11-712: NLP Lab Report

David Bamman

April 26, 2013

1 Basic Information about Akkadian

Akkadian is the oldest attested member of the Semitic language family used primarily between the 3rd and 1st millennia BCE (Huehnergard, 2005, xxi). It encompasses two central dialects: Assyrian, spoken largely in the Assyrian empire centered in northern Mesopotamia; and Babylonian, centered in the south.

Akkadian texts were written using a cuneiform script adopted from Sumerian. The cuneiform symbols largely represent syllables, though some denote logographic values (e.g., one symbol for the word "man"). The process by which a text is transcribed from a clay tablet includes the following:

1. Transliteration, in which the cuneiform signs are rendered into their syllabic values[1] (let S_n stand in for the signs):[2]

$$\underbrace{S_1}_{q\acute{a}}\underbrace{S_2}_{ra}\underbrace{S_3}_{dum}\underbrace{S_4}_{na}\underbrace{S_5}_{ra}\underbrace{S_6}_{am}\underbrace{S_7}_{i}\underbrace{S_8}_{pu}\underbrace{S_9}_{\text{uš}}$$

2. Normalization, in which the transliterated syllables are rendered into the lexical form of the word.

   qarrādum narâm īpuš

   Vowel length (long vs. short) and whether or not a consonant is doubled are important for distinguishing different words. Note that the original scribes may choose to syllabalize a word in several different ways, so that, for example, the lexical word *išarum* may be written in cuneiform as *i-ša-rum* or *i-ša-ru-um* (Huehnergard, 2005, 71).

As we have them in digitized form, Akkadian texts are generally transliterated, but not normalized. A typical text looks like the following (Kt a/k 394):

um-ma wa-ak-lúm-ma
a-na kà-ri-im
Kà-ni-iš.ki
qí-bí-ma

[1] Logograms are transliterated as capital letters, with homophones disambiguated with an index (e.g., KU$_6$)

[2] Example due to (Huehnergard, 2005, 72).

2 Past Work on the Morphology of Akkadian

The earliest computational work on Akkadian looks to be Kataja and Koskenniemi (1988), which outlines a two-level framework for Akkadian word formation, seeing regular verbs as possessing ordered slots to be filled by (in order): person, root (and flection/vocalization), gender and number, optional subjunctive indicators, and optional object markers; nouns, analogously, are defined as (in order) stem, case and number, and optional possessive markers. The authors provide about a dozen examples of useful phonological alternations, including the assimilation of N before consonants and the assimilation of dentals occuring after other dentals.

Macks (2002) describes work implementing an Akkadian morphological analyzer in Prolog (for the Babylonian dialect). The anaylzer, available online,[3] parses G, D, and N stem verbs in the preterite, perfect, imperfect, durative, precative and vetitive tenses, yielding the verb stem (but not lexical form with vocalization), tense, person, number and gender. The input verb token is required to be in normalized form, though the analyzer does retain some flexiblity by enabling wildcards for vowel length.

Barthélemy (1998) also describes work developing a two-level morphological analyzer for Old Babylonian verb forms (but without suffixes like enclitic particles or pronouns). This work is very useful as a high-level overview of the problems involved in analyzing Akkadian verbs, and offers a useful description of Akkadian *stems* (verbal structural classes, applicable to any root, marked by changes to prefixes, infixes and radical reduplication that impact the semantics of the verb, such as whether a verb is habitual, factitive, causative, passive, etc.). In a manner similar to Kataja and Koskenniemi (1988), this work decomposes a verb into nine slots, each ranging over a fixed vocabulary: the personal prefix, stem prefix, infix, first radical, infix, second radical reduplication, second radical (plus vocalization), third radical, and gender/number suffix, and then transforms the lexical form into a surface form through phonological transformations. This work list a few general trends (e.g., the dissimiliation of $bb \rightarrow mb$, $ij \rightarrow$ i) but no explicit rules. Barthélemy (2009) builds on this work by further decomposing the verb form into a tree-like structure, with the "core" (including the lexical class, voice and aspect) embedded within the personal (gender, number, etc.) affixes.

Two commonalities among all prior work is that 1.) all require normalized forms as input, not the transcribed text that we have in the form of corpora; and 2.) all focus on the more complex verbal morphology; the simpler morphology of nouns and adjectives may be good low-hanging fruit.

3 Available Resources

1. Adam Anderson (Department of Near Eastern Languages and Civilizations, Harvard), informant.

2. Akkadian lexicon divided into verbs, nouns and adjectives. The original format of these lexica are Microsoft Word documents, but can be mined for word stems.

