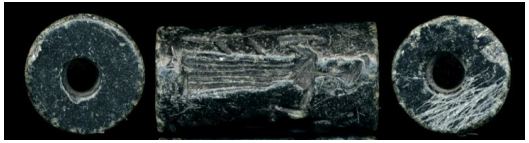
Building ETCSANS: The Electronic Text Corpus of Syntactically Annotated Neo-Sumerian

Christian Chiarcos University of Cologne, Germany
Émilie Pagé-Perron University of Oxford, UK

CDLI Cuneiform Digital Library Initiative

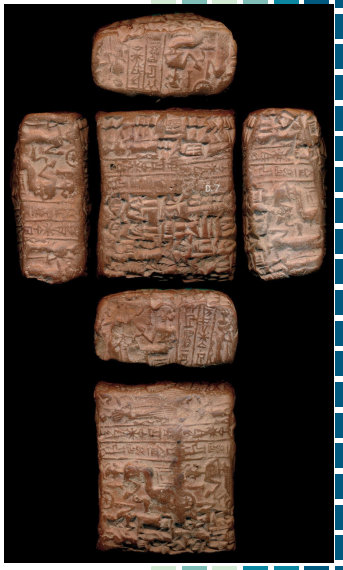


University of Cologne

WOLFSON COLLEGE UNIVERSITY OF OXFORD

UNIVERSITY OF OXFORD

- Cuneiform is script used to inscribe a dozen languages over four millennia in ancient Iraq and beyond
- An estimate of 550 000 ancient artifacts in collections around the world are inscribed using this script
- The CDLI curates, preserves, and shares metadata, text and linguistic annotations, and images of these artifacts (362 000 entries)
- The CDLI has been operating for 23 years and is an essential service for the study of cuneiform cultures



https://cdli.ucla.edu web portal, pre-LOD version

https://cdli.mpiwg-berlin.mpg.de development site

https://github.com/cdli-gh/data daily data dump

https://github.com/cdli-gh code and data

Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC)

- Trans-Atlantic Platform award
  Digging into Data Challenge
  - 2017-2020
    - University of Toronto
    - Goethe Universität Frankfurt
    - University of California, Los Angeles

Trans-Atlantic Platform
Social Sciences and Humanities

DIGGING INTO DATA CHALLENGE

Deutsche Forschungsgemeinschaft DFG

NATIONAL ENDOWMENT FOR THE HUMANITIES

SSHRC CRSH
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada

Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC)

Tools for the study of economy and society of the Neo-Sumerian period (2100-2000 BCE)

- ~100.000 texts !

MTAAC developed an innovative annotation workflow (Chiarcos et al. 2018)

=> the first syntactically annotated corpus of Sumerian

Sumerian

- the very first written language
- written in Mesopotamia (4th to 1st millenium BC)
- genres as diverse as poems and songs over mythological and historical treatises, laws and letters to contracts and administrative records.

Sumerian

agglutinative ergative language and a linguistic isolate

- Much of its syntax is **morphologically** encoded
- e.g. *Suffixanhäufung* (case stacking)

  The last word in a NP takes all the morphological case markers of its syntactic parents

Ur-{d}Nammu lugal Urim5{ki}-ma-ke4

Ur-Nammu, king Ur.**GEN**.**ERG**

"Ur-Nammu, king of Ur (did …)"

Existing Corpora for Sumerian

No corpora for Sumerian syntax

Electronic Text Corpus of Sumerian Literature (ETCSL)

- http://etcsl.orinst.ox.ac.uk
- Post-Sumerian literature, POS-tagged

Electronic Text Corpus of Sumerian Royal Inscriptions

(ETCSRI)

- http://oracc.museum.upenn.edu/etcsri/
- all periods, highly specialized domain, POS+MORPH

The Electronic Text Corpus of Syntactically Annotated Neo-Sumerian

ETCSANS Annotations

Format

The ETCSANS corpus uses a tabular format with tab-separated values (CoNLL, custom columns)

