Akkadian Treebank for early Neo-Assyrian Royal Inscriptions

Mikko Luukko
University of Helsinki
mikko.luukko@helsinki.fi

Aleksi Sahala
University of Helsinki
aleksi.sahala@helsinki.fi

Sam Hardwick
University of Helsinki
sam.hardwick@iki.fi

Krister Lindén
University of Helsinki
krister.linden@helsinki.fi

Abstract

This paper presents the first proper syntactic treebank for Akkadian, an ancient Semitic language which can only be reconstructed from its textual data. We introduce our corpus of early Neo-Assyrian royal inscriptions, present some typical syntactic constructions of this genre and discuss the morphological and syntactic choices we have made. For developing a gold standard for morphological annotations, we tested the manually annotated material against BabyFST, a morphological analyzer of Akkadian. We also tested the reproducibility of the syntactic annotations using the TurkuNLP neural parser.

1 Introduction

This first version of our Akkadian treebank consists of 22 277 words and 1845 sentences. This represents an intact subset of a total of 2211 sentences from the early Neo-Assyrian royal inscriptions[1] of the tenth and ninth centuries BCE. Because of the progressive complexity of linguistic constructions in later texts of our source material,[2] our approach is chronological and we begin with the inscriptions of Aššur-dān II (r. 934–912), published in Grayson (1991).[3] The main sub-corpus of the volume and our first version are thus the inscriptions of Ashurnasirpal II.[4] The language of the corpus is Standard Babylonian,[5] with occasional Assyrianisms,[6] whereas "Akkadian" is the umbrella term for both Assyrian and Babylonian. In the modern world, Akkadian is not as well-known as Latin, Greek, Hebrew or Egyptian languages. In the ancient world, however, Akkadian was an important cultural language with a long history of more than two-thousand-and-five-hundred years as a spoken and written language. The name of the language comes from the capital of the legendary third-millennium King Sargon of Agade or Akkade.

These royal inscriptions were extracted from Oracc (Open Richly Annotated Cuneiform Corpus),[7] where all Neo-Assyrian royal inscriptions[8] are lemmatized word-for-word. More precisely, we have made use of the bound transcription (= normalized text) of the lemmatized texts which had been previously transliterated from clay tablets. The transliteration of Akkadian is based on both syllabically and pictographically, though abstracted, written cuneiform signs. Therefore, for this Akkadian treebank, we have the advantage that we do not have to consult the original tablets or take into account the subtleties of the cuneiform script. Perforce, because of the cuneiform script (writing system), the analysis of Akkadian syntax contains more speculative interpretation than with a modern language. The factor

[1] Sometimes referred to as ARI (Assyrian Royal Inscriptions).
[2] This is a simplification, but a number of royal inscriptions from the eighth and seventh centuries BCE are syntactically much more complicated; consider, e.g., Sargon II's famous Eighth Campaign.
[3] Grayson (1991), also known as RIMA 2 (references to the volume so below; Q-numbers, also below, refer to Oracc text IDs), contains the inscriptions of Tiglath-pileser I and his successors until Tiglath-pileser II, too, but these kings are usually considered Middle Assyrian.
[4] Grayson, 1991: 189–397.
[5] Standard Babylonian is a literary variant of Babylonian dialect (for its use in Neo-Assyrian royal inscriptions, see Frahm, 2019: 144–145); it was never a spoken language.
[6] Assyrianisms in this corpus were already discussed by Deller, 1957a and b.
[7] <http://oracc.museum.upenn.edu/>. By making editions of thousands of cuneiform texts available online for everyone with an Internet connection, Oracc has laid the foundation for Digital Assyriology.
[8] <http://oracc.museum.upenn.edu/riao/>.

contributing to this is the combination of word signs – usually called logograms or Sumerograms – and syllabic signs (syllabograms). One example from each spelling category for the three main parts-of-speech will suffice here:

- Nouns:
  - syllabically written *ma-da-tu* stands for *maddattu* "tribute";
  - logographically written LUGAL stands for *šarru* "king";
  - the combination of AN-*e* stands for *šamê* "heaven".
- Verbs:
  - The normalization of *at-tu-muš* is *attumuš* and it means "I set out";
  - GUR, *utēr* "It turned into (something)";
  - KUR-*ud* stands for *akšud* and means "I conquered".
- Adjectives:
  - *dan-nu-te*, *dannūte* "strong" (masculine plural from *dannu*);
  - DUGUD, *kabta* "heavy" (in the accusative, from *kabtu*);
  - GAL-*te*, *rabīte* "great" (singular feminine in the genitive, from *rabû*).

The distribution of different types of spellings and their combinations in this corpus are provided in Table 1.

Full corpus	**Syllabic**	**Logographic**	**Logo-Syllabic**
Nouns	3739	4179	1837
Verbs	2581	34	234
Adjectives	671	243	194
Other	7030	1570	633

Table 1: Different types of spellings and their combinations in the corpus of early Neo-Assyrian royal inscriptions.

Compared with the cognate Semitic languages, for example, we are in a lucky position and rarely confront a problem of vocalic ambiguity, which in other Semitic languages results from uncertain vocalization that is not marked in the original documents.[9] Moreover, unlike in other Semitic languages, Akkadian dictionaries are based on words and not on roots.

