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
{**V.Tablan, W.Peters, D.Maynard, H.Cunningham**}@dcs.shef.ac.uk

The Sumerian Language

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)

Sumerian Morphology

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.

In its orthographic form, Sumerian is encoded in cuneiform script, as depicted in this picture, 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.

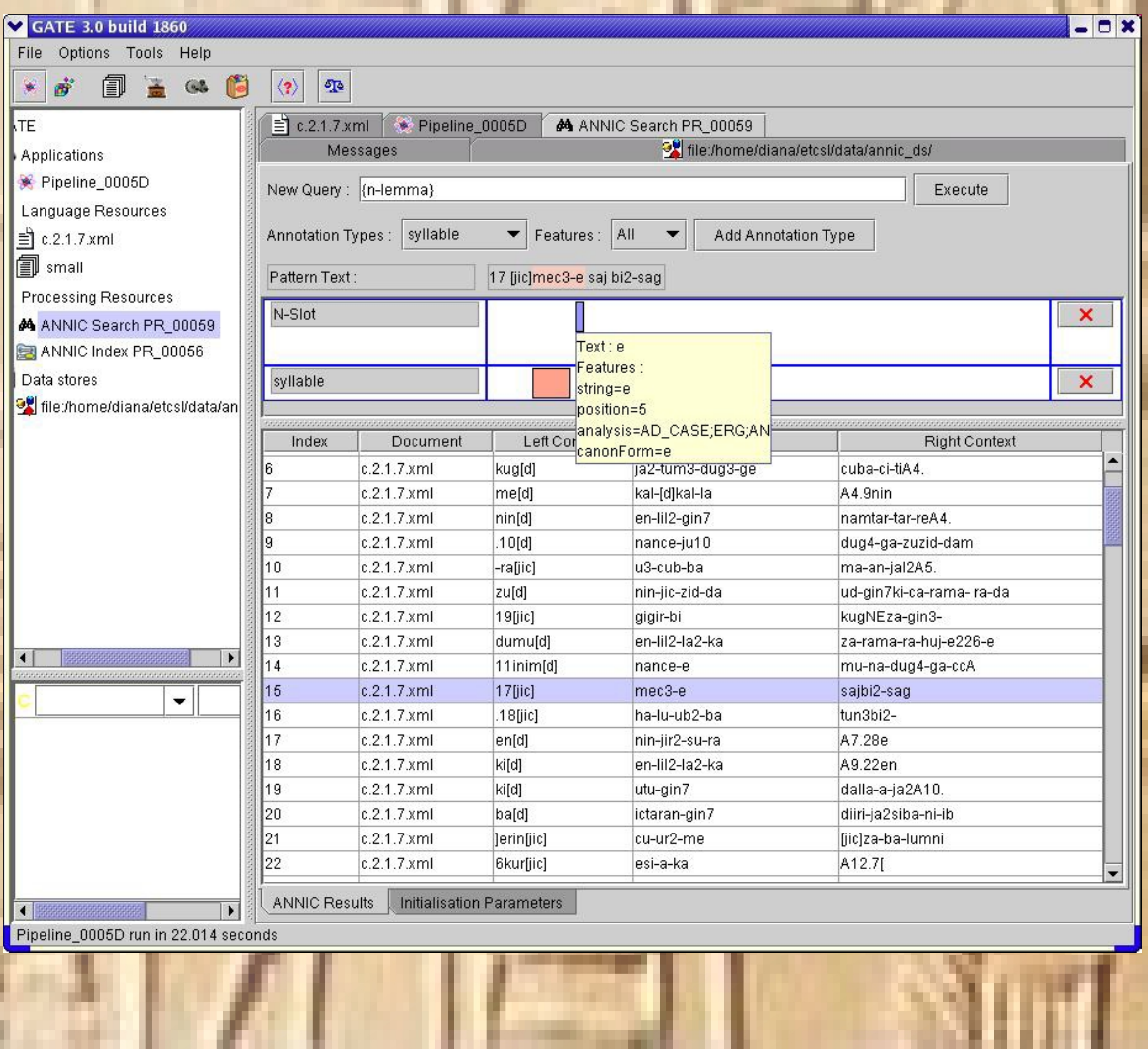
Originally, cuneiform was logographic in nature, and a sign represented a content word (a thing or an action). It gradually developed into a combined system, where the same set of signs could be used to represent logograms and phonograms or syllabograms. In texts of the period we concentrate on, i.e. late third and early second millennium BC, 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 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 (mostly 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 very large 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").



