Towards a Linked Open Data Edition of Sumerian Corpora

Christian Chiarcos*, Émilie Pagé-Perron◇, Ilya Khait*, Niko Schenk*, Lucas Reckling◇
*Goethe University Frankfurt, Germany, ◇University of Toronto, Canada
{chiarcos, khait, schenk}@informatik.uni-frankfurt.de, {e.page.perron, lucas.reckling}@mail.utoronto.ca

Abstract

Linguistic Linked Open Data (LLOD) is a flourishing line of research in the language resource community, so far mostly adopted for selected aspects of linguistics, natural language processing and the semantic web, as well as for practical applications in localization and lexicography. Yet, computational philology seems to be somewhat decoupled from the recent progress in this area: even though LOD as a concept is gaining significant popularity in Digital Humanities, existing LLOD standards and vocabularies are not widely used in this community, and philological resources are underrepresented in the LLOD cloud diagram (`http://linguistic-lod.org/llod-cloud`).

In this paper, we present an application of Linguistic Linked Open Data in Assyriology. We describe the LLOD edition of a linguistically annotated corpus of Sumerian, as well as its linking with lexical resources, repositories of annotation terminology, and the museum collections in which the artifacts bearing these texts are kept. The chosen corpus is the Electronic Text Corpus of Sumerian Royal Inscriptions, a well curated and linguistically annotated archive of Sumerian text, in preparation for the creating and linking of other corpora of cuneiform texts, such as the corpus of Ur III administrative and legal Sumerian texts, as part of the Machine Translation and Automated Analysis of Cuneiform Languages project (`https://cdli-gh.github.io/mtaac/`).

Keywords: Linked Open Data, Linguistic Linked Open Data, Sumerian, RDF, Linked Dictionaries.

1. Background

The Sumerian language is an agglutinative isolate that was written using the cuneiform script[1] in ancient Iraq; it is the first recognized written language. Assyriologists have long been painstakingly transcribing cuneiform texts for their research. These transliterations are generally published on paper, and to a lesser extent collected in electronic archives as part of perhaps a dozen projects. Unfortunately, these digital initiatives do not share the same encoding, and the computational toolset available for processing these data is limited.

As a collaboration between specialists in Assyriology, computer science and computational linguistics at the Goethe University Frankfurt, Germany, the University of California, Los Angeles (UCLA), the University of Toronto, Canada, and the Cuneiform Digital Library Initiative (CDLI)[2], our recently funded project "Machine Translation and Automated Analysis of Cuneiform Languages" (MTAAC)[3] aims to tackle natural language processing challenges presented by these ancient Mesopotamian languages (Pagé-Perron et al., 2017).

MTAAC is developing a methodology and a set of state-of-the-art NLP components geared to the processing of cuneiform text. The homogenized, annotated, and translated texts, accompanied by extracted information and prepared using this pipeline, will be made available both to designated audiences and machines to facilitate the study of the language, culture, history, economy and politics of the ancient Near East. Beyond applying statistical and neural techniques, linked data formalisms and open vocabularies will be employed to facilitate the reusability of these data, thereby contributing to interoperability, in particular with other philological portals,[4] and also to encourage research reproducibility. Additionally, because the field of Assyriology suffers from a lack of shared standards, we expect that our linked data approach will provide a serious opportunity for data integration both within the field and beyond it.

The terminological and technological foundations described in this paper represent the basis for the future publication of cuneiform corpus data and their annotation, developed within the MTAAC project. In particular, this includes the major administrative and legal Sumerian corpus of the Ur III period (2100-2000 B.C.) and the deployment of the infrastructure to enable linking all cuneiform text in the encompassing Cuneiform Digital Library Initiative (CDLI), which curates the largest corpus of digitized cuneiform artifacts.

According to Chiarcos et al. (2013), the primary objectives of linking language resources is to foster:

- Representation: a flexible representation format for research data (corpora, dictionaries, extracted information) and metadata (vocabularies);
- Interoperability: common RDF models can easily be integrated;
- Federation: data from multiple sources can be combined effortlessly;
- Ecosystem: tools for RDF and linked data are widely available under open-source licenses;

[1] The cuneiform script is formed by impressing a sharpened reed stylus into fresh clay, creating wedge-like impressions to form distinct signs. It was employed from ca. 3500 BC to the end of the first millennium BC to write texts in Sumerian, the Semitic language Akkadian, and a number of other languages spoken in the region.

[2] `https://cdli.ucla.edu`.

[3] `https://cdli-gh.github.io/mtaac`.

[4] E.g. Syriac `http://syriaca.org`, Hebrew `http://hebrew-terms.huji.ac.il/`, Indo-European, and Caucasian languages, `http://titus.fkidg1.uni-frankfurt.de/`.

– Expressivity: existing vocabularies help express linguistic resources;
– Semantics: common links express what is meant;
– Dynamicity: web data can be continuously improved.