- ETSCRI and ETCSL use JSON and XML formats

# ID	WORD	SERGM	POS	MORPH	HEAD	EDGE	MISC
1	da-da	da-da[3]	PN	PN	11	ERG	_
2	ensi2	ensi2[ruler]	N	N	1	appos	_
3	szuruppak{ki}	szuruppak{ki}[1]	SN	SN.GEN	2	GEN	_
4	ha-la-ad-da	ha-la-ad-da[1]	PN	PN	1	appos	_
5	ensi2	ensi2[ruler]	N	N	1	appos	_
6	szuruppak{ki}	szuruppak{ki}[1]	SN	SN.GEN	5	GEN	_
7	dumu-ni	dumu[child]	N	N.3-SG-H-POSS.ERG	1	appos	_
8	ad-us2	ad-us2[plank]	N	N.ABS	11	ABS	_
9	abul	abul[gate]	N	N	11	LOC	_
10	{d}sud3-da-ke4	{d}sud3[1]	DN	DN.GEN.L3-NH	9	GEN	_
11	bi2-in-us2	us2[follow]	V	3-NH.L3.3-SG-H-A.V.3-SG-P	0	_	_

POS/NER and MORPH annotations

Derived from the Electronic Corpus of Sumerian Royal Inscriptions (ETCSRI, oracc.museum.upenn.edu/etcsri)

- MORPH is slightly simplified (no slot information)

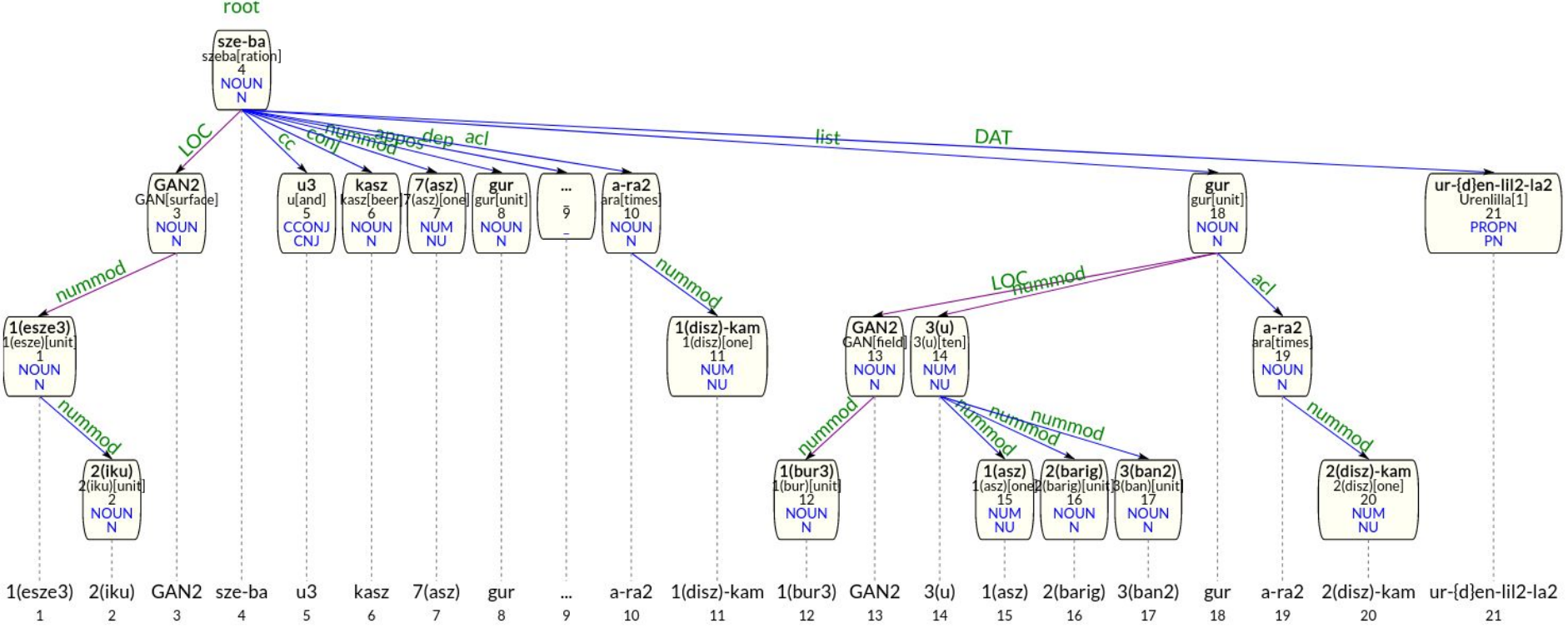
# ID	WORD	SERGM	POS	MORPH	HEAD	EDGE	MISC
1	da-da	da-da[3]	PN	PN	11	ERG	_
2	ensi2	ensi2[ruler]	N	N	1	appos	_
3	szuruppak{ki}	szuruppak{ki}[1]	SN	SN.GEN	2	GEN	_
4	ha-la-ad-da	ha-la-ad-da[1]	PN	PN	1	appos	_
5	ensi2	ensi2[ruler]	N	N	1	appos	_
6	szuruppak{ki}	szuruppak{ki}[1]	SN	SN.GEN	5	GEN	_
7	dumu-ni	dumu[child]	N	N.3-SG-H-POSS.ERG	1	appos	_
8	ad-us2	ad-us2[plank]	N	N.ABS	11	ABS	_
9	abul	abul[gate]	N	N	11	LOC	_
10	{d}sud3-da-ke4	{d}sud3[1]	DN	DN.GEN.L3-NH	9	GEN	_
11	bi2-in-us2	us2[follow]	V	3-NH.L3.3-SG-H-A.V.3-SG-P	0	_	_