1.1 Basic Characteristics of Akkadian

Akkadian is an extinct Semitic language that has not been spoken anywhere since the first century of the Common Era. It is cognate to ancient and modern languages such as Arabic, Aramaic, Hebrew, Amharic and Maltese.[10] Akkadian, written in cuneiform script, displays several distinctive features. For example, texts do not include punctuation,[11] and the language does not express definiteness by using definite or indefinite articles, but "definiteness" must always be read from the context. As is typical of Semitic languages, Akkadian has a rich (and complex) morphology.

Thanks to the durability of cuneiform tablets written on clay, Akkadian is with its hundreds of thousands of texts a well-known language,[12] but already for decades, it has been a desideratum to enhance

[9] Partially the problem relates to the wide use of different writing systems among Semitic languages; cf., e.g., Zitouni, 2014: 35. In this context, we are not concerned with the correct interpretation of the syllabic C(onsonant)V(ocal)C(onsonant) signs whose reading values do give Assyriologists some trouble.

[10] One can find treebanks of these languages at Universal Dependencies (<https://universaldependencies.org/>).

[11] Akkadian texts rarely make use of a word-divider or any other device that belongs to the area of punctuation. However, especially literary texts may occasionally leave gaps between words but, as a rule, the original texts do not delimit word boundaries by "whitespace" characters.

[12] On the size of the Akkadian text corpus, see Streck, 2010.

our understanding of its syntax. For the most part, Akkadian word order follows the S(ubject)O(bject)V(erb) structure, which probably resulted from the direct influence of the non-Semitic Sumerian language already in the third millennium BCE at the latest.[13] However, while the main tendencies of Akkadian word order are easy to sketch out, in many instances the order is relatively free, although this may signify different semantic nuances in texts. Thus, depending on the types of sentences, the "standard" word order is not always strictly followed, but there are few studies on the significance of this phenomenon. For this reason, an Akkadian treebank will enable us to study Akkadian syntax from a new and much deeper perspective.[14]

2 Current Data Set

This corpus consists of 162 royal inscriptions of four early Neo-Assyrian kings: Aššur-dan II (r. 934–912 BCE), Adad-nerari II (r. 911–891 BCE), Tukulti-Ninurta II (r. 890–884 BCE) and Ashurnasirpal II (r. 883–859 BCE). Neo-Assyrian Royal Inscriptions are rather idiosyncratic commemorative texts, which serve to self-aggrandize Assyrian kings, and distinguish themselves from the other genres of Akkadian literature. These texts often begin with a long introduction, having the king's name (usually with genealogy) or divine invocation, lengthy royal or divine titles and epithets that stress the king's bravery.[15] These epithets given to Neo-Assyrian kings or gods are usually nouns or adjectives or the combinations of the two. For example, King Adad-nerari II says that he is

(1) *hitmuṭ raggi u ṣēni*
burning wicked (person) and evil (one)
"inflamed against the evil and wicked"
RIMA 2 A.0.99.2: 17 (Q006021) and 4:4′–5′ (Q006023).

This is annotated as three nouns and a conjunction, though the latter two nouns are formally adjectives and the first one is a stative or an infinitive. The sentences are rarely complex in an introduction but mainly lengthy nominal clauses, though they may occasionally show changes in word order. After the introduction, there is usually a section on military campaigns and then a separate section on building or renovation projects. Royal inscriptions mostly close with a section on blessings for pious future rulers who will take care of their predecessor's commemorative text. If a future ruler does not respect his predecessor's wishes, curses will befall him. The long narrative texts, with list-like conquests and itemized records of received tribute, are in sharp contrast to the brief labels and epigraphs that were originally attached to objects (especially many among the inscriptions of Ashurnasirpal II). The latter type of documents numerically form a large minority of the inscriptions in this corpus.

As mentioned previously, we annotate the Neo-Assyrian Royal Inscriptions published in Grayson (1991). Since punctuation is not used in Akkadian, we arrive at sentences by syntactically annotating the unsegmented corpus, and identifying words that are head words but are not themselves dependents of other words. The corpus also contains unidentified and partly identified words, and for this reason some sentences are sentence fragments, or contains unannotated material. We excluded them from the current version of our treebank, which thereby comprises 1845 sentences with 22 277 words.

There are a total of 3398 distinct phonologically transcribed word forms in the corpus. A majority of these, 3223, have only a single analysis in terms of lemma and morphology across the corpus, with the remaining 175 receiving different analyses in different contexts. The ambiguous forms represent 4767 tokens out of a total of 22 277 tokens in the corpus, i.e. 21%, meaning that the remaining 79% of the corpus consist of tokens that have only one analysis in this corpus.