3. A corpus of Old Assyrian texts is available in the form of 2,094 letters between merchants unearthed at the colony of kārum Kaneš near Kültepe, Turkey. Since the original text documents contain annotations and metadata (e.g., line numbers, notes in English and German about the text) and, as transcriptions of cuneiform tablets, also contain word fragments (where the original tablet is missing or illegible), I create development and test corpora by extracting only those tokens with a hyphen (which denotes cuneiform sign boundaries) and exclude all

[3] `http://www.wiglaf.org/akkadian/`

words with missing/reconstructed fragments. Appendix A presents an original source text transcription along with the tokens extracted from it using this method.

From this set, I randomly created two development corpora of 5,000 tokens each (`dev1.corpus.txt` and `dev2.corpus.txt`), and one final test corpus of 10,000 tokens (`test.corpus.txt`). 59,649 remaining tokens from this corpus are saved as `unused.corpus.txt` if we need them, and the token counts of all 17,997 distinct types are saved in `vocab.counts.txt`. These counts can help prioritize work to maximize impact.

4 Survey of Phenomena in Akkadian

4.1 Nouns

4 cases (nominative, genitive, accusative, oblique), 2 genders (masculine, feminine), 3 numbers (singular, dual, plural). Inflections are all suffixes. In OB, the dual is restricted to body parts (eyes, feet, etc.). After OB, final m and n no longer appear.

4.2 Verbs.

Verbs in Akkadian exhibit a complex array of prefixes, suffixes and infixes across a number of different tenses (Preterative, Durative, Imperative and Perfect) and "stem types." (G, D, Š, N), each of which can be inflected in all tenses and which individually have different semantic interpretations (G stems verbs denote a basic, ground meaning [the *Grundstamm*]; D verbs carry a causative meaning, or "pluralistic" actions (Huehergard p. 256), actions which take place on a number of objects); Š verbs are also causative; N verbs denote the passive or middle voice). A simple example of the complexity involved in roots and patterns is the following perfect inflection of the root `šrq` ("steal"):

```
_  š _  r _ q _                  # Root (steal)
ta _ ta _ 0 _ ī                  # G Stem, Perfect, 2nd person feminine singular
```

In this case, the prefix `ta-` and the suffix `-ī` together form a 2nd person fem. sing. marker (specific to the perfect tense; analogues in this position pair are found in other tenses as well); the infix `-ta-` is a marker of the perfect tense. The thematic vowel usually found between radicals 2 and 3 is here deleted due to a dispreferred phonological sequence of two short vowels. This results in the form *taštarqī*, "you (sing. f.) stole."

5 Initial Design

The initial system design focused on 1.) the morphologically easier problem of modeling nouns, adjectives, names, pronouns, prepositions and adverbs and 2.) the problem of transforming a "normal" form (as would be seen in a textbook) to a transliterated one (as would be seen in an inscription).

Morphology. All morphological information on nouns, adjectives and names comes in the form of inflected suffixes for gender, number, case, posession and several enclitics. I modeled these by generating them all in the lexicon and running each noun type through several chains, each of which stack different information onto both the upper and lower side of the transducer (e.g., base form → person/number/case → optional possessive marker → optional enclitic `-ma` → optional enclitic

`-ta`. Rewrite rules that apply to all phonological sequences (e.g., a sequence of `a` and `u` contracting to `a`) are then performed in a subsequent stage.

To fill out the base forms, I extracted 1918 nouns and 235 adjectives from a custom Old Assyrian lexicon created by an Assyriologist; I extracted 625 names from a structured database of Old Assyrian letters (where the transliterated forms of the named entities are known); and I manually input 40 indeclinable prepositions and adverbs in consultation with my informant.

Transliteration. As observed in the data we have available, Akkadian words are written using a fixed vocabulary of cuneiform signs; if the sequence of letters in a word does not allow for any valid segmentation into signs, the word is emended (through the addition/deletion of vowels, duplication/deduplication of consonants) to fit the sign inventory. To capture this constraint, I extracted all unique sign patterns used in data (i.e., all sequences of letters delimited by the sign divider `-`) to create a vocabulary of 455 signs (e.g., `sa`, `si`, `su`, `sà`, `sá`, `sé`, `sí`, `sú`, `ta`, `tab`, `tal`, `tam`). I then required a valid word to be comprised entirely of these signs (in a manner analogous to syllabification):

```
define ValidWord [ Sign "-"]* Sign ;
```

To be sure to generate forms that can result in a valid word, I opted to overgeneralize a normalized form to a transliterated one, optionally duplicating or de-duplicating all consonants and vowels, eliding nasals `m` and `n` before other consonants, and optionally subjecting the word form to mimation (deletion of word-final `m`). This overgenerates forms but allows us to capture varability among different scribes (who may choose to engrave a word in many different ways).