Syntax Annotations

Based on Universal Dependencies (UD, universaldependencies.org)

- except: if a morphological case *defines* the type of dependency, we use the case as dependency label
  - UD inventory of nominal roles is not applicable

# ID	WORD	SERGM	POS	MORPH	HEAD	EDGE	MISC
1	da-da	da-da[3]	PN	PN	11	ERG	_
2	ensi2	ensi2[ruler]	N	N	1	appos	_
3	szuruppak{ki}	szuruppak{ki}[1]	SN	SN.GEN	2	GEN	_
4	ha-la-ad-da	ha-la-ad-da[1]	PN	PN	1	appos	_
5	ensi2	ensi2[ruler]	N	N	1	appos	_
6	szuruppak{ki}	szuruppak{ki}[1]	SN	SN.GEN	5	GEN	_
7	dumu-ni	dumu[child]	N	N.3-SG-H-POSS.ERG	1	appos	_
8	ad-us2	ad-us2[plank]	N	N.ABS	11	ABS	_
9	abul	abul[gate]	N	N	11	LOC	_
10	{d}sud3-da-ke4	{d}sud3[1]	DN	DN.GEN.L3-NH	9	GEN	_
11	bi2-in-us2	us2[follow]	V	3-NH.L3.3-SG-H-A.V.3-SG-P	0	_	_



Excerpt from a transaction text: “For 8 gan2 (~2.8 ha) [of acre] 7 gur (~2,100 l) rations and beer, first time, and for 18 gan2 (~6.5 ha) [of acre] 31.5 gur (~9,450 l), second time, for Ur-Enlilla.” (P101040)

The Electronic Text Corpus of Syntactically Annotated Neo-Sumerian

Royal Subcorpus

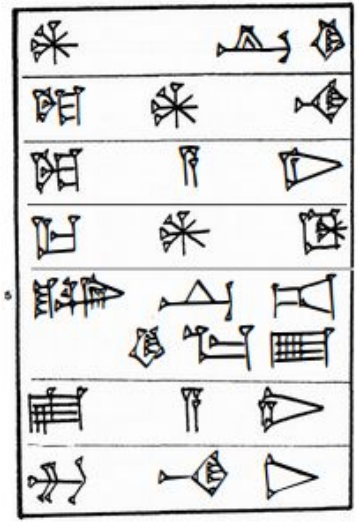
Royal Texts

612 texts, 9,133 tokens

Royal inscriptions often have a rather formulaic structure

For Nanna,
his lord,
Ur-Namma.
king of Ur,
built
his temple

1 {d}Nanna
2 lugal-a-ni
3 Ur-{d}Nammu
4 lugal
5 Urim5{ki}-ma-ke4
6 e2-a-ni
7 mu-na-du3



Morphological Annotation

project morphological annotations from ETCSRI Ur III data to CDLI editions *of the same texts*

- transcription principles differ and need to by adjusted
- sometimes, readings differ
- morphological annotations only (ETCSANS ~ ETCSRI)

Syntactic Annotation

ETCSRI does not provide syntax annotation, but *full morphology*

- Sumerian morphology explicitly encodes
  - NP structure (*Suffixanhäufung*)
  - clausal subordination (~ nominalization)

  => Rule-based conversion

  attachment ambiguitites are heuristically resolved

Systematic Errors

Within a Sumerian NP, a nominal modifier will either carry an adnominal case (e.g., *GEN*) or be an apposition (*appos*).

