Journal of Open Humanities Data

A Full Morphosyntactic Annotation of the State Archives of Assyria Letter Corpus

DATA PAPER

MATTHEW ONG

ubiquity press

ABSTRACT

The dataset consists of a full morphosyntactic annotation of the normalized letter corpus of the State Archives of Assyria online (SAAo), plus associated metadata regarding sender, recipient, estimated date of composition, script, and dialect of Akkadian (if determinable). This corpus comprises ten of the twenty-one current volumes of SAAo and contains approximately 2600 letters from the royal archives of the late Neo-Assyrian kings. Each letter features morphosyntactic annotations specifying part of speech, lemma, morphological decomposition, and syntactic dependencies of all relevant tokens in the text. The annotations were made with the help of a spaCy language model with additional human checking and completion. The annotations are available both as a set of CONLLU files (one per text) and as linked open data in a single TTL file. The associated metadata is available as a CSV file. Due to the letters' shared format, topics of concern, and historical period in which they were written, this corpus forms a natural object of study from a linguistic and social historical perspective. It is hoped this data will be of use to researchers wishing to do linguistic and sociolinguistic corpus research on these texts.

CORRESPONDING AUTHOR:
Matthew Ong
Middle Eastern Languages and Cultures, UC Berkeley, Berkeley, CA, USA
matthewcong@berkeley.edu

KEYWORDS:
morphology; syntax; Akkadian; Neo-Assyrian letters; language model; linked open data; sociolinguistics

TO CITE THIS ARTICLE:
Ong, M. (2024). A Full Morphosyntactic Annotation of the State Archives of Assyria Letter Corpus. *Journal of Open Humanities Data*, 10: 30, pp. 1–6. DOI: https://doi.org/10.5334/johd.202

1 OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.10622983.

CONTEXT

The royal archives of the late Neo-Assyrian kings (8th–7th century BCE) are an important source for our understanding of the Neo-Assyrian empire. They contain a wide variety of texts ranging from treaty tablets and legal documents to prophecies, ritual instructions, and court literature. Over the past four decades, much from these archives has been published in the State Archives of Assyria (SAA) volumes at the University of Helsinki, and in more recent years has appeared digitally under the Munich Open-access Cuneiform Corpus Initiative (LMU Munich) as the State Archives of Assyria online (SAAo).

In particular, the letters from these archives constitute the largest subgroup (some 2,600 of the more than 5,000 texts published so far) and are a valuable resource for reconstructing aspects of late imperial administration, royal ideology, biographies of notable palace authorities, and social history.[1]

The letters published in SAAo span a period from the reign of Tiglath-pileser III (r. 747–722) down to Sin-šarri-iškun (r. 627?-612). However, most are dated to the reigns of Sargon II (r. 721–705), Sennacherib (r. 704–681), Esarhaddon (r. 680–669), and Assurbanipal (r. 668–627). The letters are published across ten volumes of the SAAo series, namely SAA 1, 5, 10, 13, 15, 16, 17, 18, 19, and 21. Note that in the interests of keeping the dataset fairly homogeneous as far as form and content, we are excluding SAA 8 (astrological reports to the king) due to the differences in form and content such reports have with the letter corpus proper.[2] The small number of literary letters found in SAA 3 (Court Poetry and Literary Miscellanea) are likewise not included in the dataset due to their special use context.

Importantly, the letters in SAAo are one of the best resources for the study of the Neo-Assyrian dialect, a vernacular form of Akkadian different from the Neo-Babylonian dialect to the south as well as older forms of Akkadian such as Old Babylonian. The letters do employ various formulas and fixed expressions as a matter of genre, particularly in the introductory sections or when introducing new topics (see Ponchia 1989, and Luukko 2012). However, in comparison to texts from the same period that are written in a more conservative style (such as royal inscriptions and liturgical works), the letters feature grammatical constructions, linguistic forms, and discourse patterns more evocative of spoken discourse, in addition to reflecting influence from Aramaic (Abraham and Sokoloff 2011, Streck 2011). Moreover, the letters involve communication between individuals or groups of different social backgrounds, undertaken for a variety of aims. Combined with the fact that we have dossiers on many of these individuals and can often date the letters to a particular ruler's reign, the letter corpus offers opportunities for sociolinguistic, historical linguistic, and stylistic studies of late-stage Akkadian. Studies of this sort have been done previously by hand, but they can now also be done with the help of this dataset.

