A Pathway from Language Change to Computational Lexicography

IMS Colloquium at Stuttgart

Department of Computing Science

Dr. Wei Zhao 10/06/2024

Email: wei.zhao@abdn.ac.uk

Talk based on recent papers

- Syntactic Language Change in English and German: Metrics, Parsers, and Convergences (Arxiv-24)
  - Yanran Chen, Wei Zhao, Anne Breitbarth, Manuel Stoeckel, Alexander Mehler, Steffen Eger
- Graph-based Clustering for Detecting Semantic Change Across Time and Languages (EACL-24)
  - Xianghe Ma, Michael Strube, Wei Zhao
- Presence or Absence: Are Unknown Word Usages in Dictionaries? (Arxiv-24)
  - Xianghe Ma, Dominik Schlechtweg, Wei Zhao

Language Change

Old English
500 - 1150

A. Flexible word order
B. Rich inflectional system

Middle English
1150 - 1500

A. Loss of inflectional endings
B. Reliance on word order

Modern English
1500 - Present

A. Subject-Verb-Object
B. Simplified morphology

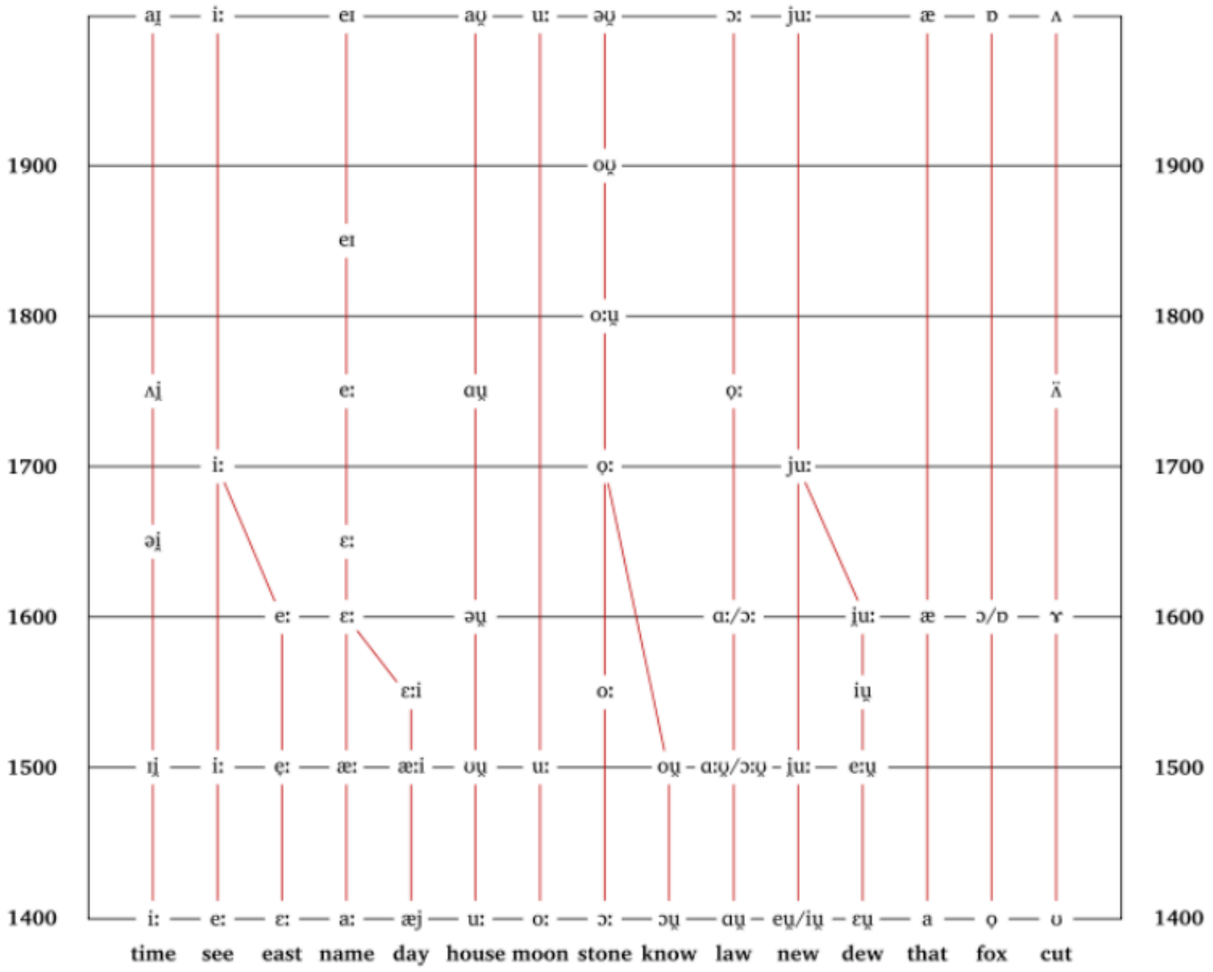
Semantic Change

1960s

1990s

Chill

Sound Change



Conservative and Innovative periods in English

Great Vowel Shift (1400-1700)

