

WHITE PAPER v.2

**Machine Translation and Automated Analysis of  
Cuneiform Languages (MTAAC)**  
2017–2020

(T-AP Digging Into Data Challenge, Round 4)

Heather D. Baker (PI, University of Toronto)

Christian Chiarcos (Co-PI, Goethe University of Frankfurt)

Robert K. Englund (Co-PI, UCLA)

Émilie Pagé-Perron (Project Coordinator; acting Co-PI, UCLA, May-June 2020)

Original submission: September 30, 2019

Updated submission: September 30, 2020

Note: This updated version incorporates results obtained by the Frankfurt and UCLA teams after the end of the Toronto team's funding in June 2019.

# Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC)

## Introduction

Ancient Mesopotamia, birthplace of writing, has produced vast numbers of cuneiform tablets that only a handful of highly specialized scholars are able to read. The task of studying them is so labor intensive that the vast majority have not yet been translated, with the result that their contents are not accessible either to historians in other fields or to the wider public. This project aimed to develop and apply new computerised methods to translate and analyse the contents of some 67,000 highly standardised administrative documents from southern Mesopotamia from the 21st century BC (the Ur III period). By automating these basic but labor-intensive processes, our goal was to free up scholars' time. The main objectives of the project were to:

- Formulate, test and evaluate methodologies for the automated analysis and machine translation (MT) of transliterated (i.e., transcribed sign-by-sign) cuneiform documents, and provide state-of-the-art technology to specialists in the field;
- Make available the translation of a specific and representative set of cuneiform documents to scholars in related disciplines and to a networked public;
- Provide new data for the study of the language, culture, history, economy and politics of the ancient Near East by harvesting the linguistic byproducts of the translation and information extraction processes;
- Formalize these new data utilizing Linked Open Data (LOD) vocabularies, and foster the practices of standardization, open data and LOD as integral to projects in digital humanities and computational philology.

In concrete terms, this project has developed methods and tools for the processing, analysis and translation of the Sumerian language. It has generated newly translated, annotated and extracted data, and provided the means to make these data available to a wider audience.

## Method, Progress, and Results

This section of the White Paper is structured according to the work packages outlined in our original proposal.

### *Preprocessing and Morphology*

The aim of this work package was to convert transliterations to a machine-readable format, correct mistakes, leverage spelling differences (this is comparable to transcription, but can be more imprecise), and to model the morphology for marking the texts, including reviewing existing specifications, defining principles, and preparing data. These deliverables were met by preparing a survey and

guidelines for transcription rules and morphology, and a collection of rule-based normalization algorithms.

### *Morphological Annotation*

The objective here was to create training data, and to evaluate and combine rule-based and machine-learning-based morphological annotation. In the first place we carried out manual annotation of the morphology of 371 Ur III texts. We also prepared a rule-based lemmatizer that is now state-of-the-art for the lemmatization of Sumerian. We developed an example-based morphological analyser with web-interface which also includes an error checker. When ready, texts are automatically uploaded to a Github repository.

### *Syntactic Annotation*

We aimed to develop an annotation scheme for syntactic annotations (dependencies), annotate 10,000 lines (sentences) with the scheme, provide rule-based syntactic parsers and create automatic syntactic annotations for the rest of the data, and link the annotations to OLiA (Ontologies of Linguistic Annotation). We succeeded in preparing an annotation scheme for Sumerian syntax and linking it to OLiA. We also prepared the tools and models required for the manual and automated annotation of Sumerian, but we were not able to prepare a substantial manually annotated corpus. This is because preparing the required tools for the annotators to edit annotations needed more time than expected. However, our rule-based parser for syntactic dependencies is now state-of-the-art in the field, and we can now produce good quality automated syntactic annotations of Sumerian, especially in the case of texts classified in the genre “royal” (ETCSRI data). For this we used a rule-based SHIFT-REDUCE algorithm implemented in CoNLL-RDF; the set of rules is defined by ordered SPARQL updates.