Developing an (L)LOD edition for Sumerian and linking representative language resources includes the application of the following ontologies:

– lemon/ontolex[5] for lexical data;
– CIDOC/CRM[6] for object metadata;
– lexvo[7] for language identification;
– Pleiades[8] for geographical information;
– OLiA[9] for linguistic annotations.

While these are established *de facto* standards in the field, editing principles for philological *corpora* are only now emerging, with different alternative vocabularies (POWLA, NIF, TELIX)[10] currently being discussed. Consequently, we focus on this aspect. Our proof-of-concept relies on the morphologically annotated Electronic Text Corpus of Sumerian Royal Inscriptions (ETCSRI) (Zólyomi et al., 2008) and describes the application of CoNLL-RDF that serves as LOD representation within the CDLI as part of the MTAAC project.

So far, only two projects currently provide open *annotated* cuneiform text. The first is the Open Richly Annotated Cuneiform Corpus (ORACC),[11] a portal hosting sub-projects with local glossaries. Most projects focus on sources in the Akkadian language.[12] Not all ORACC projects are annotated, but the platform offers a standalone lemmatizer which provides an interface to call a server service for the semi-automated annotation of lexical information. The second is the Electronic Text Corpus of Sumerian Literature (ETCSL) (Black et al., 1998–2006), which presents most known Sumerian literary compositions. A handful of other projects offer digital access to unannotated text.[13]

Linked Open Data has previously been applied to the humanities, including linguistics, NLP and other language sciences (Chiarcos, C. et al., 2012). Beyond prosopography and gazetteers on the one hand (e.g., Pelagios, perio.do), and early efforts to create addressable units for passages in texts on the other (CTS, Canonical Data Services), applications of (L)LOD to computational philology are rare, and indeed absent from the field of cuneiform studies. There are, however, projects that touch upon the classification of artifacts; for instance the Modref project (Tchienehom, 2017)[14] employs CIDOC-CRM, as is customary for the classification of museum artifacts, and connects three different collections, including the CDLI. Similarly, the British Museum provides a CIDOC-CRM-based SPARQL end point[15] that encompasses almost 22% of all CDLI artifact entries. By following explicit links within such repositories, Linked Data technology allows us to query disparate artifacts across different collections. In addition, SPARQL 1.1 federation, as described further below, allows us to access these metadata repositories remotely and to link CDLI data with them.

Two pioneering experiments on the application of *ontologies* to Sumerian are to be noted: Jaworski (2008a) presented an ontology-based approach to the semantic parsing of a domain-specific subset of Ur III administrative texts from the CDLI with the goal of tracing patterns of transfer of cattle between individuals and institutions. While he has been successful in identifying several thousand transactions, this approach is limited to a highly restricted domain. Neither the annotations nor the parser are available, but they are well documented in Jaworski (2008b). Another experiment was concerned with annotating the ETSCL corpus mentioned above with an ontology of literary concepts. The mORSuL ontology was developed to attach CIDOC-CRM to Ontomedia (Nurmikko, 2014; Nurmikko-Fuller, 2015);[16] however, this has only reached the status of a case study. No data are available from either of these applications, nor are their data being linked with the original corpora or other resources.

While these experiments show the potential interest that the scientific community would have in Sumerian corpus data being published in accordance with Semantic Web principles, neither of them actually aim to provide *Linked* (Open) Data as an end product. By bringing together corpus data, lexical data, linguistic annotations and object metadata, the MTAAC project is thus breaking new ground for the field of Assyriology, as well as computational philology in general.

2. Neo-Sumerian (Ur III) corpora

2.1. Ur III Data in CDLI

The Cuneiform Digital Library Initiative (CDLI) collects and makes available on the web metadata and, to a lesser extent, transliterations, transcriptions and translations, of all artifacts bearing cuneiform inscriptions. The project is based on the efforts of an international group of language specialists, museum curators and historians of science. World collections hold approximately 550,000 objects, and the CDLI has catalogued some 334,000 of them. Forty percent of these texts are written in Sumerian, of

[5] http://lemon-model.net/,https://www.w3.org/community/ontolex/wiki/Final_Model_Specification.

[6] http://www.cidoc-crm.org/.

[7] http://www.lexvo.org/.

[8] https://pleiades.stoa.org/.

[9] http://www.acoli.informatik.uni-frankfurt.de/resources/olia/.

[10] https://sourceforge.net/projects/powla/, http://aksw.org/Projects/NIF.html, http://ontorule-project.eu/telix.

[11] http://oracc.museum.upenn.edu.

[12] Akkadian is a Semitic language that was written using the cuneiform script. It was used from the second half of the 3rd millennium up to the end of the 1st millennium BC.

[13] Among others the CDLI, the Database of Neo-Sumerian Texts (BDTNS), a database of texts dating to the Ur III period http://bdts.filol.csic.es/, and Archibab, specializing in the Old Babylonian period (ca. 1900-1600 BC) http://www.archibab.fr.

[14] http://triplestore.modyco.fr:8080/ModRef.

