DANES 2023

Ariel University & Tel Aviv University

19–21 February 2023

digitalpasts.github.io/DANES/

DigPasts()

ARIEL UNIVERSITY

TAD
The AI and Data Science Center
Tel Aviv University

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

First Meeting Of The Digital Ancient Near Eastern Studies Network

DANES 2023

"Computational Perspectives on Ancient Near Eastern Literature, Art and Material Culture"

Israel

19th – 21st February 2023

Abstracts' Booklet

Program

Sunday, February 19

Center for Artificial Intelligence & Data Science (TAD, Tel Aviv University)
Check Point Building 002
Tel-Aviv University

9:00-10:15 Opening Session & Keynote

09:00 – 09:15	Gathering
09:15 – 09:30	Opening Remarks Yoram Cohen (Tel Aviv University), Head of the School of Jewish Studies and Archaeology Jonathan Ben-Dov (Tel Aviv University), Humanities community, TAD Center
09:30 – 10:15	Keynote: *A sketch of digital Egyptology: Exploring Texts, Language, and Scripts with 21st c. Technology* Eliese-Sophia Lincke (FU Berlin)

10:15-10:30 Coffee Break

10:30-11:45 Session 1: Optical Character Recognition

Chair: Gabriel Stanovsky (Hebrew University of Jerusalem)

Preparing Multi-layered Visualizations of Old Babylonian Cuneiform Tablets for an AI OCR Training towards Automated Sign Recognition
Hendrik Hameeuw (KU Leuven), Katrien De Graef, Gustav Ryberg Smidt, Anne Goddeeris (Ghent University), Timo Homburg (Mainz University of Applied Sciences)

Cuneiform Sign Detection & Recognition
Yunus Cobanoglu, Enrique Jiménez (Ludwig Maximilian University of Munich), Luis Sáenz (Ariel University and Heidelberg University)

"Deeplomatics": Deep learning for automatic analysis of cuneiform texts
Morris Alper (Tel Aviv University), Shai Gordin (Ariel University)

15 min. discussion

11:45-12:00 Coffee Break

12:00-12:50 Session 2: Data Visualization

Chair: Hendrik Hameeuw (KU Leuven)

Visualizing a Cuneiform Collection - Remote talk
Jon Taylor (British Museum)

Visualization Of Metadata In 2D/3D Digital Cuneiform Artifacts
Michael Wamposzyc (Edinburgh Napier University)

10 min. discussion

13:00-14:30 Lunch Break

14:30-15:45 Session 3: Critical DH and Computational Methodologies

Chair: Hubert Mara (University of Halle-Wittenberg)

Creating a Blockchain for Cultural Objects: Introducing AGUR - Remote talk
Mark Altaweel (University College London)

Combined Analysis of Ancient Near Eastern Texts and Images: Gains, Challenges and Risky Paths of a Computational Approach to Hittite Religious Atmosphere – Remote talk
Alessandra Gilibert (Università Ca' Foscari, Venice), Michele Cammarosano (Università L'Orientale, Naples), Renzo Orsini (Università Ca' Foscari, Venice)

MTAAC: Machine Translation and Automated Analysis of Cuneiform Lan-guages
Heather D. Baker (University of Toronto)

15 min. discussion

15:45-16:00 Coffee Break

16:00-16:45 Session 4: Q&A Session

Venue: Circulation Hall of the Sourasky Central Library, Tel Aviv University.
Moderator: Shai Gordin (Ariel University)

Q&A Session on Ancient Language Processing
Niek Veldhuis, Eliese-Sophia Lincke, Heather D. Beaker, Hendrik Hameeuw

16:45-17:30 Session 5: Roundtables and Posters

Moderator: Stav Klein (Tel Aviv University and TAD Center)

Poster presentations and roundtable discussions on critical topics in digital ANE studies

Monday, February 20

Hosted by the Digital Pasts Lab (Ariel University)

08:30-09:30 Bus from Tel Aviv Hotel to Ariel University

09:30-10:30 Opening Session & Keynote

09:30 – 09:45 Opening Remarks
Itzick Shai (Ariel University), Vice president & Dean of R&D

09:45 – 10:30 Keynote: *ORACC Data Acquisition*
Niek Veldhuis (UC Berkeley)

10:30-10:45 Coffee Break

10:45-12:00 Session 6: Computational Stylistics

Chair: Luis Sáenz (Heidelberg University/Ariel University)

Independent Unsupervised Examination of the Distinction Between Texts of Priestly and Non-priestly Origins in the Books of Genesis and Exodus
Gideon Yoffe (Hebrew University of Jerusalem), Axel Bühler (Collège de France), Thomas Römer (Collège de France), Nachum Dershowitz, Eli Piasetzky, Israel Finkelstein (Tel Aviv University), Barak Sober (Hebrew University of Jerusalem)

From Digital Editions to Text Analysis Queries: The Old Babylonian Example
Marine Béranger (FU Berlin)

Stylometry of First Millennium Akkadian Texts: A Method for Authorship Attribution?
Avital Romach (Yale University), Shai Gordin (Ariel University)

15 min. discussion

12:00-12:15 Coffee Break

12:15-13:30 Session 7: Computational Text Analysis

Chair: Heather D. Baker (University of Toronto)

Tracing Word Meanings in Ancient Greek and Latin: Lessons Learnt from Using Computational Methods - Remote talk
Barbara McGillivray (KCL and The Alan Turing Institute)

Computational Approach to Emesal Code-switching
Aleksi Sahala (University of Helsinki)

Establishing Quantitative and Qualitative Bibliometrics for a Library of Assyriology
Adam Anderson (UC Berkeley)

15 min. discussion

13:30-15:00 Lunch buffet

15:00-16:15 Session 8: Linguistic Annotation

Chair: Aleksi Sahala (University of Helsinki)

kīma aqbûkum! Towards an Analysis of Everyday Language as Expressed in Old Babylonian Letters Using Natural Language Processing
Katrien De Graef, Gustav Ryberg Smidt, Els Lefever, Anne Goddeeris (Ghent University)

Linguistic Annotation of Cuneiform Texts using Treebanks and Deep Learning
Matthew Ong (Ariel University and UC Berkeley)

The Cuneiform Annotator: Annotations on Cuneiform Clay Tablets in Linked Open Data
Timo Homburg (Mainz University Of Applied Sciences), Hubert Mara (University of Halle-Wittenberg), Kai-Christian Bruhn (Mainz University Of Applied Sciences)

15 min. discussion

16:15-16:30 Coffee Break

16:30-17:20 Session 9: Networks

Chair: Adam Anderson (UC Berkeley)

Three Degrees of Separation: Networks in the City of Babylon during the Reign of Darius I (522–486 BCE) - Remote Talk
Jinyan Wang (University of Toronto)

Stone Tools and Archaeological Context at Tel Burna, Israel. A Network Perspective
Shih-Hung Yang (Ariel University)

10 min. discussion

17:20-17:30 Coffee Break

17:30-18:20 Session 10: Databases, modeling, and tools

Chair: Michael Wamposzyc (Edinburgh Napier University)

Advanced Computational Tools for Qumran Scroll Research
Nachum Dershowitz (Tel Aviv University)

Digital Prosopography: Problems and Prospects - Remote Talk
Laurie Pearce (UC Berkeley)

10 min. discussion

Tuesday, February 21

Archaeological Tour: "In the footsteps of Sennacherib: A tour of the Judean Shephelah"

08:00 Bus from the Tel Aviv hotel

09:00 Tour of Bet Shemesh and Tel Azekah conducted by Boaz Gross

13:00 Lunch

14:30 Closing forum

Chairs: Shai Gordin (Ariel University), Hubert Mara (University of Halle-Wittenberg) and Gabriel Stanovsky (Hebrew University of Jerusalem)

16:00 Return to Tel Aviv

Abstracts

Establishing Quantitative and Qualitative Bibliometrics for a Library of Assyriology
Adam Anderson (University of California, Berkeley)

Beginning more broadly, and then refining the questions: How should we make use of a digital library of 150k books, journals, and datasets in Assyriology? What can the 'distant reading' approaches using ML tell us about this corpus that scholars haven't already said? Will the tools and methods from computational linguistics provide qualitative and quantitative metrics for supporting the ongoing empirical research in Assyriology? The FactGrid Cuneiform Wikibase project seeks to explore these questions and to build reproducible methods for making factual (triple) statements about the numerous named entities found in the half-a-million cuneiform sources around the world.