Most of the letters are written in the Neo-Assyrian dialect. However, a minority are written in the Neo-Babylonian dialect. This difference goes hand in hand with script difference, as the cuneiform script used by scribes trained in the Assyrian tradition versus the Babylonian also differs. Some letters written by the most educated scribes at the Assyrian court show code switching (Worthington 2006).

2 METHOD

GENERATING THE MORPHOSYNTACTIC ANNOTATIONS

The morphosyntactic annotations for the letter corpus were created using the same process described in Ong and Gordin 2024b and Ong and Gordin 2024a. At the most abstract level, this centered on a cyclic boot-strapping procedure illustrated in Figure 1. We first trained a spaCy

1 See Radner 2014.

2 For instance, the reports lack the introductory sections usually found in Neo-Assyrian letters (on which see Fales 1987: 451 and Luukko 2012), and consist only of astronomical omens and their interpretations.

language model (Honnibal, et al. 2020) on an initial batch of manually-generated annotations in CONLLU format, then applied the model to a new group of letters to yield a set of imperfect or incomplete annotations.[3] This set of annotations was corrected and completed by hand using Inception (Klie, et al. 2018), and then added to the initial training data. The language model was then retrained on the extended training set, resulting in a slight improvement in model performance, and hence faster work in hand-annotating the next batch of letters in Inception. This process was repeated numerous times until all letters were annotated. The result is a set of CONLLU files, one for each letter. For the convenience of those wishing to train their own spaCy model on this data, the CONLLU files were also converted to SPACY binary files.

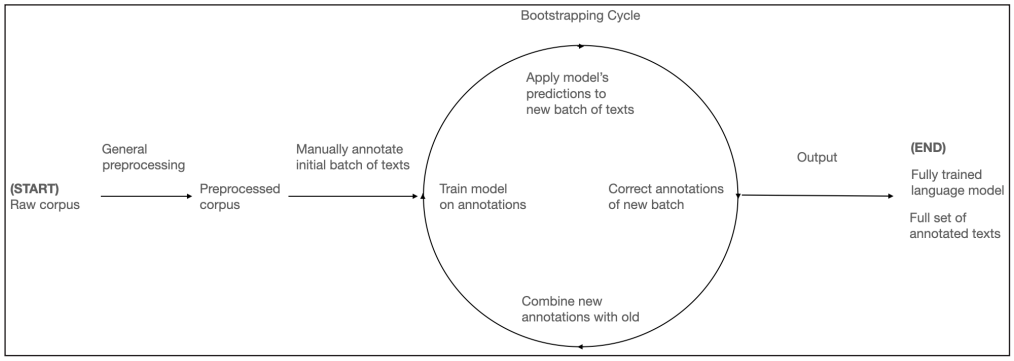

Figure 1 Abstract pipeline involving bootstrapping.

A detailed discussion of the labels and conventions used in the morphosyntactic annotations, including the labels used in the morphological parsing, are provided on the author's github account.[4] Certain points of grammatical interpretation are also covered there.

