Annotating Sumerian: A LLOD-enhanced Workflow for Cuneiform Corpora

Christian Chiarcos*, Ilya Khait*, Émilie Pagé-Perron◇, Niko Schenk*, Jayanth^λ, Lucas Reckling◇
*Goethe University Frankfurt, Germany, ◇University of Toronto, Canada,
^λUniversity of California, Los Angeles
{chiarcos|khait|schenk}@informatik.uni-frankfurt.de,
{emilie.page.perron|lucas.reckling}@mail.utoronto.ca, jayanthj@ucla.edu

Abstract

Assyriology, the discipline that studies cuneiform sources and their context, has enormous potential for the application of computational linguistics theory and method on account of the significant quantity of transcribed texts that are available in digital form but that remain as yet largely unexploited. As part of the Machine Translation and Automated Analysis of Cuneiform Languages project (https://cdli-gh.github.io/mtaac/), we aim to bring together corpus data, lexical data, linguistic annotations and object metadata in order to contribute to resolving data processing and integration challenges in the field of Assyriology as a whole, as well as for related fields of research such as linguistics and history. Data sparsity presents a challenge to our goal of the automated transliteration of the administrative texts of the Ur III period. To mitigate this situation we have undertaken to annotate the whole corpus. To this end we have developed an annotation pipeline to facilitate the annotation of our gold corpus. This toolset can be re-employed to annotate any Sumerian text and will be integrated into the Cuneiform Digital Library Initiative (https://cdli.ucla.edu) infrastructure. To share these new data, we have also mapped our data to existing LOD and LLOD ontologies and vocabularies. This article provides details on the processing of Sumerian linguistic data using our pipeline, from raw transliterations to rich and structured data in the form of (L)LOD. We describe the morphological and syntactic annotation, with a particular focus on the publication of our datasets as LOD. This application of LLOD in Assyriology is unique and involves the concept of a LLOD edition of a linguistically annotated corpus of Sumerian, as well as linking with lexical resources, repositories of annotation terminology, and finally the museum collections in which the artifacts bearing these inscribed texts are kept.

Keywords: Linked Open Data, Sumerian, Linguistic Linked Open Data, linked dictionaries, syntactic parsing, annotation pipeline, CoNLL, RDF, pre-annotation

1. Introduction

1.1. Sumerian and Cuneiform Studies

The Sumerian language, an agglutinative isolate, is the earliest language recorded in writing. It was spoken in the third millennium BC in modern southern Iraq, and continued to be written until the late first millennium BC. This language was written with cuneiform, a logo-syllabic script with around one thousand signs in its inventory, formed by impressing a sharpened reed stylus into fresh clay. This script was employed in ancient Mesopotamia and surrounding regions to inscribe many different languages, notably the East Semitic Akkadian (Babylonian and Assyrian), the Indo-European Hittite, and others.

In order to make a text available for research, Assyriologists copy and transcribe it from the artifact bearing it. The results of this labor-intensive task are usually published on paper. A dozen projects which make various cuneiform corpora available online have emerged since the early 2000s, building on digital transcriptions created as early as the 1960s. Unfortunately, these initiatives rarely use shared conventions, and the toolset available to process these data is limited, thus vasts numbers of transliterated and digitized ancient cuneiform texts remain only superficially exploited.

1.2. Linked Open Data for Sumerian

Linked Open Data (LOD) defines principles and formalisms for the publication of data on the web, with the goal of facilitating its accessibility, transparency and reusability. The application of LOD formalisms to philological resources within the field of Assyriology promises two crucial advantages. First, we shall be able to establish interoperability and exchange between distributed resources that currently persist in isolated data silos – or that provide human-readable access only, with no machine-readable content. Among other benefits that LOD provides, one should also mention its federation, ecosystem, expressivity, semantics, and dynamicity potential (Chiarcos et al., 2013). Converting out data to an RDF representation is an essential step to open up the possibility of linking with other resources and integrating content from different portals. Further, using shared vocabularies allows us to publish structured descriptions of content elements in a transparent and well-defined fashion. Ontologies play a crucial role in this regard as these define shared data models and concepts.

1.3. The MTAAC Project

The "Machine Translation and Automated Analysis of Cuneiform Languages" (MTAAC) project[1] aims to develop state-of-the-art computational linguistics tools for cuneiform languages, using internationally recognized standards to share the resulting data with the widest possible audience. (Pagé-Perron et al., 2017) This is made possible through a collaboration between the Cuneiform Digital Library Initiative (CDLI)[2] and specialists in Assyriology, computer science and computational linguistics at the Goethe University Frankfurt, Germany, the University of California, Los Angeles (UCLA) and the University of Toronto, Canada.

The project entails the preparation of a methodology and an associated NLP pipeline for the Sumerian language. The

[1] https://cdli-gh.github.io/mtaac.
[2] https://cdli.ucla.edu.

pipeline processes, annotates and translates Sumerian texts, as well as extracts additional information from the corpus. In order to facilitate the study of the language and the historical, cultural, economic and political context of the texts, these data are to be made available both to designated audiences and machines.

In order to facilitate the reusability of these data, as well as to encourage reproducibility, we use linked data and open vocabularies, thereby contributing to interoperability with other resources[3]. Another aim in the application of LOD is to set new standards for digital cuneiform studies and to contribute to resolving data integration challenges both in Assyriology and related linguistic research. The (L)LOD edition for Sumerian and the linking of representative language resources uses lemon/ontolex for lexical data, the CIDOC/CRM for object metadata, lexvo for language identification, Pleiades for geographical information, and OLiA[4] for linguistic annotations. Brigning together corpus data, lexical data, linguistic annotations and object metadata breaks new ground in the field of Assyriology, and computational philology.

