FROM SIGNS TO SEMANTICS
A PIPELINE FOR AKKADIAN TEXTS

Aleksi Sahala

University of Helsinki, Finland

TOPICS OF THE DAY

- Phonological transcription of Akkadian
  - Required for →
- Lemmatization, POS-tagging and morphological Analysis
  - Required (lemmas) for →
- Semantic analysis
  - Improving word embeddings and association measures

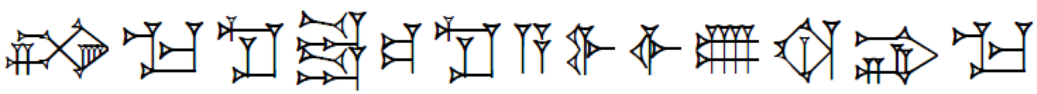
AKKADIAN LANGUAGE

- Documented from ca. 2400 BCE to 150 CE.
- An East-Semitic language
  - Old/Sargonic Akkadian (2400–2100 BCE)
  - Babylonian (2100 BCE–150 CE)
  - Assyrian (2000–612 BCE)
- Very important culture-historical language
  - Codex Hammurabi, Epic of Gilgameš, lots of information about the early days of human civilization!

Sargon of Akkad
(National Museum of Iraq)

WRITING SYSTEM

- Logo-syllabic
- About 1000 signs of which ca. 200 commonly used in Akkadian
- Highly ambiguous: signs may have up to dozens of readings!

(1) 

(2) *šum-ma* MA$_2$-LAḪ$_4$ ^{giš}MA$_2$ *a-wi-lim u$_2$-ṭe-bi-ma*

(3) *šumma mallāḫum eleppi awīlim uṭebbīma*

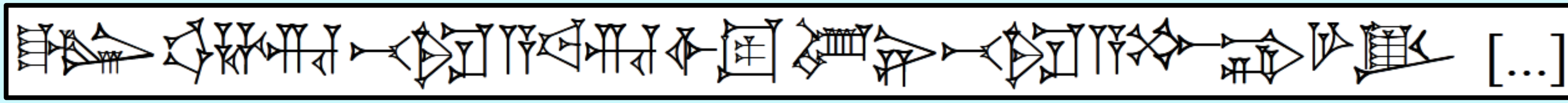
"If a sailor sank a boat of a free man (and made it refloat it, he shall give half of the boat's price in silver)"

AKKADIAN CORPUS

- Open Richly Annotated Cuneiform Corpus (Oracc)
  - 8,000 texts (1,500,000 words)
  - ca. 1,400,000 words lemmatized

AKKADIAN CORPUS

- Open Richly Annotated Cuneiform Corpus (Oracc)
  - 8,000 texts (1,500,000 words)
  - ca. 1,400,000 words lemmatized
- More data available but not digitized
  - 10M words in total (estimate by M. Streck 2011)
  - Automatic digitization and annotation tools needed

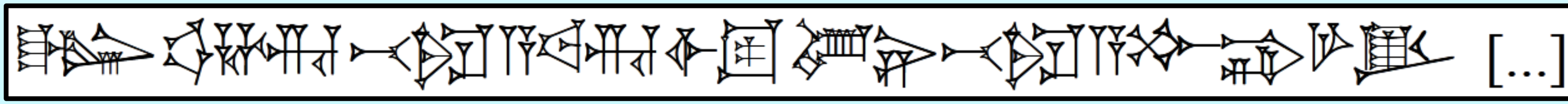


LUGAL *tam-ḫa-ri be-el a-ba-ri u$_3$ dun-ni be-el a-bu-bi ša-kin* [...]

šar tamḫāri bēl abāri u dunni bēl abūbi šakin [...]