"Deeplomatics": Deep learning for automatic analysis of cuneiform texts
Morris Alper (Tel Aviv University) and Shai Gordin (Ariel University)

Abstract: TBD

Creating a Blockchain for Cultural Objects: Introducing AGUR
Mark Altaweel (University College London)

Currently, blockchain technologies are limited in use in cultural heritage; however, many similar areas that deal with material objects collections have been adopting these technologies to facilitate knowledge, interaction, and information and object transaction. We present our efforts in developing an Ethereum-based blockchain, which we call AGUR, for cultural objects and heritage collections. The intent is to utilize blockchain digital ledgers to help promote the protection of cultural objects while also incentivising collaborative research and openness in sharing collections to the outside world. We discuss current technology used and where further developments might best be focused. Our efforts are intended to promote the use of blockchain as part of cultural object collections, such as in museums and collectors. Furthermore, we discuss the use of non-fungible tokens (NFTs) within our blockchain as incentives to share collections as well as potential drawbacks. We describe other incentives and developments to help promote the use of blockchain and wider benefits of its use for cultural collections. Challenges faced in areas such as adoption and support are also discussed. The tool is currently undergoing preliminary use by object collections and results of this are discussed. The tool is provided for free and released as an open source project.

MTAAC: Machine Translation and Automated Analysis of Cuneiform Languages
Heather D. Baker (University of Toronto)

This paper summarises the work of MTAAC, an international collaborative project involving Assyriologists, Computational Linguists and Computer Scientists from Toronto, Frankfurt and UCLA. The project was funded for two years (2017–2019) by SSHRC (Canada), DFG (Germany) and the NEH (USA) through the Trans-Atlantic Platform Digging into Data Challenge, a program that supports research projects that explore and apply new "big data" sources and methodologies to address questions in the social sciences and humanities. MTAAC aimed to develop methods and tools for the automated analysis and machine translation of cuneiform texts in transliteration, using Ur III Sumerian documents as a test corpus. These documents were chosen because of the relatively high degree of standardization of their contents, which made them particularly suitable as a test case for the application of machine translation and automated analysis. The project used Linked Open Data to formalize and make available the results of the automated data extraction, and its working method, code, and results are available in open access on the web. This ensures that the project's work can be replicated and modified as necessary, to facilitate the application of machine translation to other ancient language corpora.

From Digital Editions to Text Analysis Queries: The Old Babylonian Example
Marine Béranger (FU Berlin)

Archival texts make the substantial part of the epigraphic documentation from the Old Babylonian period (2004–1595 BC): from the 50,000 texts known for this period, ca. 34,800 are of administrative, legal, or epistolary nature. The content of these texts is a valuable source for historical, economical, and administrative studies. They also offer an important material for the study of the Old Babylonian language and script. Letters, in particular, contain colloquial and technical vocabulary, as well as metaphors and sayings; they offer a glimpse of the imaginary and thoughts of Ancient Mesopotamians. The study of this corpus is therefore also relevant for those who work on literature or art.

Since 2008, the "Archibab" database developed by Dominique Charpin (Collège de France, Paris) and his team provides access to online editions of thousands of archival texts —i.e. letters, contracts, and administrative texts— from the Old Babylonian period (www.archibab.fr). About 13,000 texts have currently been entered into this database; of them ca. 9,000 have been associated with an extensive set of metadata (place of writing, date, lemma and part-of-speech tagging, translation, etc.). In the last years, I myself have developed a project that makes it possible to export the data outside of Archibab, in order to work on the vocabulary and cuneiform signs of the texts. This study can be

conducted online, on a "txm" web portal. Txm is a text analysis software which offers different qualitative and quantitative tools. It allows to display a list of all the words attested in the corpus; to search for attestations of a word/a sequence of words; to study the spelling of a word in different cities and through the centuries; to analyze the different readings of the same cuneiform sign; to compare the contexts of a specific word/expression; and much more. This paper will present the different steps from digital editions to text analysis queries, and provide examples of possible research.

Cuneiform Sign Detection & Recognition

Yunus Cobanoglu, Enrique Jiménez (Ludwig Maximilian University of Munich), Luis Sáenz (Digital Pasts lab, Ariel University)

Optical Character Recognition (OCR) on natural languages using a supervised end-to-end approach has reached human level performance in the last decade. Nevertheless the current state of the art of OCR for Cuneiform Fragments lacks a large enough dataset for an end-to-end deep learning supervised approach similar to ones in spoken Languages.

In this article we tackle the issue of gathering a large enough data set and also present some benchmarks on the Task of Cuneiform Sign Detection (predicting bounding boxes) using state of the art Text Detection Models. On top of that we present a simple semi-supervised approach for Sign Recognition to make use of the large amount of transliterated fragments already in existence. We identify two obstacles in the task of OCR for cuneiform tablets: 1) Many different Dialects result in a number of sign classes of up to 2500 including Compound Graphemes of which ~ 900 are included in the well known Mesopotamisches Zeichenlexikon. The signs follow a power law distribution with many signs being very rare. Besides that many signs are either partially broken or completely unidentifiable. 2) Annotating Cuneiform Tablets requires experts which makes it difficult to gather large amounts within a short time. In spite of these obstacles we are still confident that even a subhuman level OCR method for Cuneiform Fragments will heavily accelerate the current research in Assyriology. As of now between half a million and two million fragments have been excavated of which less than 100,000 have been read or transliterated. OCR would enable Assyriologist to query for signs and, combined with a sign-to-reading language model, also for readings. This would make it much easier for Assyiriologist to navigate the vast space of untransliterated Fragments and find relevant Fragments for their research. To summarize in this article we present 1) a new Dataset 2) Benchmarking state of the art Text Detection Models on Cuneiform Signs 3) Proposing a semi-supervised Cuneiform Sign Recognition approach.

Advanced Computational Tools for Qumran Scroll Research
Nachum Dershowitz (Tel Aviv University)

Scripta Qumranica Electronica is the result of a five-year collaborative effort to fully integrate the high-resolution digital photographs of the Dead Sea Scrolls, imaged in recent years by the Israel Antiquities Authority, with the rich lexical database of the Qumranwörterbuch developed in Göttingen.

Qumranica provides a web-based working environment for scholars and the general public, with images and text shown side by side. Users can link regions of interest on images with segments of transcriptions. They can combine and adjust multiple images of a fragment (color, infrared, raking), and they can re-arrange virtual fragments on a digital canvas.

Tools and algorithms were designed to distinguish between the fragments appearing in a picture and the background in both old infrared images and the new multispectral ones, so users can manipulate virtual fragments without other artifacts. Algorithms have been developed to locate a fragment, based on a new image, among the many fragments in plates of fragments photographed at earlier dates and in different stages of deterioration, and to combine (register) old and new images. A pipeline has been designed to facilitate automatic alignment of transcriptions with images, involving layout analysis, line segmentation and imperfect automated text recognition, followed by algorithmic alignment, letter by letter, with the actual texts. Paleographic tools are also in development.

Qumranica is a joint project with Reinhard Kratz and Jonathan Ben-Dov. The main contributors to the development of the tools include: Taivanbat Badamdorj, Berat Barakat, Adiel Ben-Shalom, Bronson Brown-deVost, Gil Levi, Pinhas Nisnevitch, and Daniel Stökl Ben Ezra.