3 Morpho-syntactic Analysis

As a first step, we have manually annotated each token in the corpus with a lemma and a part-of-speech (POS) as well as a morphological analysis, i.e. during the manual POS tagging, we separated the morphemes and annotated them with morphological features and syntactic relations. By far the largest group

[13] Edzard, 2003: 174 and Huehnergard and Woods, 2008: 128.
[14] A preliminary treebank for Akkadian with Babylonian Royal Inscriptions of the seventh and sixth centuries BCE called PISANDUB prepared by Kamil Kopacewicz can be found at http://universaldependencies.org/ containing 101 sentences with 1852 tokens. At the time of writing, it only contained POS tags and syntactic relations and no language documentation.
[15] On the structure of Neo-Assyrian royal inscriptions, see, e.g., Frahm, 2019: 146, 149.

of bound morphemes attached to nouns, verbs or prepositions is formed by suffixes which syntactically have different functions depending on their head. In annotating nouns and other parts-of-speech, we closely follow the terminology explained and listed in Reiner (1966: 57, 137). In the morphological analysis and POS tagging, our goal is to provide as much information as is evidenced by the morphemes in context. We annotate the following subcategories of verbs:

- finiteness (finite, infinitive, stative),
- stem (G, D, Š, N etc.),
- mood (indicative, imperative, precative, prohibitive),
- tense (present, preterite, perfect), person (1, 2, 3),
- number (singular, plural) and
- gender (masculine, feminine).

Following Streck (2011: 363), we consider subordinative[16] and ventive as subcategories of their own, which we tag as boolean values. For nouns, adjectives and non-finite verbal forms the subcategories are:

- case (nominative, accusative, genitive),
- number (as above),
- gender (as above) and
- base, which can have four different values:
    - free (status rectus),
    - bound (status constructus),
    - suffixal (followed by pronominal suffixes) and
    - terminal (status absolutus).

In general, our approach to POS tagging and to the syntactic dependency relations of each word follows as closely as possible the standards created, developed and maintained by the Universal Dependencies (henceforth UD) project; these principles are elaborated on the UD website.[17] For visualizing the syntactic analysis, we use a CONLL-U viewer, a tool available on the UD website.

In this corpus, from the seventeen Universal POS tags listed on the UD website, we have used all except auxiliary (AUX; Akkadian does not have genuine auxiliaries), interjection (INTJ) and symbol (SYM). Perhaps more surprisingly, we cannot use the label punctuation (PUNCT), because cuneiform inscriptions are continuous texts without punctuation. E.g. the end of a sentence is explicitly indicated only in exceptional circumstances.[18]

As to proper nouns and ethnic names, which are often called *nisbe* in Akkadian, their morphological annotation has been simplified and does not contain as many labels as regular nouns. This is due to the fact that they were written without inflections. Thus, proper nouns are simply annotated as PROPN + gender (if a personal or a divine name) and ethnic names are labelled as NOUN + gender. The latter are not always true proper names, because they can also refer to any single person of a tribe, although often this principle is reserved for the ruler of a tribe or a town.

According to the UD principles, participles are to be annotated as verbs or adjectives, but the so-called active participles in Akkadian cannot follow this principle, since the active participles in Akkadian act as the performers of action (cf. Arabic), so they are annotated as nouns. By observing UD, we also annotate all day dates as adjectives.

The construct state of Akkadian concerns the relation between two content words (cf. Arabic *Idafa*[19] and Hebrew *smikhut*).[20] This syntactic relation between the construct state noun (possessed/governing noun) and the following noun in the genitive (possessor/governed noun) is expressed with the label nmod:poss.

The frequent determinative pronoun *ša* "of" is annotated as ADP (= preposition) in the same way as is done with "of" in English. Another frequent word *u* "and, but" also has an adverbial meaning "further(more), moreover" at the beginning of a sentence. For example,

[16] We prefer this term instead of subjunctive following von Soden (1995: 135) and Streck (2011).
[17] <http://universaldependencies.org/>.
[18] One of the few exceptions is a section ruling, i.e. a horizontal divider in an inscription, that clearly indicates the end of a sentence, section or paragraph.
[19] Cf. Zitouni, 2014: 19.
[20] On the construct state in Akkadian, see, e.g., Huehnergard, 2005: 56.

(2) *u rapšāte mātāt Nairi ana pāṭ gimriša apēl*
and broad lands Nairi to border totality-its ruled
"Moreover, I gained entire dominion over the extensive lands of Nairi" Ashurnasirpal II (passim).

Akkadian does not have definite or indefinite articles. For those familiar with the Oracc lemmatization,[21] where many determiners appear under the label XP, standing for indefinite pronouns, the UD annotation is notably different. PRON is another label for which it is appropriate to point out the difference between the UD principle and the Oracc lemmatization. In the latter, e.g., Akkadian indefinite pronouns *mamma* "somebody, anybody, (negated) nobody" and *mimmu* "something, anything, everything, (negated) nothing" bear the XP label.

3.1 BabyFST

BabyFST is a finite-state based morphological model for Babylonian, a southern dialect of the Akkadian language (Sahala et al., 2020). The model is capable of providing morphological analysis for different stages of the Babylonian dialect, including Standard Babylonian and some of its typical Assyrianisms. The model is implemented in the LEXC and XFST formalisms, which can be compiled into finite-state transducers by using compilers such as Foma (Hulden, 2009) and HFST (Lindén et al., 2009). BabyFST tags Akkadian word tokens with their morphological features (number, gender, case, construct state, mood, tense, person, verbal affixation and verbal stems including *-t-* and *-tan-* infixation) as well as lemma and part-of-speech.