[15] https://collection.britishmuseum.org/sparql.

[16] http://www.contextus.net/ontomedia.

which 2/3 were produced during the Ur III period, which refers to a dynasty of the end of the 22nd, and the whole of the 21st centuries BC.
In Ur III times, the (Neo-)Sumerian language dominated the cultural sphere: Sumerian texts were mass-produced in this era. However, this was apparently accompanied by its gradual decline as the common spoken language. Nevertheless, Sumerian remained the prevalent language not only of literature and royal texts, but also of legal, administrative, and economic documents.

Genre	Texts	/Total	Translit.	/T. total	Translated	/Translit.
Admin.	97675	96.98%	67697	96.23%	1582	2.34%
Royal	1529	1.52%	1468	2.09%	264	17.98%
Letter	744	.74%	700	1.00%	12	1.71%
Legal	451	.45%	383	.54%	9	2.35%
Other	319	.31%	100	.24%	13	34.32%
Total	100718		70348		1880	2.67%

Table 1: Ur III material in the CDLI

The Sumerian Ur III corpus available on CDLI is summarized in Table 1. The category "other" encompasses the literary, school, lexical, prayer, votive, scientific, other, uncertain, and fake genres. Although most literary and royal texts are not accompanied by a translation, the composite texts to which they are related are themselves often translated. The translations available can be viewed for all composite witnesses in the CDLI score pages, at `https://cdli.ucla.edu/tools/scores/partitur-index.html`.
As evident from the table, most of the data are of an administrative nature, but thus far untranslated. Therefore, we focus on this genre but aim to develop tools and resources to link with CDLI data in general.

2.2. Morphological Annotation in ETSCRI

The Electronic Text Corpus of Sumerian Royal Inscriptions (ETCSRI) is a sub-project of ORACC assembling all Sumerian royal inscriptions, compiled, verified and annotated by Zólyomi et al. (2008). ETCSRI is fully translated, lemmatized and morphologically annotated, and it provides transliterations and translations. Additionnaly, glossaries based on the project, which include named entities, are available for consultation.
The original texts on which this study is based are ancient inscriptions in the Sumerian language, written on diverse artifacts commemorating actions and dedications of higher elites that lived in ancient Iraq between 2900 and 1600 BC. Many of these come from the Ur III period, but they cover the history of the Sumerian language. All texts provide cross-references with the CDLI.
In addition to ETCSRI, the Electronic Text Corpus of Sumerian Literature (ETCSL)[17] also provides lemmatized, morphologically annotated and translated text in Sumerian, albeit following different (and, in parts, dated) transliteration principles. This corpus comprises a variety of literary compositions which were written down from the Ur III period onwards, mostly in the Old Babylonian period.
For our proof-of-concept, we chose the ETCSRI corpus as a basis for our efforts principally because of the reusability of the annotations. Not only does it have a substantial overlap with CDLI Ur III data, but its transliterations are up-to-date and the morphological annotations are based on a good morphological model of Sumerian. Moreover, the data are open and released to the public domain.[18]

2.3. Beyond Royal Inscriptions

The MTAAC project aims to complement the existing annotated and translated corpora of literature and royal inscriptions with a corpus of Ur III data in general, in particular its administrative texts. Due to the amount of data in question, only portions of it, however, can be manually annotated or translated. One objective of the MTAAC project is thus to provide automated analyses for this purpose.
Morphological annotations are based on ETSCRI (Zólyomi et al., 2008). ETSCRI-style part-of-speech annotations include named entity classification, for which we build on earlier efforts towards semi-automated entity annotation (Liu et al., 2015, SNER).[19]
Ur III administrative data are comparably easy in terms of morphosyntax, since morphology is often not expressed in writing (though the information may be inferred from the context). However, this also means that morphology is somewhat uninformative, so that effective querying and searching of these data requires structural analysis. We thus retrieve relational information in addition to morphosyntax as found in ETCSRI. This extends earlier work on (semantic) parsing by Jaworski (2008a) in that we do not rely on domain-specific rules, rather we employ state-of-the-art machine learning techniques. At the moment, we are evaluating the suitability of the Universal Dependency (UD)[20] schema and UD-based annotation projection for these kind of data, possibly to be augmented with an additional layer of semantics (Peterson et al., 2014): since administrative texts are not exclusively composed of grammatical sentences but also often comprise lists, semantic role labeling (SRL) annotation is considered crucial for this genre. The SRL inventory will be based on Hayes (2000) and Jaworski (2008b), yet grounded in the English PropBank.
Since further details of the annotation process will be presented elsewhere, we focus here on infrastructural measures.

3. Towards Linked Data

3.1. Corpus Representation

The **(Canonical-)ASCII Transliteration Format** ((C-)ATF) is a text encoding format developed by CDLI and the Electronic Pennsylvania Sumerian Dictionary (ePSD)[21], a text encoding scheme for cuneiform transcriptions which was designed as a human-friendly archival format to complement the usage of machine-oriented XML formats for annotations (Koslova and Damerow, 2003). Basically, it is a data entry and storage format. It first encodes the

[17] `http://etcsl.orinst.ox.ac.uk`.