Combined analysis of Ancient Near Eastern texts and images: gains, challenges and risky paths of a computational approach to Hittite religious atmosphere
Alessandra Gilibert (Università Ca' Foscari, Venice)*, Michele Cammarosano (Università L'Orientale, Naples); Renzo Orsini (Università Ca' Foscari, Venice)

Ancient Near Eastern societies produced an impressive mass of both textual and material/visual evidence, which is traditionally studied with methods typical of the philological and archaeological disciplines respectively. Although there are obvious points of contact between the two types of sources, putting them in relation presents serious difficulties. These do not derive only from terminological and disciplinary boundaries, but also from the different nature of the sources themselves, with their specific and partly

divergent modes of communication (cf. the discussion on the concept of "multimodality"). Beyond important conceptual aspects, this state of affairs also presents very concrete practical problems at the level of data modeling and management. The problem is all the more acute when we consider issues relating to data encoding and to the typically fragmentary state of preservation of the sources, with the consequent need to guarantee the maximum possible flexibility in their recording and retrievability. These difficulties have been variously addressed - both with the elaboration of so-called formal ontologies and with authority control systems and specific applications for metadata management. In this paper, a philologist, an archaeologist, and a computer scientist present their experimental attempt to merge their perspectives and use computational methods to extract meaning from a composite corpus of ancient Near Eastern sources. Specifically, we ask if it's possible to develop a computer-aided methodology to study affective dimensions of the human past. In this endeavor, our current focus is the reconstruction of Hittite religious atmospheres. We base our approach on hard data derived from images of rituals, stage architectures, and a large body of written sources on cult and cult paraphernalia. We discuss how these different sources contrast and intersect, identify the gains and challenges of a computational approach to their study and illustrate a possible methodology to access past feelings connected to place, rituals, and the sacred.

kīma aqbûkum! ***Towards an analysis of everyday language as expressed in Old Babylonian letters using Natural Language Processing***

Katrien de Graef, Gustav Ryberg Smidt, Els Lefever, Anne Goddeeris (Gent University)

The Belspo funded CUNE-IIIF-ORM project brings together an interdisciplinary consortium of Assyriologists, museum curators, digital humanities experts, computational linguists and computer scientists from the Royal Museums of Art and History (RMAH) Brussels, Ghent University (UGent) and the University of Leuven (KULeuven). The consortium will curate RMAH's collection of cuneiform tablets for both scientific exploitation and social valorisation and it will explore new digital approaches to study, publish and enrich

corpora of cuneiform texts. Within the framework of this project, UGent based Assyriologists and computational linguists will set up a pilot study to analyse the language used in Old Babylonian (OB) letters using Natural Language Processing (NLP) techniques. OB letters are particularly suitable as dataset because (1) although in part formalized, they form an invaluable source for everyday vernacular, and (2) more than 5000 have been recovered, the majority of which are accessible in transliteration and translation through the series Altbabylonische Briefe (AbB) and on CDLI. Being the closest we can come to everyday spoken language, the language used in these letters forms the ideal basis for sociocultural linguistic analyses. Based on a first batch of letters from OB Sippar, selected from AbB, and linguistically annotated in ORACC (sup-

plemented by grammatical annotation), later extended by other Akkadian letters (annotated in ORACC and supplemented by grammatical annotation), we aim to develop machine learning approaches to perform semi-automatic text analysis of the letters. As a first step, we will investigate a machine learning approach to go automatically from transliteration to grammatical annotation (including Part-of-Speech, lemma and morphological information). The result of this step can then be used for two other NLP use cases, namely (1) automatic term extraction to detect the main keywords of the letters and (2) distributional semantic analysis to cluster semantically related terms in the letters. The proposed research, resulting in (semi-)automatic grammatical and semantic analyses, will allow us to answer sociocultural linguistic questions about the everyday language used in OB letters and in a later stage to discern between the language used in these letters and that used in other textual genres.
In this contribution, we will contextualise the pilot study within our project, expand on our choice of corpus and showcase potential use cases for implementing NLP methods on it.

Keywords: Old Babylonian Akkadian letters, Natural Language Processing, grammatical annotation, historical socio(cultural)linguistics.

Preparing multi-layered visualisations of Old Babylonian cuneiform tablets for an AI OCR training towards automated sign recognition

Hendrik Hameeuw (KU Leuven), Katrien De Graef, Gustav Ryberg Smidt, Anne Goddeeris (Ghent University,) Timo Homburg (Mainz University of Applied Sciences)

In the framework of the CUNE-IIIF-ORM project, a consortium consisting of the Royal Museum of Art & History in Brussels, UGent and KU Leuven (Belgium), we aim to train an Artificial Intelligence Optic Character Recognition (AI-OCR) model that can automatically annotate Old Babylonian documentary texts. In order to train the model, we have selected 200 Old Babylonian documentary texts that are to be manually annotated based on 2D+ images. These images are still raster files and they are manually generated based on interactive Multi-Light Reflectance (MLR) datasets created with the KU Leuven Portable Light Dome system.

For each side of a clay tablet containing cuneiform characters the rich MLR information is used to generate 12 different visualisations of that surface. One with general lighting conditions, eight with varying incident raking of light at predetermined angles, two automated line drawings based on the estimated surface orientations and one normal map. These are uploaded in the Cuneiform Annotator (Homburg et al. 2022: http://dx.doi.org/10.5334/joad.92), a Gitlab-based web application that uses a modified Annotorious-OpenSeaDragon instance (https://github.com/recogito/annotorious-openseadragon), in order to annotate areas of raster renderings depicting cuneiform

signs using the W3C Web Annotation Data Model Standard in RDF (https://www.w3.org/TR/annotation-model/).

The Cuneiform signs are identified by their Unicode codepoint and reading. By annotating the surface, line, character rotation and character index, the sign images can be related to positions in the transliteration; important for data quality control actions. A python script that ensures continuous integration will crop out the annotations and creates a corpus of sign images for all the twelve underlying visualisations. The corpus can be created on-demand from a Gitlab environment or on any other system executing the same generation script. All the individual sign forms with their IDs will be the core dataset for the further to developed and finetune AI-OCR training modal.

Title: A sketch of digital Egyptology: Exploring texts, language, and scripts with 21 st c. technology

Eliese-Sophia Lincke (FU Berlin)

In recent years, the integration of digital research methods and the growing availability of digitized texts has had a profound impact on the field of Egyptology, as well as on many other areas of Ancient Studies and the Humanities. During this talk, I will focus on digital research methods in Egyptian philology, i.e., textual scholarship and language studies.

I will highlight recent advancements in the encoding and display of Egyptian scripts and explore ongoing efforts to establish standards for metadata. Not only the potential of but also the challenges to these efforts will be mentioned. Thereafter, an overview of the main corpus projects for Egyptian and Coptic texts will be presented, including their respective annotation schemes and the NLP tools they provide. Additionally, I will delve into the preparation and (re)modeling of a Coptic legacy dataset for one of these corpora.

The utilization of machine learning techniques and their applications in text digitization (such as OCR and HTR) and automatic transcription, as well as paleography (pattern recognition) and text analysis (stylometry), will be explored as a rapidly growing and exciting branch of research. Lastly, I will briefly touch upon the integration of digital methods in university courses in Egyptology.

Tracing word meanings in Ancient Greek and Latin: lessons learnt from using computational methods

Barbara McGillivray (KCL and The Alan Turing Institute)

Over time, new words enter the language, others become obsolete, and existing words acquire new meanings. These phenomena are grounded in a fascinatingly complex mix of cognitive, social, and contextual factors, responding to language contact, emerging circumstances, cultural and socio-political changes, stylistic choices, and different communicative needs (Meillet 1958[1905-06]); Bréal 1964[1900]; Munat 2007, among many others). The past decade has seen a growing interest in automatic methods for semantic change (i.e. meaning change) detection from large corpus data (Tahmasebi et al. 2018; Armaselu et al. 2021), which have made it possible to conduct quantitative studies aimed at detecting broad patterns in the data. Most of these automatic detection methods rely on distributional semantics methods and trace the computational representation of a corpus-driven word's semantic profile (via vector embeddings) over time to identify if and when a potential change in the semantic profile may have occurred. Such systems perform well on large datasets spanning the last few centuries and have been tested on various modern languages (Hamilton et al. 2016; Kutuzov et al. 2021). Despite these encouraging results, there is still a lot of work to do to gain new in-depth insights into the mechanisms of semantic change using these methods.

