



A Survey of Body Part Construction Metaphors in the Neo-Assyrian Letter Corpus

MATTHEW ONG 

SHAI GORDIN 

*Author affiliations can be found in the back matter of this article

DATA PAPER

 ubiquity press

ABSTRACT

The dataset consists of approximately 2,400 examples of metaphors in Akkadian of what we term Body Part Constructions (BPCs) within the letter sub-corpus of the State Archives of Assyria online (SAAo). The dataset was generated by a multi-step process involving the training and application of a language model to the SAAo letter sub-corpus, converting the resulting annotations to linked open data format amenable to searching for BPCs, and manually adding metalinguistic data to the search results; these files, in CONLLU and TTL formats, are also made available in this publication. The BPC dataset is stored as a CSV file, and can serve as an easy starting place for other scholars interested in finding socio-linguistic usage patterns of this construction.

CORRESPONDING AUTHOR:

Matthew Ong

Middle Eastern Languages and Cultures, UC Berkeley, Berkeley, CA, USA

matthewcong@berkeley.edu

KEYWORDS:

metaphor; Akkadian; body parts; Neo-Assyrian letters; language model; linked open data

TO CITE THIS ARTICLE:

Ong, M., & Gordin, S. (2024). A Survey of Body Part Construction Metaphors in the Neo-Assyrian Letter Corpus. *Journal of Open Humanities Data*, 10: 10, pp. 1–6. DOI: <https://doi.org/10.5334/johd.142>

Repository location <https://doi.org/10.5281/zenodo.8289986>

CONTEXT

The royal archives of the late Neo-Assyrian kings (8th–7th century BCE) constitute an important source for understanding many facets of the Neo-Assyrian empire. Ranging from treaty tablets and legal documents to prophecies, ritual instructions, and even court literature, the approximately five thousand texts in this corpus primarily come from the palatial complex at Nineveh and document the reigns of Sargon II (r. 721–705), Sennacherib (r. 704–681), Esarhaddon (r. 680–669), and Assurbanipal (r. 668–627). 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).

Within this corpus, the set of letters constitutes a sizeable sub-corpus that is valuable not only in terms of reconstructing social history, but also as a representative of vernacular late Akkadian. It is this linguistic fact that motivates our dataset, discussed more extensively in Ong and Gordin ([under review](#)). Here we provide a summary of the dataset's contents. It is a CSV file with approximately 2,400 examples of what we term Body Part Constructions (BPCs) in Akkadian, where a BPC is defined as a verb with a compound prepositional phrase based on a body part term. For instance, *X ina qat Y šûlû* literally means 'to lift X from the hand of Y', but colloquially means 'to estrange X from Y'. These constructions are interesting as they are a productive vehicle for metaphors in Akkadian as well as a socio-linguistic feature of various subgroups within the Neo-Assyrian letter corpus.

For a given BPC, the CSV file lists the lexical item representing each syntactic component of the BPC (verb, body part term, direct object, etc.), the lemmas associated with these lexical items, and the CDLI P-number of the text that the BPC appears in.¹ Most BPCs also come with a translation in context, as well as a number of fields describing the letter they appear in, like sender, receiver, sender's location, date of composition, dialect and script of composition, genre, and provenience, alongside additional linguistic and rhetorical properties of the BPC. Most of these fields are described in detail under Ong and Gordin ([under review](#)).

2 METHOD

The dataset was generated via a three-step process:

LANGUAGE MODEL TRAINING

We first sought to train a spaCy language model² on a subset of manually annotated, normalized texts drawn from a variety of Oracc projects (consisting both of letters and other genres).³ The training set included: all texts in SAAo 1, 2, 5, 9, 15, and 21, a small set of texts from each of SAAo 8, 10, 13, 16, 17, 18, and 19, SB Anzu, a few extispicy texts from the Corpus of Ancient Mesopotamian Scholarship (CAMS)/Barutu project, selected royal inscriptions of Esarhaddon found in the Royal Inscriptions of Assyria Online project (RIAO), a few Middle Assyrian letters from the Text Corpus of Middle Assyrian project (TCMA), and about two hundred and fifty synthetic training sentences generated manually.⁴ The model files and training, development, and test data are available from the above-listed Zenodo repository as well as the primary

¹ The Cuneiform Digital Library Initiative (CDLI) P-number is a conventional ID in the field of cuneiform studies.