2. Corpus Data and Data Formats

2.1. Ur III Data in CDLI

One objective of our project is to complement the range of cuneiform corpora with morphologic, syntactic and semantic annotations for an extensive, but currently undertranslated genre, namely the administrative texts, especially for the Neo-Sumerian language of the Ur III period.

The Cuneiform Digital Library Initiative (CDLI) is a major Assyriological project which aims to provide information on all objects bearing cuneiform inscriptions kept in museums and collections around the world. The images, metadata, transliterations, transcriptions, translations and bibliography are made available online. At the moment the CDLI catalog contains entries for about 334,000 objects out of an estimated total of around 550,000.

The corpus we chose is a subset of these entries: 69,070 administrative and legal texts produced during the Ur III period (2100-2000 BC). These texts are available in transliteration but only 1,966 have parallel English translation. Textual data in the ATF format are presented as follows:[5]

```
&P142051 = WO 11, 21
#atf: lang sux
@tablet
@obverse
1. 2(gesz2) 2(u) 4(disz) udu bar-gal2
#tr.en: 144 sheep with fleece,
2. 4(disz) sila4 bar-gal2
#tr.en: 4 lambs with fleece,
3. 7(disz) udu bar-su-ga
#tr.en: 7 sheep without fleece,
4. 3(gesz2) 1(u) 2(disz) ud5 masz2 hi#-[a]
#tr.en: 192 mixed nanny and billy goats,
5. ki kas4-ta
#tr.en: from Kas
6. lu2-dsuen i3-dab5#
#tr.en: Lu-Suen took;
$ blank space
@reverse
$ blank space
1. mu us2-sa ki-maszki# ba-hul
#tr.en: year after: "Kimaš was destroyed".
```

These data are composed of lines of transliteration that start with a number; they also include structure tags, translation and comments which complement the content of each textual entry.

2.2. Other Sumerian Corpora

Previous research on Sumerian text has produced two corpora; of literary texts (ETCSL) (Black et al., 1998–2006) and royal inscriptions (ETCSRI, within ORACC)[6] respectively, but both corpora were limited to morphosyntactic annotation. To the best of our knowledge, this also corresponds to the state of the art in other branches of Assyriology, where representative morphosyntactic annotations (glosses) have been assembled, for example, within the ORACC[7] portal. Additionally, some other projects offer digital access to unannotated texts.[8]

2.3. Automated Annotation and Analysis

Experiments in automated syntactic annotation have been described by Jaworski (2008) and Smith (2010), but both focused on extracting automatically annotated fragments rather than on providing a coherently annotated corpus. The mORSuL ontology[9], developed to attach CIDOC-CRM to Ontomedia (Nurmikko, 2014; Nurmikko-Fuller, 2015),[10] has only reached the status of a case study. These experiments show the potential interest in Sumerian corpus data published in accordance with Semantic Web principles, but neither of these projects actually aims to provide Linked Data as an end product.

With respect to semantics, current research focuses on shallow techniques such as named entity recognition (SNER[11] on Sumerian), or entity linking and prosopography (Darmstadt on Hittite) – to the best of our knowledge, the annotation of cuneiform corpora with syntactic relations is limited to experiments[12], and semantic relations annotating has not

[3] E.g., Syriac http://syriaca.org, Hebrew http://tinyurl.com/guwe8kr, and Indo-European and Caucasian languages http://titus.fkidg1.uni-frankfurt.de/.

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

[5] Text published by Hruška (1980), CDLI entry prepared by Robert K. Englund. https://cdli.ucla.edu/P142051.

[6] http://oracc.museum.upenn.edu/etcsri/; as with all ORACC projects, ETCSRI uses a slightly different version of ATF as its core format.

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

[8] Apart from the CDLI, it is important to mention the Database of Neo-Sumerian Texts (BDTNS), a database of texts dating to the Ur III period http://bdts.filol.csic.es/.

[9] https://github.com/terhinurmikko/morsul.

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

[11] https://github.com/wwunlp.

[12] Karahashi and Tinney have previously worked on a rule based syntax annotator from which we expect to reuse some rules in the further development of our tool. https://github.com/oracc/oracc/tree/master/misc/ssa3. Unfortunately the documentation written by Karahashi is not available for con-

previously been attempted.

2.4. The CDLI-CoNLL Format

The CDLI-CoNLL format is an abridged version of the CoNLL-U format.[13] Because of the scarcity of specialists in the Sumerian language, our format was designed with ease and speed of annotation in mind.

The SEGM field contains information on the lemma, comprising a dictionary word and its sense, appended and in square brackets, e.g. udu[sheep] or dab[seize]. Affixes are standardized in conformity to a list of morphemes, following the ETCSRI project's morphological scheme. These morphemes are separated by a dash placed before the morpheme, except for the first element in the chain. When the analysis of the word demands a morpheme that is not explicit in the writing of the form, it is enclosed in square brackets.

The XPOSTAG field contains the part-of-speech tag associated with the morpheme present in the SEGM column. If the form represents a named entity, the named entity tag will take the place of the POS tag. The tags we employ are again those of the ETCSRI project. These tags are separated using a period placed before the morpheme, except in the case of the first element in the chain.

The information in these fields can easily be converted to the CoNLL-U format following rules and maps. Our converter uses maps to create the UPOSTAG from our domain-specific POS tags and for the conversion of morphemes to the verbose CoNLL-U FEATS column.

