Creating Tools for Morphological Analysis of Sumerian

Valentin Tablan, Wim Peters, Diana Maynard

October 11, 2005

1 Introduction

Sumerian is a long-extinct language documented throughout the ancient Middle East, in particular in the south of modern Iraq. It is arguably the first language for which we have written evidence, the rival candidate being ancient Egyptian. Sumerian is a language isolate, i.e. 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.

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.

2 Creation of Linguistic 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 complex cases), 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.

An application has been developed in GATE (General Architecture for Language Engineering) [Cunningham et al., 2002] for performing morphological analysis on Sumerian. The application performs the following high-level tasks:

a. 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.

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

c. Slot fillers look-up: identifies syllables in the input that might denote morphological inflection.

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

e. Slot identification: finds sets of candidate morphological slots by grouping syllables identified at step c) and assigns labels to such slots.

f. Part-of-speech analysis: identifies nouns and verbs.

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

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, 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 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.

3 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. 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:

- recognition of nouns: Precision 59%, Recall 84%, F- Measure 69%
- recognition of verbs: Precision 65%, Recall 67%, F-Measure 66%
- morphological analysis of nouns: Precision 52%, Recall 73%, F-Measure 61%.

The only other system for analysing Sumerian automatically 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. However, onsidering 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.

4 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). 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). 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 annoations 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.

5 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. Sumerian is a very difficult language to analyse because there is much ambiguity, because so little work has been previously done on it and there are few resources available, and because the rules are extremely complicated. In the full paper we shall give more details of the components and some further results.

References

[Cunningham et al., 2002] Cunningham, H., D. Maynard, K. Bontcheva, and V. Tablan, 2002. 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).*