Default. As a default guesser, all words ending in the suffixes `-um`, `-am` or `-im` are assigned the class `UNKNOWN_NOUN_ADJ`; all other words are assigned the class `UNKNOWN`. If a word is recognized as an instance of any morphological analysis, the `UNKNOWN` designation is dropped (i.e., the union betweeen analyses is a priority union on the lower [output] side, with valid morphological analyses ranking higher than unknown defaults).

6 System Analysis on Corpus A

Development corpus 1 contains 5,000 tokens and 2,437 types. The initial system generates a non-guessed morphological analysis for 3,038 of these tokens and 940 types, resulting in coverage (of at least one form) of 60.8% and 38.6%, respectively.

7 Lessons Learned and Revised Design

Lessons. One clear limitation of the initial design (beyond coverage for verbs) is the extremely limited lexicon for all parts of speech. To help mitigate this, I sorted the output of the first development test by the number of occurences of missing forms for each word type and prompted my informant for the lexical lemma from which it was derived; many of the most common missing forms were closed-class prepositions and adverbs. Given that this problem is likely to persist beyond the end of the project, it also forced me to reconsider the design to make the addition of lemmas by Assyriologists easy.

Verbs. In the revised design, I tackled the far more complex problem of verbal morphology. To model the complex combinations described above while also striving to create as simple an organization as possible (to faciliate lexicon expansion), I designed the bulk of verb operations not in the lexicon but rather in the subsequent transformational rules. Each verb lemma is represented in the lexicon as a base form paired with the sequence of root radicals, interradical positions (into which templatic morphological inflecitons can be inserted) and the thematic vowels used in the durative and preterite tenses (which must be specified for each verb). This information fully specifies a verb and allows us to generate a complete verb paradigm. For example, the verb šarāqum ("to steal") is represented as the following:

```
šarāqum: _ š _ r _ q _ i/i
```

By marking the positions of the root consonants and the spaces in between, we have the foundation on which to lay the affix patterns required by each specific verb type. To generate the 3rd person feminine plural perfect form of this verb, for example, requires the following transformations of the (often empty) elements at particular positions:

- Slot 1 → Prefix *i-*
- Slot 2 → Infix *-ta-*
- Slot 3 → $\emptyset$
- Slot 4 → Suffix *-ā*

Yielding the form *ištarqā*.

To fill out the verbal lexicon, I extracted 255 verbs (defined as base forms, root radicals and theme vowels) from the hand-created Old Assyrian lexicon described above.

8 System Analysis on Corpus B

Development corpus 2 contains 5,000 tokens and 2,344 types. The revised system, with an updated lexicon and verbal morphology, generates a non-guess morphological analysis for 3328 tokens and 1042 types, yielding coverage of 66.6% and 44.5%, respectively (approximately a 10% relative improvement at the token level and 15% relative improvement at the type level from the first system design).

9 Final Revisions

Lexical coverage is still clearly a problem for the revised system. To deal with this, I implemented a revised default guesser for verbal forms: since most verbs are comprised of a root with 3 radicals, it is straightforward to generate all possible combinations of three consonants, define a `GUESSED_VERB` as this random combination, and then subject it to the standard transformations that apply to known verbs. While this approach greatly improves coverage, it overgenerates to such a degree as to be impractical (nearly all lexical forms observed in the data can be seen as originating in one of these combinations.

Final revisions included manually analzing the errors made during the system analysis on corpus B and adding both new transformational rules and new lexical entries not in the earlier system. In a final evaluation on the test corpus of 10,000 tokens (corresponding to 4,101 types), the revised system generates a non-guessed analysis for 6,757 tokens (yielding a coverage of 67.6%) and 1,712 types (for a coverage of 41.7%). To estimate the practical impact of coverage, my informant manually analyzed the 50 most frequent tokens in the test data, which together comprise 34.2% (3419 tokens) of the overall dataset; of this subset, a correct morphological analysis was among those generated 93.6% of the time, which suggests when a morphological analysis is generated, it is likely to be correct; the long tail of work to come is in expanding the lexical coverage.