Figure 1 illustrates the CDLI-CoNLL format in comparison with the CoNLL-U format as far as morphology and morphosyntax are concerned:[14] the FORM column provides the transliteration of the original cuneiform signs, but without elements marking the state of the text on the medium (breaks, omissions, etc). The original SEGM column provides segmentation into morphological (rather than graphemic) segments. Because of the characteristics of the Sumerian noun, the LEMMA directly follows from this segmentation as its first substring. However, CoNLL-U does not allow us to preserve full SEGM information, so the LEMMA is used instead. The original XPOSTAG includes information about the part-of-speech and named entities categories (SN), as well as grammatical features. However, UD conventions allow us to preserve only parts of the morphological information in CoNLL-U: the last word of a Sumerian noun phrase aggregates all case morphology (its own as well as that of its – preceding – head), a phenomenon known as *Affixanhäufung*. In this case, the place name *Shuruppak* is a genitive attribute of an ergative argument. It is thus inflected for *both* genitive (*-ak*) and ergative (*-e*). In CoNLL-U, multiple case marking is not foreseen, so that here, a language-specific aggregate feature for multiple cases is introduced.[15] In addition, the SN tag marks Shuruppak as a site name, and we derive non-human animacy.

For the mapping between our morphological tags, the Universal dependencies tags and features (as well as Unimorph categories features), we adopt a Linked Open Data approach: we provide an ontological representation of the CDLI annotation scheme, and link its concepts via *skos:broader* (etc.) statements with the UD and Unimorph ontologies provided as part of the OLiA ontologies.[16] CDLI-CoNLL can also be converted to the Brat Standoff format through our pipeline described below, for further syntactic annotation, visualization, or using other tools geared to processing data in this format.

2.5. Linked Open Data Representations

Linked Open Data in Assyriology is limited at the moment to metadata on artifacts, which, however, seems practical when working on cuneiform corpora. The Modref project (Tchienehom, 2017)[17] is used in the classification of museum artifacts and employs CIDOC-CRM for that purpose. CDLI is among the three collections it connects[18]. Additionally, almost 22% of all CDLI artifacts are encompassed by the CIDOC-CRM-based SPARQL end point of the British Museum[19]. Linked Data technology allows us to query disparate artifacts across different collections using explicit links within such repositories. The SPARQL 1.1 federation allows us to query these metadata repositories and to link CDLI data with them.

Edition principles for philological corpora are only just emerging, with different alternative vocabularies (POWLA, NIF, TELIX) currently being discussed. In the MTAAC project, we generally base our proof-of-concept on the morphologically annotated ETCSRI corpus; the application of CoNLL-RDF serves as LOD representation within the CDLI.

2.6. CoNLL-RDF

CoNLL-RDF (Chiarcos and Fäth, 2017) is a rendering of RDF in CoNLL's tab-separated value format. It represents a convenient and human-readable data model that is close to conventional representations and can be serialized in RDF format. Crucially, it is comparably easy to read and parse as CoNLL: it provides the direct means to string-based manipulations that CoNLL is praised for, but in addition it allows

sultation. The only existing cuneiform corpus with (manual) annotation of syntax is the Annotated Corpus of Hittite Clauses, see (Molina, 2017);

[13] `http://universaldependencies.org/format.html`

[14] Both the CoNLL-U and CDLI-CoNLL formats have additional fields to handle relationships between words, such as syntax.

[15] This solution is problematic in that long chains of case markers can arise, and it is no longer possible to generalize over the resulting multitude of case features. Case combinatorics in the ETCSRI corpus yield 47 case chains resulting from only 15 case labels.

[16] `http://purl.org/olia`, for Unimorph, see `http://purl.org/olia/owl/experimental/unimorph/`..

[17] `http://triplestore.modyco.fr:8080/ModRef`.

[18] The other two are the ObjMythArcheo database `http://www.limc-france.fr` and `http://medaillesetantiques.bnf.fr`, a corpus of archaeological objects related to mythological iconography, and BiblioNum, a DL about France in the 20th century.

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

# ID	FORM	SEGM	XPOSTAG
o.0.4	szuruppak{ki}-ga-ke4	Szuruppag[1]-ak-e	SN.GEN.ERG

# ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS
o.0.4	szuruppak{ki}-ga-ke4	Szuruppag[1]	PROPN	SN	Number=Sing|Case=Gen.Erg|Animacy=Nhum

Figure 1: CDLI-CoNLL annotation compared to CoNLL-U

us to seamlessly integrate LOD resources to process, manage, and manipulate CoNLL data with off-the-shelf technologies (Chiarcos and Schenk, 2018).

We argue for the use of CoNLL-RDF in our setting because of its suitability for LLOD integration. In fact, it is directly processable with Semantic Web technology insofar as it facilitates interoperability, interpretability, linkability, queryability, transformability, database support, and integration with web technologies. In the context of our corpus annotation workflow, CoNLL-RDF is used as an internal format for parsing and for the pre-annotation of syntax using SPARQL (Buil Aranda et al., 2013), cf. Section 3.3., but it can also serve as a future release format, cf. Mazziotta (2010) for Old French.

3. Annotation

3.1. Annotation Workflow

As explained in the corpus section (2.), the raw data entering the pipeline comprise unannotated textual data in the ATF format. Before conversion, this text is validated against structure rules and content. Structure is defined in the ATF format[20] specifications. Content is checked for word tokens and sign tokens against the existing data available at the CDLI.