English Revolution

Peasants' Revolt

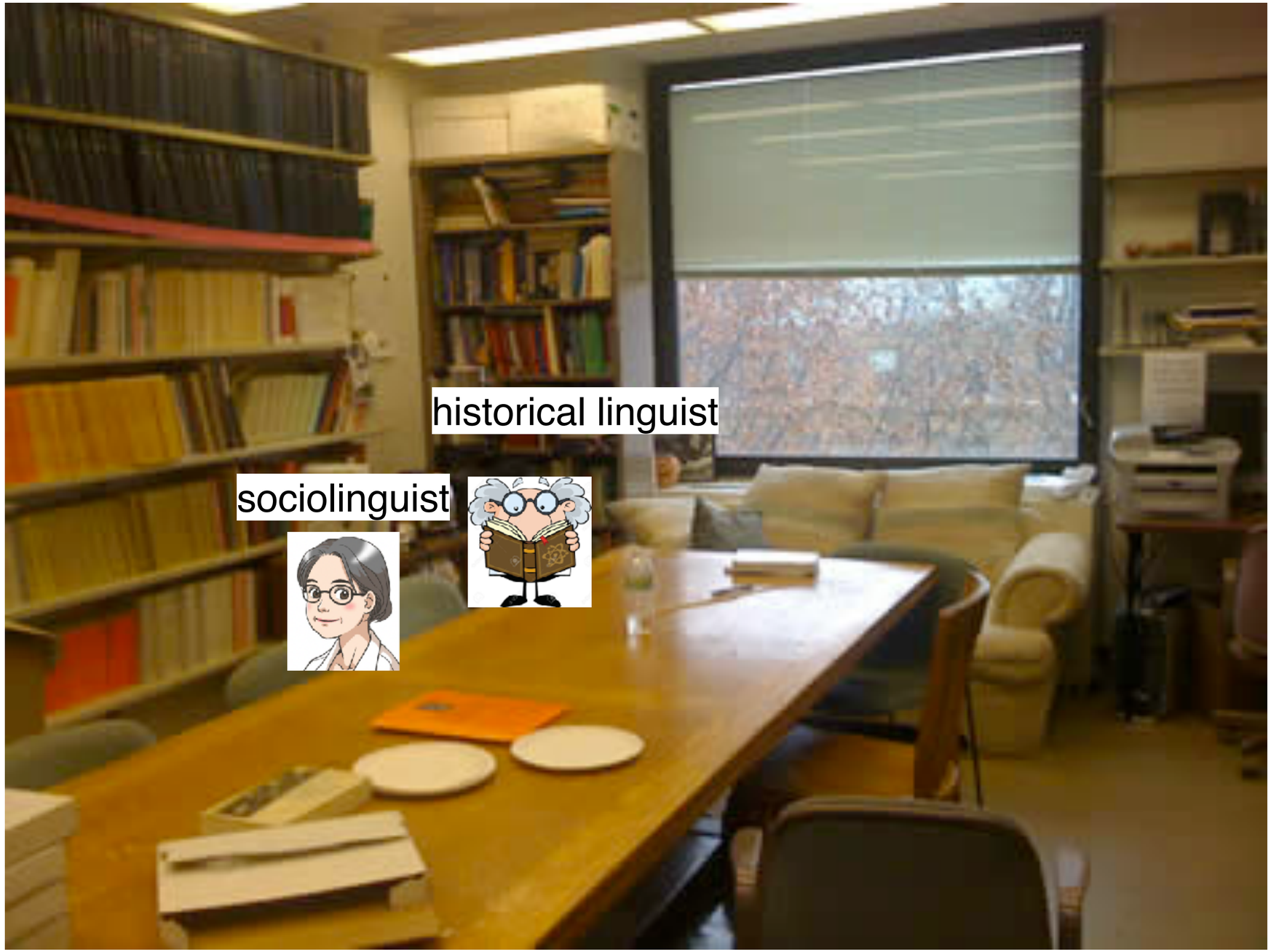
Syntactic Change

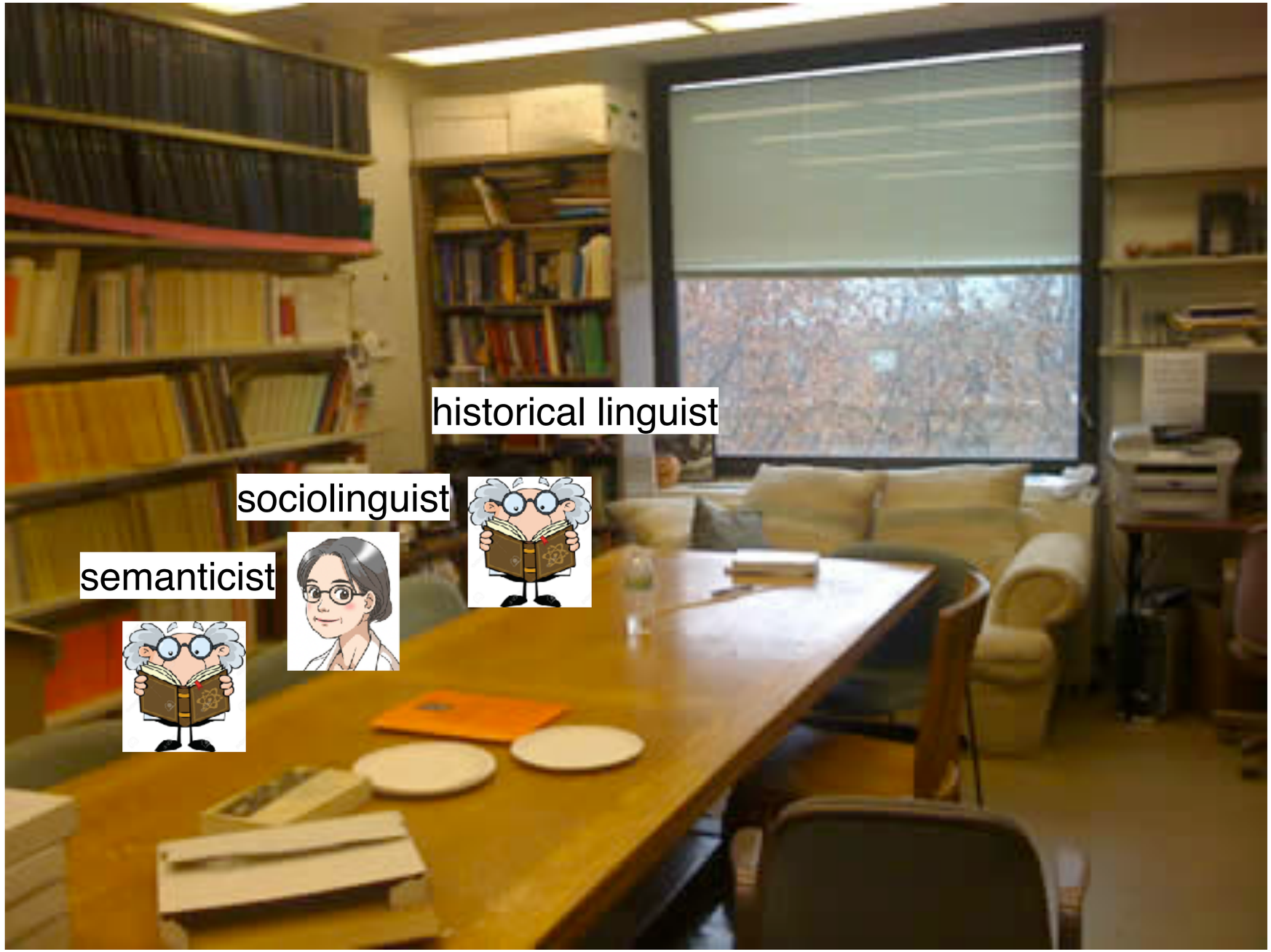
...þæt ic **ðas boc** of Ledenum gereorde to Engliscre spræce **awende**.
...that I this book from Latin language to English tongue translate
"...that I translate this book from the Latin language to the English tongue."
(AHTh, I, pref, 6; van Kemenade 1987: 16)