šarru+N+masc+nom+sg; tamḫāru+N+masc+gen+sg; bēlu+N+masc+nom+sg+construct

šarru; šarru; 1.000; šarru; tamḫāru; 0.254; šarru; bēlu; 0.642

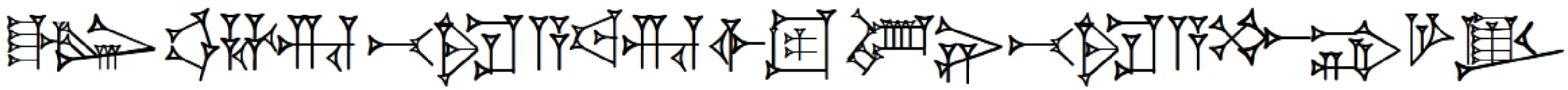


LUGAL *tam-ḫa-ri be-el a-ba-ri u$_3$ dun-ni be-el a-bu-bi ša-kin* [...]

***šar tamḫāri bēl abāri u dunni bēl abūbi šakin* [...]**

šarru+N+masc+nom+sg; tamḫāru+N+masc+gen+sg; bēlu+N+masc+nom+sg+construct

šarru; šarru; 1.000; šarru; tamḫāru; 0.254; šarru; bēlu; 0.642



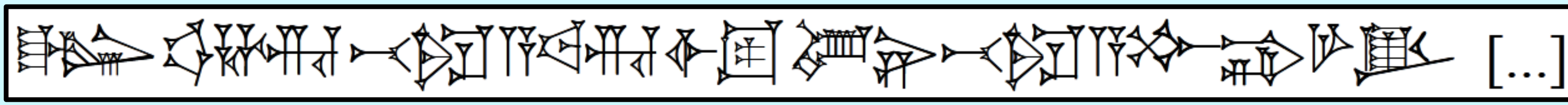
[...]

LUGAL *tam-ḫa-ri be-el a-ba-ri u*$_3$ *dun-ni be-el a-bu-bi ša-kin* [...]

***šar tamḫāri bēl abāri u dunni bēl abūbi šakin* [...]**

šarru+N+masc+nom+sg; tamḫāru+N+masc+gen+sg; bēlu+N+masc+nom+sg+construct

šarru; šarru; 1.000; šarru; tamḫāru; 0.254; šarru; bēlu; 0.642

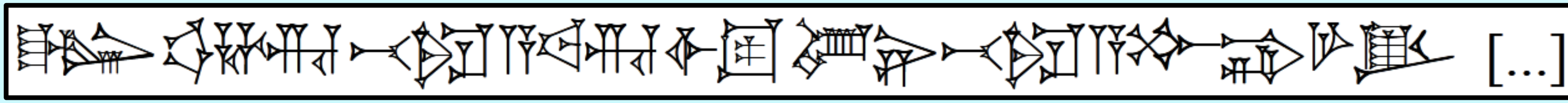


LUGAL *tam-ḫa-ri be-el a-ba-ri u*$_3$ *dun-ni be-el a-bu-bi ša-kin* [...]