[18] See ORACC `http://oracc.museum.upenn.edu/doc/opendata/index.html` "Open Data". Until recently, the data were available only under a Creative Commons Share alike license but the release of the data in JSON format was done under the public domain.

[19] `https://wwunlp.github.io/sner/`.

[20] `http://universaldependencies.org/`.

[21] `http://psd.museum.upenn.edu/`.

CDLI text-ID along with its designation, some feature tags to encode language and medium type, and a line-by-line numbered transliteration, augmented with interlinear annotations such as normalization, translation, and comments on structure and content. It also uses specific conventions to annotate structure. Transliteration lines are restricted to the ASCII character range.

See, as an example, one of the exemplars of a royal inscription of king Amar-Suen[22]:

```
&P226657 = RIME 3/2.01.03.01, ex.  07
#atf:  lang sux
@object brick
@surface a
1.  {d}amar-{d}suen
2.  nibru{ki}-a
3.  {d}en-lil2-le
4.  mu pa3-da
5.  sag-us2
6.  e2 {d}en-lil2-ka
7.  nita kal-ga
8.  lugal uri5{ki}-ma
9.  lugal an ub-da limmu2-ba
@surface b
1.  {d}amar-{d}suen
2.  nibru{ki}-a
3.  {d}en-lil2-le
4.  mu pa3-da
5.  sag-us2
6.  e2 {d}en-lil2-ka
7.  nita kal-ga
8.  lugal uri5{ki}-ma
9.  lugal an ub-da limmu2-ba
```

The format was subsequently extended in ORACC to provide support for additional annotation layers. ORACC-ATF[23] uses Unicode characters in the transliteration lines. Additionally, annotations are stored in comment lines in between lines of text.[24]

The ETCSRI edition of the Sumerian royal inscription above and its morphological annotation are available from ORACC in XHTML and JSON formats [25]. The ATF and XML versions are available only to privileged users.

ORACC uses comment lines to store more information about the text, such as lemma information, but it is impossible to add another layer of annotation, such as syntax, for instance, into the ATF format. Aligning with the initial philosophy of the C-ATF format, we do not intend to extend the specification but instead we will supplement it with community standards, in this case, the CoNLL TSV format. Due to the specifics of our data, we define our own inventory of columns for information storage, although we convert this CDLI-CoNLL format to fully-fledged CoNLL-U for further processing.[26] In the case of the proof-of-concept presented here, the original text and annotations are extracted from XHTML and converted to CoNLL. As for the full MTAAC project, text and annotations are stored as follows: C-ATF file for the textual and text structure data, and CDLI-CoNLL files for the annotation. A validation script to check the alignment between both files is under preparation.

These data formats are primary *annotation* formats; they do not provide data structures that can be easily queried or transversed using off-the-shelf technology. As such, for the sake of a web publication, we thus provide an RDF converter which is built on the CoNLL2RDF tool[27], whose output serves as the basis for linking the corpus. We will, in a subsequent step, make the corpus available in other machine friendly formats such as (RDF-) XML and JSON.

3.2. CoNLL-RDF

While the development of vocabularies for lexical data has progressed significantly and was recently aggregated in the lemon community standard (see below), the representation of linguistically annotated text is a more heterogeneous area, with highly generic models for *richly annotated corpora* on the one hand (Chiarcos, 2012), and problem-specific models on the other, such as *NLP web services* (Hellmann et al., 2013) or semantic annotation (Sanderson et al., 2017). All these data models can be serialized in different RDF formats – which are, however, generally verbose and not intended for human consumption nor direct (string-based) manipulation.

Finally, CoNLL-RDF (Chiarcos and Fäth, 2017) fills in this gap by providing a middle ground that accounts for the needs of NLP specialists: easy to read, easy to parse, and close to conventional representations. Important is the format's *potential for LLOD integration*: it is directly processable using Semantic Web technology, thereby facilitating interoperability, interpretability, linkability, queryability, transformability, database support, and integration with web technologies. In addition, CoNLL-RDF complements its data model with layout conventions (and a formatter) to facilitate the easy low-level access to the CoNLL TSV format.

An example of RDF rendering of an ETCSRI excerpt[28] is

[22] https://cdli.ucla.edu/P226657.

[23] On the differences between the ATF dialects see http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/index.html.

[24] http://oracc.museum.upenn.edu/doc/help/editinginatf/primer/structuretutorial/index.html, http://oracc.museum.upenn.edu/doc/help/lemmatising/primer/.

[25] http://oracc.museum.upenn.edu/etcsri/Q000981.

[26] This is a characteristic of most members of the CoNLL format family; tool support for CoNLL thus typically involves routines for reordering, merging or dropping columns.