When entering the pipeline, the text is first converted from ATF to the CDLI-CoNLL format. Like most members of the CoNLL format family, this is a TSV format with one word per line, newline-separated sentences. In comparison to, e.g., the widely used CoNLL-U format, it does come with project-specific columns. It is both more compact and more informative, but tailored to our specific use.

The CDLI-CoNLL file is then fed into morphological pre-annotation. A dictionary-based pre-annotation tool fills most of the morphological information for each form present in the text. The human editor goes over the result, filling the lines left incomplete, and verifying that the annotations are correct. Before storing the annotated CDLI-CoNLL text alongside the ATF text in the database, the content is again validated, both for content and conformity to the CDLI-CoNLL format. The resulting CDLI-CoNLL data are then stored in the database.

For syntactic parsing, the CDLI-CoNLL data are subject to the syntax pre-annotation tool described below, cf. Section 3.3.. The resulting data are serialized as CoNLL-U, but part of the conversion process is to replace CDLI-specific annotations with those conforming to the Universal Dependencies. For this purpose, we provide and consult an OWL representation of the CDLI annotation scheme and its linking with UD POS, feature and dependency labels. Using SPARQL update, these ontologies are loaded, their hierarchical structure traversed by property paths, and the corresponding tags replaced. We argue that the clear separation of (SPARQL) code and (OWL) data of different provenience (CDLI annotation model, UD annotation models, linking between both) facilitates the transparency, reproducibility and reversibility of our mapping in comparison to direct replacement rules.

Finally, the CoNLL-U data are converted to the Brat standoff format[21]; the human editor can thus verify and finalize the syntactic annotation of the text using the CDLI Brat server interface.

The completed Brat Standoff file is exported and converted back to CDLI-CoNLL. At this point, the novel annotations need to be merged with the original CDLI data. Although conflicts should not occur as long as the data was not *manually* manipulated, we need a robust merging routine in case such corrections have been applied. For this purpose, we employ CoNLL-Merge.[22] CoNLL-Merge performs a word-level diff on the FORM column. Beyond merely identifying mismatches, it also provides heuristic but robust merging strategies in case a mismatch occurred, e.g., if a word has been split, two words have been merged, or deletions or additions occurred.

Only the ATF and CDLI-CoNLL versions of the data are kept in the datastore as we can easily convert the CDLI-CoNLL format to CoNLL-U and CoNLL-RDF formats, according to need. While both will be important publication formats to facilitate usability and re-usability of our data, they will only be generated on demand. We are, however, exploring options to offer CoNLL-RDF as a dynamic view on the internal (relational) database via technologies such as R2RML (Das et al., 2012).

An illustration of the annotation workflow, including intermediate data formats, is shown in Fig. 2[23]

3.2. Morphological Pre-Annotation

As part of the pipeline, we have designed a morphological pre-annotation tool[24] to make the manual annotation process more efficient in respect to speed of annotation as well as consistency and actual morphological analysis correct-

[20] http://oracc.museum.upenn.edu/doc/help/editinginatf/primer/index.html.

[21] http://brat.nlplab.org/standoff.html.

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

[23] Cuneiform text of the Ur III period from the settlement of Garshana, Mesopotamia (Owen, 2011, no. 851) and its transliteration as stored in the Cuneiform Digital Library Initiative (CDLI) database https://cdli.ucla.edu/P322539 (picture reproduced here with the kind permission of David I. Owen).

[24] The code for this tool and all the other tools we are designing for this pipeline are available in repositories kept under the CDLI organization page on Github https://github.com/cdli-gh.