² Available at <https://github.com/megamattc/Akkadian-language-models> (last accessed: 20 November 2023).

³ For a complete list of Oracc projects see <http://oracc.museum.upenn.edu/projectlist.html> (last accessed: 20 November 2023).

⁴ The Oracc pages for these texts are:

SB Anzu: <http://oracc.museum.upenn.edu/cams/anzu/corpus>.

CAMS/Barutu: <http://oracc.museum.upenn.edu/cams/barutu/corpus>.

TCMA: <http://oracc.museum.upenn.edu/tcma/>.

RIAO: <http://oracc.museum.upenn.edu/riao/corpus/>.

author's GitHub account.⁵ We then used INCEPTION to make the manual annotations for these texts, which were then encoded in CONLLU format (Figure 1).⁶ We also incorporated into our training set the annotated Neo-Assyrian royal inscriptions from Luukko, Sahala, Hardwick, and Lindén (2020), slightly modified to match our own annotation format.

```

http://oracc.org/saao/saa01/P334729
# text = ana šarri bēliya urdaka Ina-šar-Bel-allak lū šulmu ana šarri...
1   ana          ana          ADP    ADP    Case=Gen|Gender=Masc|Number=Sing 2   case  -  -
2   šarri        šarru       NOUN  NOUN  Case=Gen|Gender=Masc|Number=Sing 0   ROOT  -  -
3   bēliya       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 1 Section of SAAo letter in CONLLU format.

We then used the resulting language model to generate automatic parses of the remainder of the SAAo letter sub-corpus and combined them with the original training set to yield a complete set of annotations for the SAAo letters in CONLLU format. More information about the accuracy of the automatic parses can be found in Ong and Gordin (under review).

CONVERSION TO LINKED OPEN DATA

We converted the CONLLU format annotations from the previous step to RDF turtle format (TTL) using a Java package named conll-rdf (Figure 2).⁷ Each SAAo volume was associated with a single TTL file forming part of our dataset. Both the CONLLU and TTL files of SAAo texts, broken down by individual SAA volume, are made available with the current BPC dataset within the above-listed Zenodo repository.⁸ The TTL files were then uploaded to TriplyDB, a data hosting service that enables users to easily query their own linked open data projects through a variety of APIs.⁹ We then used a SPARQL query to search through the various SAAo letter volumes individually for the exact syntactic and lexical features defining BPCs.¹⁰ The resulting attestations were then amalgamated into a CSV file.

```

@prefix conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
...

:s1_1 rdf:type nif:Word, :s1_1 conll:WORD "ana", :s1_1 conll:EDGE "case", :s1_1 conll:HEAD :s1_2, conll:ID "1", :s1_1 conll:LEMMA "ana", :s1_1 conll:POS "ADP",
:s1_1 conll:UPOS "ADP", :s1_1 nif:nextWord :s1_2

:s1_2 rdf:type nif:Word, :s1_2 conll:WORD "šarri", :s1_2 conll:EDGE "ROOT", :s1_2 conll:FEAT "Case=Gen|Gender=Masc|Number=Sing", :s1_2 conll:HEAD :s1_0, :s1_2 conll:ID
"2", :s1_2 conll:LEMMA "šarru", :s1_2 conll:POS "NOUN", :s1_2 conll:UPOS "NOUN", :s1_2 nif:nextWord :s1_3