Using the CoNLL format has the additional advantage that annotated corpora stored in other formats can easily be converted to a standard format without requiring extensions of a proprietary archival representation. This was especially helpful to facilitate working with the ETCSRI Corpus which is internally stored both as ATF and TEI/XML but published only as XHTML and JSON; their ORACC-ATF files are not open and might not contain all of the annotations since the ORACC website gathers data both from the ATF and the glossary to generate the final rendering of the text output.

[27] https://github.com/acoli-repo/conll-rdf.

[28] The example uses a fragment from a short votive inscription by Ur-Namma, the first ruler of the Ur III dynasty (ETCSRI's Ur-Namma 2, line 3), see

```
@prefix : <http://oracc.museum.upenn.edu/etcsri/Q000935#> .
@prefix conll: <http://ufal.mff.cuni.cz/conll2009-st/task-description.html#> .
@prefix rdf: <http://www.w3.Org/1999/02/22-rdf-syntax-ns#> .
@prefix terms: <http://purl.org/acoli/open-ie/> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix rdfs: <http://www.w3.Org/2000/01/rdf-schema#> .

:s2_0 nif:nextSentence :s3_0 .

:s3_0 a nif:Sentence .

:s3_1 a nif:Word; conll:WORD "lu2";
                  terms:lemma <http://psd.museum.upenn.edu/epsd/epsd/e3356>;   conll:BASE "lu2";
                  conll:CF "lu";   conll:EPOS "n"; conll:FORM "lu2";
                  conll:GW "person";
                  conll:HEAD :s3_0; conll:ID "1"; conll:LANG "sux"; conll:MORPH "N1=lu";
                  conll:MORPH2 "N1=stem"; conll:NORM "lu"; conll:POS "N"; conll:SENSE "person";
                  nif:nextWord :s3_2 .

:s3_2 a nif:Word; conll:WORD "e2";
                  terms:lemma <http://psd.museum.upenn.edu/epsd/epsd/ell66>;   conll:BASE "e2 ";
                  conll:CF "e";    conll:EPOS "n"; conll:FORM "e2";
                  conll:GW "house";
                  conll:HEAD :s3_0; conll:ID "2"; conll:LANG "sux"; conll:MORPH "N1=e";
                  conll:MORPH2 "N1=STEM"; conll:NORM "e"; conll:POS "N"; conll:SENSE "house, temple";
                  nif:nextWord :s3_3 .

:s3_3 a nif:Word; conll:WORD "{d}nanna";
                                                                   conll:BASE "{d}nanna";
                  conll:CF "Nanna"; conll:EPOS "DN"; conll:FORM "{d}nanna\\gen\\abs";
                  conll:GW "1";
                  conll:HEAD :s3_0; conll:ID "3"; conll:LANG "sux"; conll:MORPH "N1=Nanna.N5=ak.N5=Ø";
                  conll:MORPH2 "N1=name.N5=gen.N5=abs"; conll:NORM "Nanna.ak.Ø"; conll:POS "DN"; conll:SENSE "1" .
```

Figure 1: Illustration of a CoNLL-RDF representation of an ETCSRI excerpt. (One word per line format expanded for better legibility.)

provided in Figure 1. The first line in a sentence refers to the sentence URI and defines it as a `nif:Sentence`. The second line holds the first content word and defines it as a `nif:Word`, followed by its `conll:WORD`, other annotations in alphabetical order of their properties, concluding with a `nif:next` statement pointing to the next word in the sentence (if available). The relation between words and sentences is established via `conll:HEAD`. `conll:WORD`, `conll:HEAD`, etc., are properties that express the annotations found in the respective columns in the original CoNLL file. Turtle provides different triple separators: "." separates independent triples, ";" separates one from the next that shares the same (omitted) subject, "," enumerates multiple objects for one property of the same subject. In CoNLL-RDF, all of these are written in one line, so that the CoNLL convention of WPL annotations is respected; different properties are separated using ";", different values enumerated with ",". Finally, the sentence is concluded with a `nif:nextSentence` statement (if more sentences follow).

Figure 2 illustrates the overall conversion workflow including different data formats that were involved up to this point, i.e. from raw (ATF) text to (CoNLL)RDF-compliant corpus representations. A final step involves in particular the linking of external knowledge sources, such as annotations, lexical data, and meta data, which we describe in the following section.

3.3. Annotation Pipeline

Linking the annotations is the last task in our pipeline and it is done automatically. The MTAAC project comprises two major steps, the semi-automated annotation of our gold corpus and the automated annotation performed based on this gold corpus. The first step starts with the validation of the ATF text data, which are then morphologically pre-annotated using a dictionary-based tool that is fueled by a database of forms and their associated morphological analyses[29]. A human annotator then verifies the morphological annotations and advances the text in the pipeline. The text is then pre-annotated using a rule-based syntax annotation. The syntactic annotations are generated using our RDF-

`http://oracc.museum.upenn.edu/etcsri/Q000935`. Note that the fragment under consideration, translated as "the man who (built) the temple of Nanna", is in fact part of a longer sentence that we have shortened for the sake of illustration.