- The rule-based syntax cannot disambiguate this
- Heuristic disambiguation: Use case labelled dependencies for lower dependencies

The Electronic Text Corpus of Syntactically Annotated Neo-Sumerian

Transaction Subcorpus

Transactional Texts

The vast majority of Ur III texts is administrative in nature

22,276 texts, 1,742,634 tokens

Mostly concerned with the transfer of commodities

Writing of morphology is largely defective, but structure is usually very consistent

- good conditions for a pattern-based approach for annotation

Transactional Texts

P249089, Umma, Ur III period

obverse

1. 5(asz) dabin gur lugal
   5 royal gur flour

2. 3(asz) zi3 sig15 gur
   3 gur rough-ground flour,

3. 5(asz) munu4 si-e3gur
   5 gur sprouted malt,

4. 3(asz) SZIM du gur
   3 gur ŠIM, regular (quality),

reverse

1. 2(asz) SZIM saga gur
   2 gur ŠIM, high (quality),

2. ki ad-da-ta
   from Adda

3. lu2-du10-ga szu ba-ti
   (did) Lu-duga receive;

4. iti min-esz3
   month "mineš."

amount product unit

transaction information

persons involved

date

recording the transfer of ingredients for approx. 7,000 l beer

Morphological Annotation

- manual annotations for morphology for a subcorpus created in the MTAAC project
- automated annotations for POS and NER
  - CRF tagger trained on MTAAC + ETCSRI + ETCSL annotations
- lookup-based annotation of MORPH
  - based on MTAAC + ETSCRI
  - uses partial matches to analyze unseen forms
  - disambiguation by frequency
    - only applied if consistent with POS/NER annotation

Syntactic Annotation

Syntactic annotation is inspired by Jaworski's (2010) pattern-based *semantic* annotation of parts of this data

- Lookup-based annotation of commodities
- CFG parse of transaction agents and transactions
- Rule-based conversion to ETCSANS dependencies

  Recipient => *DAT*, *Supplier* => *ERG*,

  Commodity => *ABS*, etc.

Systematic Errors

- annotation is incomplete
  - covers approx. 75% of all tokens in a transactional text

=> where an English translation is available, complement it with results of annotation projection (cf. parallel subcorpus)

The Electronic Text Corpus of Syntactically Annotated Neo-Sumerian

Parallel Subcorpus

Parallel Subcorpus

1,572 texts, 46,321 tokens

- all Ur-III texts that have English translations in CDLI
- morphological annotation like for transaction corpus
- syntax annotation projected from English translation
  - Stanford Core parser, producing UD v.1 annotations
  - rule-based mapping from UD to ETCSANS
    - Sumerian cases heuristically extrapolated from English prepositions

Systematic Errors

Annotation projection often suffers from alignment errors

- Normally, one Sumerian noun is expressed by the noun and associated function words (determiner, adposition) in English (e.g., *to the workers*)
- If the noun is aligned with a function word rather than with the noun, projected annotations will be incorrect

The Electronic Text Corpus of Syntactically Annotated Neo-Sumerian

Extended ETCSANS Corpus

Extended ETCSANS Corpus

47,476 texts (1,775,582 tokens)

- automated annotations for POS/NER and MORPH
- generic, rule-based annotator that exploits the explicit morphological marking of phrase structure boundaries and clausal subordination

~ royal subcorpus, but *less precise*

(not on manual morphology annotation)

Conclusion

Evaluation

- automated annotation
  - POS: 99.1% (F-score, MTAAC gold corpus)
  - MORPH: 83.2% (accuracy, ETCSRI)

- for syntax, yet to come
  - we created a small (11,220 tokens) test corpus from textbook examples
    - no evaluation yet, because textbook examples need to be aligned with CDLI texts

Release

- developer access via (https://github.com/cdli-gh).
  - CDLI-CoNLL: native tabular format
  - Linked Data edition