The UCLA and Frankfurt teams continued to work on syntactic annotation during their respective cost-neutral extensions, and explored alternative routes to create annotations. For a sub-corpus of 15,000 lines, syntactic dependencies have been projected from the English translations, transformed in accordance with the MTAAC annotation scheme, and partially manually verified. Furthermore, rule-based components have been developed to detect commodities and transactions as frequently found in administrative texts (see “Information Extraction,” below). These produce partial annotations only, but their analyses have also been transformed to the MTAAC scheme and increase the amount of syntactically-annotated Sumerian data. Additionally, we also have improved and implemented a context-free grammar parser based on a model designed to handle transactions of sheep (Jaworski 2008). We expect to combine syntactic annotations resulting from these various tools to further enrich our dataset. We have exceeded the original goal in quantitative terms and provide high-quality (rule-based) parsers. The development of automated parsers and the refinement of the existing annotations are left as a topic for subsequent research.

### *Machine Translation*

This work package aimed to evaluate and refine the statistical and neural machine translation system. Statistical machine translation experiments using Moses and Giza++ alignment and drawing on

Sumerian lemmatized data from ETCSRI and ETCSL yielded poor results and were abandoned early on. We have successfully experimented with various neural machine translation models as planned, using the translations and transliterations available and data augmentation techniques. Because of the small amount of annotations produced manually, it was not possible to compare our machine translation pipelines using supervised methods with the unsupervised methods, but we have had unexpectedly excellent results with unsupervised methods.

### *Ur III Translation and Evaluation*

Here we aimed to apply the methods described in the preceding package to Sumerian texts and to evaluate the results. Our translation results using neural machine translation were evaluated using the BLEU scale and achieved scores ranging between 18.5 and 44. Human evaluation of our best models matched the BLEU score (2.5/3). Some of these results will be soon published in an accepted paper at the Computational Linguistics international conference. We have not yet prepared a full translation of our Ur III corpus; we are currently integrating our numeral and Named Entities translation modules into our pipeline for better accuracy.

### *Machine-learning-based Information Extraction (IE)*

This work package aimed to extract selected pieces of information for qualitative and quantitative analysis. Progress in this area has been less advanced because we decided to prioritise translation and data sharing. As part of this package we have prepared a state-of-the-art survey of IE and we have worked on defining information to be extracted based on its pertinence for humanities research.

Due to the great challenges related to data sparsity in our domain, we have experimented with semi-supervised ML-based information extraction methods: in the context of a Google Summer of Code project, we successfully released the first Sumerian semantic role-labelling system. The underlying idea here has been (i) to use a supervision signal in the form of annotations which are available in one language; (ii) to project them onto our Sumerian texts using machine-translation-based alignment models, and, finally, (iii) to train a statistical model on the projected annotations. Successful pilots were reported and we could demonstrate that it is feasible to train our algorithm to identify coarse-grained semantic patterns.

We have also prepared algorithms for the automated extraction and translation of numerals (rule-based), Part of Speech (POS) annotations, and Named Entities detection (including which type) using Hidden Markov Model and Conditional Random Fields (CRF), Bi-directional Long short-term memory CRF, FLAIR and RoBERTa (a specialization of Bert). The best performing models were Rules + CRF for POS, giving us a 0.991 F1 score, and BBPE + RoBERTa for NER, with a F1 score of 0.9537.

As part of their cost-neutral extension, the Frankfurt team developed a rule-based semantic parser to detect and retrieve transactions in administrative texts. It provides structured semantic frames as output, as well as a mapping to MTAAC syntactic annotation. It is currently extended to process the annotations produced by the aforementioned components for numeral and named-entity annotation. Because of delays in the creation of syntactically annotated gold data, the information extraction

systems employ machine learning for entity-level annotations subtasks only, whereas relation extraction was largely implemented in a rule-based fashion.

### *Interoperability*

The aim here was to publish data and tools in an interoperable fashion. We have prepared a survey of representation formalisms and designed guidelines for Linked Open Data editions of data and tools in cuneiform philology. We have prepared a model for linking metadata, textual data, and linguistic information. We have also designed a LOD-interface for CDLI data, and have published tools and generated data. The generated data are at this time limited to selected examples, but the database is ready to accept the new data in the appropriate formats. We have also developed an API that can deliver text metadata, inscriptions, and annotations of texts in RDF compatible formats (ttl, xml-rdf, json-rdf) and we expect to also offer all the data live through our SPARQL endpoint. We have also opened up the range of open formats in which we offer the data; although not all formats are linked data, they strongly increase accessibility and interoperability between projects, for easy injection and export of data (e.g., CSV, TSV, CoNLL-U, Brat).