First, they only focus on semantic change over a few centuries and do not account for semantic variation, with the exception of Perrone et al. (2019) who propose a genre-aware topic model for semantic change in ancient Greek and extend it to Latin data (Perrone et al. 2021). Second, they generally are only able to detect whether a word's semantics may have changed and when (Basile & McGillivray 2018), but do not provide further details about the mechanisms for or explanations of the change detected. Third, they are subject to detecting noise (Dubossarsky et al. 2017) and, because they are based on distributional data, they tend to conflate topical change and sense change (Shoemark et al. 2019).

In this talk I will present my research on developing computational models for semantic change detection in historical texts, particularly on ancient Greek and Latin. I will share my experience of working at different scales and in a range of interdisciplinary projects and the lessons learnt.

References

Armaselu, F., Apostol, E.S., Khan, A.F., Liebeskind, C., McGillivray, B., Truică, C.O., Utka, A., Valūnaitė Oleškevičienė, G. and van Erp, M. (2022). LL (O) D and NLP perspectives on semantic change for humanities research. *Semantic Web Journal*, 13:6, pp.1051-1080.

Basile, P. & McGillivray, B. (2018). Exploiting the Web for Semantic Change Detection. *21st International Conference, DS 2018, Limassol, Cyprus, October 29–31, 2018, Proceedings*, LNCS, volume 11198 (DS2018), 194-208. https://doi.org/10.1007/978-3-030-01771-2

Bréal, Michel (1964) [1900]. Semantics: Studies in the Science of Meaning. Trans. By Mrs. Henry Cust. New York: Dover.

Dubossarsky, H., Grossman, E., & Weinshall, D. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. *Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark*, 1, 1147–1156.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of ACL 2016* (pp. 1489–1501). http://arxiv.org/abs/1605.09096

Kutuzov, A., & Pivovarova, L. (2021). RuShiftEval: a shared task on semantic shift detection for Russian. In Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue. Redkollegija sbornika.

Meillet, Antoine (1958 [1905-06]). Comment les mots changent de sens. In Meillet, Antoine, Linguistique historique et linguistique Générale, 230-280. Paris: Champion.

Munat, J. (2007) Lexical Creativity, Texts and Contexts. Amsterdam: John Benjamins.

Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J. Q., & McGillivray, B. (2019). GASC: Genre-Aware Semantic Change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy. Association for Computational Linguistics.

Perrone, V., Hengchen, S., Palma, M., Vatri, A., Smith, J. Q., & McGillivray, B. (2021). Lexical semantic change for Ancient Greek and Latin. Computational approaches to semantic change, 287-310.

Shoemark, P., Ferdousi Liza, F., Nguyen, D., Hale, S., & McGillivray, B. (2019) Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

Tahmasebi, N., Borin, L. and Jatowt, A. (2018). Survey of Computational Approaches to Lexical Semantic Change, arXiv: Computation and Language.

Digital Prosopography: Problems and Prospects
Laurie Pearce (UC Berkeley)

Abstract: TBD

Stylometry of First Millennium Akkadian Texts: A Method for Authorship Attribution?
Avital Romach (Yale University), Shai Gordin (Ariel University)

Abstract: TBD

Computational Approach to Emesal Code-switching
Aleksi Sahala (University of Helsinki)

One of the great remaining mysteries in the study of the Sumerian language is the nature and origin of its only known variety, Emesal, which made a somewhat counterintuitive appearance in ancient Mesopotamian texts only after the extinction of Sumerian as a spoken language around 2000 BCE. Although it is well known that Emesal words mostly occur in liturgical texts and lamentations, it is not yet understood what conditions triggered the code-switching from Sumerian into Emesal within certain parts of these texts, and why Emesal became a part of certain Sumerian compositions in the first place. With the most comprehensive digital collection of Emesal texts now available in the Open Richly Annotated Cuneiform Corpus, we aim to analyze the Emesal texts with Natural Language Processing methods to shed light on these questions.

Based on our team's philological analysis, at least in one isolated case the code-switching between the language varieties is not spontaneous, but rather constrained by semantic differences between the Emesal and Standard Sumerian words, which may indicate that at least some Emesal words were used in restricted contexts. We aim to determine the extent of regularity behind the code-switching over the entire Emesal vocabulary by using word embedding models, which we use to measure how predictable the choice of the language variant is in certain linguistic contexts.

Our preliminary results indicate varying degree of predictability in the use of Emesal words, and we hope that through further philological and statistical analysis of these results we are able to discover and explain previously unnoticed regularities behind the code-switching. By identifying and explaining the contexts in which code-switching between Sumerian and Emesal takes place, and especially by studying whether those contexts varied in different time periods, text genres, or geographical regions, we aim to understand this long dead language on its own terms, and to improve our knowledge on the origins and development of Emesal.

Visualising a Cuneiform Collection
Jon Taylor (British Museum)

The collection of the British Museum is estimated at over 8 million objects, acquired over more than 250 years of history. About 130,000 of these objects are tablets inscribed with cuneiform. As such, it is the largest collection of cuneiform outside of Iraq. This part of the collection dates back as far as 1821, and extends into the 21st century. It thus represents two centuries of collecting; a longer history than most other collections. Its scope ranges over 3,000 years, all kinds of genres, and sites in what are now Iran, Iraq, Turkey, Syria, and Egypt. The Museum often serves as an archetype of western museum collecting. Many assumptions are made about the collection and its acquisition.

This presentation analyses the cuneiform sub-collection, providing an evidence base for understanding it correctly. I compare the patterns observed with those recently described for the Museum collection as a whole, highlighting the specificities of this part of the collection. How much was excavated, donated, or purchased? Who from? What are the temporal patterns and how can they be explained? Which sites are represented and how do purchases relate to excavations? This work informs our understanding of a museum collection, and will also form part of future efforts to understand collections around the globe, and the connections between them.

Title: ORACC Data Acquisition
Niek Veldhuis (UC Berkeley)

This talk will discuss techniques for acquiring data that are freely available from ORACC. Different types of projects need different data structures. A typical Bag of Words approach, such as word clouds, distance metrics based on a Document Term matrix, or Topic Modeling, will need a fairly straightforward data representation that can be extracted from the ORACC JSON files with a few lines of code. Other types of projects, such as those based on word embeddings, Social Network Analysis, tree-banking, or any other approach that involves tokens understood in context, will require a more fine-grained data representation. ORACC JSON represents such data, but extracting them from the JSON files is somewhat more challenging. A deeper understanding of how textual data is represented in ORACC JSON will enable a researcher to access the data in the format she needs.
A few words about the future of cuneiform data representation in the light of the FAIR data principles (http://go-fair.org) will conclude this talk.

Visualisation of Metadata in 2D/3D Digital Cuneiform Artefacts

Michael Wamposzyc (Edinburgh Napier University)

As postulated by Bertin (1967) and later Ware (2010) – a necessary process of systematic reduction and simplification takes place in the production of any aesthetic form of scientific visualisation. From an epistemic perspective the 2D/3D data translation and representation utilised by artificial intelligence, machine learning, and natural language processing, presents a new set of challenges for the disciplines of Information Design and Data Visualisation. The proposed exploration addresses those challenges by introducing a concept of 'augmented reification' i.e. the constructive or generative aspect of perception, by which the experienced percept contains more explicit spatial information than the sensory stimulus on which it is based and which it is augmenting.