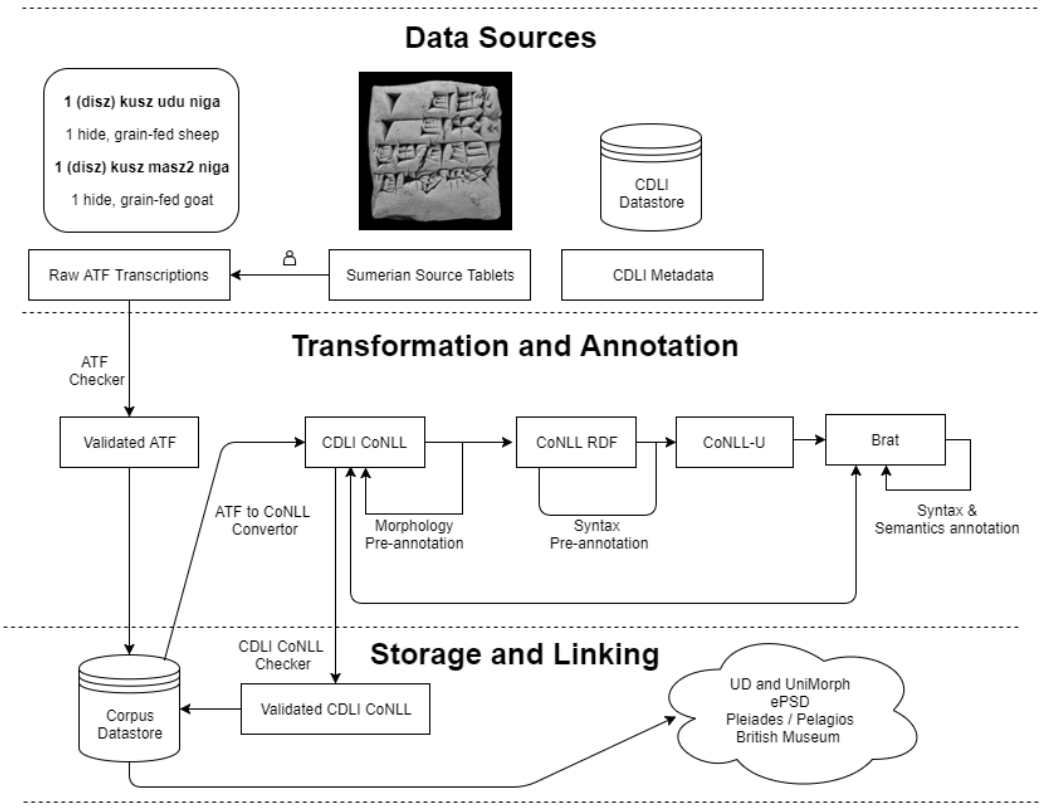

Figure 2: Corpus annotation pipeline: from ATF to RDF

ness. The tool is applied after the ATF text has been converted to the CDLI-CoNLL format. It uses the principle of a dictionary lookup to provide the most frequent annotation associated with the form it is annotating. For example, a text could contain the form ensi2 ("ruler"), without attached morphemes. All variant analyses of the form encountered to date, and their frequency, are stored in the dictionary.

```
"ensi2": [
  {
    "annotation": [
      "ensik[ruler]",
      "N"
    ],
    "count": 3
  },
  {
    "annotation": [
      "ensik[ruler][-ak][-ø]",
      "N.GEN.ABS"
    ],
    "count": 1
  },
  {
    "annotation": [
      "ensik[ruler][-ø]",
      "N.ABS"
    ],
    "count": 1
  },
  {
    "annotation": [
      "ensik[ruler][-ra]",
      "N.DAT-H"
    ],
    "count": 1
  }
],
```

When the pre-annotation tool encounters the form ensi2, it will add the first option of the example in the appropriate SEGM and XPOSTAG fields, based on frequency. The other choices are appended in subsequent columns so the human editor can easily copy and paste another option in the appropriate fields, if required. The additional columns will be destroyed while validating the contents. The pre-annotation tool can add new entries to the dictionary on demand, so it is best to perform this operation frequently to augment the accuracy of the tool.

The CDLI-CoNLL validator is integrated into the pre-annotation morphological tool. It performs checks on the syntax of the ID field, the existence of the lemmata in our dictionary, and the parallelism of the SEGM and XPOSTAG fields, based on a mapping of morphemes that can appear in the SEGM field and the morphological tags employed in the XPOSTAG field.

3.3. RDF-Based Pre-Annotation

CoNLL-RDF has been developed with the goal of flexible transformation of annotated corpora for the output of state-of-the-art NLP tools: every CoNLL sentence constitutes a graph, and parsing rules can be formulated as rewriting operations on this graph.

While this approach is qualitatively different from conventional parsing, we adopt the terminology of classical Shift-Reduce parsing (Nivre et al., 2007, 100-104). However, we model SHIFT and REDUCE as RDF properties that result from parsing operations rather than these parsing operations themselves. Along with that, parsing is no more sequential, and data structures such as QUEUE and STACK are no longer necessary; instead, both the 'queue' of tokens and the 'stack' of partial parses are marked by explicit SHIFT relations that represent their sequential order.

The method is initialized by adding a SHIFT relation for ev-

ery *nif:nextWord* property in the graph, i.e., the 'queue' of partial parses corresponds to the sequence of words. During parsing, language-specific rules are applied. Unlike classical Shift-reduce parsing, the words are not processed from left to right, but bottom-up. If an attachment rule applies for a word/partial parse X, it is removed from the 'queue' of words (which is no longer distinguished from the 'stack' of partial parses) by dropping its SHIFT relations. Instead, a REDUCE relation with its head is established, and the sequence of SHIFTS is restored by connecting the head of the partial parse with its SHIFT-precedent, or successor.

With any remaining SHIFT relations of the reduced elements being transferred to the (partial) parse, the sequence of SHIFTs takes over the functions of the traditional 'queue' and the traditional 'stack' at the same time, but elements are processed regardless of their sequential order; instead, the order of parsing rules plays a decisive role in the parsing process.

Parsing rules can be expressed as SPARQL Update statements, which are applied and iterated in a predefined order until there are no more transformations, i.e., because a single root for the sentence has been established. Finally, the SHIFT transitions are removed, whereas the REDUCE transitions are replaced by *conll:HEAD* properties.

Parsing, as defined here, is deterministic and greedy, and more or less context-insensitive. However, this is enough to provide a convenient means of implementing 'default' rules for syntactic attachment, which can be corrected afterwards during manual annotation.

In this sense, our basic rule-based parser provides a satisfactory syntactic *pre*-annotation with only 7 rules[25]:

1. Reduce adjective to preceding noun with adjectival modifier relation:
$\text{NOUN}_0\ \text{ADJ} \Rightarrow \text{NOUN} \xleftarrow{\text{amod}} \text{ADJ}$
E.g. nita $\xleftarrow{\text{amod}}$ kalag-ga "strong male".

2. Reduce noun in the genitive to preceding noun with appositional modifier relation:
$\text{NOUN}\ \text{NOUN}_{\text{GEN}} \Rightarrow \text{NOUN} \xleftarrow{\text{GEN}} \text{NOUN}$
E.g. lugal $\xleftarrow{\text{GEN}}$ $\text{urim}_5^{\text{ki}}$-ma "king of Ur".

3. Reduce noun with case marker to preceding noun with no case marker with appositional modifier relation:
$\text{NOUN}_0\ \text{NOUN}_{\text{CASE}} \Rightarrow \text{NOUN}_{\text{CASE}} \xleftarrow{\text{appos}} \text{NOUN}$
E.g. $^{\text{d}}\text{inana}_{\text{DAT}} \xleftarrow{\text{appos}}$ nin-a-ni "to Inanna, his lady".