To facilitate the uptake of LOD technology in Assyriology beyond our project, we also developed an LOD specification for the Electronic Penn Sumerian Dictionary (ePSD2). In parts, this has been delayed by the need to define community-approved specifications to represent attestations in an electronic dictionary. Under the lead of members of the Frankfurt team, these have been developed in the W3C Community Group “Ontology-Lexica” and were published in May 2020. In addition, a concept for the lexical data-modelling for Akkadian, the other major cuneiform language, under the lead of MTAAC project members was developed at the 3rd Summer Datathon on Linguistic Linked Open Data.

### *Methodology*

The objectives of this work package were to reflect, develop and evaluate methods for automatically analysing, translating, and operating with cuneiform data. The original plan was to produce six-monthly reports, but in order to streamline our work we decided early on in the project to publish our work instead in methodological articles, so as to increase access to and visibility of our research. We also have prepared extensive practical documentation which also documents our methodology.

### *Applied Use of Data*

Here we aimed to study the dynamics of social mobility in the Ur III period as a test case. For reasons of time and resources we were not able to accomplish this goal in full. We have looked into the distribution of usage in grain types over time, showing an increase of about 10% of barley over time which seems to confirm the existing but quantitatively unverified theory concerning salination of the soils that had been put forward in the past (Jacobsen 1982). We expect that our colleagues, and ourselves, will produce this type of statistical analysis more frequently in the future. Other research groups have been working with the data we have produced. We expect that the PIs on this project and other researchers will prepare incubation research projects using our data and tools. Once the new

CDLI Framework is released, the MTAAC Project’s branded portal on CDLI, displaying our newly minted grammatical and syntactical annotations, will be available to scholars and the public.

### *Infrastructure*

This work package aimed to address infrastructure optimisation and development. In collaboration with the NEH-funded CDLI Framework Update project, we contributed towards refreshing and optimizing the existing CDLI structure and assisted in the conceptualization and implementation of the infrastructure to host our new project data. We have also implemented the basis of a REST API (discussed in the “Interoperability” section). We also benefited from an award of resources from Compute Canada via their Research Platforms and Portals scheme to establish a backup of all CDLI and MTAAC data, including the new data we have produced, and a virtual machine to test the new CDLI Framework and establish a mirror site once the testing phase is completed.

### *Dissemination*

The aim was to document and present the project and its results, and to develop the interface for users to access the data in different ways. We worked in conjunction with the CDLI Framework Update project to prepare the new CDLI interface to host and display our data and tools. We assumed responsibility for design, including audience profiles, research into ergonomics and usage research, taking into account accessibility for differently abled people. We had an especially fruitful collaboration with Google through the Google Summer of Code (2018, 2019, 2020), whereby students have produced a map-based browsing interface and data visualization graphs of five different types for the texts’ meta-data. One of our sub-projects was concerned with the analysis of commodities in our chosen corpus; the tools help users to visualize concordance and similar items. As an integral part of the project, we have developed a website that fully documents our research process and presents guides to understanding our data formats and how to use the tools we have developed. All of the data we produced is released to the public domain and all of our tools and code snippets are licensed as open source, generally under the MIT license, and are publicly available. The project’s publications, presentations, and tools are listed in the Appendix. MTAAC data and code are also available from our GitHub repository.

## **Problems encountered and lessons learned**

One of the main problems encountered was staff turnover and the resulting disruption and need to transfer knowledge. Another challenge was posed by the difficulty of recruiting suitably qualified personnel in some instances, especially technical staff. Structural differences in the various national research and funding environments meant that the Toronto and UCLA teams relied on skilled graduate students, with inevitable limits on the amount of time available for project work, while the Frankfurt team could employ (part-time) Postdoctoral researchers.

We also faced other challenges that we had anticipated in our planning of the project: manual annotation of the texts proceeded too slowly to be able to fulfil all of our annotation objectives, and there was only a small amount of translation available for supervised machine translation.

In this project—as with others—a recurring challenge has always been the long-term preservation of project data and interface. Working in the humanities with large amounts of data generally means not having the appropriate institutional support and not having adequate funding to profit from support comparable to that offered in the sciences. The Cuneiform Digital Library Initiative, being a long term endeavour, had built up inter-institutional support over the past twenty years and we used this network to preserve our own data and interface. We were also able to extend and strengthen this network using resources granted by Compute Canada.