The use of visualisation of the metadata in this context, hopes to expand the invisible digital datasets and visually indicate the exploratory data connections and translation processes involved. The core research question asks about the theoretical implications and practical consequences of reification during visual perception of the Interfaces and the visual representations developed within the Edinburgh's Creative Informatics Research Grant for 2D cuneiform script and 3D cuneiform artefacts. The aim is to explore and discuss what epistemic opportunities and challenges the 'augmented reification' of metadata visualisation in AR/VR/MR environments could offer for the wider disciplines of Digital Humanities and Cultural Heritage Studies.

Three Degrees of Separation: Networks in the City of Babylon during the Reign of Darius I (522–486 BCE)

Jinyan Wang (University of Toronto)

Babylonian private archives from the sixth and fifth centuries BCE record the economic interactions between people with various legal and social statuses. Social Network Analysis provides a useful framework and method to understand these interactions. In this presentation, I reconstruct the networks of Babylonian urban dwellers during the reign of Darius I (522–486 BCE) based on 803 tablets from ten private archives from Babylon. I aim to examine the patterns of different social groups' economic relationships and the impact of the emergence of entrepreneurs in the society through three steps of network reconstruction and analysis. First, I reconstruct the whole network and provide an overview concerning the distribution and connectivity of different social groups, such as women, slaves, foreigners, and officials. Second, I reconstruct the subnetworks of these social groups, and examine the frequency and patterns of economic interactions within each subnetwork. These analyses suggest that members of each social group were not particularly interconnected within their own groups. Third, I use community detection to partition the network into clusters and examine the underlying

structure of the network and the different roles people played in the business activities. The results reveal the bridging impact of entrepreneurial collaborations between economic players with various legal and social statuses.

Stone Tools and Archaeological Context at Tel Burna, Israel. A Network Perspective

Shih-Hung (Benjamin) Yang (Ariel University)

Archaeologists have been applying Network Analysis to understand ancient societies since the late 1980s. The development of this method usually combines with GIS technology to describe the interaction between sites, routes, and the environment. This paper applies the same concept on a smaller scale, which aims to view the relations between archaeological remains and its context based on a network perspective. The site which will be studied here is Tel Burna, Israel. Tel Burna is located in the region of the southern Shephelah, about 35 kilometers southwest of Jerusalem. Since 2011, the excavation has exposed twenty squares on the western slope and a Canaanite cultic enclosure dated to the thirteen century BCE (Building 29305) was revealed. Over a hundred stone tools were found among twenty squares with varying functions, including grinding stones, pounding stones, perforated stones, etc. Each stone tool is given a set of approximate coordinate value to calculate the spatial distance between stones. Afterward, a spatial network was generated accordingly. Position, typology, functionality are the three factors to define the node and the edge is determined by the distance and context between stone tools. The values of the betweenness centrality of stone tools were calculated as well, in which, the space of the enclosure can be better understood from the quantitative viewpoint. Python NetworkX is the compiler in this research along with other digital application. The analyzing result demonstrates the network analysis can reflect on the archaeological context and the spatial structure at the certain level. In addition, this research may bring the idea of how the network analysis can possibly be applied within a few excavating squares.

The Cuneiform Annotator: Annotations on cuneiform clay tablets in linked open data

Timo Homburg[1], Hubert Mara[2], and Kai-Christian Bruhn[1]

[1] *Mainz University Of Applied Sciences, Germany*
[2] *University of Halle-Wittenberg, Germany*

December 7, 2022

Motivated by the demands of assyriologists for digitality, we develop tools for digital scholarly editions as known in Digital Humanities (Sahle, 2016). These modern editions approach will also unlock significant potential for studying cuneiform texts. In this digital turn, the provision of transliterations increasingly relies on interoperable specifications, and many texts are accessible accordingly in corpora such as ORACC[1]. However, linking observations and transliterations by specialists with regions in digital images of the object has not yet been uniformly solved. The fact that cuneiform tablets are objects also being captured in $3D$ poses a challenge for the linkage of $1D$ text with $2D$ images and $3D$ datasets. Besides photographs, we create $2D$ images from $3D$ models. Furthermore, acquisition devices like the *Leuven dome* create $2D$ image stacks, which can also be rendered using different algorithms (Hameeuw and Willems, 2011).

With the Cuneiform Annotator, a Git-based tool, we propose a response to the demand for concise linkage between $1D$, $2D$, and $3D$. We choose openly specified standards to achieve the interconnectivity of image resources to transliterations and other relevant details on the surface of the established archaeological object. Those guarantee a stable and sustainable digital affiliation and enable the digital outcomes to become available in different scholarly contexts.

Our Cuneiform Annotator allows annotating renderings of 3D meshes in a web-based environment and relating them to transliteration contents. Annotations are created using the W3C Web Annotation Data Model (Sanderson et al., 2017) in RDF (Cyganiak et al., 2014) using modified versions of the open source JavaScript tools Annotorious-OpenSeaDragon[2] and recogito-js[3] (Simon et al., 2017). In a postprocessing stage, using continuous integration, the annotated information can and has been used to create digital artifacts such as cropped image information and 3D annotations (Homburg et al., 2022).

The annotator comprises two parts as shown in Figure 1. The left-hand side part allows annotating areas of interest (firing holes, broken areas, cuneiform signs, words, lines, cuneiform wedges) on images of 3D renderings. The right-hand side part allows entering transliterations with semantic and linguistic annotations. Annotations on images are linked to transliterations using JavaScript and in the RDF representation of the annotations.

Figure 2 shows text annotations in the given transliteration environment. Annotations may link to semantic concepts in knowledge graphs such as Wikidata (Vrandečić and Krötzsch, 2014), linguistic concepts like part of speech tags in linked data or lexemes from semantic dictionaries (McCrae et al., 2017).

We will show the LOD models developed, the capabilities of our Cuneiform Annotator, and their contributions to developing a cuneiform LOD cloud using test cases with single Akkadian tablets from

[1] http://oracc.museum.upenn.edu/projectlist.html
[2] https://github.com/recogito/annotorious-openseadragon
[3] https://github.com/recogito/recogito-js

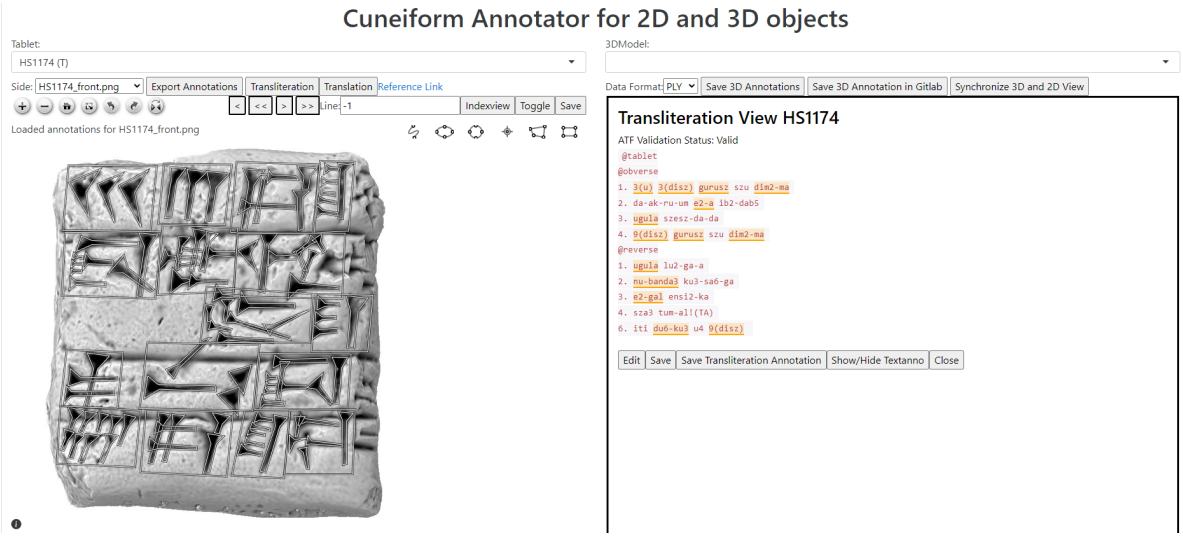


Figure 1: Annotations in the Cuneiform Annotator: Sign and wedge annotations on the left-hand side and a transliteration text with text annotations on the right-hand side on renderings of cuneiform tablet HS 1174