TE
Applications
Pipeline_0005D
Language Resources
c.2.1.7.xml
small
Processing Resources
ANNIC Search PR_00059
ANNIC Index PR_00056
Data stores
file:/home/diana/etcsl/data/an

c.2.1.7.xml | Pipeline_0005D | ANNIC Search PR_00059
Messages | file:/home/diana/etcsl/data/annic_ds/

New Query : {n-lemma} Execute
Annotation Types : syllable Features : All Add Annotation Type
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[

ANNIC Results | Initialisation Parameters

Pipeline_0005D run in 22.014 seconds

The ETCSL Project

The Electronic Text Corpus of Sumerian Literature - ETCSL , based at the University of Oxford, is the largest available corpus of literary Sumerian. It aims to make accessible on the web over 350 works composed during the late third and early second millennia BC. 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.

(http://www-etcsl.orient.ox.ac.uk/)

Morphological Analysis Tools

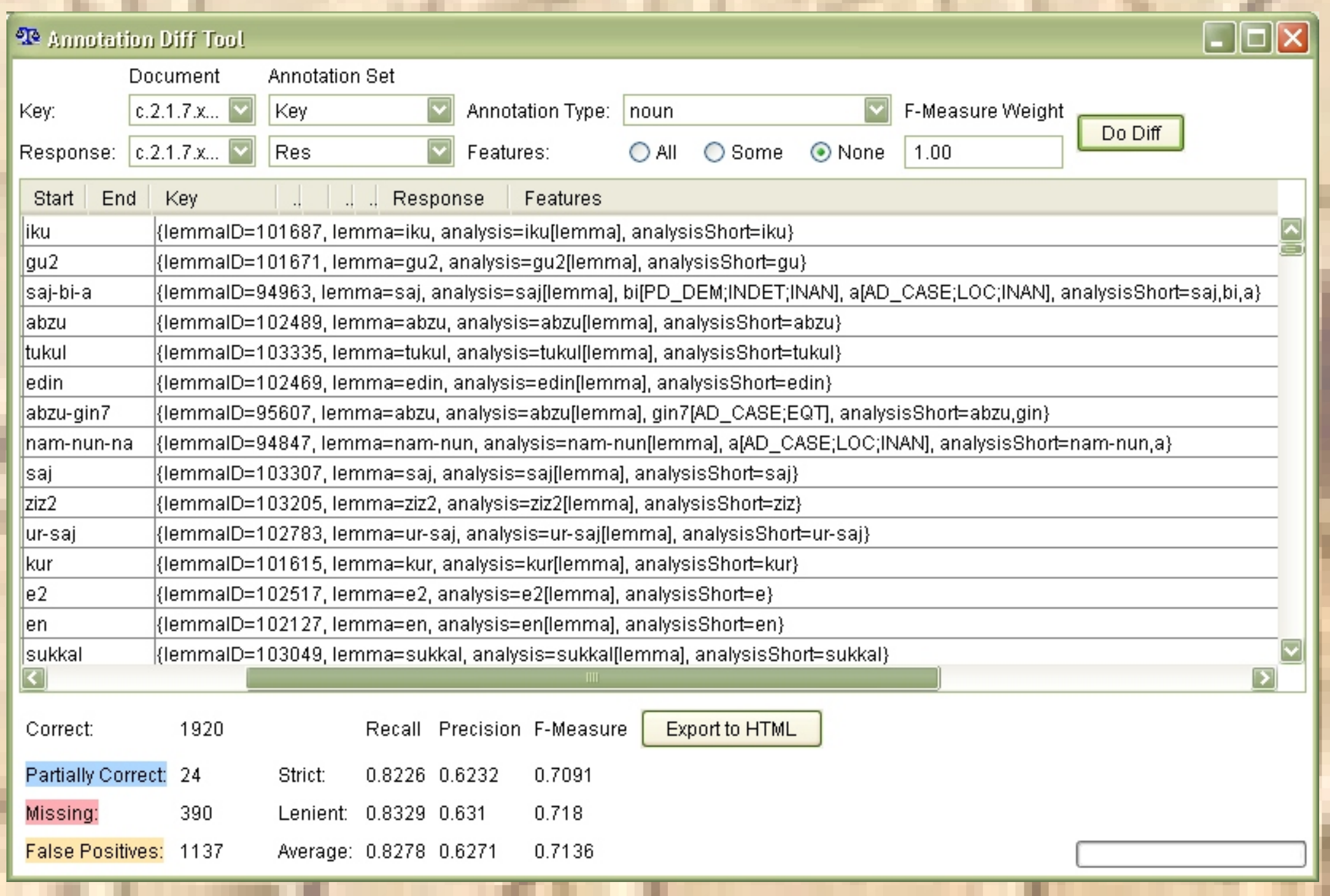
The main aim of our work is to create a set of tools for performing automatic morphological analysis of Sumerian. This essentially entails identifying the part of speech for each word in the corpus (technically, this only involves nouns and verbs which are the only categories that are inflected), separating 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 from the Unviersity of Oxford, 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.