[29] See our morphology pre-annotation tool here: `https://github.com/cdli-gh/morphology-pre-annotation-tool`.

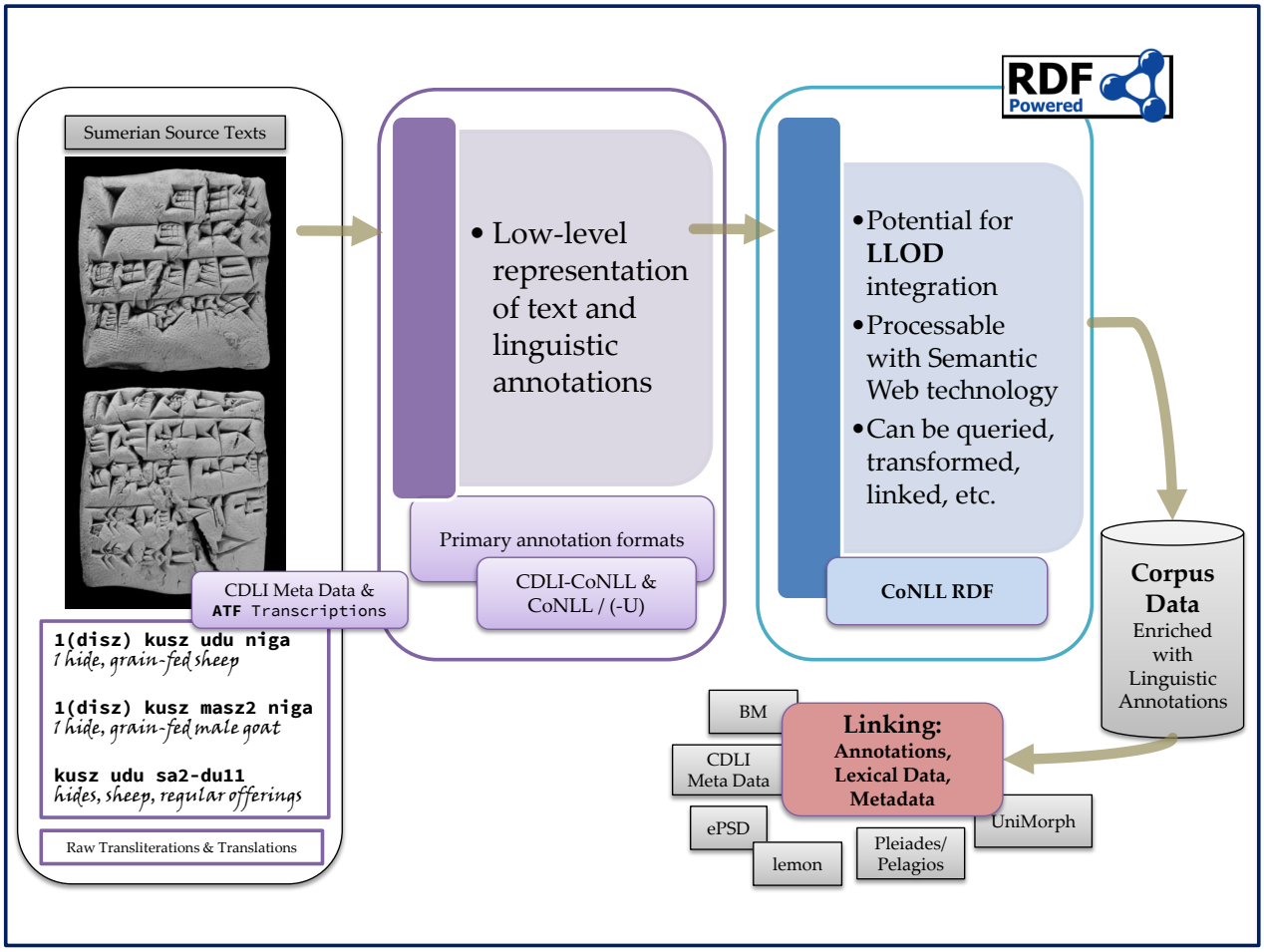

Figure 2: Workflow illustration for the LOD edition of Sumerian texts

based pre-annotation tool[30] and then manually adjusted, when needed. This pipeline is more thoroughly described and evaluated in our LDL2018 paper[31].

The annotated texts produced in this semi-automated pipeline are in the exact same format at those exiting the automated pipeline, thus our linking process applies to both, while it also handles the ETCSRI data we use for our proof-of-concept.

3.4. Linking Annotations

In the case of morphology, and as part of the proof-of-concept we have developed, we map the existing ETCSRI morphological annotation scheme[32] of which we use the "MORPH 2" tag, with the Universal Morphological Feature Schema (UniMorph) specifications[33], though a Turtle RDF mapping of the ORACC:ETCSRI morphological tags inventory with UniMorph[34]. UniMorph is able to define morphological features in language-independent terms, allowing for facile translation between languages employing the schema (Sylak-Glassman, 2016, 3); this effectively brings Sumerian for the first time into the language corpora that are linked by their linguistic annotations, making it now available for richer cross-linguistic research.