We verified and normalized the manually produced morphological annotations by using BabyFST to ensure that the human-produced annotations were consistent and formally in line with BabyFST's output, which for the tokens is true for 94.6% of the lemmas, and 85.5% of the morphological analysis. The seemingly low score on morphological annotation is due to underspecification in the manual annotation, local variation in the gender of a few frequent nouns and local spelling variants and Assyrianisms, i.e. Babylonian words with Assyrian influences. We can compare Akkadian writing standards to current writing conventions in social media discussion forums.

Used as a morphological gold standard, the treebank contains fully-specified morphological analyses for 3012 nouns, 2053 verbs and 555 adjectives. The analyses are underspecified for 5317 nouns, 136 verbs and 338 adjectives, often because a word is written using a logogram and the inflected form is not explicitly indicated. Underspecification may also occur in the construct state, where the case endings are not marked. In such instances, one or more subcategories are marked as undefined.

The morphological annotation in the treebank will allow using the current annotation as a gold standard for morphological analysis, e.g. using BabyFST with disambiguation or using neural networks to predict both lemma and annotation in context.

4 Syntax

4.1 Language-specific Remarks

Traditionally, the study of Akkadian grammar has been dominated by morphological and lexical studies, and syntactic studies have been more peripheral. Standard Akkadian grammars, such as von Soden 1995, have been the mainstay of Akkadian syntax and monographs on the topic are still rare. However, relatively recently there has been a clear increase in the large-scale syntactic studies of Akkadian (e.g. Deutscher, 2000 and Cohen, 2012), although mainly syntactic studies are published in articles.

UD lists thirty-seven different syntactic relations. We have used twenty-five of them in our annotation; the following relations have so far not been applied partly due to the nature of the text genre: aux (auxiliary); clf (classifier);[22] compound; cop (copula); csubj (clausal subject); dislocated; expl (expletive); fixed; flat; orphan; punct (punctuation); reparandum.

[21] <http://oracc.museum.upenn.edu/doc/help/languages/akkadian/index.html>.
[22] The original texts include determinatives, but they were omitted, when the bound transcription of a text was prepared.

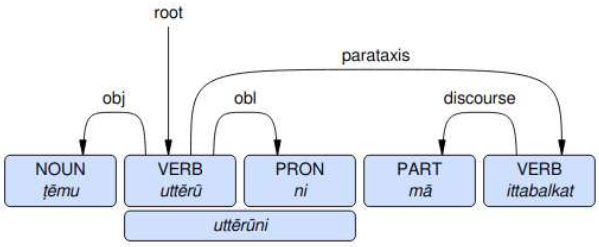
We have used the following UD relations: acl (adjectival clause); advcl (adverbial clause modifier); advmod (adverbial modifier); amod (adjectival modifier); appos (appositional modifier); case; cc (coordinating conjunction); ccomp (clausal complement); conj (conjunct); dep (unspecified dependency); det (determiner); discourse; goeswith; iobj; list; mark (marker); nmod (nominal modifier); nsubj (nominal subject); nummod (numeric modifier); obj (object); obl (oblique nominal); parataxis; root; vocative;[23] xcomp (open clausal complement).

The used relations, discourse, goeswith, list and vocative are rare in this corpus: For the relation goeswith, which mends erroneously split words, we have only one attested case in which a single concept made out of the negation *lā* and the following noun *salīma* has been split over two separate lines in the original: *lā salīma* not peace "truceless" RIMA 2 A.0.101.17 V 101–102 (Q004471).

In a way, early Neo-Assyrian royal inscriptions contain many different types of "lists" (e.g., of conquered cities or received tribute from foreign rulers or of various dishes offered at a special inaugural banquet for a new palace or of exotic plants, trees and animals, etc.). Nevertheless, for the most part we have chosen to tag the items enumerated in such lists with the conj relation.

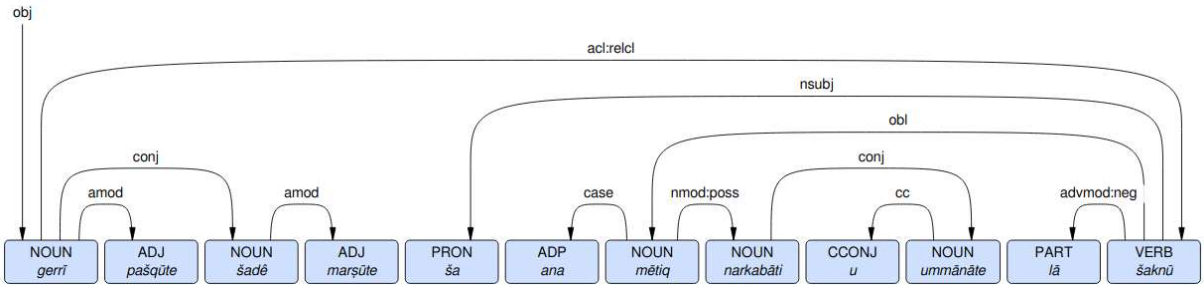
The only exceptional case which we tag as discourse is the use of *mā*[24] to indicate a direct speech quotation, a rare phenomenon in this genre, in a way it equals a colon:

(3) *ṭēmu* *uttērūni* *mā …* *ittabalkat*
report returned-me saying … crossed over
"A report was brought back to me: 'It (= a city) … has rebelled'"
RIMA 2 A.0.101.1 I 75 (Q004455).