šar tamḫāri bēl abāri u dunni bēl abūbi šakin **[...]**

šarru+N+masc+nom+sg; tamḫāru+N+masc+gen+sg; bēlu+N+masc+nom+sg+construct

šarru; šarru; 1.000; šarru; tamḫāru; 0.254; šarru; bēlu; 0.642

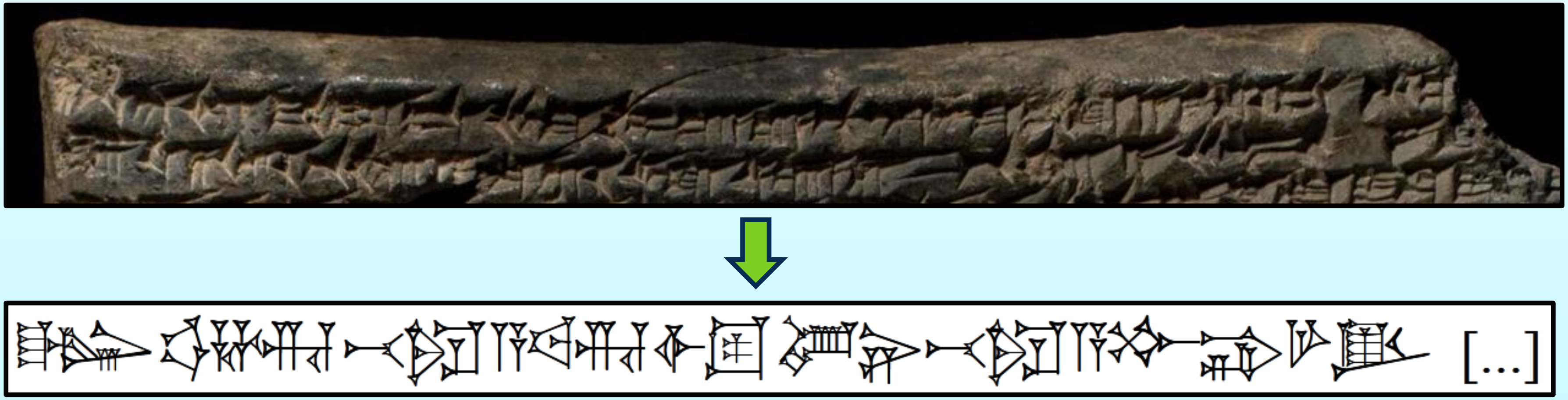


LUGAL *tam-ḫa-ri be-el a-ba-ri* u_3 *dun-ni be-el a-bu-bi ša-kin* [...]

šar tamḫāri bēl abāri u dunni bēl abūbi šakin [...]

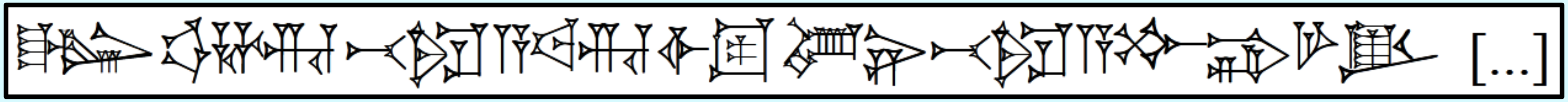
šarru+N+masc+nom+sg; tamḫāru+N+masc+gen+sg; bēlu+N+masc+nom+sg+construct

šarru; šarru; 1.000; šarru; tamḫāru; 0.354; šarru; bēlu; 0.642



OCR of Cuneiform

- Over 50 research papers published since 1980s
- Many papers focus on improving the 3D/2D-representations of tablets
  - Vectorized, rasterized, graph representations etc. etc.
- Incredibly difficult task
  - Inconsistent source data, segmentation etc.
- State-of-the-art sign spotters can reach 90% accuracy in restricted in-domain settings. Full-scale evaluations do not exist.



LUGAL *tam-ḫa-ri be-el a-ba-ri ù dun-ni be-el a-bu-bi ša-kin* […]

Transliteration and tokenization

- State-of-the-art transliteration from OCR has an accuracy of 10%
  (Bogacz et al. 2017)
- From unicode ca. 97% in-domain, 70% out-of-domain accuracy
  (Gordin et al. 2020)
- Models used in Chinese and Japanese do not perform very well
  (Homburg 2016)
- Challenges:
  - Exponentially growing ambiguity
  - Sign segmentation if done from OCR: signs lack fixed lenght and may overlap!
  - Lack of sign-by-sing labeled training data

AUTOMATIC PHONOLOGICAL TRANSCRIPTION

Sahala, Silfverberg, Arppe & Lindén (2020). *Automated phonological transcription of Akkadian cuneiform text. Proceedings of The 12th Language Resources and Evaluation Conference,* pp. *3528-3534.*

PHONOLOGICAL TRANSCRIPTION
THE TASK

- Task
  - Assign correct consonant and vowel quantities, e.g.
    - ***i-be-el*** → *ibēl* ’he ruled’ vs. *ibêl* ’he rules’
    - ***i-di-in*** → *idin* ’give!’ vs. *iddin* ’he gave’
    - ***a-na-ku*** → *anāku* ’I’ vs. *annaku* ’tin’