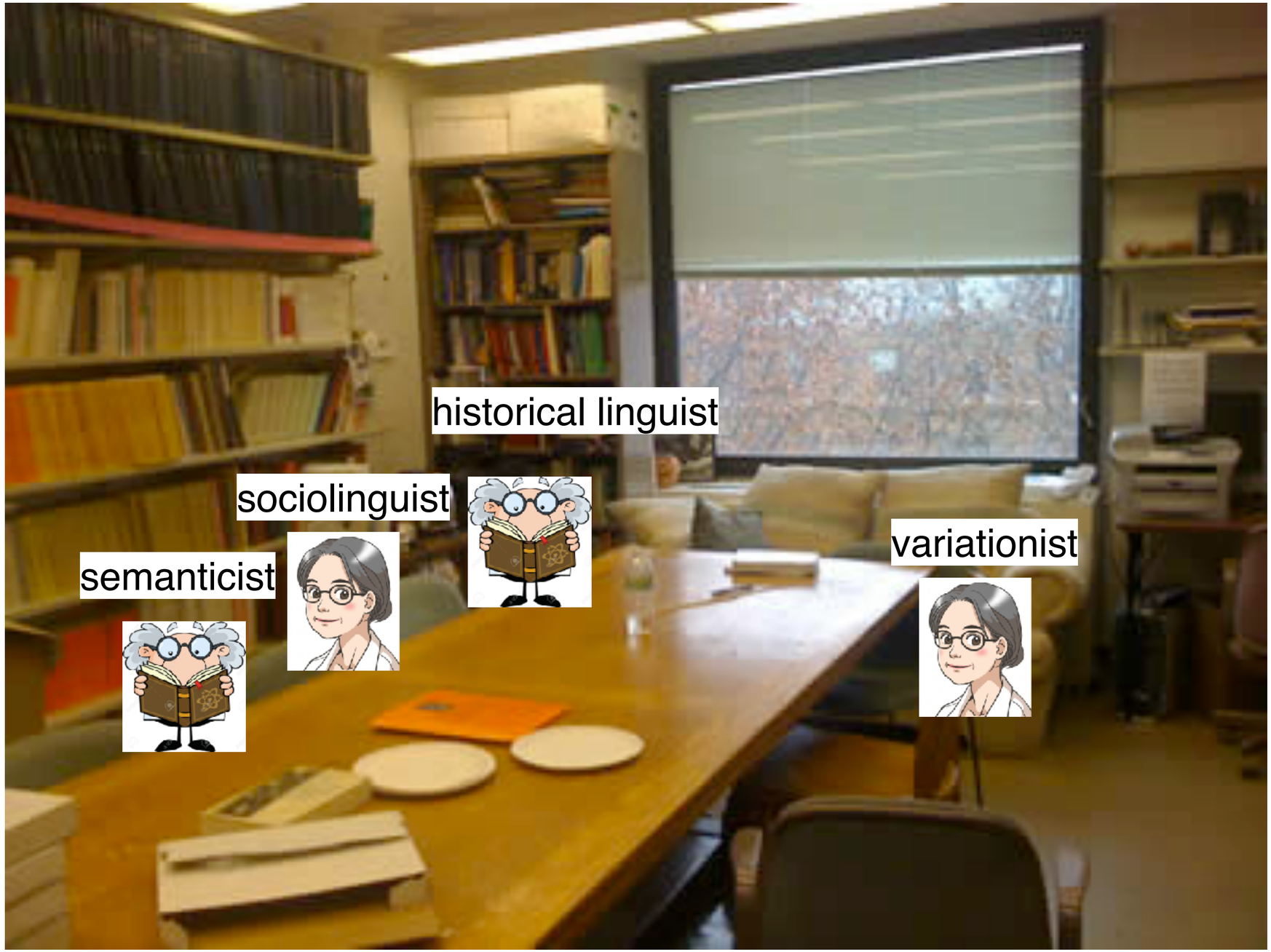
...þæt he **his stefne** up **ahof**.
...that he his voice up raised
"...that he raised up his voice."
(Bede 154.28)

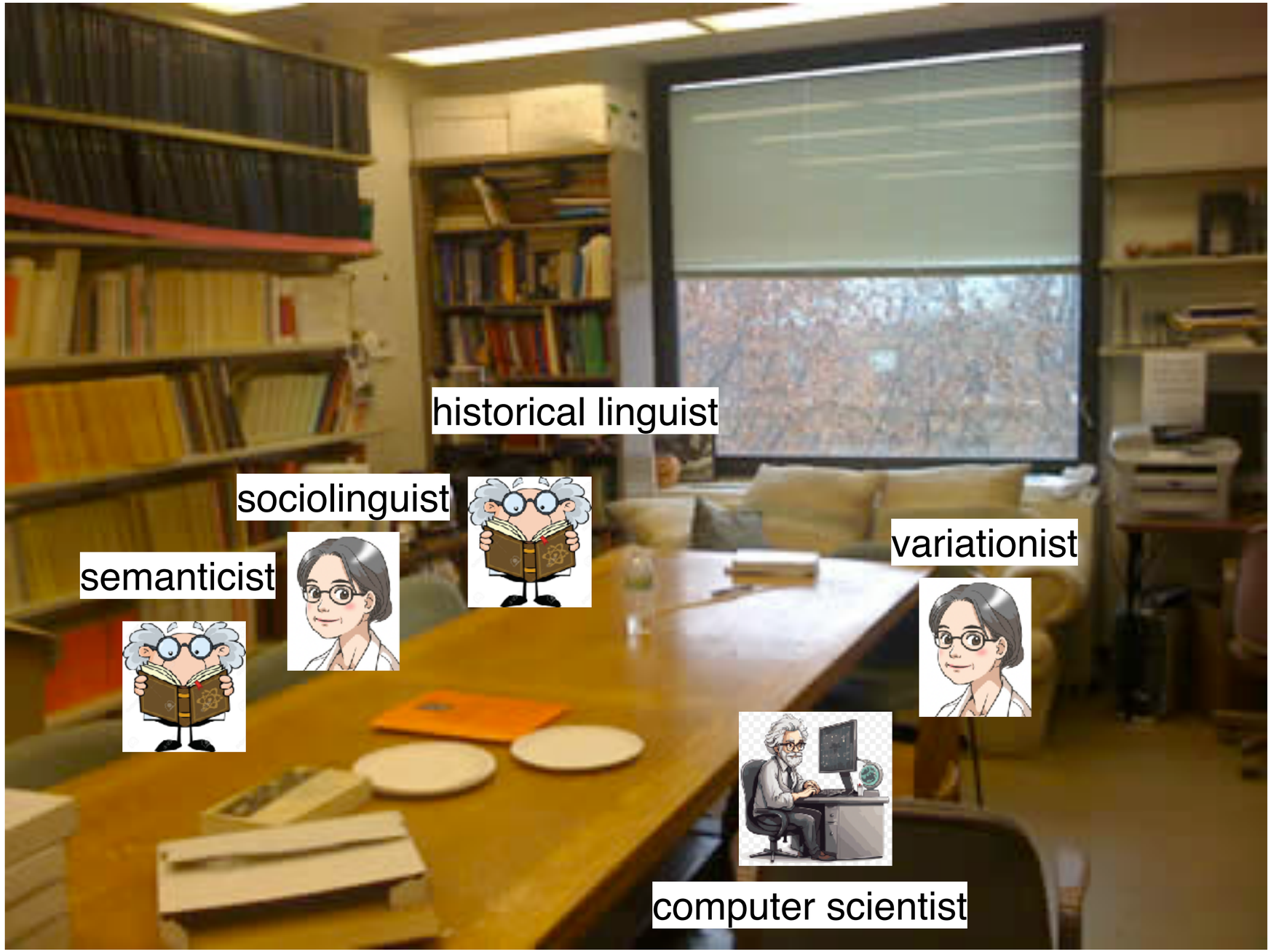
...forþon of Breotone **nædran** on scippe **lædde wæron**.
...because from Britain adders on ships brought were
"...because vipers were brought on ships from Britain."
(Bede 30.1-2; Pintzuk 1991: 117)