The following relation subtypes have been used: acl:relcl for relative clauses, advmod:emph,[25] advmod:neg for the negation particles *lā* and *ul*, det:poss for possessive determiners and nmod:poss for the construct state. They are all frequent in Akkadian.

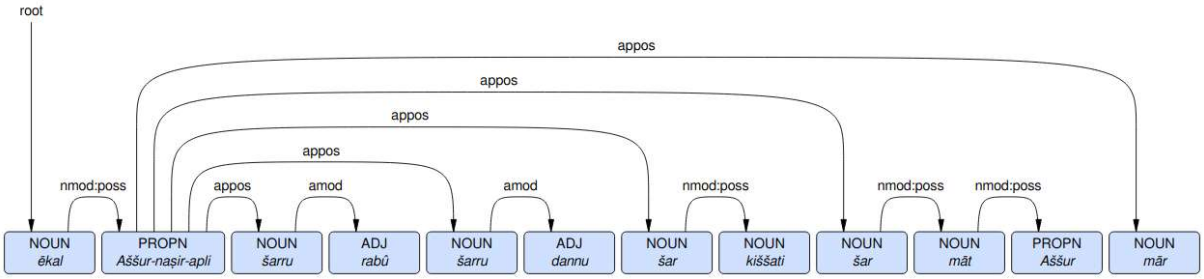
(4) *gerrī pašqūte šadê* *marṣūte ša* *ana mētiq narkabāti u* *ummānāte lā* *šaknū*
ways narrow mountains difficult which for route chariots and troops not put
"Difficult paths (and) rugged mountains which were unsuitable for chariotry and troops"
RIMA 2 A.0.101.17 I 65–66 (Q004471).



For example, a typical, brief label in this corpus does not include a verbal clause, but it enumerates the ruler (owner) and his immediate ancestors, and begins like this with several nmod:poss cases:

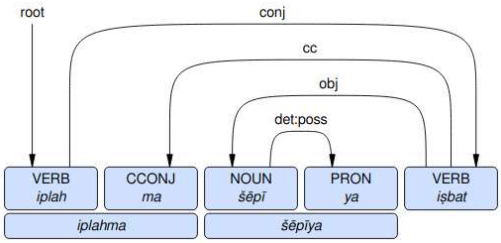
[23] For the only example tagged as vocative, see Adad-nerari II, RIMA 2 A.0.99.2: 77–78 (Q006021-5), in which the king addresses himself in public in front of his magnates in the third person.
[24] For example, Deutscher (2000: 66–91) calls the related Babylonian umma "a quotative marker".
[25] This subtype concerns the particle *lū* (or *lu*) in its asseverative function (Kouwenberg, 2017: 640–43), although no strict attempt has been made here to keep it distinct from the precative *lū*. Hence most of the cases in which *lū* is separate from the verb has got the advmod:emph relation.

(5) *ēkal Ashurnasirpal šarru rabû šarru dannu šar kiššati šar māt Aššur mār …*
palace Ashurnasirpal king great king strong king totality king land Aššur son ...
"(Property of the) palace of Ashurnasirpal, great king, strong king, king of the universe, king of Assyria, son of … (followed by a short genealogy of the king's father and grand-father)"
RIMA 2 A.0.101.102 (Q004556 and passim).



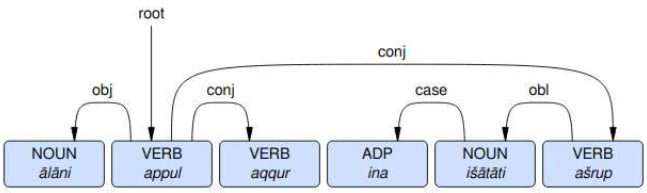
In the contemporary Neo-Assyrian letters the enclitic *-ma* particle has become obsolete in coordinating verbal clauses (Hämeen-Anttila, 2000: 66, 122; Luukko, 2004: 108), and in this corpus its use is also clearly on the decline, though several verbs still take *-ma*. For example,

(6) *iplah-ma … šēpīya iṣbat*
be(come) afraid-*ma* feet-my seized
"He took fright and submitted to me."
Ashurnasirpal II, RIMA 2 A.0.101.1 III 73 (Q004455).



However, verbal clauses are mainly coordinated asyndetically:

(7) *ālāni appul aqqur ina išātāti ašrup*
cities demolished destroyed in fires burnt
"I razed, destroyed, (and) burnt the cities." Passim



Along similar lines, in nominal clauses the phrases with the conjunctive *u* "and", such as *biltu u maddattu* "tribute and tax", and its equivalent *biltu maddattu* "tribute (and) tax" without a coordination conjunctive appear in free variation with one another.[26]

In this corpus, written in Standard Babylonian, the verbal subordinative is either the Babylonian *-u* or the Assyrian *-(ū…)ni*.[27] We label the relation of the Assyrian subordinative marker *-ni* with its main