PHONOLOGICAL TRANSCRIPTION
THE TASK

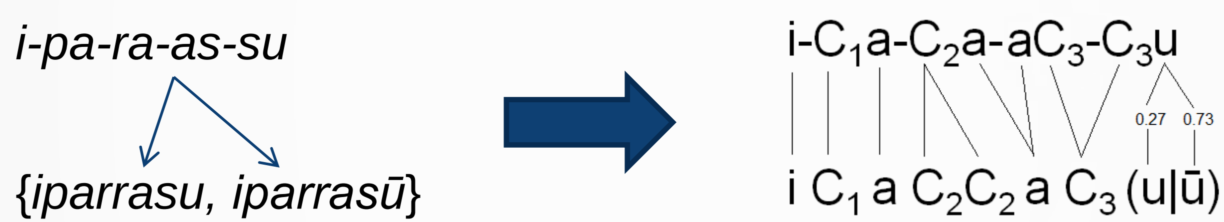
- Task
  - Assign correct consonant and vowel quantities, e.g.
    – ***i-be-el*** → *ibēl* 'he ruled' vs. *ibêl* 'he rules'
    – ***i-di-in*** → *idin* 'give!' vs. *iddin* 'he gave'
    – ***a-na-ku*** → *anāku* 'I' vs. *annaku* 'tin'
- Transcribe logograms into wordforms
  - Relation is suppletive, e.g. DU → *alāku*, *illik*..., DU_3 → *banû*, *ibni* ...
  - Extreme (theoretical) ambiguity:
    – **IGI** → *pān*, *pānu*, *pāni*... 'front', *maḫar*, *maḫru*, *maḫri*... 'before'; *amāru*, *īmur*, *immar*, *ītamar*, *innamir*... 'to see'

PHONOLOGICAL TRANSCRIPTION
METHODS

- Training data 337k tokens divided into 80/10/10 training/dev/test sets
- Baseline: dictionary lookup that chooses the most common transcription
  - {”i-pa-ar-ra-su” : ”iparrasū”, ...}
- Statistical-heuristic model that learns abstract relations and their mapping probabilities (just a Python script, nothing fancy)
- LSTM attentional encoder-decoder with context-awareness

PHONOLOGICAL TRANSCRIPTION
METHODS

- Statistical-heuristic mapping (Abstract Pattern Maps)
  - Exploit the Semitic root-pattern morphology of Akkadian
  - Learn mappings between transliteration and transcription and their probabilities from a corpus (Oracc)



  - Can generalize correct phoneme quantities for all words that belong to the same conjugational class (if they have the same spelling).
    – *i-ga-ma-ar-ru* → *igammarū*, *igammaru* and *i-ša-pa-ar-ru* → *išapparū, išapparu*

PHONOLOGICAL TRANSCRIPTION
METHODS

- LSTM attentional encoder-decoder
  - Input sequence as character embeddings
  - One hidden layer
- Three models
  - non-context aware
  - context-aware (character based context)
  - context-aware (token based context)

non-context aware		b e - e l	
context-aware (character based context)	i - n a	b e - e l	E_2
context-aware (token based context)	<i-na>	b e - e l	<E_2>

PHONOLOGICAL TRANSCRIPTION
EVALUATION

- Intrinsic
  - Test how often the model produces the wanted phonological form
- Extrinsic
  - Feed 2000 auto-transcribed outputs into morphological analyzer
  - Test only if they produce correct lemmata and POS-tag
    - – No morphological gold standard available to evaluate morph. labels.

PHONOLOGICAL TRANSCRIPTION
INTRINSIC

Syllabic

	Baseline	Stat	Enc-Dec	Enc-Dec+Context	Enc-Dec+Char-Context
Recall @ 1	81.37	87.25	89.44	**90.01**	89.59
Recall @ 3	83.74	91.93	**96.65**	96.19	95.91
Recall @ 10	83.75	92.55	**98.14**	97.58	97.33

Logograms / Logo-syllabic

	Baseline	Stat	Enc-Dec	Enc-Dec+Context	Enc-Dec+Char-Context
Recall @ 1	60.70	60.64	57.72	**69.10**	68.70
Recall @ 3	82.15	**82.16**	81.14	81.97	81.86
Recall @ 10	**88.90**	**88.90**	88.79	86.09	86.17