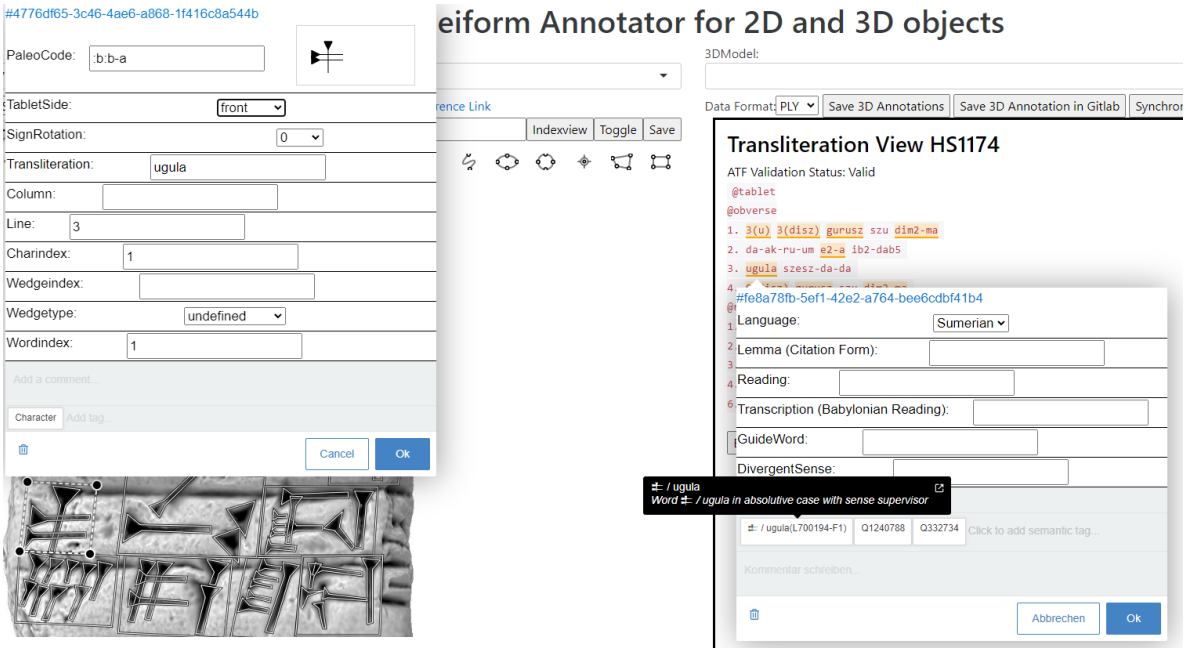


Figure 2: Annotation contents of the image and textual annotations: On the left, the cuneiform sign "ugula" is annotated on the 2D rendering, assigned the *PaleoCode :b:b-a* and indexed. On the right, in the transliteration, the word form L700194-F1 (ugula) of lemma L700194 (ugula) with the sense L700194-S2 (overseer) linked to the Wikidata concept Q1240788 (supervisor) is annotated. In the same way, any linguistic annotation attached to the word form L700194-F1 (ugula), here the Q332734 (absolutive case) can be added.

the HT project[4], and Cune-IIIF-orm[5] project and two Sumerian tablets from the Hilprecht Collection and the Tell Chuera excavation[6].

Acknowledgements

This work was partially funded by German Science Foundation (DFG) under grant number 424957759[7] We are grateful for all the feedback and discussions with our colleagues in Assyriology from the Universities of Mainz, Ghent, and Heidelberg.

References

Cyganiak, R., Wood, D., and Lanthaler, M. (2014). RDF 1.1 concepts and abstract syntax. W3C recommendation, W3C. https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

Hameeuw, H. and Willems, G. (2011). New visualization techniques for cuneiform texts and sealings. *Akkadica*, 132(2):163–178.

Homburg, T., Zwick, R., Mara, H., and Bruhn, K.-C. (2022). Annotated 3d-models of cuneiform tablets. *Journal of Open Archaeology Data*, 10.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Sahle, P. (2016). What is a scholarly digital edition? *Digital scholarly editing: Theories and practices*, 1:19–39.

Sanderson, R., Ciccarese, P., and Young, B. (2017). Web annotation data model. *W3C recommendation*, 23.

Simon, R., Barker, E., Isaksen, L., and de Soto Cañamares, P. (2017). Linked data annotation without the pointy brackets: Introducing recogito 2. *Journal of Map & Geography Libraries*, 13(1):111–132.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

[4] https://gepris.dfg.de/gepris/projekt/424957759
[5] https://www.ugent.be/lw/en/services/library/rd/cuneiiiform.htm
[6] https://www.orientarch.uni-halle.de/digs/chuera/chu96_e.htm
[7] https://gepris.dfg.de/gepris/projekt/424957759

Draft version December 3, 2022
Typeset using LaTeX default style in AASTeX631

An Independent Unsupervised Examination of the Distinction Between Texts of Priestly and Non-priestly Origins in the Books of Genesis and Exodus

Gideon Yoffe [1,2] Axel Bühler [3,4] Thomas Römer [3] Nachum Dershowitz [5] Eli Piasetzky [6] Israel Finkelstein [7] and Barak Sober [1]

[1] *Dept. of Statistics and Data Science, the Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel*
[2] *Dept. of Archaeology and Ancient Near Eastern Cultures, Tel Aviv University, Ramat Aviv 6997801, Israel*
[3] *Collège de France, 11 Pl. Marcelin Berthelot, 75231 Paris, France*
[4] *Faculté de théologie, Université de Genève, 1205 Geneva, Switzerland*
[5] *School of Computer Science, Tel Aviv University, Tel Aviv University, Tel Aviv 6997845, Israel*
[6] *School of Physics and Astronomy, Tel Aviv University, Tel Aviv University, Tel Aviv 6997845, Israel*
[7] *Dept. of Archaeology and Ancient Near Eastern Cultures, Tel Aviv University, Tel Aviv University, Tel Aviv 6997845, Israel*

1. INTRODUCTION

We examine the hypothetical distinction between texts of priestly (P) and non-priestly (nonP) origin in the books of Genesis and Exodus, for which exists a surprisingly large agreement amongst biblical scholars (e.g., 8; 9; 5). Examining this distinction with an independent, unsupervised computational methodology would establish a measure of confidence therein and encourage its application to additional instances of biblical texts, especially those of greater controversy, where our approach could help tilt the scale in favor of one hypothesis over another.

2. METHODOLOGY

We intertwine descriptive and inferential statistics. The first is used in text classification and interpretability analyses, whereas the latter quantifies uncertainty through hypothesis testing. While descriptive statistics were successfully applied to specific texts (e.g., 7; 10), we are unaware of similar studies where uncertainty quantification was considered. Furthermore, identification of literary features *responsible* for the classification, as opposed to cluster-wise significant feature detection (e.g., 6; 2; 11), is novel for stylometry.

2.1. *Corpus*

We use STEPBible[1] (digitized Leningrad codex), with its morphological and semantic tags for all words, prefixes, and suffixes. We consider two representations of the text: word-wise and a grammatical representation by phrase-dependent parts-of-speech (pdps).

We obtained a scholarly labeling assigning each verse in Genesis and Exodus as P/nonP.

2.2. *Parameterization and Embedding*

Our underlying assumption is that significant literary differences between texts manifest in simple linguistic parameters. Therefore, we consider three parameters, distinct combinations of which result in different classifications. These are: (**1**) word-/pdp-wise representations, (**2**) n-gram size, the length of sequences of consecutive words/pdps, and (**3**) running-window size, the number of verses surrounding the original, providing additional context.

We use tf-idf to encode each verse, assigning a relevance score to each feature in the context (1). The critical consideration behind choosing this traditional embedding is that it allows interpretability of the results, unlike neural-net-based language models, which are convoluted (e.g., 3; 4).