## **Project management and communications**

It proved impossible for all three PIs to meet during the lifespan of the project. However, two key meetings were held at which all teams were represented. The initial planning meeting took place in Bellingham in July 2017, attended by C. Chiarcos and Maria Sukhareva (Frankfurt), R. Englund (UCLA) and É. Pagé-Perron (Toronto). Also present were local advisors from Western Washington University, Jim Hearne and YùDōng Liu (Computer Science, working on named entity recognition), Steven Garfinkle (Assyriology), as well as Saurabh Trikande (Amazon). The second key meeting, for the purpose of mid-term review, took place in Frankfurt in July 2018, attended by H.D. Baker (Toronto), C. Chiarcos and I. Khait (Frankfurt), and É. Pagé-Perron (UCLA). We relied heavily on electronic communications (Slack, Google Hangouts, Skype), with weekly video meetings supplemented by one-on-one meetings as needed.

## **Contribution towards the training of graduate students and ECRs**

### *Frankfurt*

The ACoLi lab in Frankfurt employed two postdoctoral researchers who defended their PhDs during the project. They have both developed state-of-the-art technology for the processing, translation and linking of the Sumerian language, and their work has been published in peer-reviewed journals. One BSc assistant and several students worked in supporting the team in developing our translation and linking pipelines. Three BSc projects resulted in new technology that was used in the overall project. For one of these projects the BSc student wrote a javascript module to display syntactic relationships of annotated texts which interprets CDLI-CoNLL format data. This work will be integrated into the CDLI Framework to display the MTAAC annotations. The other BSc project was concerned with developing an OntoLex model for the Sumerian language which was required for us to build our linked data model for this language. A third BSc project was concerned with a rule-based morphological parser that can build on the OntoLex data.

Beyond the MTAAC core staff, we also collaborated with a graduate student from the Research Group “Linked Open Dictionaries” on applications of CoNLL-RDF as developed in that project and applied within MTAAC.

Furthermore, MTAAC members co-organized the 3rd Summer Datathon on Linguistic Linked Open Data in May 2019 at the Leibniz Center for Informatics, Schloss Dagstuhl, Wadern, Germany. For an audience of 42 international participants (graduate students and ERCs) with different scientific

backgrounds, we presented and discussed a use-case for the modelling of lexical resources in cuneiform languages.

### *Toronto*

At the University of Toronto's Department of Near and Middle Eastern Civilizations our Assyriology PhD students became immersed in digital humanities approaches to the linguistic analysis of Sumerian and were able to contribute to some of the project's publications and presentations. More specifically, they learned to use specialized software for linguistic annotation, such as Brat. They were instrumental in modelling Sumerian grammar and syntax in order to map it to universal linguistic classifications, and they wrote much of the online documentation of our manual annotation pipeline. This work also greatly enhanced their knowledge and understanding of Ur III Sumerian administrative texts, leading to a better grasp of how the administration worked. Another PhD student was tasked with overall project coordination; she also worked in close collaboration with the Frankfurt team in developing new technologies. An MA student from the Information School was able to gain experience of the entire process involved in developing a web interface, from audience analysis and accessibility questions to deciding on which libraries and functionalities would be implemented to address the needs of the interface. His work also included the development of a graphic line including a new logo, choice of colours, and typography.

### *UCLA*

Nine Computer Science MSc students participated in the project at UCLA or were employed by the CDLI Framework Update project and collaborated with us. Their tasks were varied and revolved mostly around computational linguistics and infrastructure work. For instance, each conversion tool in our processing pipeline was designed by a different UCLA student. They gained hands-on experience with real research challenges and real, imperfect data. They also became intimately acquainted with the technologies used, whether database software, programming language (PHP, Python, Java, JS), programming patterns for web and for data processing. Each of them had to improve their communications skills to explain their work to non-computer science colleagues. Through the Google Summer of Code program in which the CDLI participates yearly, additional international students (mostly from India), undergraduates, MAs, and PhDs, collaborated with the project. They each addressed a specific challenge, formulating and implementing their own development plan, and some of the work was published in peer-reviewed journals.