historical linguist


historical linguist
sociolinguist


historical linguist
sociolinguist
semanticist


historical linguist
sociolinguist
semanticist
variationist


historical linguist
sociolinguist
semanticist
variationist
computer scientist

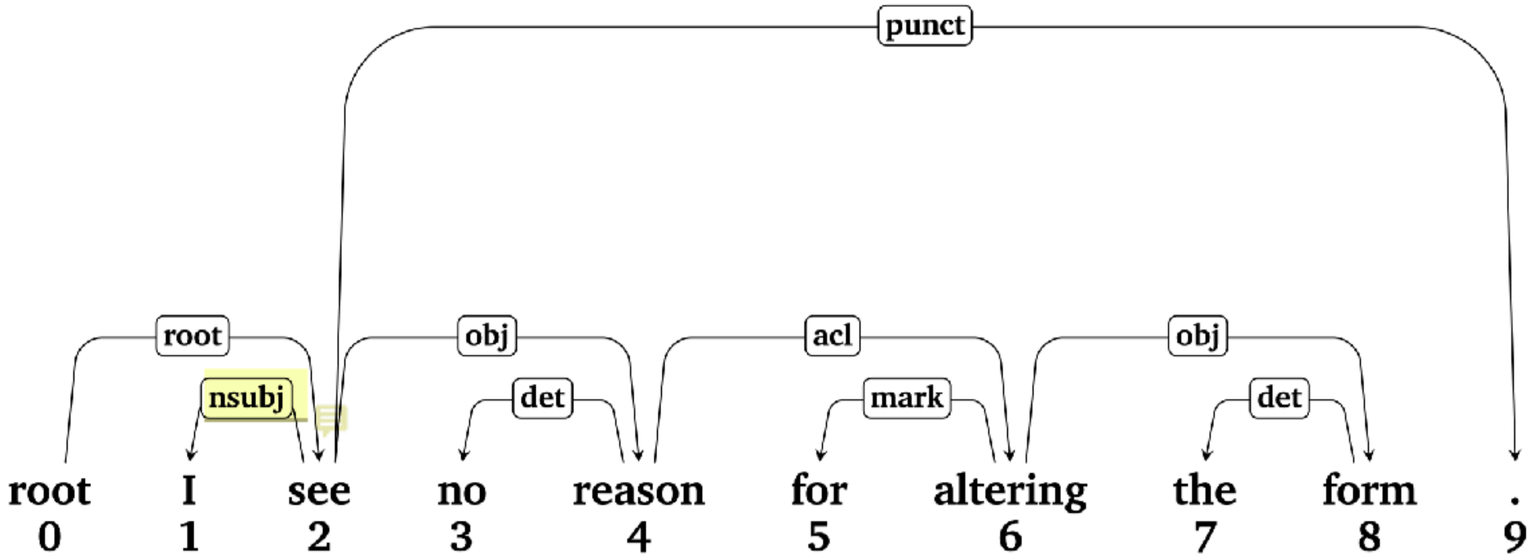
Syntactic Language Change in English and German: Metrics, Parsers, and Convergences

Yanran Chen, Wei Zhao, Anne Breitbarth, Manuel Stoeckel, Alexander Mehler, Steffen Eger

Motivation

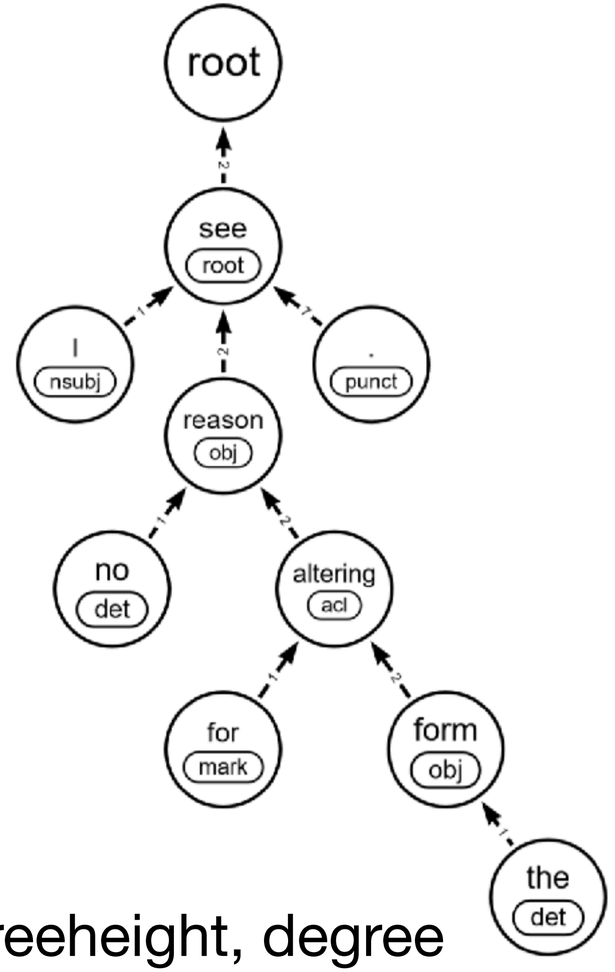
- Syntactic change is a beacon of language complexity over time
  - Human languages are optimized towards low complexity
    - Over time, sentences become syntactically simple and easy-to-understand
    - Dependency distance minimization (DDM): syntactically related words are placed closer to one another over time
  - But this result is based on
    - Stanford CoreNLP parser - invented 9 years ago
      - Trained on modern treebank; Issues in historical corpora
    - DDM (linear dependency distance). What about graph properties?