    (CoNLL-RDF, Chiarcos & Fäth 2017)
  - TEI edition

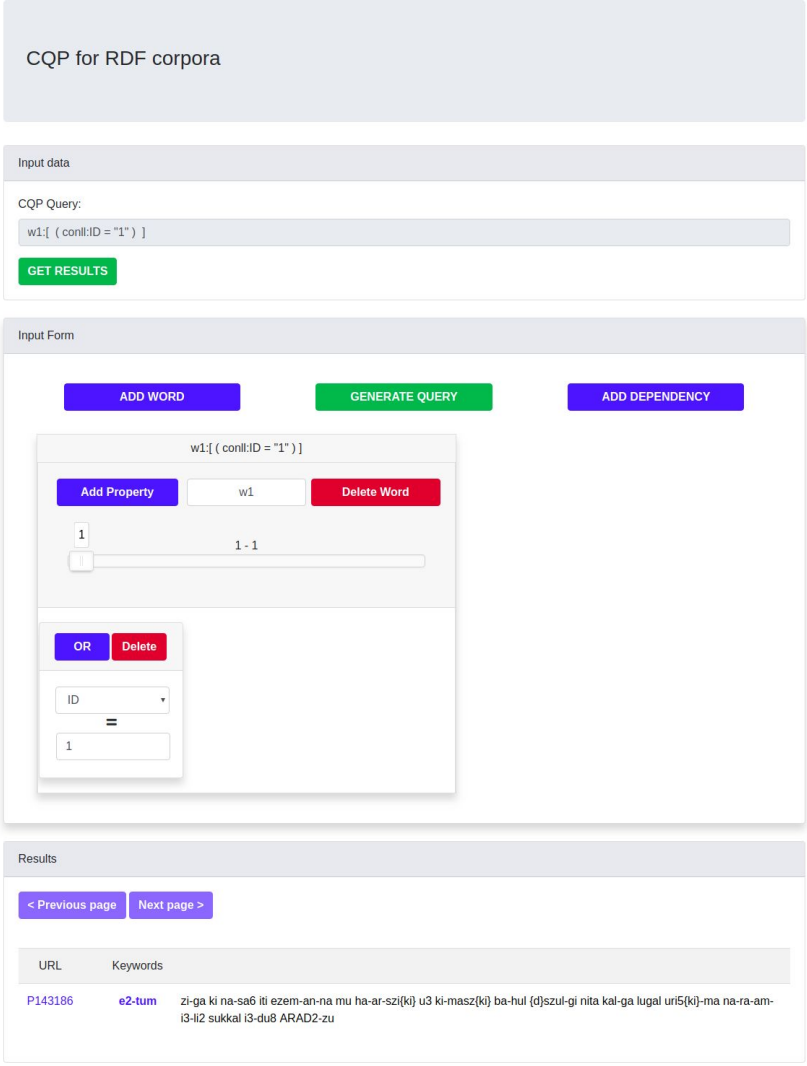
    (for local querying with TEITOK, Janssen 2018)
- New CDLI Framework:
  - all three formats accessible from the CDLI API along with
    - the associated inscribed artifacts, object metadata, bibliography, etc. (native formats and RDF)

Release: Search

- developer access via (https://githu
  - CDLI-CoNLL: native tabular fo
  - Linked Data edition

    (CoNLL-RDF, (
  - TEI edition

    (for local querying with T
- New CDLI Framework:
  - all three formats accessible fr
    along with
    - the associated inscribed artif
      bibliography, etc. (native



Crowdsourcing Annotation Verification

Current annotations are to a large extent automated and will be continuously improved

We plan to achieve this by mobilizing the CDLI community

- transcription is currently crowd-sourced to acknowledged experts
- moderated by CDLI team

The new CDLI Framework is being extended with

- annotation view, annotation validation and annotation updates

Take-Away

- The conventional way to build a syntax corpus doesn't work for Sumerian
  - we have the data, but we cannot hire annotators to build a training corpus
    - Sumerian is a *rare* skill
- Instead, we
  - aggregated and transformed heterogeneous source data **(core technology: RDF!)**
  - developed rule-based annotators **(SPARQL!)**
  - extend existing crowd-sourcing workflow to annotation validation

NATIONAL ENDOWMENT FOR THE HUMANITIES

UCLA CENTER FOR DIGITAL HUMANITIES

HUMTECH
HUMANIT. TECH-KNOWLEDGY. IT STARTS HERE!