## **External funding and resource contributions**

### *Compute Canada*

We received a three-year Research Platforms and Portals award under Compute Canada's Resource Allocation scheme to establish a Toronto mirror and backup for the CDLI website; this award was renewed for a further three years, beginning in 2020. We were also able to use Compute Canada resources for running CPU-intensive machine learning algorithms in the later phase of the project.

### *CDLI and CDLI Framework Project*

The CDLI Framework Project has proceeded in synergy with the MTAAC project. While MTAAC is concerned with developing state-of-the-art tools and producing new data based on research inquiry, the Framework project is concerned specifically with developing and sustaining the CDLI infrastructure. CDLI itself is integral in our sustainability plan as it provides long term hosting and backup of our newly produced data; it will display the data and will offer processing services using the new tools available.

### *University of Toronto Work Study Program*

The University of Toronto's Work Study Program provided 70% of the funding to employ one of our Assyriology PhD students between September 2018 and February 2019 in order to carry out linguistic annotation.

### *Google Summer of Code*

For three summers now, the Google Summer of Code (GSoC) program has provided support for full-time students working on specific projects related to MTAAC. For instance, contributions towards our multi-level annotations search tool and its interface were produced in this context. Our semantic role labeller and our pipeline for neural machine translation of the Sumerian language are also deliverables arising from GSoC.

### *Dean of the Humanities, UCLA*

The Dean of the Humanities at UCLA generously provided the funding of Graduate Student Assistants for the project for 12 quarters at 25%, including resident tuition.

### *Volunteers*

Many volunteers contributed to the project. For instance, previous computer science student assistants from UCLA contributed volunteer work after their tenure, most senior personnel in the project have invested their free time outside of the initial commitment to the project, and colleagues volunteered their time to mentor our students for GSoC. The GSoC prospective participants and the participants themselves have volunteered countless hours, mostly in helping to refine our new web platform but

also processing data for Machine Translation, testing algorithms and furthering our work in the avenue of Machine Learning.

### *HumTech*

HumTech (formerly the Center for Digital Humanities) at UCLA has supported the CDLI since the beginning of Prof. Englund's tenure. They have been actively participating in the design and implementation of our hardware infrastructure plan. Aside from offering free resources (virtual machines and network maintenance), CDLI has been granted 4 hours of IT support on a monthly basis.

### *University of Oxford*

The University of Oxford has provided us with access to their supercomputing resources "Advanced Research Computing".

## **The Teams**

The project involved three teams, from the University of Frankfurt, University of Toronto, and the University of California Los Angeles.

### *Frankfurt*

#### **Team**

Christian Chiarcos (PI, Computational Linguistics)

Niko Schenk (PhD & Post-Doc, Computational Linguistics)

Ilya Khait (PhD & PostDoc, Assyriology)

Max Ionov (PhD Candidate, Computational Linguistics; associated, Google Summer of Code Mentor)

Mohamed Boudan (Undergraduate assistant, Computer Science)

#### **Associates (Unfunded)**

Maria Sukhareva (PhD Candidate, Computational Linguistics; associated)

Florian Stein (BSc student, Computer Science)

Robin Diemar (BSc student, undergraduate final project, Computer Science)

Katrin Peikert (BSc student, undergraduate final project, Linguistics)

Julius Steuer (BSc student, undergraduate final project, Linguistics)

### *Toronto*

Heather D. Baker (PI, Assyriology)

Émilie Pagé-Perron (Project Coordinator, PhD candidate in Assyriology)

Lukas Reckling (PhD student, Assyriology)

William McGrath (PhD student, Assyriology)  
Jinyan Wang (PhD student, Assyriology)  
Pouya Lajevardi (BSc student, Computer Science)  
Aman Biswas (MA student, Information)  
Clare Kim (BSc student, Computer Science)

### *UCLA*

Robert K. Englund (PI, Assyriology)  
Émilie Pagé-Perron (Project Coordinator)  
Jayanth (MSCS)  
Anoosa Sagar (MSCS)  
Aoxuan (Douglas) Li (MSCS)  
Prashant Rajput (MSCS)  
Shraddha Manchekar (MSCS)  
Jiachen (Julight) Zhong (MSCS)  
Lars Willighagen (Google Summer of Code participant)  
Himanshu Choudhary (Google Summer of Code participant)  
Logan Born (Google Summer of Code participant)  
Rachit Bansal (CDLI Intern)  
Swati Sharma (MSCS)  
Qiwen (Nelly) Liu (MSCS)  
Tim Bellefleur (PhD candidate, Sanskrit)  
Bakhtiyar Syed (Google Summer of Code participant)  
Sagar Sagar (Google Summer of Code participant)  
Samarth Sharma (Google Summer of Code participant)  
Ravneet Punia (Google Summer of Code participant)

This report was prepared by H.D. Baker, É. Pagé-Perron, and C. Chiarcos, with input from I. Khait and N. Schenk.