PHONOLOGICAL TRANSCRIPTION
EXTRINSIC

	Baseline	Stat	Enc-Dec	Enc-Dec+Context	Enc-Dec+Char-Context
Recall @ 1	76.66	84.40	87.25	**89.85**	89.30
Precision @ 1	38.75	38.33	37.22	**38.82**	38.79
Recall @ 3	80.05	89.70	**94.31**	93.70	93.45
Precision @ 3	**35.10**	34.73	31.54	30.49	29.64
Recall @ 10	80.60	90.50	**96.50**	95.55	95.80
Precision @ 10	**31.42**	31.12	26.34	22.24	22.19

- With human-transcribed unambiguous inputs we got
  - Recall 96.6
  - Precision 41.2
- The neural model works pretty well!

PHONOLOGICAL TRANSCRIPTION
ISSUES

- Readings of unseen logograms cannot be predicted!
  - Suppletive relation:
    - DU <-> *alāku* ”to go”
    - DU$_3$ <-> *banû* ”to build”
  - Ways to guess the reading based on syllabic example?
    - Probably not enough training data

I took a bus to home	I took a  to 

”BABYFST” MORPHOLOGICAL ANALYZER

Sahala, Silfverberg, Arppe & Lindén (2020). BabyFST - Towards a Finite-State Based Computational Model of Ancient Babylonian. *Proceedings of the 12th Conference on Language Resources and Evaluation,* pp. 3528–3534. European.

Luukko, Sahala, Hardwick & Lindén (2020). Akkadian Treebank for early Neo-Assyrian Royal Inscriptions. *Proceedings of the 19th Workshop on Treebanks and Linguistic Theories.* pp. 124-134.

”BABYFST” - MORPHOLOGICAL ANALYZER

- Only one comprehensive Akkadian morph. analyzer exists for the Old Assyrian dialect (Bamman 2012)
- **BabyFST** is optimized for Babylonian
- Covers language stages over a timespan of 2000 years
  - This is necessary as the Standard Babylonian literary language is based on Old Babylonian (ca. 2000-1600 BCE) but it has some residue from the contemporary dialects.
  - Can be modified for individual dialects easily.

”BABYFST” - MORPHOLOGICAL ANALYZER

- Written in XFST (Beesley & Karttunen 2003), compiled in Foma (Hulden 2009)
- Verb lexicon (350k items)
  - Stems enumerated from Sahala (2011, 2014)
  - ca. 2000 roots and 1400 patterns
- Other parts-of-speech lexicons (ca. 50k items)
  - Semi-automatically generated from Oracc lemmata
- 15MB, 550k states, 1M arcs, 4.77×10^{12} paths

”BABYFST” PERFORMANCE

	OB	MB	SB	NB	LB
Nouns	96.3%	96.3%	96.7%	96.4%	97.6%
Verbs	89.8%	89.0%	92.1%	87.8%	88.4%
Adjectives	97.9%	98.6%	98.0%	97.5%	95.5%
Adverbs	98.6%	98.6%	99.1%	98.1%	98.8%
Pronouns	92.0%	90.8%	95.0%	92.5%	95.6%
AVG	**94.9%**	**94.7%**	**96.2%**	**94.5%**	**95.2%**

- Task: produce correct lemma and POS tag for 1M tokens.
- Average precision ca. 40%

NEXT STEPS

- Disambiguation for lemmatization + POS-tagging
- Disambiguation for morphology
  - Gold standard in the making by the Akkadian Treebanking Project.
- Lemmatize texts from transliterated corpora

CSW
CONTEXT SIMILARITY WEIGHTED
WORD ASSOCIATION MEASURES AND WORD EMBEDDINGS

Sahala & Lindén (2020). Improving Word Association Measures in Repetitive Corpora with Context Similarity Weighting*. Proceedings of the 12th International Conference on Knowledge Discovery and Information Retrieval.*

Svärd, Alstola, Jauhiainen, Sahala, & Lindén (2021). Fear in Akkadian Texts: New Digital Perspectives on Lexical Semantics. *The Expression of Emotions in Ancient Egypt and Mesopotamia, ed. by Hsu, S.-W. & Llop-Raduà, J. Leiden: Brill, pp. 470-502. Culture and History of the Ancient Near East 116* (in press).