Motivation



distance (form, altering) = 2

distance (., see) = 7



Treeheight, degree

Research Questions

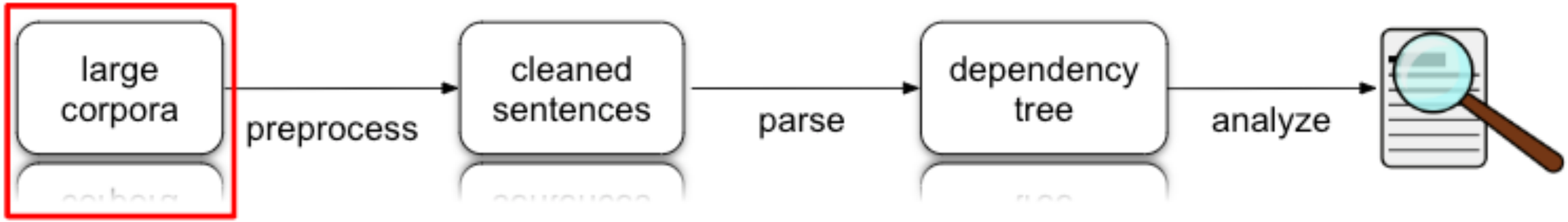
- **Language:** Are syntactic changes in English and German similar?
- **Parser:** Are parsers trained on modern treebanks reliable to parse historical data?
- **Parser:** Are trend predictions of syntactic change from different parsers consistent?
- **Metric:** Can we capture syntactic change based on graph properties instead of dependency distance? What are trend predictions?

Schematic Overview

1. Draw sentences from a corpus (decade-wise)
2. Parse them
3. Compute statistics/metrics (mean dependency distance, tree height, etc.) per decade
4. Build time-series and calculate trend

- Repeat for different (2) parsers and (3) metrics
- Compare two languages, English and German

Data Sources



Corpora in political debate domain:

- German: DeuParl
  - Plenary protols from the German Reichstag and Bundestag
  - 1860s-2020s (17 decades)
- English: Hansard
  - The official report of all British Parliamentary debates
  - 1800s-2020s (23 decades); focus on data from 1860

Rule-based filtering

- Sentences must start with a capitalized character
- Sentences must end with a period, or a question mark, or an exclamation mark
- Sentences must contain a verb (based on the part-of-speech tags)
- The number of (double) quotation marks must be even
- The number of left brackets must be equal to that of right brackets

Validation & correction of the extracted sentences:

- Human evaluation (Issues: e.g., OCR errors, historic spelling, …)

