GHENT
UNIVERSITY

FACULTY OF ARTS AND PHILOSOPHY

AT THE CROSSROAD OF CUNEIFORM & NLP

Challenges for Fine-grained Part-of-Speech Tagging – Gustav Ryberg Smidt, Katrien de Graef and Els Lefever

GHENT UNIVERSITY

AssyrUGent

language and translation technology team

OUTLINE

- Introduction and questions
- What is Cuneiform and Akkadian?
- Cuneiform NLP
  - Input data and goals
  - Hurdles
  - Experiments
  - Results
- Perspectives

GHENT UNIVERSITY

INTRODUCTION

GHENT
UNIVERSITY

CUNE-IIIF-ORM & NLP

- Assyriologists and computational linguists

- Old Babylonian Akkadian letters
  - Narrow but “big-ish” corpus
  - Close to everyday language
  - Varied content
  - Many layers of society represented – Sociolects?
  - Mainly Akkadian with occasional Sumerian

- Questions
  - What is the reality of language vs prescriptive grammars?
  - How do we best provide data for the research community?

GHENT UNIVERSITY

WHAT IS CUNEIFORM?

GHENT
UNIVERSITY

CUNEIFORM

- Sumerian pictographic script from modern day Iraq
- Wedges pressed in clay combined into signs
- Logo-syllabic script with homophony, polyphony and classifiers
- No punctuation
- Used for many languages

Homophony	Polyphony
𒀀 = u_1	𒀭 = il_3
𒌋 = u_2	𒀭 = an
𒌑 = u_3	𒀭 = sa_8

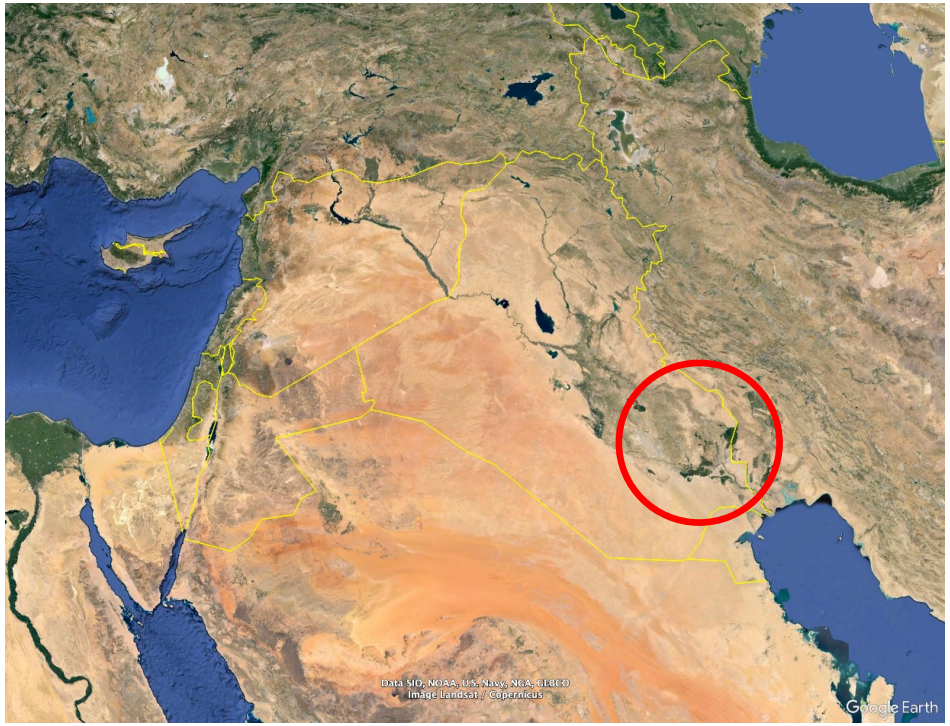

Image of the Middle East from Google Earth, highlighting southern Iraq from where Cuneiform originated.

~3500 B.C.E.

~100 C.E.

GHENT UNIVERSITY

AKKADIAN

- Semitic language
- Greatly influenced by Sumerian (linguistic isolate)
- Subject – Object – Verb
- Two main dialects – Babylonian (South) & Assyrian (North)
- Old Babylonian Akkadian – 2000 to 1600 B.C.E.

Picture of the back of the tablet AO 1781 from the Louvre collection website [2].