10 Future Work

As noted throughout, the most important element of future work is expanding the lexicon of root forms from which morphological analyzes can be derived. While the existing manually created lexical resources still offer some room for more elaborate mining, another important resource will be the Assyriologists who will be the end-users of this analyzer. Beyond this lexical work, we can add coverage for the remaining (less frequently encountered) verb stem types (Gt, Dt, etc.) and revise existing rules to ensure adequate treatment important exception classes, such as weak roots.

A Data

Example Old Assyrian text transcription (for text CCT 1, 11b). Brackets (e.g., "[a-ší]" in line 8) denote text that is missing or illegible in the original cuneiform tablet but has been reconstructed by scholars in the process of transcription. Notes are scattered throughout (e.g., *obv.* denotes the text of the front side [obverse] of a tablet, *rev.* the text on the back [reverse]).

```
CCT 1, 11b
#3 BM 113574a
#6 EL 24;
        15-16: Larsen, OACP 31
#10 - OIP XXVII 59, 22-30 ##


obv.!        (Siegelabrollung B)
1        KIŠIB DINGIR-ma-lá-ak DUMU Sú-en6-SIPA
        KIŠIB Bé-lá-nim DUMU Šu-Ku-bi-im
        KIŠIB E-ni-ba-áš DUMU A-šùr-DU10
         (Siegelabrollung C)
        [KIŠIB] Wa!-wa!-lá
lo.e.         (Siegelabrollung C)
rev.!        (2 Stempel C)
5        [1/3 ma]-na KÙ.BABBAR ṣa-ru-pá-am
        [i-ṣé-e]r DINGIR-ma-lá-ak
        [DUMU Sú-en6]-SIPA ù Wa-wa-lá
        [a-ší]-tí-šu d.En-líl-ba-ni i-šu
        [i]š-tù ha-muš-tim ša kà-ší-im
         (Siegelabrollung A)
10        ša qá-tí E-na-nim <ITU>.KAM
        a-lá-na-tim li-mu-um
u.e.         (Siegelabrollung B)
        A-gu5-tum 1/2 GÍN.TA i-ITU.KAM
        ṣí-ib-tám ú-ṣú-bu KÙ.BABBAR
        i-qá-qá-ad šal-mì-šu-nu ra-ki-is
le.e.
15        qá-té d.En-líl-ba-[ni]
         (Siegelabrollung B)
        Wa-wa-lá [ú-kà-al]
r.e.         (Siegelabrollung A)
@END_FILE
```

Tokens extracted from this text by the method described in section 3 are:

```
DINGIR-ma-lá-ak Sú-en6-SIPA
Bé-lá-nim Šu-Ku-bi-im
E-ni-ba-áš A-šùr-DU10
Wa!-wa!-lá
ṣa-ru-pá-am
DINGIR-ma-lá-ak
Wa-wa-lá
d.En-líl-ba-ni i-šu
ha-muš-tim kà-ší-im
qá-tí E-na-nim
a-lá-na-tim li-mu-um
A-gu5-tum i-ITU.KAM
ṣí-ib-tám ú-ṣú-bu
i-qá-qá-ad šal-mì-šu-nu ra-ki-is
qá-té
Wa-wa-lá
```

References

François Barthélemy. A morphological analyzer for akkadian verbal forms with a model of phonetic transformations. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, Semitic '98, pages 73–81, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1621753.1621766.

François Barthélemy. Une description morphologique structurée en arbre du verbe akkadien qui utilise des structures de traits et des transducteurs multirubans, 2009.

Richard Caplice. *Introduction to Akkadian.* Editrice Pontifico Istituto Biblico, Roma, 2002.

John Huehnergard. *A Grammar of Akkadian.* Eisenbrauns, Winona Lake, Indiana, 2005.

Laura Kataja and Kimmo Koskenniemi. Finite-state description of semitic morphology: a case study of Ancient Akkadian. In *Proceedings of the 12th conference on Computational linguistics - Volume 1*, COLING '88, pages 313–315, Stroudsburg, PA, USA, 1988. Association for Computational Linguistics. ISBN 963 8431 56 3. doi: 10.3115/991635.991699. URL http://dx.doi.org/10.3115/991635.991699.

Aaron Macks. Parsing Akkadian verbs with prolog. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, SEMITIC '02, pages 1–6, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118637.1118638. URL http://dx.doi.org/10.3115/1118637.1118638.