In brief, the annotations provide, for each interpretable form in a text, the Universal Dependencies part of speech tag (UPOS), the lemma, and morphological decomposition expressed as a string of feature-value pairs.[5] The morphological analysis describes, in the case of nouns and adjectives, the gender, number, and case of the form, and whether the form was in the bound or free state. In the case of finite verbs, the person, number, gender, tense, stem, and mood are specified. Suffixes and enclitics attached to head forms are encoded as additional feature-value pairs within the morphological analysis of the head form. Sentence breaks, in most cases, were not marked. The morphological analysis of a form at the level of part of speech came directly from Oracc metadata. Other features such as verb stem and suffix patterns required annotator judgment, although the Oracc translation of the text was almost always accepted and used in making a decision. Grammatical analyses were largely based on Hämeen-Anttila 2000, save that verbal adjective forms in the stative were labeled as verbs.[6]

The most valuable part of the annotations is perhaps the universal dependency relations between interpretable forms. As discussed in Ong and Gordin 2024b, syntactic dependencies are the most difficult task for our spaCy model, and the area where most of the manual correction and completion takes place. At the same time, it is still rare to find digitized Akkadian corpora marked for syntactic parsing or even language models trained to perform such a task (Luukko, et al. 2020, Ong and Gordin 2024b).

Illustrations of how these features were marked within Inception and the underlying CONLLU file are given in Figures 2 and 3.

CONVERTING THE ANNOTATIONS TO LINKED OPEN DATA

All the CONLLU files representing letter annotations were concatenated into a single file, with a comment line above each section indicating which text it represents. This block file was

3 Details concerning model performance can be found in Section 4 of Ong and Gordin 2024a.

4 https://github.com/megamattc/Akkadian-language-models (last accessed: 13 March 2024).

5 Uninterpretable tokens are given dummy values for UPOS and lemma.

6 Following arguments of Kouwenberg 2000, Kouwenberg 2010, and Kamil 2023.

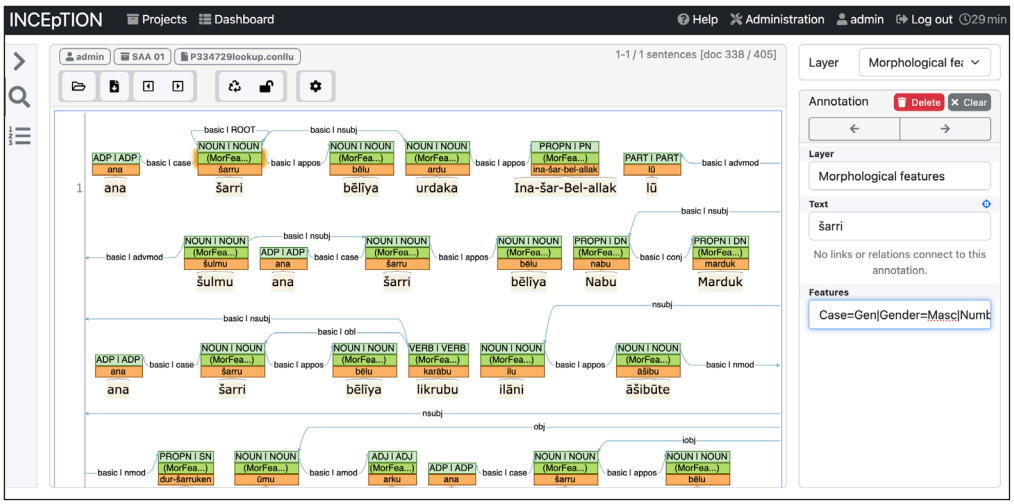

Figure 2 Morphosyntactic annotation of SAAo letter in Inception.

```
http://oracc.org/saao/saa01/P334729
# text = ana šarri bēlīya urdaka Ina-šar-Bel-allak lū šulmu ana šarri...

1       ana                     ana                     ADP     ADP     _                                               2       case    _       _
2       šarri                   šarru                   NOUN    NOUN    Case=Gen|Gender=Masc|Number=Sing                 0       ROOT    _       _
3       bēlīya                  bēlu                    NOUN    NOUN    Case=Gen|Gender=Masc|Number=Sing|PossSuffGen=Com...      2       appos   _       _
4       urdaka                  ardu                    NOUN    NOUN    Case=Nom|Gender=Masc|Number=Sing|PossSuffGen=Masc...     2       nsubj   _       _
5       Ina-šar-Bel-allak       ina-šar-bel-allak       PROPN   PN      Case=Nom|Gender=Masc|Number=Sing                 4       appos   _       _
...
```