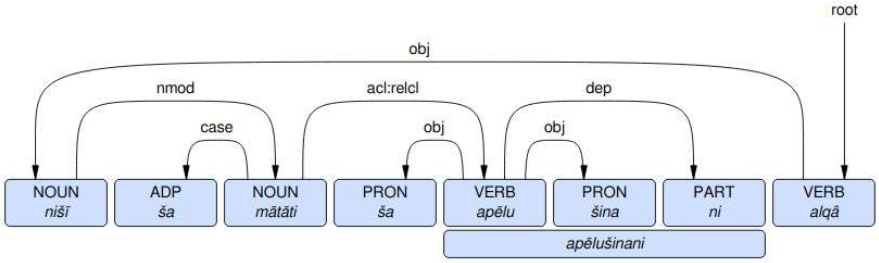
[26] Both these variants occur even on the same line in RIMA 2 A.0.99.2: 115 (Q006021). In the print edition of the two longest texts of the corpus, RIMA 2 A.0.101.1 and 17 (Q004455 and Q004471), there are altogether, i.e., both in nominal and verbal clauses, 389 and 209 restored "(and)" cases respectively!

[27] Some Assyrian examples in this corpus were already given in Deller, 1957a: 153–54 and id. 1957b: 272. For the use of the term subordinative instead of the subjunctive, see (also above) now Bjøru and Pat-El (2020: 71, n. 1) and already von Soden (1973).

word as dep. When *-ma*, which is attached to a verb, appears in clause-final position, we tag its relation similarly to the verb with dep.

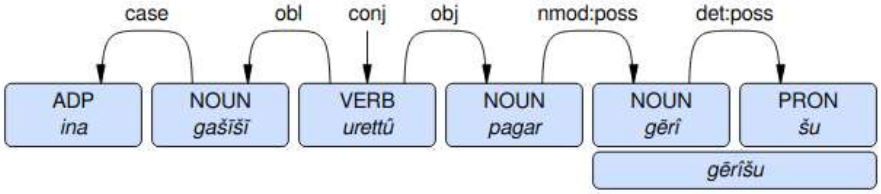
Subordinate clauses precede the main clause but relative clauses, introduced by *ša*, immediately follow the main word (in a main clause) which they qualify:

(8) *nišī ... ša mātāti ša apēlušinani ... alqâ*
people of lands which ruled (over)-them took
"I took people … from the lands over which I had gained dominion."
Ashurnasirpal II, e.g., in RIMA 2 A.0.101.2: 53–55 (Q004456) and 23: 15–17 (Q004477).



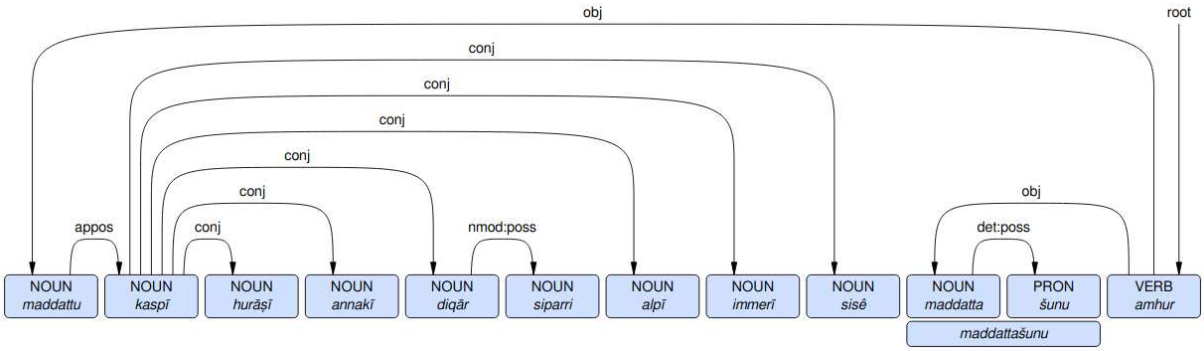
Occasionally, unlike in the standard word order, an object may follow the main predicate:

(9) *ina gašīšī urettû pagar gērî šu*
on stakes installed body enemy-his
"He hung the corpses of his enemies on stakes."
Ashurnasirpal II, RIMA 2 A.0.101.1 I 29 (Q004455)



If there are several tribute bearers and many items, then the main object (tribute) may be repeated before the predicate (unlike here, it is usually left untranslated in editions):

(10) *maddattu ... kaspī hurāṣī annakī diqār siparri alpī immerī sisê maddatta*
tribute silver gold tin large bowl bronze oxen sheep horses tribute
šunu amḫur
their received
"I received the tribute … silver, gold, tin, bronze casseroles, oxen, sheep, (and) horses, their tribute" Ashurnasirpal II, RIMA 2 A.0.101.1 II 21–23 (Q004455).



4.2 Parser Experiment

The TurkuNLP neural parser (Kanerva et al. 2018) is a processing pipeline for segmentation, morphological analysis, dependency parsing and lemmatization. Each of these tasks is implemented by separate neural models, and when combined, the parser is able to produce fully annotated CoNLL-U files from

raw text. It was overall a top-ranked parser in the CoNLL-18 shared task for multilingual parsing from raw text to universal dependencies.

The parser is provided with two human-annotated CoNLL-U files: a training set, which is used for adjusting the neural weights, and a development set which is used for observing the performance of the parser during training. In addition, to evaluate its final performance, a test set, unused during training, is annotated by the trained parser.