2.3. *Optimization*

We use k-means to classify the embedded verses and use an unbalanced accuracy measure to quantify the goodness of classification. We perform cross-validated grid-search on a range of running-window and n-gram sizes for words/pdps, identifying the combination that yields the highest accuracy (Fig. 1).

[1] https://github.com/STEPBible/STEPBible-Data

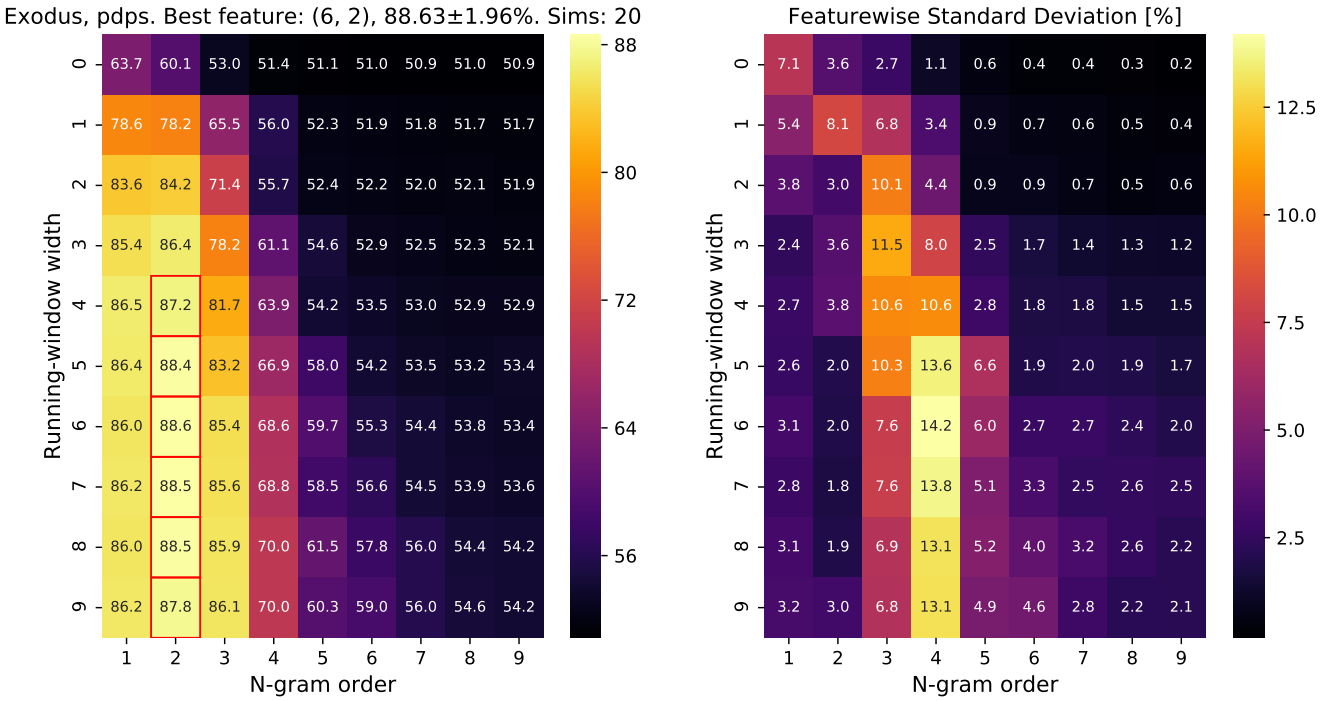


Figure 1. Cross-validated grid-search optimization for Exodus (pdps). **Left Matrix**: averaged accuracy over 10 simulations with respect to combinations of window sizes (y-axis) and n-gram sizes (x-axis). The best-fit combination yields $\approx 89.51.96\%$ accuracy, and the cells of feature combinations within 1σ thereof are marked with red. **Right Matrix**: standard deviation of the left matrix.

2.4. *Testing and Validating*

Through hypothesis testing, we aim to establish statistical significance of the distinction. We perform two tests, under the null hypothesis that our labels were randomly assigned: (**1**) arbitrary permutations, and (**2**) *cyclic* permutations, where we generate the null by shifting the labeling cyclically, seeking to conserve implicit correlations between consecutive verses.

2.5. *Feature Importance*

Minimizing k-means loss is equivalent to maximizing *inter-cluster* variances. Leveraging this, we extract a vector of feature-wise importance that maximizes the inter-cluster variance found by 2-means. This vector allows us to trace the features most responsible for the classification (Fig. 2).

3. CONCLUSIONS

We examined the hypothesized P/nonP distinction in Genesis and Exodus and introduced a novel computational and statistical methodology for text stylometry that is essentially independent of-, but in synergy with- established philological practices. We sought an optimal single feature—a combination of running-window and n-gram sizes, and extracted features that contribute most to classification and their respective proportions. We achieve a 73% and 90% (balanced) accuracy for Genesis and Exodus. The difference in accuracy between the two seems to arise from the more sporadic distribution of P in Genesis, as opposed to a more formulaic one in Exodus.

REFERENCES

[1]Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[2]Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 333–342, 2010.

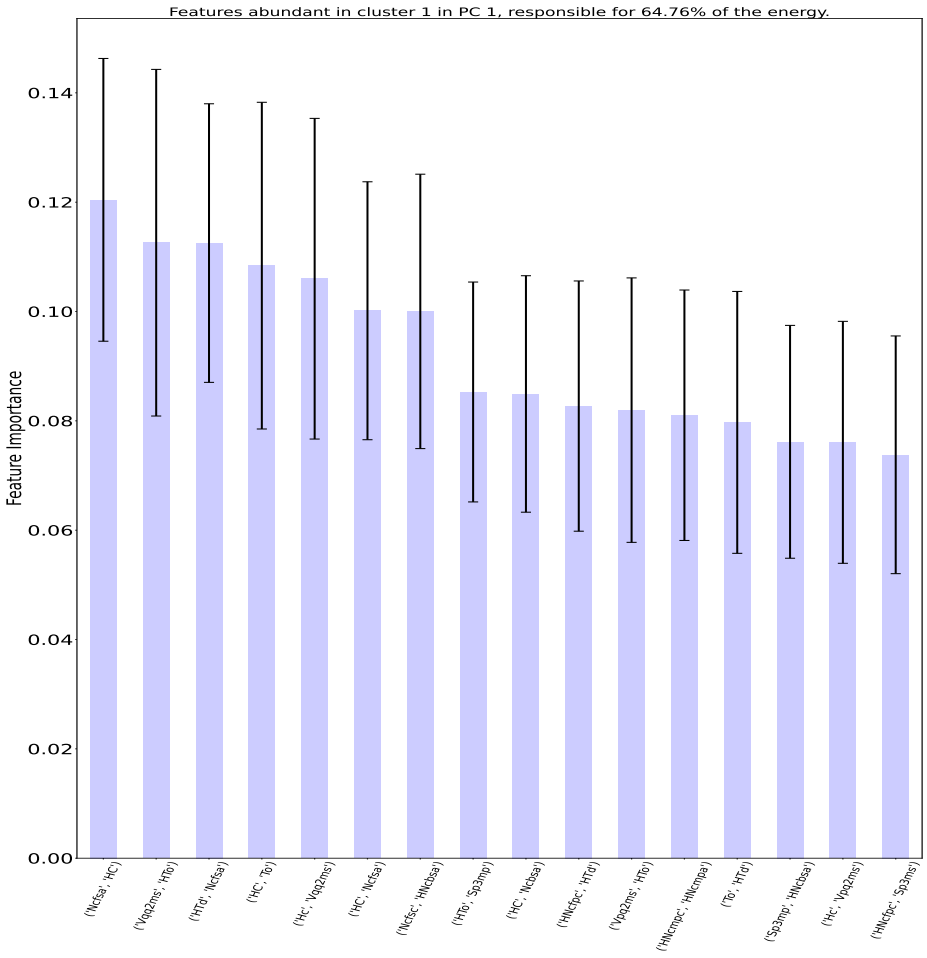


Figure 2. Feature importance bar plot for Exodus embedded as bigrams of pdps with window width 4. Here, 100% of the distinction is made by features that are abundant in the nonP cluster.