UCLA College Humanities

Deutsche Forschungsgemeinschaft
DFG

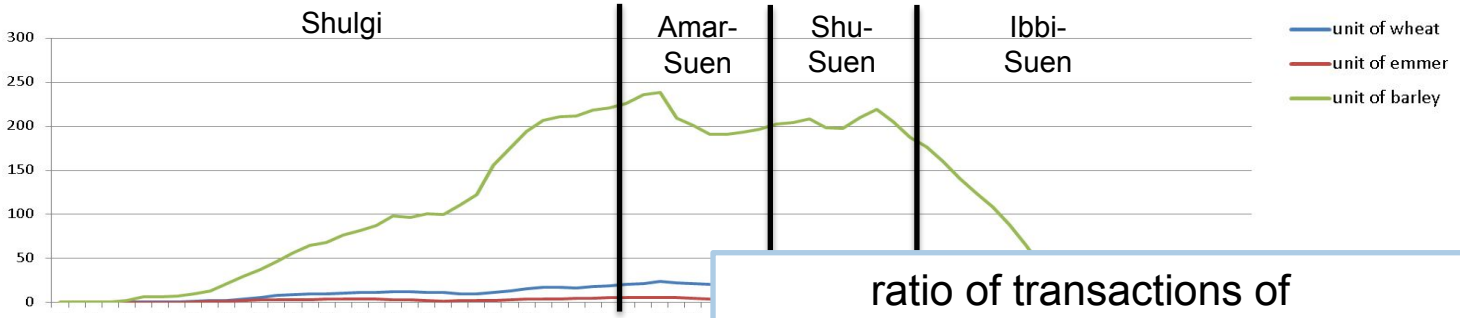
SSHRC CRSH
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada

Google
Summer of Code

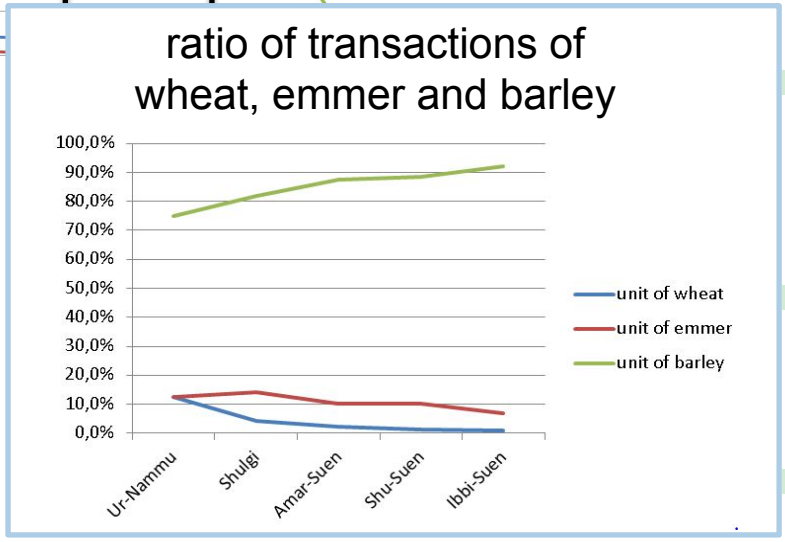
Thank you!

and thanks to our supporters ;)

Insights: Changes in Transactions



- changes in the agrarian sector
  - abandonment of wheat (einkorn, emmer) in favour of barley?
- salination hypothesis (Jacobsen 1982)
  - barley is more salt-resistant than wheat



LOD elements of the annotation pipeline

RDF-Based Syntactic Pre-Annotation

- simple, deterministic and non-lexicalized
- Using SHIFT and REDUCE terminology
- ordered rules to execute SPARQL updates
- REDUCE relations connects tokens with their respective HEAD

$\text{NOUN}_0\ \text{ADJ}_{\text{CASE}} \Rightarrow \text{NOUN}_{\text{CASE}} \xleftarrow{\text{amod}} \text{ADJ}$

e.g., nita $\xleftarrow{\text{amod}}$ kalag-ga "strong male".

$\text{NOUN}\ \text{NOUN}_{\text{GEN}} \Rightarrow \text{NOUN} \xleftarrow{\text{GEN}} \text{NOUN}$

e.g., lugal $\xleftarrow{\text{GEN}}$ $\text{urim}_5^{\text{ki}}$-ma "king of Ur".

$\text{NOUN}_0\ \text{NOUN}_{\text{CASE}} \Rightarrow \text{NOUN}_{\text{CASE}} \xleftarrow{\text{appos}} \text{NOUN}$

e.g., $^{\text{d}}\text{inana}_{\text{DAT}}$ $\xleftarrow{\text{appos}}$ nin-a-ni "to Inanna, his lady".

$\text{NOUN}_0\ \text{NOUN}_{\text{CASE1+CASE2}} \Rightarrow \text{NOUN}_{\text{CASE1}}$ $\xleftarrow{\text{CASE 2}}$ NOUN

https://www.mdpi.com/2078-2489/9/11/290