Manual syntactic annotation of the corpus had resulted in unsegmented running text with dependency markings. We used this annotation to automatically split the texts into sentences. In a perfect situation, this should have required nothing more than allocating each dependency tree into its own sentence resulting in a segmentation of the entire text (i.e. all the tokens). However, parts of the text that were possible to transcribe only in part, or not at all, resulted in incomplete tree structures.

We first attempted to segment the 162-text corpus in its entirety, allocating unidentified or partly labelled tokens to nearby sentences, and use this data to train the parser. The parser both received training data and outputted parsing results that contained tokens with blank fields for lemmas, morphology and syntax. The result of this experiment in terms of numeric scores was, however, disappointing, the system not being designed with this sentence fragment scenario in mind.

We then produced a set of sentences which did not have structural problems resulting from unidentified or partly labelled tokens. These numbered 1845 out of a total of 2211 possible trees. Here, "possible trees" means tokens that could be syntactic roots, i.e. they have dependents but do not depend on other tokens, and are in effect an upper bound. These sentences were randomly shuffled and split into the previously mentioned training (80% of sentences), development (10%) and test sets (10%). We deemed shuffling to be preferable to assigning sentences in running order, as the corpus is rather heterogeneous, a few long texts dominating the word count.

On the test set, we tested both the case where we provided segmentation cues, which in most other treebanks are present in the form of punctuation or formatting, and the case where all the shuffled sentences occurred as a consecutive string of words. In the latter case, the parser infers sentence and token boundaries. Errors in these tasks contribute to lower scores in the parsing task. We calculated the scores with the CoNLL 2018 shared task evaluation script (SIGNLL 2018).

For the segmented case, we obtained a LAS (labeled attachment) score F1 of 93.29, an MLAS (morphology-aware attachment) score of 87.53 and a BLEX (bi-lexical dependency) score of 91.71. These are the main metrics used in the CoNLL 2018 shared task. LAS is a reflection of how well the dependency relations (arc and label) matched between the parser's output and the gold standard; MLAS includes the requirement that the morphological analysis is also matching; BLEX the requirement that the lemmatization matches. These results are surprisingly good relative to the automatic parsing of most languages, and probably reflects the rather repetitive nature of this corpus, and of course the segmentation provided by us.

When no segmenting cues were provided, we obtained a LAS of 69.95, an MLAS of 58.97 and a BLEX of 62.44. This is on par with that obtained in the "small treebanks" subtask in CoNLL 2018.

5 Discussion and Conclusion

The fragmentary state of cuneiform texts is a frequent problem and it concerns this sub-corpus of Assyrian royal inscriptions too. In Assyriology, indiscernible words in the transcription are indicated with an *x* in the transcription. Sometimes our standard text editions exacerbate the problem by providing too few (or too many) *x*s, making restorations and the syntactical annotation of a text difficult or even impossible in many cases when the *x*s distort the syntactic flow of the text if the number of *x*s given in transliterations or bound transcriptions does not correspond to the situation on the original text carrier. The issue is aggravated in the current text genre which consists of many relatively large artefacts; the shorter and the more standardized the texts are, the easier it is to restore and assign the length of the gaps relatively reliably.[28]

As to restorations in general, we have adhered to the suggestions given in Grayson (1991) to the extent that restorations now appear without brackets, which is the usual way to indicate broken passages in Akkadian texts. Methodologically, this will probably not do much harm when studying Akkadian

[28] On the challenges of preparing a treebank of a language originally written in the cuneiform script according to the UD model, see Inglese (2015).

syntax, but in text research that may delve deeper into the details of a passage, some of the restorations could be questioned.

We have briefly described the Akkadian language, with some of its characteristics, and defined our corpus of early Neo-Assyrian royal inscriptions for building a treebank, the first proper treebank for Akkadian (comprising Assyrian and Babylonian). The manual annotation process is thus far the work of a single expert annotator (Mikko Luukko), who first used the Brat rapid annotation tool but later switched to WebAnno. To achieve a consistent morphological gold standard, the morphological annotation was checked against BabyFST, a morphological analyzer. The syntactic annotation consistency has been tested with the TurkuNLP parser.

Our first treebank will be released under the Universal Dependencies scheme with 1845 out of a total of 2211 possible sentences. When testing a parser on the pre-segmented sentences, we obtained a LAS score F1 of 93.29, an MLAS score of 87.53. When no segmenting cues were provided, we obtained LAS 69.95 and MLAS 58.97, which is on par with that obtained in the "small treebanks" subtask in CoNLL 2018. In the near future, our main challenge is to generalize the annotation to new material from other text genres.

Acknowledgements

The research for this article was carried out as part of the Centre of Excellence in Ancient Near Eastern Empires (ANEE) in cooperation with FIN-CLARIN and the Language Bank of Finland, taking place in Helsinki, and funded by the Academy of Finland. We would like to thank Niek Veldhuis (Berkeley), the initiator of the project, and David Bamman (Berkeley), who helped in setting up the first attempt for annotation on the Brat rapid annotation tool. For making this project possible, we are also indebted to Karen Radner, Jamie Novotny and Nathan Morello (all three LMU, Munich) and to Grant Frame and Steve Tinney (both UPenn, Philadelphia).