Performance evaluation for the system in its current state, shows an F-Measure of 69% for noun identification, 66% for recognising verbs and 61% for morphological analysis of nouns.



Document	Annotation Set
Key: c.2.1.7.x...	Key
Response: c.2.1.7.x...	Res

Annotation Type: noun
Features: ○ All ○ Some ● None
F-Measure Weight 1.00
Do Diff

Start End	Key .. .. .. Response Features
iku	{lemmaID=101687, lemma=iku, analysis=iku[lemma], analysisShort=iku}
gu2	{lemmaID=101671, lemma=gu2, analysis=gu2[lemma], analysisShort=gu}
saj-bi-a	{lemmaID=94963, lemma=saj, analysis=saj[lemma], bi[PD_DEM;INDET;INAN], a[AD_CASE;LOC;INAN], analysisShort=saj,bi,a}
abzu	{lemmaID=102489, lemma=abzu, analysis=abzu[lemma], analysisShort=abzu}
tukul	{lemmaID=103335, lemma=tukul, analysis=tukul[lemma], analysisShort=tukul}
edin	{lemmaID=102469, lemma=edin, analysis=edin[lemma], analysisShort=edin}
abzu-gin7	{lemmaID=95607, lemma=abzu, analysis=abzu[lemma], gin7[AD_CASE;EQT], analysisShort=abzu,gin}
nam-nun-na	{lemmaID=94847, lemma=nam-nun, analysis=nam-nun[lemma], a[AD_CASE;LOC;INAN], analysisShort=nam-nun,a}
saj	{lemmaID=103307, lemma=saj, analysis=saj[lemma], analysisShort=saj}
ziz2	{lemmaID=103205, lemma=ziz2, analysis=ziz2[lemma], analysisShort=ziz}
ur-saj	{lemmaID=102783, lemma=ur-saj, analysis=ur-saj[lemma], analysisShort=ur-saj}
kur	{lemmaID=101615, lemma=kur, analysis=kur[lemma], analysisShort=kur}
e2	{lemmaID=102517, lemma=e2, analysis=e2[lemma], analysisShort=e}
en	{lemmaID=102127, lemma=en, analysis=en[lemma], analysisShort=en}
sukkal	{lemmaID=103049, lemma=sukkal, analysis=sukkal[lemma], analysisShort=sukkal}

			Recall	Precision	F-Measure
Correct:	1920				
Partially Correct:	24	Strict:	0.8226	0.6232	0.7091
Missing:	390	Lenient:	0.8329	0.631	0.718
False Positives:	1137	Average:	0.8278	0.6271	0.7136

Export to HTML

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).

ANNIC is based on Jakarta Lucene, but extends the model to allow users to query on annotations and their features by providing patterns that are similar to the JAPE rules in GATE (http://gate.ac.uk).

This work has been supported by the UK Arts and Humanities Research Council

Arts & Humanities Research Council