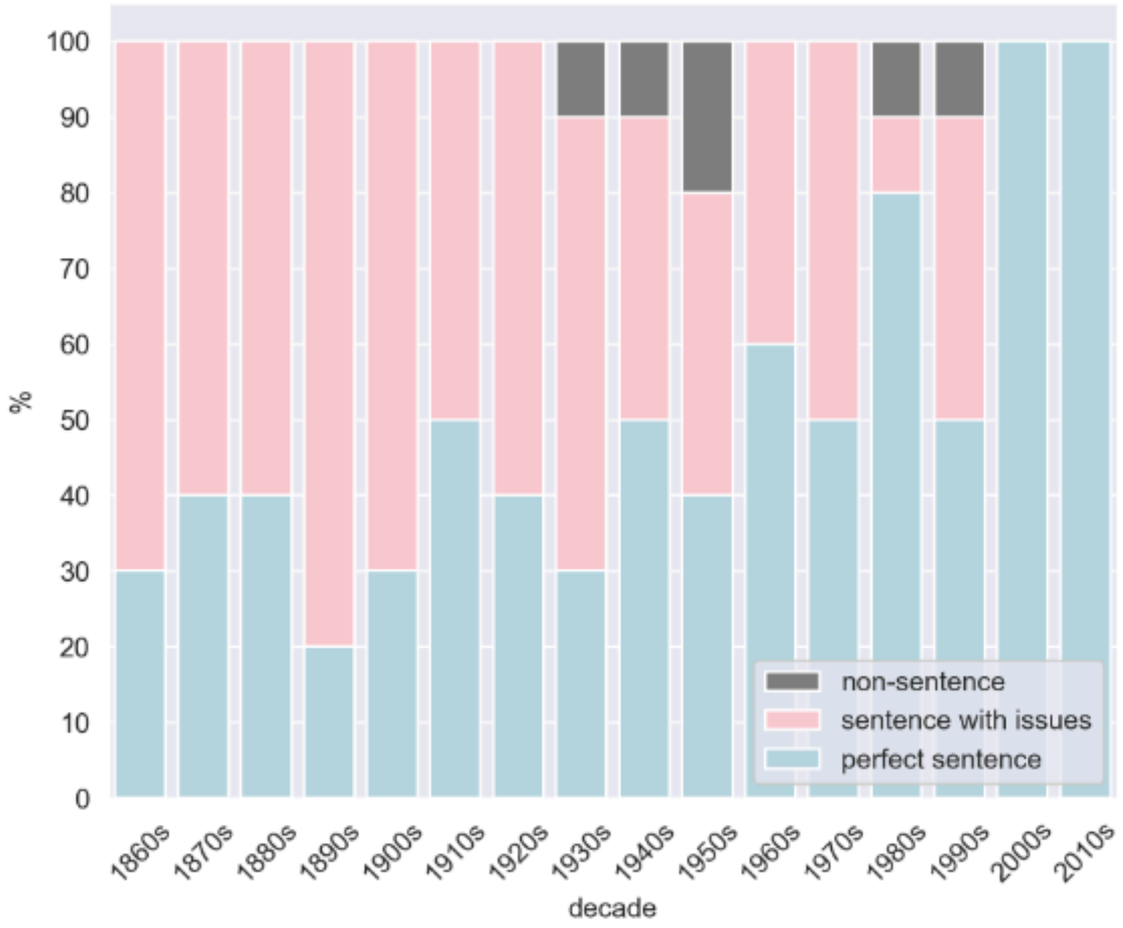
Error analysis in data

Category	Example	Correction
Spelling	Das ist die Mehrheit; die Diskussion ist geschlof-fen.	[...] geschlossen. OCR
Space	Der EVG-Vertrag spreche von der „westlichen Verteidigung", und im Protokoll der NATO-Staaten sei vom Zusammenschluß der w e s t europäischen Länder die Rede.	[...] westeuropäischen [...] OCR
Missing Material	Ich bin der Meinung — und viele an uns herangekommene Klagen lassen auch darauf schließen — , dass die Auffassung des Reichss-parkommissars, dass das Personal der Deutschen Reichspost voll ausgelastet, aber nicht überlastet sei, nicht richtig ist.	add "—" OCR
Extra Material	[...] daß durch den Bund zweierlei Recht für die Norddeutschen geschaffen werden soll, (Sehr richtig!) daß gewissermaßen zweierlei Klassen von Norddeutschen geschaffen werden sollen, (Sehr gut!) eine Selekta, die vermöge ihrer Gesit-tung [...]	delete "(Sehr gut!)" and "(Sehr richtig!)" Interjections
Punctuation & Symbol	Ich bitte, daß diejenigen Herren, welche für den Fall der Annahme des Z 33b in demselben die Worte „oder an Druck m anderen öffentlichen Orten" aufrecht erhalten wollen, sich von ihren Plätzen erheben.	[...] § 33b [...] OCR

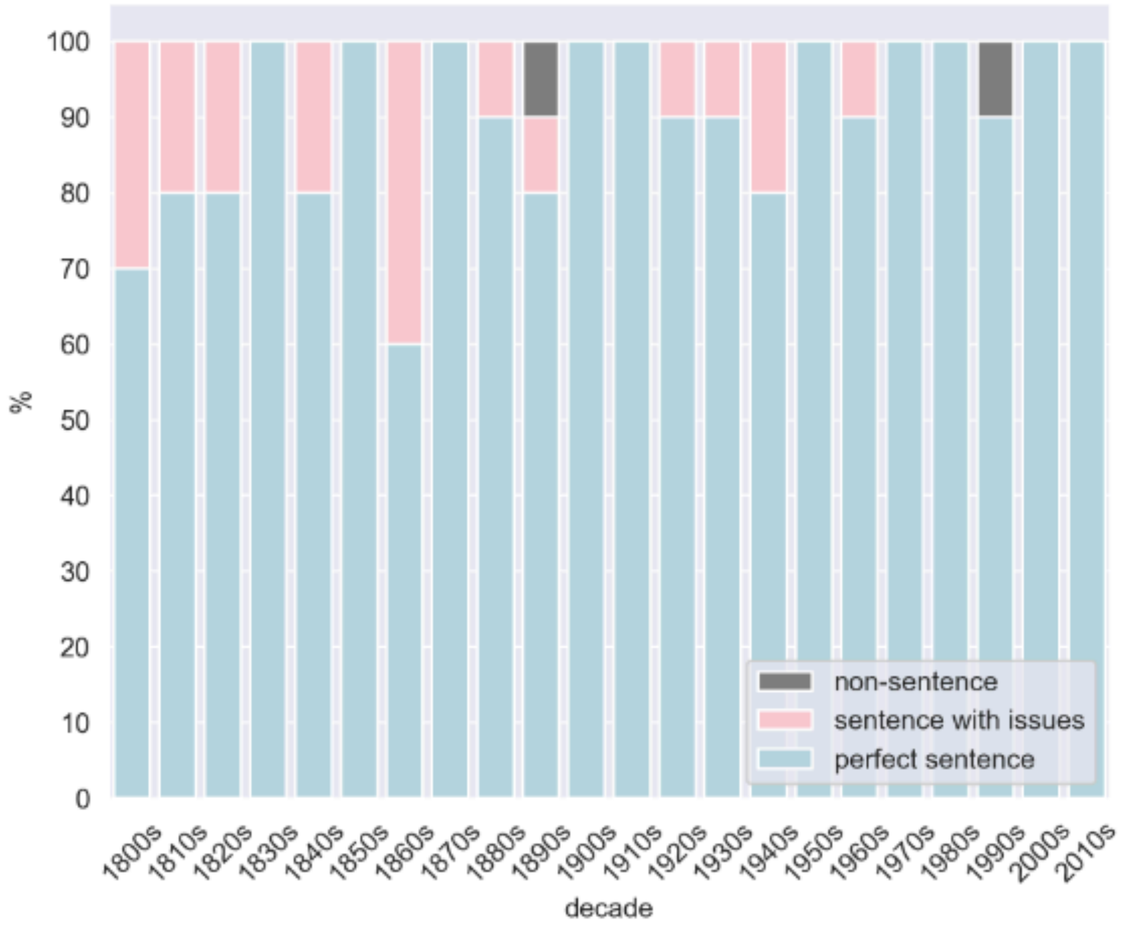
"Spelling" issue with "Historic" origin:

daß → dass

Error analysis in data

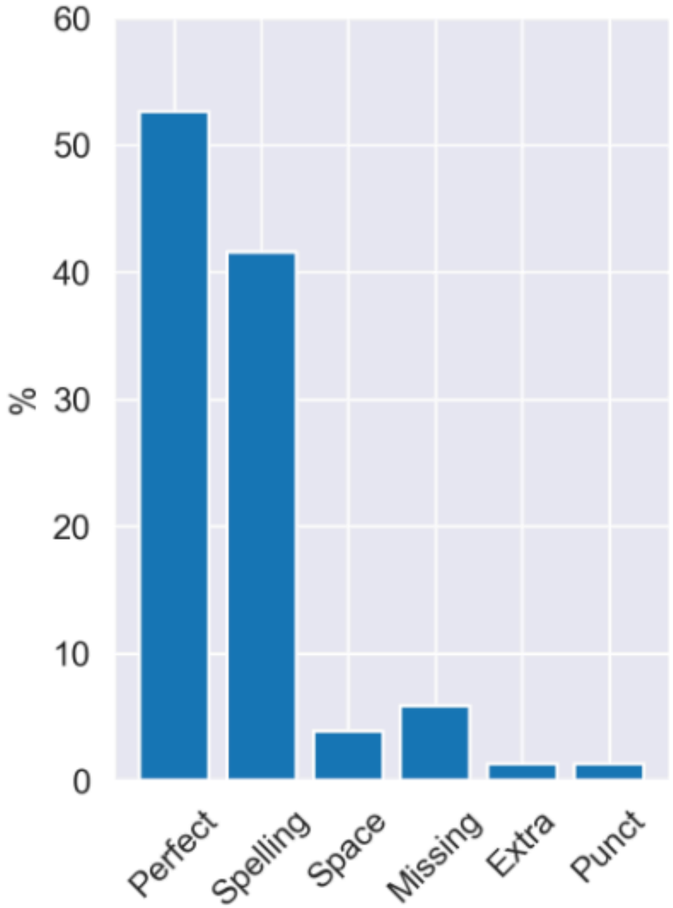


(a) Deuparl

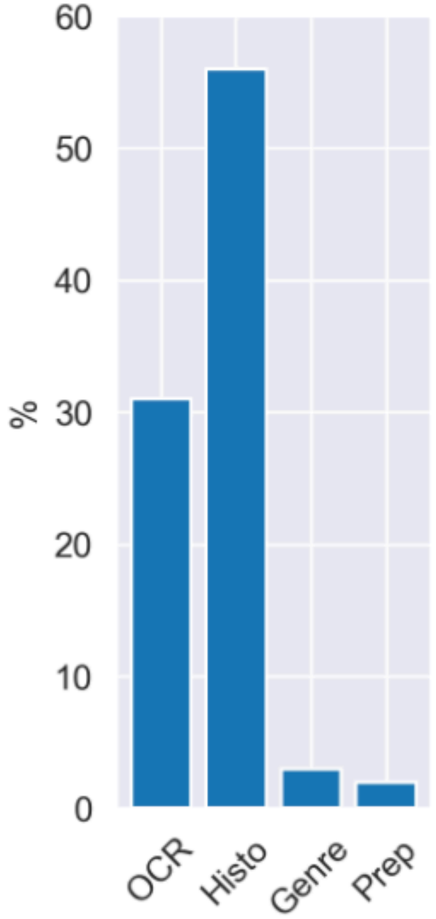


(b) Hansard

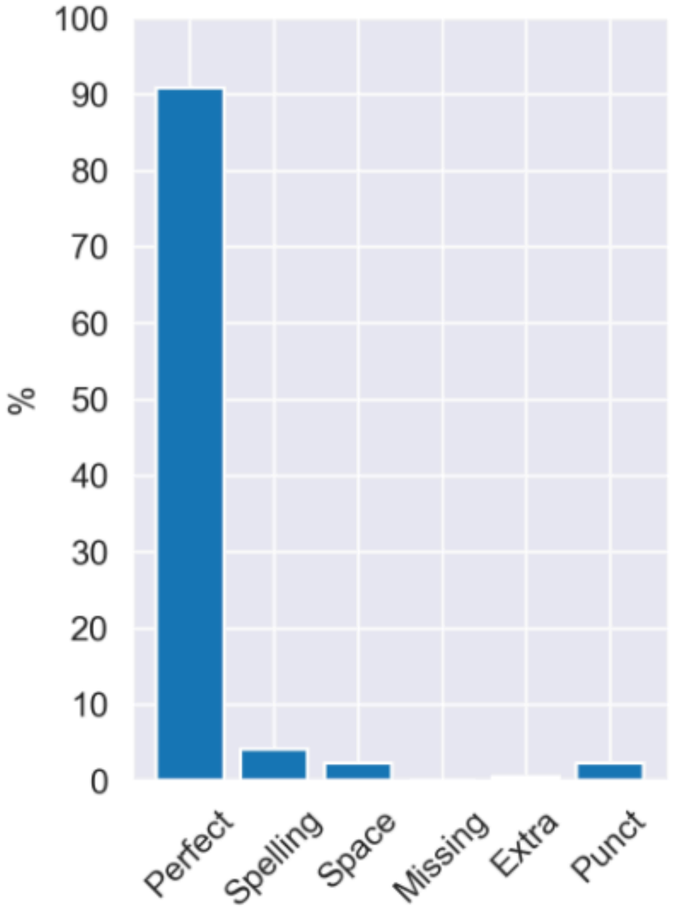
Error analysis in data



(a) **Deuparl** - issues

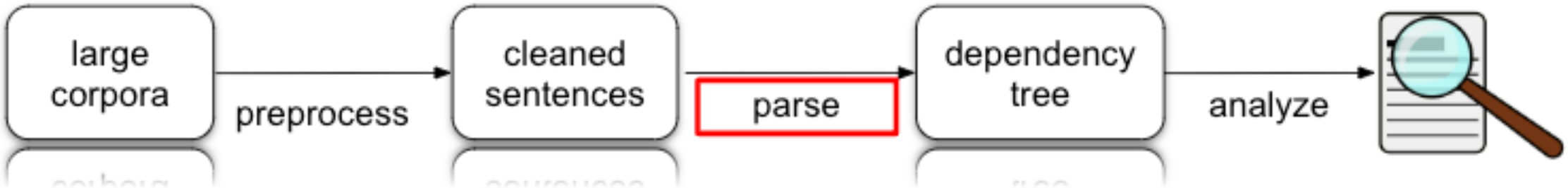


(b) **Deuparl** - origins



(c) **Hansard** - issues

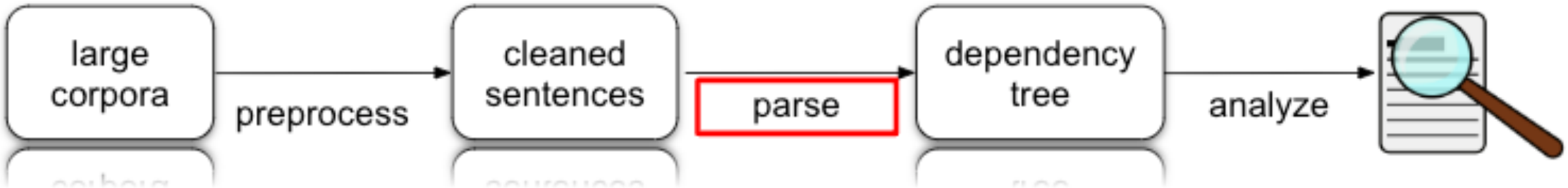
Transition-based vs. graph-based



Parsers with the two most popular design choices

- Transition-based parsers
  - Predict edges step by step based on the current state
  - Stanford CoreNLP (Manning et al. 2014), StackPointer (Ma et al. 2018)
- Graph-based parsers
  - Globally optimized the dependency tree, aiming to find the highest-scored one
  - Deep Biaffine (Dozat and Manning 2017), Stanford Stanza (Qi et al. 2020), CRF2O (Zhang et al. 2020), TowerParse (Glavasˇ and Vulic´ 2021)

Evaluation



- Modern treebanks – Universal Dependencies
- Target treebanks – 111 sentences from Hansard, 163 sentences from DeuParl
  - Human annotation with unknown reliability

RQ: What is the parser reliability on historical data?

