ba[d]	ictaran-gin7	diiri-ja2siba-ni-ib
21	c.2.1.7.xml	]erin[jic]	cu-ur2-me	[jic]za-ba-lumni
22	c.2.1.7.xml	6kur[jic]	esi-a-ka	A12.7[

Pipeline_0005D run in 22.014 seconds

Figure 2: Using ANNIC on Sumerian text

words, ANNIC additionally provides linguistic information (or other annotations) in context, in response to a search for particular linguistic patterns. ANNIC is based on Jakarta Lucene, but extends the model to index large corpora and allow users to query on annotations and their features by providing JAPE patterns (LHS rules), as described in Section 3.. This enables the retrieval of information in the form of Annotations in Context, and provides a useful interface to assist the creation of new JAPE rules. The functionality is provided as a plugin in GATE.

ANNIC consists of two processing resources (index and search) and a visual resource (viewer). The index processing resource creates an index that is required for the search process. A corpus can be indexed on words, morphs or other annotations as appropriate – these are the segments that will be searched on in the second stage. The search processing resource takes as input a pattern on which to search, which can consist of annotations and regular expressions: for example, one can search on specific combinations of morphs or whole words. A context size parameter is also set, determining how large a context window should be used. The viewer is the interface which displays the results. Given a query to the ANNIC Search engine, it returns the list of documents that contain the specific pattern, and for each document it returns the patterns and contexts. Users have the option of viewing the results in different ways according to their needs. Figure 2 shows a screenshot of ANNIC with Sumerian text.

6. Conclusions

In this paper we have described the development of tools for the linguistic analysis of Sumerian, including a facility to search a corpus of annotated text for morphological patterns. Work is still ongoing but current results are very promising. The complexity of Sumerian and the scarcity of available textual material will, however, remain obstacles for the success of automatic analysis.

The additional search facilities embodied in the ANNIC tool will enable the user to perform manual checks and detect additional combinatorial phenomena in the text corpus.

7. References

N. Aswani, V. Tablan, K. Bontcheva, and H. Cunningham. 2005. Indexing and Querying Linguistic Metadata and Document Content. In *Proceedings of Fifth International Conference on Recent Advances in Natural Language Processing (RANLP2005)*, Borovets, Bulgaria.

Gregory Crane. 1996. Building a digital library: the perseus project as a case study in the humanities. In *Proceedings of the first ACM international conference on Digital libraries*, pages 3–10. ACM Press.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002a. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, and C. Ursu. 2002b. *The GATE User Guide*. `http://gate.ac.uk/`.

Wikipedia. 2006. Sumerian language — wikipedia, the free encyclopedia.