Some challenges have emerged in the mapping, since Sumerian is a language isolate. The first challenge is the verb modality expressed through a series of prefixes. Both for the LLOD mapping using UniMorph and for Universal Dependencies mapping when converting our home CoNLL format to ConLL-U Feats field, these prefixes cannot all find an adequate home. We are currently preparing proposals to include adequate tags in both schemes. The second hurdle is the Sumerian enclitic copula, which also has no equivalent analysis in UniMorph. But overall, the impressive flexibility of UniMorph made it possible to combine tags to account for the exact meaning of certain morphemes, for example in the case of the locative morphemes, that we represent with "IN+ESS", "ON+ESS" and "APUD+ESS".

We use CoNLL-RDF as a working format to leverage the capabilities of SPARQL for syntactic annotation. Linking to the syntactic data is made possible through CDLI-CoNLL, the main format used in our corpusto store annotations. For this purpose, we provide and consult an OWL representation of the CDLI annotation scheme and its link-

[30] Seehttps://github.com/cdli-gh/mtaac_work/tree/master/parse

[31] http://ldl2018.linguistic-lod.org/.

[32] vhttp://oracc.museum.upenn.edu/etcsri/parsing/index.html

[33] http://unimorph.org/.

[34] https://github.com/cdli-gh/mtaac_work/blob/master/lod/annotations/um-link.ttl.

ing with UD POS, feature and dependency labels.

3.5. Linking Lexical Data

The ePSD is the only comprehensive and available digital resource for Sumerian vocabulary. We converted the ePSD data to an index of deep links, expressed as a lemon dictionary. Our pipeline consults (or constructs, if it is not found) a lemon/ontolex compliant index for the ePSD, whose URIs are provided to the linking. Because of the structure of the ePSD, links point to guide word entries. While we prepare the MTAAC Ur III research corpus, additional local lexical resources for the Sumerian language will be provided.

3.6. Linking Metadata

Cuneiform text is inscribed on objects about which specialists gather a set of metadata that is useful when integrated into a method for text analysis. Information such as provenience, period, and size of the artifact are examples of such characteristics. Other information such as the condition of the object, the museum in which it is kept, and the publications mentioning the text in question are all helpful for the discovery and study of the artifact.

The CDLI catalogue data live in a MySQL database and is exported daily in CSV format. For this exercise we used the CSV data package. Fields of interest are: museum no (here BM: British Museum), id_text (CDLI object id), composite, height, thickness, width, material, and finally period. The composite number regroups witnesses of a composition that was copied on different artifacts.

Our modeling approach follows that of the British Museum: the composite field translates to `?composite a crm:E34_Inscription` and id_text to `?object URI ? object crm:P65_shows_visual_item ?composite` [35] and lastly, the museum_no maps with `owl:sameAs` which is a resolved museum number.

Using this model, we convert the data to RDF with the csv2rdf tool supplemented with embedded custom turtle templates that we prepared for the occasion.[36] We link to external metadata repositories: the Modref project[37] and the British Museum.[38] The ModRef project's goal is to "move heterogeneous data into triplestores also called data warehouses or collections of RDF files in order to improve the sharing, exchange and discovery of new knowledge" (Tchienehom, 2017). Their model formalizes three different collections in a coherent model: the CDLI catalogue, the ObjMythArcheo database,[39] a corpus of archaeological objects related to mythological iconography, and BibllioNum, a DL about France in the 20th century. Because text is so much more meaningful with its context, linking catalog information of the artifacts on which the texts are inscribed greatly enriches our linking model for Assyriologists, who need this information to understand the texts as well as to compare these artifacts with other classes of artifacts that possess similar characteristics, such as provenience, period, size, and the collection in which they are kept. Cuneiform objects are for the most part studied only in the field of Assyriology. Making them available in the semantic web increases the possibility of including them in larger-scale studies, thus overcoming this limitation in the scope of research.

4. Summary and Outlook

In this paper, we described the (L)LOD edition and linking of corpora for Assyriology. We apply our model to ETCSRI as a proof-of-concept for future application to the Neo-Sumerian (Ur III) administrative corpus targeted by the MTAAC project. We have successfully integrated these diverse and distributed knowledge sources (linguistic and non-linguistic):

- CDLI (local resource; CoNLL-RDF plus CIDOC-CRM)
- ORACC:ETSCRI (by conversion; CoNLL-RDF)
- ePSD (by conversion and links to HTML; lemon)
- ModRef & BM (by federation; CIDOC-CRM)

With this experiment we also demonstrated the applicability and usefulness of (L)LOD standards to Assyriology. Other vocabularies such as Pleiades, Snap dragon and perio.do, among others, can be added analogously. The annotation data on Sumerian morphology and syntax will be produced in the future by the (semi-)automatic annotation pipeline under development, which we see as a crucial step towards an (L)LOD edition of cuneiform corpora. Lastly, this proof-of-concept is now our tested and refined template for the infrastructure that will be integrated into the CDLI as part of the MTAAC project. As such, we welcome feedback to further strengthen our model.