References

Øyvind Bjøru and Na'ama Pat-El. 2020. The Historical Syntax of the Subordinative Morphemes in Assyrian Akkadian. *Zeitschrift für Assyriologie und Vorderasiatische Archäologie*, 110(1):71–83.

Eran Cohen. 2012. *Conditional Structures in Mesopotamian Old Babylonian* (Languages of the Ancient Near East, 4). Eisenbrauns, Winona Lake, IN.

Karlheinz Deller. 1957a. Zur sprachlichen Einordnung der Inschriften Aššurnaṣirpals II. (883–859). *Orientalia Nova Series* 26(2):144–156.

Karlheinz Deller. 1957b. Assyrisches Sprachgut bei Tukulti-Ninurta II (888–884). *Orientalia Nova Series* 26(3):268–272.

Guy Deutscher. 2000. *Syntactic Change in Akkadian: The Evolution of Sentential Complementation*. Oxford University Press, Oxford and New York.

Dietz Otto Edzard. 2003. *Sumerian Grammar* (Handbuch der Orientalistik: Der Nahe und der Mittlere Osten, 71). Brill, Leiden and Boston.

Eckart Frahm. 2019. The Neo-Assyrian Royal Inscriptions as Text: History, Ideology, and Intertextuality. In *Writing Neo-Assyrian History: Sources, Problems, and Approaches* (State Archives of Assyria Studies, 29), edited by Giovanni B. Lanfranchi, Raija Mattila, and Robert Rollinger, 139–159. The Neo-Assyrian Text Corpus Project, Helsinki.

A. Kirk Grayson. 1991. *Assyrian Rulers of the Early First Millennium B.C. I (1114 – 859 B.C.)* (Royal Inscriptions of Mesopotamia, Assyrian Periods, 2). University of Toronto Press, Toronto.

Jaakko Hämeen-Anttila. 2000. *A Sketch of Neo-Assyrian Grammar* (State Archives of Assyria Studies, 13). The Neo-Assyrian Text Corpus Project, Helsinki.

John Huehnergard. [2]2005. *A Grammar of Akkadian* (Harvard Semitic Museum Studies, 45). Eisenbrauns, Winona Lake, IN.

John Huehnergard and Chris Woods. 2008. Akkadian and Eblaite, In *The Ancient Languages of Mesopotamia, Egypt and Aksum*, edited by Roger D. Woodard. 83–153. Cambridge University Press, Cambridge.

Mans Hulden. 2009. Foma: A Finite-State Compiler and Library. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL*): Demonstrations Session, 29–32. Association for Computational Linguistics, Athens.

Guglielmo Inglese. 2015. Towards a Hittite Treebank. Basic Challenges and Methodological Remarks. In *Proceedings of the Workshop on Corpus-Based Research in the Humanities (CRH) 10 December 2015, Warsaw, Poland*, edited by Francesco Mambrini, Marco Passarotti, and Caroline Sporleder, 59–68.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, edited by Daniel Zeman and Jan Hajič, 133–142. Association for Computational Linguistics, Brussels.

N.J.C. Kouwenberg. 2017. *A Grammar of Old Assyrian* (Handbuch der Orientalistik, 1/118). Brill, Leiden and Boston.

Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. 2009. HFST tool for morphology: An efficient open-source package for construction of morphological analyzers. In *State of the Art in Computational Morphology*, edited by Cerstin Mahlow and Michael Piotrowski, 28–47. Communications in Computer and Information Science, Berlin and Heidelberg.

Mikko Luukko. 2004. *Grammatical Variation in Neo-Assyrian* (State Archives of Assyria Studies, 16). The Neo-Assyrian Text Corpus Project, Helsinki.

Erica Reiner. 1966. *A Linguistic Analysis of Akkadian* (Janua Linguarum, Series Practica, 21). Mouton, The Hague.

Aleksi Sahala, Miikka Silfverberg, Antti Arppe, and Krister Lindén. 2020. BabyFST – Towards a Finite-State Based Computational Model of Ancient Babylonian. In *Proceedings of the 12th Language Resources and Evaluation Conference* (*LREC*), 3886–3894.

SIGNLL = Special Interest Group on Natural Language Learning of the Association for Computational Linguistics. 2018. Evaluation script for the 2018 CoNLL shared task, web description. http://universaldependencies.org/conll18/evaluation.html

Wolfram von Soden. 1973. Der akkadische Subordinativ-Subjunktiv. *Zeitschrift für Assyriologie und Vorderasiatische Archäologie* 63(1):56–58.

Wolfram von Soden. [3]1995. *Grundriss der akkadischen Grammatik* (Analecta Orientalia, 33/47). Pontificium Institutum Biblicum, Rome.

Michael P. Streck. 2010. Großes Fach Altorientalistik: Der Umfang des keilschriftlichen Textkorpus. *Mitteilungen der Deutschen Orient-Gesellscha*ft, 142:35–58.

Michael P. Streck. 2011. Babylonian and Assyrian. In *The Semitic Languages: An International Handbook,* edited by Stefan Weninger, 359–396. De Gruyter Mouton, Berlin and Boston.

Imed Zitouni (ed.). 2014. *Natural Language Processing of Semitic Languages*. Springer, Berlin and Heidelberg.