Figure 3 Morphosyntactic annotation of SAAo letter in CONLLU format.

converted to RDF turtle format (TTL)[7] using the Java package conll-rdf.[8] The value of converting the annotations to linked open data is that they may then be easily searched via SPARQL queries for various morphological and syntactic features (see Ong and Gordin 2024b for an example). They can also be converted to other knowledge graph representations such as Neo4j, which enable even more sophisticated graph queries.

GENERATING THE METADATA

The process for providing metadata for the letters is similar to that in Ong and Gordin 2024b. Preexisting metadata was extracted from SAAo catalogue files and combined in a CSV file. This metadata includes sender, recipient, estimated date of composition, script, and dialect of Akkadian (if determinable).

3 DATASET DESCRIPTION

OBJECT NAME

A Full Morphosyntactic Annotation of the State Archives of Assyria Letter Corpus.

FORMAT NAMES AND VERSIONS

CSV, CONLLU, SPACY, TTL, TXT

CREATION DATES

2022-09-01–2024-02-01

DATASET CREATOR

Matthew Ong (UC Berkeley) was responsible for all aspects of the project.

LANGUAGE

English, Akkadian

LICENSE

7 See Bizer, et al. 2018.

8 Available at https://github.com/acoli-repo/conll-rdf (last accessed: 5 February 2024).

REPOSITORY NAME

Zenodo

PUBLICATION DATE

2023-08-28

4 REUSE POTENTIAL

The dataset (both morphsyntactic annotations and metadata) can be used in a variety of ways. Scholars may search the normalized letter corpus for text patterns conforming to any number of syntactic, morphological, or even phonological features (i.e. spelling) provided those features are marked in the annotations and the query itself can be expressed in SPARQL (or other knowledge graph query format such as Neo4j). This goes beyond the current search capabilities of the letter corpus on Oracc, which is largely based on keywords.[9] When the metadata is incorporated into this search, one may also begin to search for sociolectal, ideolectal, topolectal, and other linguistic patterns in the letters that have so far escaped human readers.

The methods used to generate the annotations (as described in Ong and Gordin 2024a) can be applied to other lemmatized Akkadian corpora on Oracc. Researchers wishing to annotate such texts in conjunction with spaCy model training may benefit from using the spaCy Akkadian language package in development by the author.[10] This package is still under development and currently only works for normalized texts. However, one of its most useful features currently is the able to correctly tokenize a large number of lexicalized construct phrases in Oracc (encoded by a 'long dash' in the online edition).

Finally, the process of converting CONLLU annotations to RDF triples allows the data to be integrated into other linked open data projects, particularly those involving other annotated corpora or Mesopotamian culture more generally.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR ROLE

Matthew Ong: All roles

AUTHOR AFFILIATIONS

Matthew Ong orcid.org/0000-0003-2566-9205
Middle Eastern Languages and Cultures, UC Berkeley, Berkeley, CA, USA

REFERENCES

Abraham, K., & **Sokoloff, M.** (2011). Aramaic Loanwords in Akkadian – A Reassessment of the Proposals. *Archiv für Orientforschung*, *52*, 22–76. Retrieved 2024-02-06, from https://www.jstor.org/stable/24595102

Bizer, C., Vidal, M.-E., & **Skaf-Molli, H.** (2018). Linked open data. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 2096–2101). Springer New York. DOI: https://doi.org/10.1007/978-1-4614-8265-9_80603

Fales, F. M. (1987). Aramaic Letters and Neo-Assyrian Letters: Philological and Methodological Notes. *Journal of the American Oriental Society*, *107*(3), 451–469. DOI: https://doi.org/10.2307/603465

Hämeen-Anttila, J. (2000). *A Sketch of Neo-Assyrian Grammar (SAAS 13)*. Neo-Assyrian Text Corpus Project.