4. Reduce noun to preceding noun with case relation:
$\text{NOUN}_0\ \ \text{NOUN}_{\text{CASE1+CASE2}} \Rightarrow \text{NOUN}_{\text{CASE1}} \xleftarrow{\text{CASE2}} \text{NOUN}$
This rule is applicable mostly for complex genitive chains.
E.g. $\text{lugal}_{\text{ERG}} \xleftarrow{\text{GEN}}$ $\text{urim}_5^{\text{ki}}$-ma-$\text{ke}_4$ "king of Ur".

5. Reduce noun to preceding numeral with numeral modifier relation:
$\text{NUM}_0\ \text{NOUN}_{(\text{CASE})} \Rightarrow \text{NUM}_{(\text{CASE})} \xleftarrow{\text{nummod}} \text{NOUN}$
E.g. 3(u) $\xleftarrow{\text{nummod}}$ sila_3 "thirty sila (measuring unit)"

6. Reduce noun in case to following verb with absolutive relation:
$\text{NOUN}_{\text{ABS}}\ \text{VERB} \Rightarrow \text{NOUN} \xrightarrow{\text{ABS}} \text{VERB}$
E.g. numun-na-ni $\xrightarrow{\text{ABS}}$ he_2-eb-til-le-ne "may they end his lineage".

7. Reduce noun in case to following verb with case relation:
$\text{NOUN}_{\text{CASE}}\ \text{VERB} \Rightarrow \text{NOUN} \xrightarrow{\text{CASE}} \text{VERB}$

In part, these rules employ grammatical case features as dependency labels. After pre-annotation, however, these internal labels are to be mapped to CoNLL-U relations.

The graph-rewriting rules are implemented in SPARQL Update,[26] as illustrated by the example below, which matches the noun in the absolutive case to verb reduction rule (No. 6).

```
DELETE {
        ?last conll:SHIFT ?noun.
        ?noun conll:SHIFT ?verb.
} INSERT {
        ?noun conll:REDUCE ?verb; conll:EDGE ?case.
        ?last conll:SHIFT ?verb.
} WHERE {
        ?noun conll:POS ?pos FILTER(strends(str(?pos),'N')).
        ?noun conll:CASE ?case FILTER(?case="ABS")
        ?noun conll:SHIFT ?verb.
        ?verb conll:POS ?vPos FILTER(strstarts(str(?vPos),'V'))
        OPTIONAL {?last conll:SHIFT ?noun. }
}
```

Figure 3: SPARQL query for rule 6

An example of the output of the syntactic pre-annotation for a Sumerian royal inscription of Ur-Namma of Ur (approx. 2112-2095 B.C.)[27] is provided below.

We estimate that this method can be efficiently used for pre-annotation in order to enhance the syntactic annotation process; however, one cannot fully rely on its unsupervised result: mistakes and ambiguities are expected and these have to be resolved manually.

3.4. Manual Annotation

Manual annotation of the syntax is greatly simplified with the application of the pre-annotation tool. Using our Brat server[28], the human annotator must first verify that annotations generated by the pre-annotation tool are correct. When an annotation is faulty, the annotator removes the annotation and creates the appropriate one instead. Navigating the Brat interface is made easy as we modified the GUI to necessitate fewer clicks for each task. Finally, missing relationships must be added. The pre-annotation tool is

[25] Abbreviations follow Universal Dependencies; SHIFT and REDUCE relation are designated by whitespace (left) and arrow (right) respectively.

[26] The full code is available from `https://github.com/cdli-gh/mtaac_work/tree/master/parse`.

[27] See `http://oracc.museum.upenn.edu/etcsri/Q000937`.

[28] `http://brat.nlplab.org/`.

```
s1_1  . . / DAT-H---- ang          BASE an GW 1 ID 1 MORPH2 N1=NAME POS DN
s1_2  . . . \ appos-- lugal        BASE lugal GW king ID 2 MORPH2 N1=STEM POS N
s1_3  . . . . \ GEN-- dinjir-re-ne BASE dinjir GW deity ID 3 MORPH2
                                   N1=STEM.N4=PL.N5=GEN POS N
s1_4  . . . \ appos-- lugal-a-ni   BASE lugal GW king ID 4 MORPH2
                                   N1=STEM.N3=3-SG-H-POSS.N5=DAT-H POS N
s1_5  . . / ERG------ ur-{d}namma  BASE ur-{d}namma GW 1 ID 5 MORPH2 N1=NAME
                                   POS RN
s1_6  . . . \ appos-- lugal        BASE lugal GW king ID 6 MORPH2 N1=STEM POS N
s1_7  . . . . \ GEN-- urim₅{ki}-ma-ke₄ BASE urim₅{ki} GW 1 ID 7 MORPH2
                                   N1=NAME.N5=GEN.N5=ERG POS SN
s1_8  . . / ABS------ kiri₆        BASE kiri₆ GW orchard ID 8 MORPH2 N1=STEM POS
                                   N
s1_9  . . . \ amod--- mah          BASE mah GW great ID 9 MORPH2
                                   NV2=mah.N5=ABS POS V/i
s1_10 . . \       mu-na-gub        BASE gub GW stand ID 10 MORPH2
                                   V4=VEN.V6=3-SG-H.V7=DAT.V11=3-SG-H-A.V12=gu
                                   b.V14=3-SG-P POS V/i

s1_11 . . / ABS------ barag        BASE barag GW dais ID 11 MORPH2
                                   N1=STEM.N5=ABS POS N
s1_12 . . / L2-NH---- ki           BASE ki GW place ID 12 MORPH2 N1=STEM POS N
s1_13 . . . \ amod--- sikil-la     BASE sikil GW pure ID 13 MORPH2
                                   NV2=STEM.N5=L2-NH POS V/i
s1_14 . . \       mu-na-du₃        BASE du₃ GW build ID 14 MORPH2
                                   V4=VEN.V6=3-SG-H.V7=DAT.V11=3-SG-H-A.V12=ST
                                   EM.V14=3-SG-P POS V/t
```