## Appendix: List of Deliverables

### *Publications*

1. Pagé-Perron, Émilie, Sukhareva, Maria, Khait, Ilya, and Chiarcos, Christian. 2017. [Machine Translation and Automated Analysis of the Sumerian Language](#). *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics Anthology*.
2. Chiarcos, Christian, Pagé-Perron, Émilie, Khait, Ilya, Schenk, Niko, and Reckling, Lucas. 2018. [Towards a Linked Open Data Edition of Sumerian Corpora](#). *Proceedings of the Language Resources and Evaluation conference*.
3. Chiarcos, Christian, Khait, Ilya, Pagé-Perron, Émilie, Schenk, Niko, Jayanth, and Reckling, Lucas. 2018. [Annotating Sumerian: A LLOD-enhanced Workflow for Cuneiform Corpora](#). *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2018): Towards Linguistic Data Science*.
4. Chiarcos, Christian, Khait, Ilya, Pagé-Perron, Émilie, Schenk, Niko, Jayanth, Fäth, Christian, Steuer, Julius, McGrath, William, and Wang, Jinyan. 2018. [Annotating a Low-Resource Language with LLOD Technology: Sumerian Morphology and Syntax](#). *Information 9(11): 290*.
5. Chiarcos, Christian, Ionov, Maxim, de Does, Jesse, Depuydt, Katrien Khan, Anas Fahad, Stolk, Sander, Declerck, Thierry and McCrae, John Philip. 2020. Modelling Frequency and Attestations for OntoLex-Lemon. *Proceedings of Globalex Workshop on Linked Lexicography (GlobaLex@LREC-2020), Marseille, France, May 2020*.
6. Punia, Ravneet, Schenk, Niko, Chiarcos, Christian, and Pagé-Perron, Émilie. 2020. Towards the First Machine Translation System for Sumerian Transliterations: Annotation Guidelines. *Proceedings of the 28th International Conference on Computational Linguistics*.

### *Papers presented*

1. Pagé-Perron, Émilie, Sukhareva, Maria, Khait, Ilya, and Chiarcos, Christian. “Machine Translation and Automated Analysis of the Sumerian Language” (LaTeCH-CLfL Workshop, Vancouver, August 4, 2017).
2. Pagé-Perron, Émilie, and Reckling, Lucas. “Introducing the Machine Translation and Automated Analysis of Cuneiform Languages Project” (University of Toronto Digital Humanities Network Event, Toronto, August 29, 2017).
3. Pagé-Perron, Émilie, and Reckling, Lucas. “MTAAC: Machine Translation and Automated Analysis of Cuneiform Languages” (Canadian Society for Mesopotamian Studies Symposium “Digital Cuneiform: New Approaches to the Study of Ancient Near Eastern Written Sources,” Toronto, September 30, 2017).
4. Pagé-Perron, Émilie, and Nurmikko-Fuller, Terhi. “Getting LOADed: Practical Considerations, Tools, and Workflows for Producing Linked Open Assyriological Data” (American Schools of Oriental Research Annual Meeting, Boston, November 15–18, 2017).