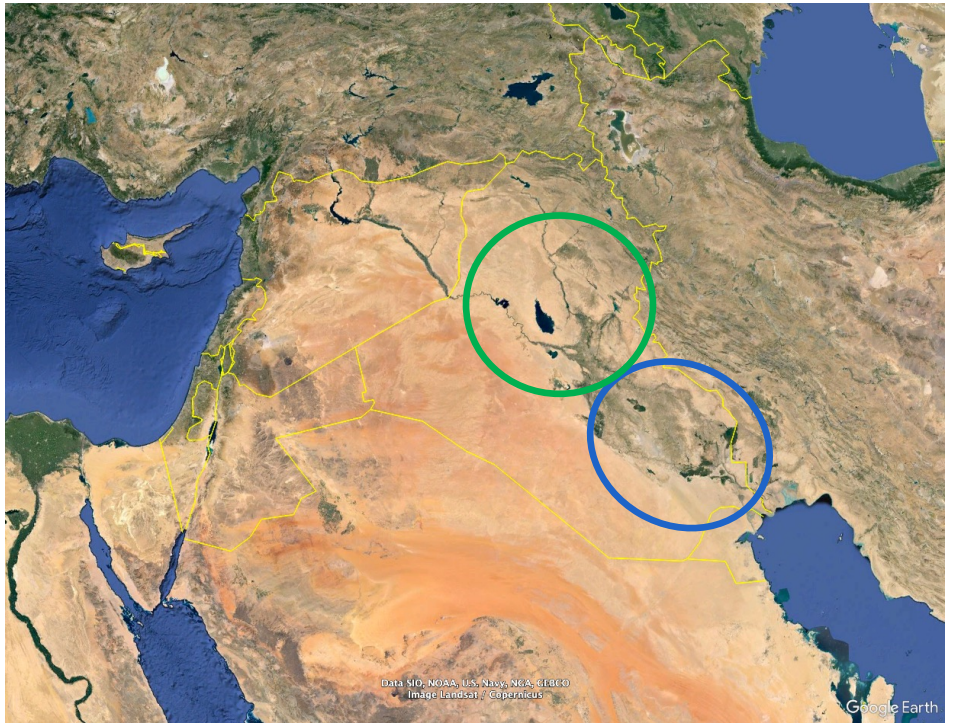

Image of the Middle East from Google Earth, highlighting Babylonia (Blue) and Assyria (Green).

~2500 B.C.E.

~100 C.E.

GHENT UNIVERSITY

CUNEIFORM NLP

GHENT
UNIVERSITY

INPUT DATA & GOALS

- Input data
  - 121 letters / ~10.000 words
  - Lemmatized in Open Richly Annotated Cuneiform Corpus
  - Morphological annotations with regular expressions and manual corrections
- Goals
  - Semi-automatic morphological tagger
    - Corpus to contain all Old Babylonian Akkadian letters
  - Computational analyses of the corpus

GHENT UNIVERSITY

INPUT DATA & GOALS

- Data format
  - Written form in transliteration
    - Preserves spelling variation
  - Morphological analyses with all features
    - Make definitions and labels more explicit
  - No syntactical information

HURDLES

1. Small datasets
   [6]: c. 10.000.000 Akkadian words

2. Imperfect understanding of Akkadian syntax

3. How do we present an accurate picture of Akkadian, with Unicode cuneiform? [8]

EXPERIMENTS

1. Pre-trained models to predict labels
   - Predict effective PoS and simple grammatical features
   - Training data: transliterated text, verbs define sentence splits

2. Sentence segmentation
   - Predict effective PoS
   - Training data: transliterated text, various sentence splits

3. Unicode cuneiform
   - Predict effective PoS
   - Training data: transliterated or Unicode cuneiform text, verbs define sentence splits

EXPERIMENTS

- 80/20 training/test split
- 5 folds
- Each sentence randomly distributed between folds

- Flair contextual string embeddings [9]
  - Multilingual [10]
  - Arabic [11]
  - Japanese [12]
  - Spanish [13]

- Data and scripts available on Github: https://github.com/assyrugent/LREC-Coling2024 [14]

GHENT UNIVERSITY

RESULTS – 1. PRE-TRAINED MODELS

- Data input
  - u_2-na-ah-hi-id-ma<tab_sep>V_D_preterite_1st_communalis_singular

- No big difference between results
- Arabic (Semitic) performed best
- Multilingual contains many different language families

Embeddings	Average accuracy
Multilingual	71,0 %
Arabic	**76,2 %**
Spanish	74,1 %
Japanese	72,6 %

GHENT UNIVERSITY

RESULTS – 2. SENTENCE SEGMENTATION

- Data input
  - Automatic insertion of newline when reached a condition, e.g. PoS = Verb or word index + 1 is a new text

Seperation type	Average Accuracy
Text	88,2 %
Line	92,5 %
Verb	**94,8 %**
Verb – u	93,9 %
Verb – u & ma	94,1 %
Verb – u, ma & conjunction	93,5 %