Figure 4: Syntactic pre-annotation of Ur-Namma 5

improved from the feedback of the human annotators along the way. Generally, annotations will be correct as they are created using the rules described in 3.3.; more complex cases are not covered by the rules, so they are to be created by the annotator. Figure 5 shows a screenshot of three examples of relationships between words. Clicking on one term and then another one opens up a panel for choosing the nature of the relationship and creates it on confirmation; selecting a word or a relationship and pressing DEL removes the annotation.

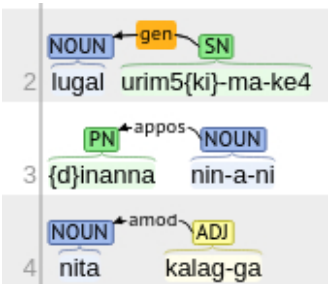

Figure 5: Brat annotation example

3.5. Linking Resources

As described above, morphosyntactic and syntactic annotations of the CDLI corpus have been linked with models of UD parts-of-speech, features and dependency labels, and this information is actively used during syntactic pre-annotation. To facilitate interpretability of our data, it can also be provided as part of the RDF edition of the annotated CDLI corpus.

In addition, morphological features have also been defined in language-independent terms, by linking the existing CDLI/ETCSRI morphological annotation scheme[29] with an ontological model of the UniMorph specifications available as part of the OLiA ontologies[30]. This effectively positions Sumerian among the language corpora that are linked by their linguistic annotations, and employing this schema will also facilitate translation since Unimorph is able to define morphological features in language-independent terms (Sylak-Glassman, 2016, 3). The only digital resource for Sumerian vocabulary is the ePSD[31]. We have prepared an index of deep links, modeled as a lemon dictionary[32]. Until the proper integration of Linked Data into the anticipated upcoming ePSD2 edition, this acts as a placeholder. Despite being a preliminary resource, at best, this index, comprising lemon-compliant lexical entries, forms and senses, already serves to illustrate linking with lexical resources. Additional local lexical resources will abe provided as we prepare the Ur III research corpus.

The CDLI catalog provides metadata on objects bearing cuneiform inscriptions which, especially when integrated into the text analysis method, can prove to be useful for the discovery and study of the artifact. It is stored in a MySQL database and is exported daily in CSV format. We convert the data to RDF with the csv2rdf tool[33] supplemented with embedded custom turtle templates, and link to external metadata repositories: the Modref project and the British Museum.

4. Discussion and Outlook

4.1. Limits of Morphological Pre-Annotation

The first limitation of morphological pre-annotation concerns word identification. Since a word can have different meanings, identifying the right one requires an awareness of the context. The same problem occurs when dealing with forms where case markers were not inscribed; they must be inferred based on the analysis of the whole sentence, or in the case of the Ur III administrative texts, the order of words, since it is often stereotyped. To counteract those limitations, the human annotator analyses the text and corrects and refines the generated annotations.

Because of the sheer quantity of the texts to annotate, semi-automated annotation using the morphological pre-annotation tool coupled with the input of an annotator to prepare all texts is not feasible. As discussed elsewhere, we are developing a machine-learning pipeline for the automated annotation and translation of texts, based on the translation and annotations prepared to form the required gold corpus, using the method described in section 3.1.

4.2. Limits of Syntactic Pre-Annotation

The implementation of syntactic pre-annotation is not a fully-featured parser, but a simple deterministic and greedy algorithm to assist manual annotation. This process, based on 'default rule', allows us to automatically pre-annotate *most* of the material and then correct it, rather than manually annotate everything from scratch.

[29] `http://oracc.museum.upenn.edu/etcsri/parsing/index.html`.

[30] `purl.org/olia/owl/experimental/unimorph/`, also cf. `http://unimorph.org/`.

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

[32] Our index is hosted on the Oracc server, home of the ePSD: `http://oracc.museum.upenn.edu/ttl/epsd1.ttl`.

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

Some examples in which the syntactic pre-annotation analysis will likely be incorrect, are presented below (from Jagersma (2010)):

1. Nominal clause. Clauses that does not contain an independent verbal form might not be parsed correctly in some cases, e.g.:

   urdu$_2$ lu$_2$-še lugal-zu-u$_3$
   urdu2.d lu$_2$ =še =Ø lugal =zu =Ø
   slave man=that=ABS master=your=ABS
   'Slave! Is that man your master?'
   (Jagersma, 2010, 716, no. 7)

2. Word order. Sumerian normally has a SOV word order, with the verb at the final position. However, exceptional right-dislocated clauses are known, e.g.:

   i$_3$-ĝu$_{10}$ i$_3$-gu$_7$-e d nisaba-ke$_4$
   ì =ĝu =Ø 'i -gu$_7$-e nisaba.k=e
   fat=my=ABS VP-eat -3SG.A:IPFV Nisaba =ERG
   'She will eat my cream, Nisaba.'
   (Jagersma, 2010, 300, no. 27)

   Clause boundaries will not be correctly recognized in such cases.