CONTEXT OF RESEARCH: LEXICOGRAPHY

- Semantic Domains in Akkadian Texts Project (2017-2020)
- Center of Excellence in Ancient Near Eastern Empires (2018-2025)
- How to study lexicography of a long-extinct language?
  - No informants

    –> Emic approach

CONTEXT OF RESEARCH: LEXICOGRAPHY

- Semantic Domains in Akkadian Texts Project (2017-2020)
- Center of Excellence in Ancient Near Eastern Empires (2018-2025)
- How to study lexicography of a long-extinct language?
  - No informants

    –> Emic approach
- Syntagmatic relationships
  - Word association measures
- Paradigmatic relationships
  - Word embeddings

ISSUES WITH AKKADIAN DATA

- Sparse data (total ca. 1.4M lemmatized words)
  - Count-based embeddings > word2vec, fastText
    - PPMI+SVD, PMI_{δ}+SVD, $PPMI_{\lambda}$+SVD (Bullinaria & Levy 2007, Jungmaier et al. 2020 etc.)

ISSUES WITH AKKADIAN DATA

- Sparse data (total ca. 1.4M lemmatized words)
  - Count-based embeddings > word2vec, fastText
    - PPMI+SVD, PMI_{δ}+SVD, $PPMI_{\lambda}$+SVD (Bullinaria & Levy 2007, Jungmaier et al. 2020 etc.)
- Lots of partial and full duplication
  - Formulaic way of writing
  - Stylistic repetition
  - Fragments or more or less different versions same texts

ISSUES WITH AKKADIAN DATA

- Sparse data (total ca. 1.4M lemmatized words)
  - Count-based embeddings > word2vec, fastText
    - PPMI+SVD, PMI_{δ}+SVD, $PPMI_{\lambda}$+SVD (Bullinaria & Levy 2007, Jungmaier et al. 2020 etc.)
- Lots of partial and full duplication
  - Formulaic way of writing
  - Stylistic repetition
  - Fragments or more or less different versions same texts
- Problem
  - Words in repetititeve passages are statistically over-represented
  - Word embeddings produce very similar results with association measures

REDUCING THE EFFECT OF (PARTIAL) DUPLICATION

- We want to reduce the impact of duplication consistently
- We do not want to alter the source data manually
  - Avoid having to explain why text/part/fragment X was removed instead of Y
  - Reproducibility

HOW CSW WORKS?

$$PMI(a;b) = \log_2 \frac{p(a,b)}{p(a)p(b)}$$

$$p(a,b) = \frac{\varphi(a,b) \cdot f(a,b)}{N}$$

$$\varphi(a,b) = \left(\frac{1}{m} \sum_{i=1}^{w} \frac{|V_i|}{\max(|W_i|,1)} \right)^k$$

- Algorithm
  1. Store co-occurrence windows of words **a** and **b**, aligned by **a**
  2. Count the proportion of unique context words $w \notin \{a, b\}$ at each window position i
     1. Ignore words *a* and *b*
  3. Calculate average proportion ignoring zero-values to get $\varphi(a,b)$

[**rome**	is	the	capital	of	**italy**		]
[**rome**	is	the	capital	of	**italy**		]
[**rome**	is	the	largest	city	in	**italy**]	

HOW CSW WORKS?

$$PMI(a;b) = \log_2 \frac{p(a,b)}{p(a)p(b)}$$

$$p(a,b) = \frac{\varphi(a,b) \cdot f(a,b)}{N}$$

$$\varphi(a,b) = \left(\frac{1}{m}\sum_{i=1}^{w}\frac{|V_i|}{\max(|W_i|,1)}\right)^k$$

- Algorithm
  1. Store co-occurrence windows of words **a** and **b**, aligned by **a**
  2. Count the proportion of unique context words $w \notin \{a, b\}$ at each window position i
     1. Ignore words a and b (= Rome and Italy)
  3. Calculate average proportion ignoring zero-values to get $\varphi(a,b)$