5. Pagé-Perron, Émilie. “Le projet MTAAC: traduction et analyse automatique de textes cunéiformes” (Conférence de l’Association des études du Proche-Orient ancien, Université du Québec à Montréal, March 13, 2018).
6. Pagé-Perron, Émilie. “Recent Developments in Natural Language Processing for Cuneiform Languages” (workshop “Thinking Digital in Cuneiform Studies: Methods, Problems, Perspectives,” Venice, March 27–28, 2018).
7. Pagé-Perron, Émilie. “Pre-requisites and Workflow for the Machine Translation of the Sumerian Language” (workshop “Future Philologies: Digital Directions in Ancient World Text,” Institute for the Study of the Ancient World, New York, April 20, 2018).
8. Chiarcos, Christian, Pagé-Perron, Émilie, Khait, Ilya, Schenk, Niko, and Reckling, Lucas. 2018. "Towards a Linked Open Data Edition of Sumerian Corpora." (Language Resources and Evaluation conference, Miyazaki, May 7–12, 2018).
9. Chiarcos, Christian, Khait, Ilya, Pagé-Perron, Émilie, Schenk, Niko, Jayanth, and Reckling, Lucas. 2018. "Annotating Sumerian: A LLOD-enhanced Workflow for Cuneiform Corpora." (6th Workshop on Linked Data in Linguistics (LDL-2018): Towards Linguistic Data Science, Miyazaki, May 7–12, 2018).
10. Baker, Heather D. and Pagé-Perron, E. “Machine Translation and Automated Analysis of Cuneiform Languages” (contributions to workshop “Building International Bridges Through Digital Scholarship: The Trans-Atlantic Platform Digging Into Data Challenge Experience,” DH2018 Conference, Mexico City, June 25, 2018).
11. Baker, Heather D. “Introducing the MTAAC project: Machine Translation and Automated Analysis of Cuneiform Languages” (64th Rencontre Assyriologique Internationale, Innsbruck, July 20, 2018).
12. Khait, Ilya (in Russian). “Annotating Sumerian: (Semi)automatic Tagging of Morphology and Syntax for a Low-Resource Language” (first session of the Seminar for Comparative-Historical Linguistics, Institute of Linguistics of the Russian Academy of Sciences, Moscow, October 4, 2018).
13. Pagé-Perron, Émilie. “Cuneiform Collections as (Linked) Data” (workshop “Collections as Data,” Digital Libraries Forum, Las Vegas, October 15–17, 2018).
14. Pagé-Perron, Émilie. “New Technologies for Cuneiform Studies: Processing & Linking Textual Data” (University of California Berkeley, April 8, 2019).
15. Pagé-Perron, Émilie, Ilya Khait and Jinyan Wang. “Sumerian Annotation Workshop” (University of California Berkeley, April 8, 2019).
16. Khait, Ilya and Pagé-Perron, Émilie. “The MTAAC Project: Computational Analysis for Sumerian” (Broadening Horizons 6 Conference, Freie Universität Berlin, June 24–28, 2019).
17. Chiarcos, Christian, Ionov, Maxim, de Does, Jesse, Depuydt, Katrien Khan, Anas Fahad, Stolk, Sander, Declerck, Thierry and McCrae, John Philip. (2020), “Modelling Frequency and Attestations for OntoLex-Lemon” (Globalex Workshop on Linked Lexicography (GlobaLex@LREC-2020), Marseille, France, May 12, 2020.)
18. Pagé-Perron, Émilie “Open Science at the CDLI, Focus on Collaboration and Accessibility” (Recent Developments in Digital Assyriology, University of Helsinki, August 26–27, 2020).

19. Baker, Heather D. “MTAAC: Machine Translation and Automated Analysis of Cuneiform Languages” (Recent Developments in Digital Assyriology, University of Helsinki, August 26–27, 2020).
20. Punia, Ravneet, Schenk, Niko, Chiarcos, Christian, and Pagé-Perron, Émilie “Towards the First Machine Translation System for Sumerian Transliterations: Annotation Guidelines” (The 28th International Conference on Computational Linguistics, December 8–13, 2020.)
21. CDLI Google Summer of Code projects demo 2019 <https://www.youtube.com/watch?v=MHUN4pODmBo>
22. CDLI Google Summer of Code projects demo 2020 <https://www.youtube.com/watch?v=UvnkVGaDMU8>

## *Code and other Materials*

### *Data*

- Manually annotated morphology gold corpus [https://github.com/cdli-gh/mtaac\\_gold\\_corpus/tree/workflow/morph/to\\_dict](https://github.com/cdli-gh/mtaac_gold_corpus/tree/workflow/morph/to_dict)
- Syntax guidelines and annotations [https://github.com/cdli-gh/mtaac\\_syntax\\_corpus](https://github.com/cdli-gh/mtaac_syntax_corpus)
- Stable Ur III corpus (for statistical machine translation and [https://github.com/cdli-gh/mtaac\\_cdli\\_ur3\\_corpus](https://github.com/cdli-gh/mtaac_cdli_ur3_corpus)
- RDF Ur III texts [https://github.com/cdli-gh/rdf\\_converted\\_data](https://github.com/cdli-gh/rdf_converted_data)