:s1_3 rdf:type nif:Word, :s1_3 conll:WORD "bēliya", :s1_3 conll:EDGE "appos", :s1_3 conll:FEAT "Case=Gen|Gender=Masc|Number=Sing|PossSuffGen=Com|PossSuffNum=Sing|
PossSuffPer=1", :s1_3 conll:HEAD :s1_2, :s1_3 conll:ID "3", :s1_3 conll:LEMMA "bēlu", :s1_3 conll:POS "NOUN", :s1_3 conll:UPOS "NOUN", :s1_3 nif:nextWord :s1_4

```

Figure 2 Section of SAAo letter in TTL format.

ENRICHING THE METADATA

The last step involved enriching the search results from Step 2 in two ways. First, we extracted certain metadata from the Oracc JSON files associated with the SAAo letter sub-corpus and merged it into the list of attestations. The result was an enlarged CSV file which provided, for a BPC within a given letter, the dialect of Akkadian, likely date and ruler under which that letter

5 The model distribution is provided under `ak_AkkParser_Norm_1_2_5_8_9_10_13_15_16_17_18_19_21_anzu_barutu_rinap4_tcmaassur-0.0.0.tar.gz` in the Zenodo repository. The training, development, and test data is available in CONLLU format under the `ak_norm_conllu.zip` file. Given that spaCy requires one to first convert all training data to a special binary format before it can be used for model training, we have also included these pre-compiled binaries under the `ak_norm_spacy.zip` file of the Zenodo archive. The CONLLU, model, and binary files may also be found at the primary author's GitHub repository https://github.com/megamatc/Akkadian-language-models/tree/main/ak_norm_model/assets/UD_Akkadian, [https://github.com/megamatc/Akkadian-language-models/tree/main/ak_norm_model/corpus/UD_Akkadian](https://github.com/megamatc/Akkadian-language-models/tree/main/ak_norm_model/packages/ak_AkkParser_Norm_1_2_5_8_9_10_13_15_16_17_18_19_21_anzu_barutu_rinap4_tcmaassur-0.0.0/ak_AkkParser_Norm_1_2_5_8_9_10_13_15_16_17_18_19_21_anzu_barutu_rinap4_tcmaassur-0.0.0), respectively (last accessed: 20 November 2023). Readers wishing to train the model from scratch should also consult the documentation available at <https://github.com/megamatc/Akkadian-language-models/tree/main> (last accessed: 20 November 2023).

6 Available for download at <https://inception-project.github.io/> (last accessed: 20 November 2023).

7 Available at <https://github.com/acoli-repo/conll-rdf> (last accessed: 20 November 2023).

8 Available under `akk_mcong-ud-saa.zip`.

9 <https://triplydb.com/> (last accessed: 20 November 2023).

10 This query is found as part of the file `sparql_Cxn_BCP.txt` within the Zenodo repository.

was written in, as well as the letter's sender and receiver, provenience, and sender's location (provided such things were known). Second, the primary author made a manual evaluation of the list of BPC attestations, clearing the data for false results and describing the semantic and discourse properties of most of the BPCs, as detailed in Ong and Gordin ([under review](#)).¹¹

3 DATASET DESCRIPTION

OBJECT NAME

Body part construction metaphors in the Neo-Assyrian letter corpus.

FORMAT NAMES AND VERSIONS

CSV, CONLLU, SPACY, TTL, TXT

CREATION DATES

2022-09-01 – 2023-08-10

DATASET CREATORS

Matthew Ong (UC Berkeley) was responsible for conceptualization, research design, data extraction and validation, dataset creation, and software development.

Shai Gordin (Ariel University and the Open University) was responsible for funding acquisition, project administration, resources, supervision, data validation, and editing of the manuscript.

LANGUAGE

English, Akkadian

LICENSE

Creative Commons Attribution-Share-Alike 4.0

REPOSITORY NAME

Zenodo

PUBLICATION DATE

2023-08-28

4 REUSE POTENTIAL