[rome	is	the	capital	of	italy		]
[rome	is	the	capital	of	italy		]
[rome	is	the	largest	city	in	italy]	
[0.0	0.33	0.33	0.67	0.67	1.0	0.0	]

HOW CSW WORKS?

$$PMI(a;b) = \log_2 \frac{p(a,b)}{p(a)p(b)}$$

$$p(a,b) = \frac{\varphi(a,b) \cdot f(a,b)}{N}$$

$$\varphi(a,b) = \left(\frac{1}{m} \sum_{i=1}^{w} \frac{|V_i|}{\max(|W_i|,1)} \right)^k$$

- Algorithm

1. Store co-occurrence windows of words **a** and **b**, aligned by **a**
2. Count the proportion of unique context words $w \notin \{a, b\}$ at each window position i
    1. Ignore words a and b
3. Calculate average proportion ignoring zero-values to get $\varphi(a,b)$

[rome	is	the	capital	of	italy		]
[rome	is	the	capital	of	italy		]
[rome	is	the	largest	city	in	italy]	
[0.0	0.33	0.33	0.67	0.67	1.0	0.0	] = 0.6

HOW CSW WORKS?

- Raise weight to the power of *k* → Better results $\varphi(a,b)^k$ k=2 or k=3
- Element-wise multiply sparse co-occurrence matrix with the weight matrix
- Calculate desired PMI-variant
  - Use as they are for association measures
  - Truncate with SVD for word embeddings

HOW CSW WORKS?

- Raise weight to the power of k → Better results $\varphi(a,b)^k$ k=2 or k=3
- Element-wise multiply sparse co-occurrence matrix with the weight matrix
- Calculate desired PMI-variant
  - Use as they are for association measures
  - Truncate with SVD for word embeddings
- Issues
  - $O(n \times m^2)$ space complexity (n = corpus size, m = window size)

    –> Not feasible for very large corpora; can be optimized to $O(\frac{n \times m^2}{2})$

BOUNDS

- If all contexts are perfectly unique: $\varphi(a,b) = 1.0^k$
- If all contexts are prefectly similar: $\varphi(a,b) = (1/n)^k$

BOUNDS

- If all contexts are perfectly unique: $\varphi(a,b) = 1.0^k$
- If all contexts are prefectly similar: $\varphi(a,b) = (1/n)^k$
- Redefines PMI as follows:
  - Maximum score is achieved when all co-occurrences convey previously unseen information and the words are in perfect statistical dependency

BOUNDS

- If all contexts are perfectly unique: $\varphi(a,b) = 1.0^k$
- If all contexts are prefectly similar: $\varphi(a,b) = (1/n)^k$
- Redefines PMI as follows:
  - Maximum score is achieved when all co-occurrences convey previously unseen information and the words are in perfect statistical dependency
- Consider CSW as a re-ordering operation
  - Move uninteresting co-occurrences out of the window
  - Thus we do not adjust the marginal probabilities or the corpus size (i.e. we won't remove anything from the corpus)!

OBSERVATIONS IN AKKADIAN

	$k = 0$	$k = 1$	$k = 3$
1	dangerous	attack	attack
2	attack	enemy	to attack
3	enemy	army	enemy
4	army	to attack	army
5	weapon	downfall	downfall
6	*gall bladder	*gall bladder	*gall bladder
7	*bright	to kill	to kill
8	to overthrow	to overthrow	border (of land)
9	*frost	weapon	stranger, outsider
10	people	*bright	to bind

Royal Inscriptions of the Neo-Assyrian Period

ba-hu-la-te {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣu	ša ul-tu ul-la a-na
ba-hu-la-ti {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣu	ša ul-tu ul-la a-na
ba-hu-la-ti {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣu	ša ul-tu ul-la a-na
ba-hu-la-ti {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣi	i-na {GIŠ}TUKUL
ba-hu-la-te {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣi	i-na {GIŠ}TUKUL.
ba-hu-la-te {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣi	i-na {GIŠ}TUKUL.
ba-hu-la-ti {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣi	i-na {GIŠ}TUKUL.
ba-hu-la-a-ti {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣi	i-na {GIŠ}TUKUL.
ba-hu-la-ti {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣi	i-na {GIŠ}TUKUL.
ba-hu-la-ti {URU}hi-rim-me {LU$_2$}KUR$_2$ ak-ṣu	ša ul-tu ul-la a-na