### *Software*

#### **Converters and Checkers**

- metadata converter (CSV to TTL-RDF) [https://github.com/cdli-gh/mtaac\\_work/tree/master/lod/metadata](https://github.com/cdli-gh/mtaac_work/tree/master/lod/metadata)
- C-ATF to CDLI-CoNLL converter <https://github.com/cdli-gh/atf2conll-converter>
- ATF to TEI <https://github.com/cdli-gh/atf2tei>
- CDLI-CoNLL to CoNLL-U converter <https://github.com/cdli-gh/CDLI-CoNLL-to-CoNLLU-Converter>
- CoNLL-U to Brat standalone converter <https://github.com/cdli-gh/conllu.py>
- Brat standalone to CDLI-CoNLL converter [https://github.com/cdli-gh/brat\\_to\\_cdli\\_conll\\_converter](https://github.com/cdli-gh/brat_to_cdli_conll_converter)
- Conll2rdf <https://github.com/acoli-repo/conll-rdf>
- CDLI version of PyOracc <https://github.com/cdli-gh/pyoracc> (unfinished)
- JTF ATF checker
- ATF Normalizer [https://github.com/cdli-gh/mtaac\\_work/tree/master/ATF\\_transliteration\\_processor](https://github.com/cdli-gh/mtaac_work/tree/master/ATF_transliteration_processor)
- ETCSRI corpus (metadata and annotations) to RDF [https://github.com/cdli-gh/mtaac\\_work/tree/master/lod](https://github.com/cdli-gh/mtaac_work/tree/master/lod)

- CDLI data conversion for MT ingestion <https://github.com/cdli-gh/cdli-data-extractor>
- CoNLL merge <https://github.com/acoli-repo/conll-merge>

### Pre-annotators

- Morphology pre-annotation tool <https://github.com/cdli-gh/morphology-pre-annotation-tool>
- Syntax parser [https://github.com/cdli-gh/mtaac\\_work/tree/master/parse](https://github.com/cdli-gh/mtaac_work/tree/master/parse)
- Context Free Grammar Parser <https://github.com/cdli-gh/cfg-parser>
- Syntax pre-annotator [https://github.com/cdli-gh/mtaac\\_syntax\\_pipeline](https://github.com/cdli-gh/mtaac_syntax_pipeline)
- Dictionary based glosser <https://github.com/acoli-repo/acoli-glosser>

### Manual Annotation and Visualization

- Morphology annotation [https://github.com/cdli-gh/annotation\\_assistant](https://github.com/cdli-gh/annotation_assistant)
- Syntax annotation and visualization <https://github.com/cdli-gh/sumerian-syntax-tree>

### ML and Other Tools

- Annotator assistant for Sumerian Morphology [https://github.com/cdli-gh/annotation\\_assistant](https://github.com/cdli-gh/annotation_assistant)
- ATF parser <https://github.com/cdli-gh/mtaac-package>
- Semantic role labeling tool for Sumerian <https://github.com/cdli-gh/Semantic-Role-Labeler>
- Translation pipeline (Including POS and NE tagging)  
<https://github.com/cdli-gh/Sumerian-Translation-Pipeline>
- NMT (Summer 2019) <https://github.com/cdli-gh/Machine-Translation>
- New MT models (Summer 2020)  
<https://github.com/cdli-gh/Semi-Supervised-NMT-for-Sumerian-English>

### Visualization, Search, Browse

- CDLI new framework (alpha) <https://gitlab.com/cdli/framework>
- Multilayer annotation query tool CQP4RDF <https://github.com/cdli-gh/cqp4rdf>
- Commodities visualization <https://github.com/cdli-gh/cdli-accounting-viz>
- Numerals translator API <https://github.com/cdli-gh/cdli-accounting-viz>
- CTS Server <https://github.com/cdli-gh/cdli-cts-server>
- Scaife viewer <https://github.com/cdli-gh/scaife>
- CDLI Framework API client <https://github.com/cdli-gh/framework-api-client>

### Documentation

- MTAAC documentation website <https://cdli-gh.github.io/>
- Readme files in each tool repository
- Additional annotation documentation <https://github.com/cdli-gh/annodoc>