3. Enclitic copula. The Sumerian copula *me* can be both independent and enclitic. In the latter case the analysis of the token in context of other words is ambiguous, as it contains both nominal and verbal annotation, e.g.:

   še dub-sar-ne-kam
   še dub.sar=ene=ak =Ø ='am
   barley scribe =PL =GEN=ABS=be:3N.S
   'This is barley of the scribes.'

   nagar-me-eš$_2$
   nagar =Ø =me-eš
   carpenter=ABS=be -3PL.S
   'They are carpenters.'
   (Jagersma, 2010, 681-2, nos. 24 and 27)

4. Enclitic possessive pronouns and dimensional prefixes. To facilitate subsequent dependency parsing, enclitic possessives are analyzed in terms of their *morphosyntactic* characteristics, not on grounds of their *semantics*: In their function, enclitic possessives are referential and this could be explicitly expressed with explicit links between possessor and possessum within UD using the language-specific but popular `nmod:poss` relation. However, such links cannot be easily integrated into UD-compliant syntactic annotation as it may easily lead to non-projective trees (i.e., crossing edges):

   sipa-de$_3$-ne / gu$_2$-ne-ne-a / e-ne-ĝar
   sipa.d =enē=r(a) gu$_2$ =anēnē='a 'i -nnē -n -ĝar -Ø
   shepherd=PL =DAT neck=their =LOC VP-3PL.OO-3SG.A-place-3N.S/DO
   'He placed this (as a burden) on the shepherds, on their necks.'
   (Jagersma, 2010, 686, no. 21a) In this example, the locative argument syntactically depends on the verb; at the same time, the enclitic possessive (glossed as 'their') refers to the preceding argument. Therefore, these semantic relations are to be captured in a subsequent processing step akin to anaphor resolution in other languages.

This incomplete list gives examples of cases where the analysis by the pre-annotation tool would be incorrect at this time in the development of the tool. But the bulk of these grammatical elements occur very rarely in Ur III administrative texts and royal inscriptions. Still, the pre-annotation algorithm will be extended with more elaborate rules in the future to improve its performance and to incorporate more complex features and constructions since we aim to make this tool useful to annotate all genres of the Sumerian language.

4.3. Conclusions

The workflow that brings ATF raw textual data to publication as Linked Open Data, and the pipeline for text annotation–in particular the annotation of morphology and syntax–described in this paper, draws a roadmap for further development in the processing and analysis of ancient cuneiform languages. Improving and automating the annotation process for Sumerian sources is foundational for future work on cuneiform corpora, while the generation of annotations using a semi-automated annotation process for Sumerian syntax is generally unprecedented and innovative. We find the implementation of new standards for Assyriology as a digital discipline hardly meaningful without compatibility with existing LLOD standards on the one hand, and their adaptation to the particular languages and the material under scrutiny on the other, hence the choice of the CoNLL formats, RDF, UD, and the CIDOC-CRM. Building the machine translation pipeline for Sumerian, the ultimate goal of the MTAAC project, is greatly dependent on this work.
These altogether are crucial steps towards LLOD editions of Sumerian and other cuneiform languages. We hope that our work will help to provide Assyriogists and researchers from other fields with new open access annotated textual datasets, and reusable infrastructure that can significantly contribute to the study of ancient languages and cultures.

Acknowledgements

The Machine Translation and Automated Analysis of Cuneiform Languages project 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.[34]
Our appreciation goes to Heather D. Baker and Robert K. Englund for their insights and suggestions.

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

5. Bibliographical References

Black, J. A., Cunningham, G., Ebeling, 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.

Buil Aranda, C., Corby, O., Das, S., Feigenbaum, L., Gearon, P., Glimm, B., Harris, S., Hawke, S., Herman, I., Humfrey, N., Michaelis, N., Ogbuji, C., Perry, M., Passant, A., Polleres, A., Prud'hommeaux, E., Seaborne, A., and Williams, G. (2013). SPARQL 1.1 overview. https://www.w3.org/TR/sparql11-overview.

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. and Schenk, N. (2018). The ACoLi CoNLL Libraries: Beyond tab-separated values. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

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, Berlin, Heidelberg.

Das, S., Sundara, S., and Cyganiak, R. (2012). R2RML: RDB to RDF mapping language. Technical report.

Hruška, B. (1980). Drei neusumerische Texte aus Drehem. *Die Welt des Orients*, 11:27.

Jagersma, A. H. (2010). *A descriptive grammar of Sumerian*. Ph.D. thesis, Faculty of the Humanities, Leiden University.

Jaworski, W. (2008). *Ontology-based knowledge discovery from documents in natural language*. Ph.D. thesis, Warszawa: Uniwersytet Warszawski.

Mazziotta, N. (2010). Building the Syntactic Reference corpus of Medieval French using NotaBene RDF annotation tool. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 142–146. Association for Computational Linguistics.

Molina, M. (2017). Syntactic annotation for a Hittite corpus: problems and principles.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser: a language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):9 5–135.

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.

Owen, D. I. (2011). *Garshana studies*. CDL Press.

Pagé-Perron, É., Sukhareva, M., Khait, I., and Chiarcos, C. (2017). Machine Translation and Automated Analysis of the Sumerian Language.

Smith, E. (2010). *Query-based annotation and the Sumerian verbal prefixes*. Ph.D. thesis, University of Toronto.

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

Tchienehom, P. L. (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)* , Barcelona, Spain, May.