The language model and training data it is based on can be used to perform other high-level morpho-syntactic queries on any normalized Akkadian texts formatted according to Oracc standards, including the Neo-Assyrian letter corpus used for this project. One may simply take the trained model and modify the SPARQL query appropriate to one's desired target structure. Following the instructions at the end of the second step of our method ('Conversion to linked open data') will generate a list of search results for the corpus.

In Ong and Gordin ([under review](#)), for example, we used this dataset to illustrate a prominent dialectal difference between Neo-Assyrian and Neo-Babylonian letters with respect to a natural class of BPCs. Specifically, BPCs involving directed motion use *ina* in the Neo-Assyrian letters while those written in Neo-Babylonian use *ana*. In addition, we came up with an approximate distribution of the BPCs in the letter corpus with respect to rhetorical purpose and metaphorical salience ([Figure 3](#)).

¹¹ This data is available at `BPC_saa_01_05_10_13_15_16_17_18_19_21_28082023.csv` in the Zenodo repository.

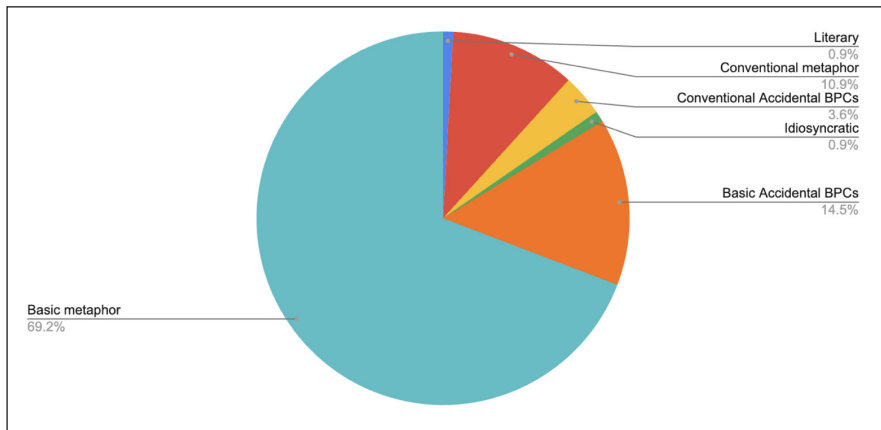


Figure 3 Distribution of BPCs in the Neo-Assyrian letter corpus according to metaphorical salience.

Other scholars may use this dataset as a reference list of BPC examples they wish to consult. They can also conduct further analysis on this dataset, discovering interesting patterns in the distribution of BPCs and related metaphors beyond those discussed in Ong and Gordin ([under review](#)). Scholars can also extend the dataset to include information about other BPCs beyond the Neo-Assyrian letter sub-corpus, or follow our methodology to generate a dataset of other grammatical constructions and merge it with our data.

The fact that much of the dataset is automatically generated by a neural language model imposes certain limits on its use. First, a small percentage of BPCs in the letter sub-corpus is not contained in our data. Our dataset is thus not an exhaustive list of BPCs. Secondly, efforts to improve the language model and reapply it may result in a larger, but slightly different list of BPCs from what we provide here. Thus independent improvements (or alternatives) to our model by different scholars may lead to a need to aggregate the various search results.

ACKNOWLEDGEMENTS

We wish to thank Eve Sweetser (UC Berkeley) for her linguistic comments on this project.

FUNDING INFORMATION

Matthew Ong's work on this project was supported by the PhD Sandwich Fellowship Program of the Planning and Budgeting Committee (PBC) of Israel.

COMPETING INTERESTS

The authors have no competing interests to declare.


AUTHOR CONTRIBUTIONS

Matthew Ong: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft

Shai Gordin: Funding acquisition, Project administration, Methodology, Resources, Validation, Supervision, Writing – review and editing

AUTHOR AFFILIATIONS

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

Shai Gordin  orcid.org/0000-0002-8359-382X
Land of Israel and Archaeology, Digital Pasts Lab, Ariel University, Ariel, Israel; Digital Humanities and Social Sciences Hub, Open University, Ra'anana, Israel

REFERENCES

- 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). Stroudsburg, PA: Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2020.tlt-1.11>
- Ong, M., & Gordin, S.** (under review). Neo-Assyrian Metaphors through the Telescope: Linguistic Patterns involving Body Part Constructions in the State Archives Letter Corpus.

Ong and Gordin
*Journal of Open
Humanities Data*
DOI: 10.5334/johd.142

6

TO CITE THIS ARTICLE:

Ong, M., & Gordin, S. (2024). A Survey of Body Part Construction Metaphors in the Neo-Assyrian Letter Corpus. *Journal of Open Humanities Data*, 10: 10, pp. 1–6. DOI: <https://doi.org/10.5334/johd.142>

Submitted: 30 August 2023

Accepted: 17 October 2023

Published: 22 January 2024

COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

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