[3]Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. Interpretability of deep learning models: A survey of results. In *2017 IEEE Smartworld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, pages 1–6. IEEE, 2017.

[4]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5]Avraham Faust. The world of P: The material realm of priestly writings. *Vetus Testamentum*, 69(2):173–218, 2019.

[6]Eduardo R. Hruschka and Thiago F. Covoes. Feature selection for cluster analysis: an approach based on the simplified silhouette criterion. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, volume 1, pages 32–38. IEEE, 2005.

[7]Mike Kestemont, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, 63:86–96, 2016.

[8]Israel Knohl. *The Sanctuary of Silence: The Priestly Torah and the Holiness School*. Eisenbrauns, 2007.

[9]Thomas Römer. From the call of Moses to the parting of the sea: Reflections on the priestly version of the Exodus narrative. In *The Book of Exodus*, pages 121–150. Brill, 2014.

[10]Mayuri Verma. Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools. *International Journal of Computer Applications*, 168(8):39–45, June 2017. doi:10.5120/ijca2017914486.

[11]Guo-Niu Zhu, Jie Hu, Jin Qi, Jin Ma, and Ying-Hong Peng. An integrated feature selection and cluster analysis techniques for case-based reasoning. *Engineering Applications of Artificial Intelligence*, 39:14–22, 2015.

FactGrid Cuneiform Discovery Project: Building Linked Open Data Repositories

- **FactGrid AWCA Google Drive & Google Colab**
- **FactGrid Cuneiform AWCA GitHub Org. Repos**

"Data is a precious thing and will last longer than the systems themselves."

Tim Berners-Lee, Inventor of the World Wide Web

This project is inspired by the durability of the data preserved in the oldest writing system known to mankind, called **cuneiform**. There are approximately a half-million artifacts with cuneiform writing spread all over the planet. Many of these objects are not even photographed, let alone translated. Scholars in this field have made a number of relational text databases, in order to identify these objects housed in museums and private collections, and while these databases have helped create a system of identificaiton and textual analysis, they have yet to be **linked** together to each other and to the existing scholarship.

FactGrid is a Wikibase triplestore designed for historical research, which makes this the ideal hub for linking the existing scholarship, both primary and secondary sources, for every cuneiform artifact in publication. As a graph database (using RDF triples), this system allows for an expansive number of relationships to exist between any objects or entities, and for these relationships to be classified for greater discoverability. Additionally, the structure of a graph database provides simple solutions to duplication, disambiguation, and alternative editions of the same entity, which allows for a more comprehensive approach for linking the different datasets in various languages. Because the database is in Wiki(data) format, this provides an open forum for editing and collaborating internationally, which makes the data more robust, timely, and sustainable.

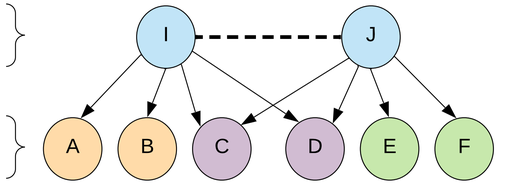
BERT & D2V

Entities I and J can be related based on shared **references**.

docs A, B, C, & D have similarities based on shared **contexts**.

Concepts of Linked Data and bibliographic coupling

Entities I and J are items with URIs, for example in Wikidata. Such items can be attested bibliographically and cited by documents C and D. The documents which cite these entities can be bibliographically coupled and their relational values can be measured by a count-frequency and cosine similarity scores.



Topic Model (LDA)

Entities I and J are assigned the same topic number. Topic weight determines directedness of edges.

Sources are more representative of a topic or genre, based on TF-IDF + LDA

The goal for this project is to link every cuneiform artifact on record to the primary and secondary sources, which make explicit reference to the given object, and to extend this referencial system to include the entities named on each artifact
(i.e. people, places, and things)

PDF (OCR)

html (no OCR)

Catalog for Drive (CSV), fields:
- (google) 'id'
- (file) 'name'
- 'mimeType'
- 'md5Checksum'
- 'webViewLink'
- 'createdTime'
- 'modifiedTime'
- 'size'

Page Splitting Orientation

Google ID integration

Batch processing

Language detection

OCR
tesseract
- raw TXT
- Google ID
- Doc ID
- Page ID
- Page count
- File name
- Catalog info

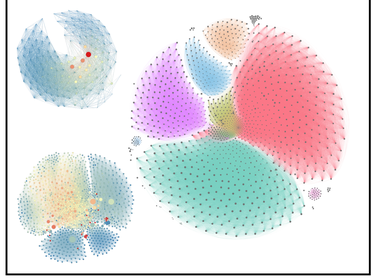
ocr-output (CSV), fields:
- Google ID
- File name
- Page ID
- Page Count
- Catalog meta
- Language(s)
  - Scripts

Trans-lations
- ocr-output
- segment
- google translate

NER
- PNs
- RNs
- GNs
- Bib

NLP networks
- TopicModel
- D2V
- W2V
- BERT+
- GPT+

Linked Data



Discovery Students

Aidan Curran	2022
Giselle Fuchs	2022
Minoo Kim	2022
Qianlin Wang	2022
Win Moe	2022
Ziyue Wang	2022
Circle Chen	2021-22
Conner Mi	2021-22
Daisy Wang	2021-22
Floyd Fang	2021-22
Kevin Gao	2021-22
Tina Chen	2021-22
Zaid Maayah	2021-22

Our Cuneiform project in FactGrid is building language support for all languages written in cuneiform, a writing system used for about 4000 years (from 3200 BCE to 50 CE). We are working with more than 350k documents to build dictionaries for these languages and social network graphs for the people, places, other entities named in these texts.

The main challenge we're working on is how to build reproducible workflows for linking four online databases of cuneiform sources (each with more than 100k documents) with two datasets of secondary sources. We're using Python notebooks (ipynb) to harmonize these open source databases, and we are linking the results using FactGrid Cuneiform, which is a triple store (or database for RDF triple statements). Our work helps us deepen our knowledge of Python for NLP and SparQL, the query language for Wikidata.

cdli

Oracc

ETCSL

BDTNS

KeiBi

AfO

URL + URI
Linking
- CDLI
  - catalog
  - text 350k
- ORACC
  - catalog
  - text 200k?
- BDTNS 150k
- ETCSL
- KeiBi
- AfO Register

PNs
names
- NER match
- ocr-output
- CDLI [ATF]
- ORACC [PN]
- edge list
- Wikidata
- FactGrid
- Import (CSV)

RNs
rulers
- NER match
- ocr-output
- CDLI [dates]
- ORACC [RN]
- Period
- Dynasty
- Wikidata
- FactGrid
- Import (CSV)

GNs
places
- NER match
- ocr-output
- CDLI [provenience]
- ORACC [GN]
- Wikidata
- FactGrid
- Import (CSV)

Bib
citations
- ocr-output
- Author
- Title
- Date
- **Pub**
- **Abb.**
- **oclc**
- **isbn**
- **doi**

zenodo

GitHub

WIKIDATA

FactGrid a database for historians

QuickStatements (CSV) import to Wikidata & FactGrid

SparQL Query for exporting whole datasets

End-to-end pipeline completion for participating databases

FactGrid:Cuneiform Project
Factgrid

FactGrid Cuneiform & Wikidata Lexemes

WIKIDATA LEXEMES
Wikidata:Lexicographical data
Wikidata

cuneiform clay tablet
Wikidata

personal name
Wikidata

historical period
Wikidata

ancient city
Wikidata

toponym
Wikidata

dynasty
Wikidata

archaeological site
Wikidata

Abbreviations for Assyriology
Ox

WIKIDATA BOOKS
Wikidata:WikiProject Books
Wikidata