Although creating new linguistic data and tools to manipulate this data should improve the research outcomes for Assyriologists, we realize that knowledge circulation is directly dependent on access, classification and discoverability. As such, the linking of linguistic and other resources has been built in as an essential part of the MTAAC project. We also share our linking workflow under publicly `https://github.com/cdli-gh/mtaac_work/tree/master/lod`.

Acknowledgements

The project Machine Translation and Automated Analysis of Cuneiform Languages is generously funded by the German Research Foundation, the Canadian Social Sciences and Humanities Research Council, and the American National Endowment for the Humanities through the T-AP Digging into Data Challenge.[40]

Our appreciation goes to Heather D. Baker and Robert K. Englund for their insights and suggestions.

[35] From the CRM documentation, the superproperty crm:P128_carries would suit too, see `http://www.cidoc-crm.org/html/5.0.4/cidoc-crm.html#P128` hence, it must be a physical thing, in our case a `E84_Information_Carrier` However, if no `?composite` is found, then a separate `crm:E34_Inscription` must be created.

[36] `http://clarkparsia.github.io/csv2rdf/`.

[37] `http://modref-labexpassespresent.huma-num.fr`.

[38] `https://collection.britishmuseum.org/sparql`.

[39] `http://www.limc-france.fr` and `http://medaillesetantiques.bnf.fr`.

[40] `https://diggingintodata.org/`.

5. Bibliographical References

Black, J. A., Cunningham, G., Eberling, G., Flückiger-Hawker, J., Robson, E., Taylor, J., and Zólyomi, G. (1998–2006). The Electronic Text Corpus of Sumerian Literature. `http://etcsl.orinst.ox.ac.uk`.

Chiarcos, C. and Fäth, C. (2017). CoNLL-RDF: Linked corpora done in an NLP-friendly way. In *International Conference on Language, Data and Knowledge*, pages 74–88. Springer.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In Alessandro Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*, pages 7–25. Springer, Berlin, Heidelberg.

Chiarcos, C., et al., editors. (2012). *Linked Data in Linguistics*. Springer, Berlin, Heidelberg.

Chiarcos, C. (2012). POWLA: Modeling linguistic corpora in OWL/DL. In *Extended Semantic Web Conference*, pages 225–239. Springer.

Hayes, J. L. (2000). *A Manual of Sumerian Grammar and Texts. Second Revised and Expanded Edition*. Number 5 in ARTANES. Undena Publications, Malibu.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using linked data. In *International semantic web conference*, pages 98–113. Springer.

Jaworski, W. (2008a). Contents modelling of neo-Sumerian Ur III economic text corpus. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 369–376, Stroudsburg, PA. Association for Computational Linguistics.

Jaworski, W. (2008b). *Ontology-Based Knowledge Discovery from Documents in Natural Language*. Ph.D. thesis, University of Warsaw.

Koslova, N. and Damerow, P. (2003). From cuneiform archives to digital libraries: The Hermitage Museum joins the Cuneiform Digital Library Initiative. In *Proceedings of the 5th Russian Conference on Digital Libraries RCDL 2003*, pages 1–8. St.-Petersburg, Russia.

Liu, Y., Burkhart, C., Hearne, J., and Luo, L. (2015). Enhancing Sumerian lemmatization by unsupervised named-entity recognition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1446–1451. Association for Computational Linguistics.

Nurmikko-Fuller, T. (2015). *Telling ancient tales to modern machines: ontological representation of Sumerian literary narratives*. Ph.D. thesis, University of Southampton.

Nurmikko, T. (2014). Assessing the suitability of existing OWL ontologies for the representation of narrative structures in Sumerian literature. *ISAW Papers*, 7(1):1–9.

Pagé-Perron, É., Sukhareva, M., Khait, I., and Chiarcos, C. (2017). Machine translation and automated analysis of the Sumerian language. In *Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, ACL*, pages 10–16. Association for Computational Linguistics.

Peterson, D., Palmer, M., and Wu, S. (2014). Focusing annotation for semantic role labeling. In *LREC*, pages 4467–4471.

Sanderson, R., Ciccarese, P., and Young, B. (2017). Web Annotation data model. Technical report, W3C Recommendation.

Sylak-Glassman, J. (2016). The composition and use of the universal morphological feature schema (UniMorph schema). Technical report, Department of Computer Science, Johns Hopkins University.

Tchienehom, P. (2017). ModRef Project: from creation to exploitation of CIDOC-CRM triplestores. In *The Fifth International Conference on Building and Exploring Web Based Environments (WEB 2017)*.

Zólyomi, G., Tanos, B., and Sövegjártó, S. (2008). The Electronic Text Corpus of Sumerian Royal Inscriptions. `http://oracc.museum.upenn.edu/etcsri/introduction/index.html`.