- Allows us to take a look on the freer use of language beyond formulaic litanies
- PMI(king, X), X = good things;
- PMI(king, X)+CSW, X = not only good things

EVALUATION

- Calculate average repetition in Akkadian corpus
  - Take 1000 random pairs of words and calculate average window similarity (1- φ)
- Artificially duplicate English Wikipedia corpus.
  - 10% repetition, 17% repepetition, 25% repetition (as in Akkadian)
- Bootsrap by sampling 100 random 2M and 10M word corpora from the duplicated base corpus
- Test with symmetric window sizes of 3, 5 and 7 by using eight different PMI variants and *k*-values between 0 (= no CSW) and 6.
- Calculate average Spearman correlation with Wordsim353 relatedness set (Agirre et al. 2009) and Mturk771 (Halawi et al. 2012)
- Compare results with and without CSW

RESULTS (WORD RELATEDNESS)

	No CSW	$k = 1$	$k = 2$	$k = 3$	Target
Low repetitiveness (< 0.1)					
10M-3	0.39	0.40	0.40	**0.40**	0.40
10M-5	0.48	0.50	0.52	**0.52**	0.52
10M-7	0.52	0.54	0.55	**0.56**	0.56
Moderate repetitiveness (< 0.17)					
10M-3	0.37	0.38	**0.39**	0.39	0.40
10M-5	0.46	0.50	0.51	**0.52**	0.52
10M-7	0.50	0.53	0.55	**0.56**	0.56
High repetitiveness (< 0.25)					
10M-3	0.34	0.36	**0.37**	0.37	0.40
10M-5	0.42	0.46	0.48	**0.49**	0.52
10M-7	0.45	0.50	0.53	**0.54**	0.56

- CSW+PMIδ (Pantell & Lin 2002), 10M setting average spearman correlations with the gold standard

RESULTS (WORD RELATEDNESS)

- Observations
  - PMI measures with low-frequency bias benefit less of CSW
    - PMI (Church & Hanks 1990), NPMI (Bouma 2009), PMIα (Omer & Levy 2015)
  - Freq-balanced PMI measures benefit more
    - PMI^2, PMI^3 (Daille 1994), PMIδ (Pantell & Linn 2002), $NPPMI^2$ (Sahala 2020)
- CSW also consistently improves results in corpora without artificial repetition (a little)
  - Wikipedia 10M w=7: 0.54 → 0.56 at k=3
  - Reducing the impact of uninteresting information matters

PRELIMINARY RESULTS (WORD SIMILARITY)

- Preliminary tests with Akkadian word similarity gold standard
  - Developed as a joint-project by University of Helsinki, LMU Munich and UC Berkeley
  - At the moment we have 300 word pairs ranked by five independently working Assyriologists

PRELIMINARY RESULTS (WORD SIMILARITY)

- Best achieved scores with different embeddings
- 1.5M word Akkadian corpus (OOV rate = 0.0)

Embeddings	**Spearman's σ**
PPMI+SVD+CSW	0.344
fastText	0.232
PPMI+SVD	0.225 (Bullinaria & Levy 2007)

CONCLUSIONS + FUTURE WORK

- Problems tackled:
  - Phonological transcription
  - Automatic lemmatization, POS-tagging and morphological analysis
  - Problems with word association measures and word-embeddings

CONCLUSIONS + FUTURE WORK

- Problems tackled:
  - Phonological transcription
  - Automatic lemmatization, POS-tagging and morphological analysis
  - Problems with word association measures and word-embeddings
- Things to do
  - Disambiguate BabyFST output
  - Finish morphological gold standard
  - Finish Akkadian word similarity gold standard
  - Lemmatize lots of texts and morphologically annotate Oracc

THANK YOU!