Honnibal, M., Montani, I., Landeghem, S. V., & **Boyd, A.** (2020). spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.1212303

9 See https://oracc.museum.upenn.edu/doc/search/index.html (last accessed: 5 February 2024).

10 Available at https://github.com/megamattc/Akkadian-language-models/tree/main/ak (last accessed: 13 March 2024).

Kamil, I. (2023). T-Forms of the Akkadian Stative. *Brill's Journal of Afroasiatic Languages and Linguistics*, *15*(1), 262–290. DOI: https://doi.org/10.1163/18776930-01501008

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., & **Gurevych, I.** (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: System demonstrations* (pp. 5–9). Association for Computational Linguistics. Retrieved from http://tubiblio.ulb.tu-darmstadt.de/106270/

Kouwenberg, N. (2000). Nouns as Verbs: The Verbal Nature of the Akkadian Stative. *Orientalia Nova Series*, *69*(1), 21–71.

Kouwenberg, N. (2010). *The Akkadian verb and its Semitic Background*. Eisenbrauns. DOI: https://doi.org/10.1515/9781575066240

Luukko, M. (2012). On Standardisation and Variation in the Introductory Formulae of Neo-Assyrian Letters. *Iraq*, *74*, 97–115. DOI: https://doi.org/10.1017/S0021088900000292

Luukko, M., Sahala, A., Hardwick, S., & **Lindén, K.** (2020). Akkadian Treebank for early Neo-Assyrian Royal Inscriptions. In K. Evang, L. Kallmeyer, R. Ehren, S. Petitjean, E. Seyffarth, & D. Seddah (Eds.), *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories* (pp. 124–134). Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.tlt-1.11

Ong, M., & **Gordin, S.** (2024a). Linguistic annotation of cuneiform texts using treebanks and deep learning. *Digital Scholarship in the Humanities*, fqae002. DOI: https://doi.org/10.1093/llc/fqae002

Ong, M., & **Gordin, S.** (2024b). A Survey of Body Part Construction Metaphors in the Neo-Assyrian Letter Corpus. *Journal of Open Humanities Data*, *10*(1), 10. DOI: https://doi.org/10.5334/johd.142

Ponchia, S. (1989). Royal Decisions and Courtiers' Compliance: On some Formulae in Neo-Assyrian and Neo-Babylonian Letters. *State Archives of Assyria Bulletin*, *3*, 115.

Radner, K. (2014). An Imperial Communication Network: The State Correspondence of the Neo-Assyrian Empire. In K. Radner (Ed.), *State Correspondence in the Ancient World: From New Kingdom Egypt to the Roman Empire* (pp. 64–93). Ludwig-Maximilians-Universität München. DOI: https://doi.org/10.1093/acprof:oso/9780199354771.003.0004

Streck, M. P. (2011). Akkadian and Aramaic Language Contact. In S. Weninger (Ed.), *The Semitic Languages: An International Handbook* (pp. 416–424). De Gruyter Mouton. DOI: https://doi.org/10.1515/9783110251586.416

Worthington, M. (2006). Dialect Admixture of Babylonian and Assyrian in SAA VIII, X, XII, XVII and XVIII. *Iraq*, *68*, 59–84. DOI: https://doi.org/10.1017/S0021088900001169

TO CITE THIS ARTICLE:
Ong, M. (2024). A Full Morphosyntactic Annotation of the State Archives of Assyria Letter Corpus. *Journal of Open Humanities Data*, 10: 30, pp. 1–6. DOI: https://doi.org/10.5334/johd.202

Submitted: 10 February 2024
Accepted: 18 March 2024
Published: 12 April 2024

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.