- Formatting our data works
- No clear winner out of the four verb-formatted sentences
- Syntatic information can help, but it's time consuming

GHENT UNIVERSITY

RESULTS – 3. UNICODE CUNEIFORM

- Data input
  - wa-aš-bu<tab_sep>Verb vs 𒉿𒍪𒁍<tap_sep>Verb

    𒉿 = pi | pe | wa | wi | we | wu | tal_2

    𒍪 = aš | rum | dil | ṭil | dal_3 | ina

    𒁍 = bu | pu | sir_2 | $šir_2$ | gid_2 | qid_2

- Japanese script is logo-syllabic like Cuneiform
- A mix of script and language familiy relevant data could be beneficial, e.g. Japanese and Arabic

Embeddings	Transliteration	Unicode	Loss
Multilingual	91,3 %	61,3 %	30 %pt.
Arabic	**94,1 %**	69,4 %	24,7 %pt.
Spanish	93,7 %	63,0 %	30,7 %pt.
Japanese	93,4 %	**73,9 %**	19,5 %pt.

GHENT UNIVERSITY

PERSPECTIVES

GHENT
UNIVERSITY

PERSPECTIVES

- What data can we acquire and build on?
- Can we emulate syntactic relations with morphological data and are there differences between periods and genres?
- Can we get close to the “real” cuneiform Akkadian?

GHENT
UNIVERSITY

REFERENCES

[1] = Louvre. 2024. "Code de Hammurabi: -1792 / -1750 (1ère dynastie de Babylone : Hammurabi)," Louvre Collection, https://collections.louvre.fr/en/ark:/53355/cl010174436.

[2] = Louvre. 2024. "tablette: -1894 / -1595 (1ère dynastie de Babylone)," Louvre Collection, https://collections.louvre.fr/en/ark:/53355/cl010125719.

[3] = CDLI contributors. 2024. “Home,” Cuneiform Digital Library Initiative, https://cdli.mpiwg-berlin.mpg.de/

[4] = E. Robson. 2019. "AKK: Oracc Linguistic Annotation for Akkadian," Oracc: The Open Richly Annotated Cuneiform Corpus, Oracc, http://oracc.museum.upenn.edu/doc/help/languages/akkadian/.

[5] = A. Sahala and K. Lindén. 2023. "Babylemmatizer 2.0 – a neural pipeline for pos-tagging and lemmatizing cuneiform languages," In Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pp. 203–212.

[6] = M.P. Streck. 2010. "Großes fach altorientalistik: Der umfang des keilschriftlichen textkorpus," Mitteilungen der Deutschen Orient-Gesellschaft zu Berlin, 142:35–58.

[7] = G.R. Smidt, C. Johansson, N. Melin-Kronsell, S. Nett and R. Rattenborg. 2023. "Cuneiform Inscriptions Geographical Site - Assemblage Estimates (CIGS-AE) (1.0)," Zenodo, https://doi.org/10.5281/zenodo.8379793.

[8] = G. Gutherz, S. Gordin, L. Sáenz, O. Levy, and J. Berant. 2023. "Translating akkadian to english with neural machine translation," PNAS Nexus, 2(5):1–10.

[9] = Anonymous. n.d. "Flair embeddings | flair," Flair, accessed on 28th of April 2024 at 15:50, https://flairnlp.github.io/docs/tutorial-embeddings/flair-embeddings.

[10] = Ž. Agić and I. Vulić. 2019. "JW300: A wide-coverage parallel corpus for low-resource languages," In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

[11] = by @stefan-it: https://github.com/flairNLP/flair/issues/614.

[12] = by @iamyiwha: https://github.com/flairNLP/flair/issues/80.

[13] = by @alanakbik: https://github.com/flairNLP/flair/issues/527.

[14] = G.R. Smidt, K.D. Graef and E. Lefever. 2024. "assyrugent / LREC-Coling2024,” https://github.com/assyrugent/LREC-Coling2024.

[15] = F. Thureau-Dangin. 1910. "Lettres et Contrats de l’époque de La Première Dynastie Babylonienne,” Paris: Librairie Paul Geuthner.

GHENT UNIVERSITY

FACULTY OF ARTS AND PHILOSOPHY

GustavRyberg.Smidt@Ugent.be

@assyrugent & @gurysm

@assyr_u

Thanks to our research partners at The Royal Museums of Art and History in Brussels and KU Leuven.

Funded by Belspo (BR/212/P/CUNE-IIIF-ORM)

GHENT UNIVERSITY

belspo

A&H

KU LEUVEN