	UAS		LAS	
	UD	TARGET	UD	TARGET
CoreNLP*	78.6±5.5	82.9 (+4.3)	73.4±6.5	78.7 (+5.3)
Stanza*	88.2±6.2	88.0 (-0.2)	84.9±8.0	85.0 (+0.1)
TowerParse*	87.1±4.5	**92.8 (+5.7)**	82.8±5.9	**90.3 (+7.5)**
StackPointer	85.7±5.2	84.2 (-1.5)	81.3±6.2	80.2 (-1.1)
CRF2O	87.2±4.6	86.6 (-0.6)	83.5±5.4	82.8 (-0.7)
Biaffine	**91.6±2.6**	90.1 (-1.5)	**88.5±3.3**	87.2 (-1.3)

(a) English

	UAS		LAS	
	UD	TARGET	UD	TARGET
CoreNLP*	74.0±2.9	74.4 (+0.4)	67.3±2.9	69.6 (+2.3)
Stanza*	84.3±4.1	87.6 (+3.3)	78.8±4.4	82.0 (+3.2)
TowerParse*	86.2±2.0	89.7 (+3.5)	80.3±1.3	84.5 (+4.2)
StackPointer	83.6±1.3	86.9 (+3.3)	78.2±1.6	80.8 (+2.6)
CRF2O	78.5±1.5	81.9 (+3.4)	70.2±2.4	72.9 (+2.7)
Biaffine	**87.0±2.5**	**90.8 (+3.8)**	**81.7±1.8**	**84.5 (+2.8)**

(b) German

Metrics

Metrics based on linear dependence distance

- Mean Dependency Distance (MDD)
- Normalized Mean Dependency Distance (NDD)
- Root Distance
- Number of Crossings (#Crossing)
- Head final Distance

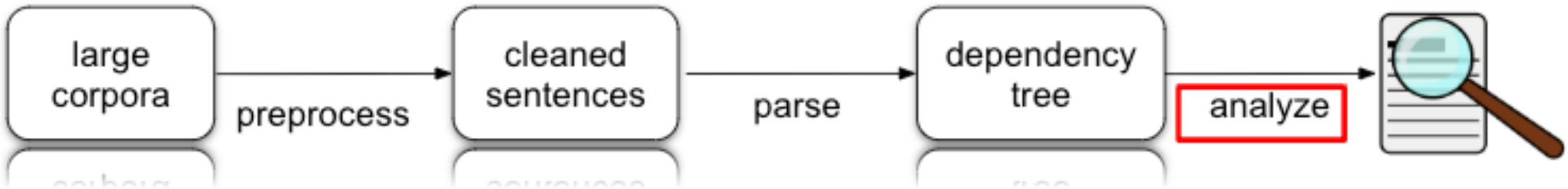
$$mDD = \frac{1}{n} \sum_{(i,j) \in D} |i - j|$$

$$nDD = \left| \log \left(\frac{mDD}{\sqrt{d_{root} \times \text{length}}} \right) \right|$$

based on graph properties

- Tree Height
- Depth Variance & Depth Mean
- Arity/Tree Degree
- Degree Variance & Degree Mean
- Number of leaves (#Leaves)
- Head final ratio
- Longest Path Distance
- Random Tree Distance

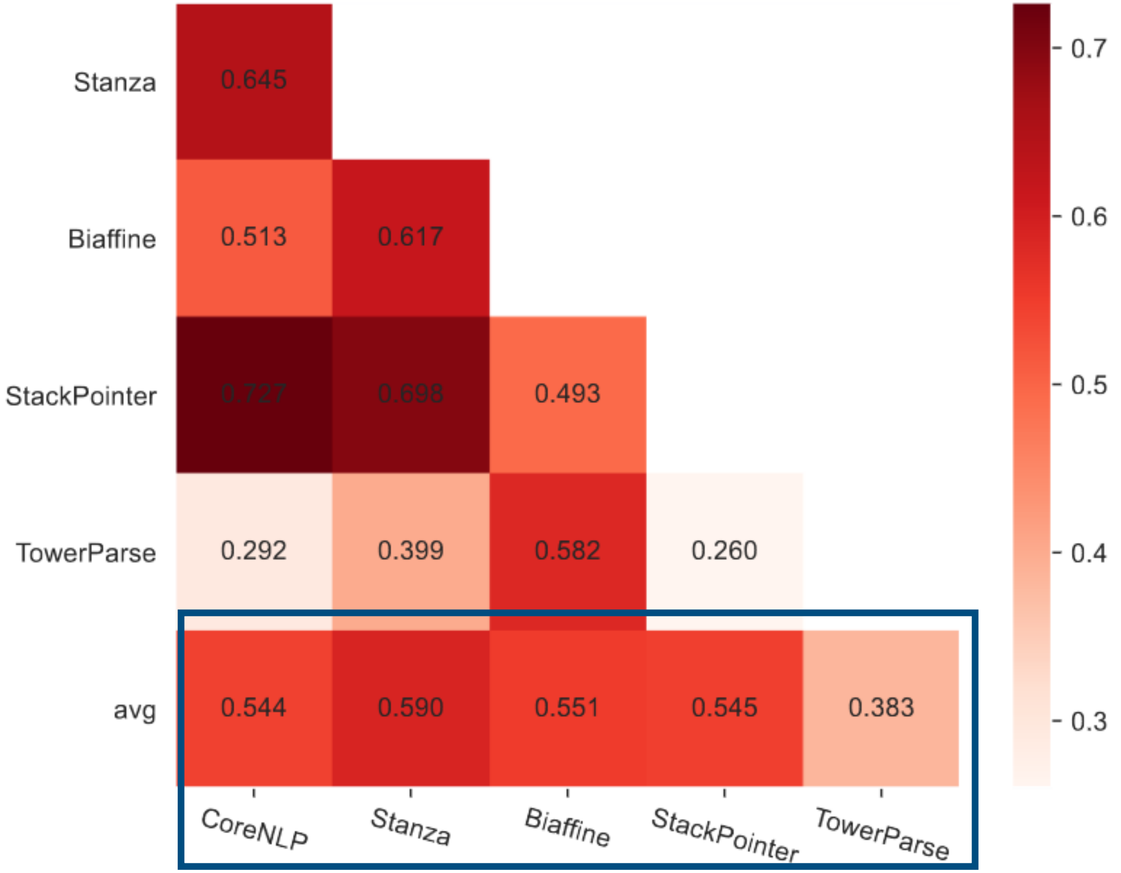
Trend prediction



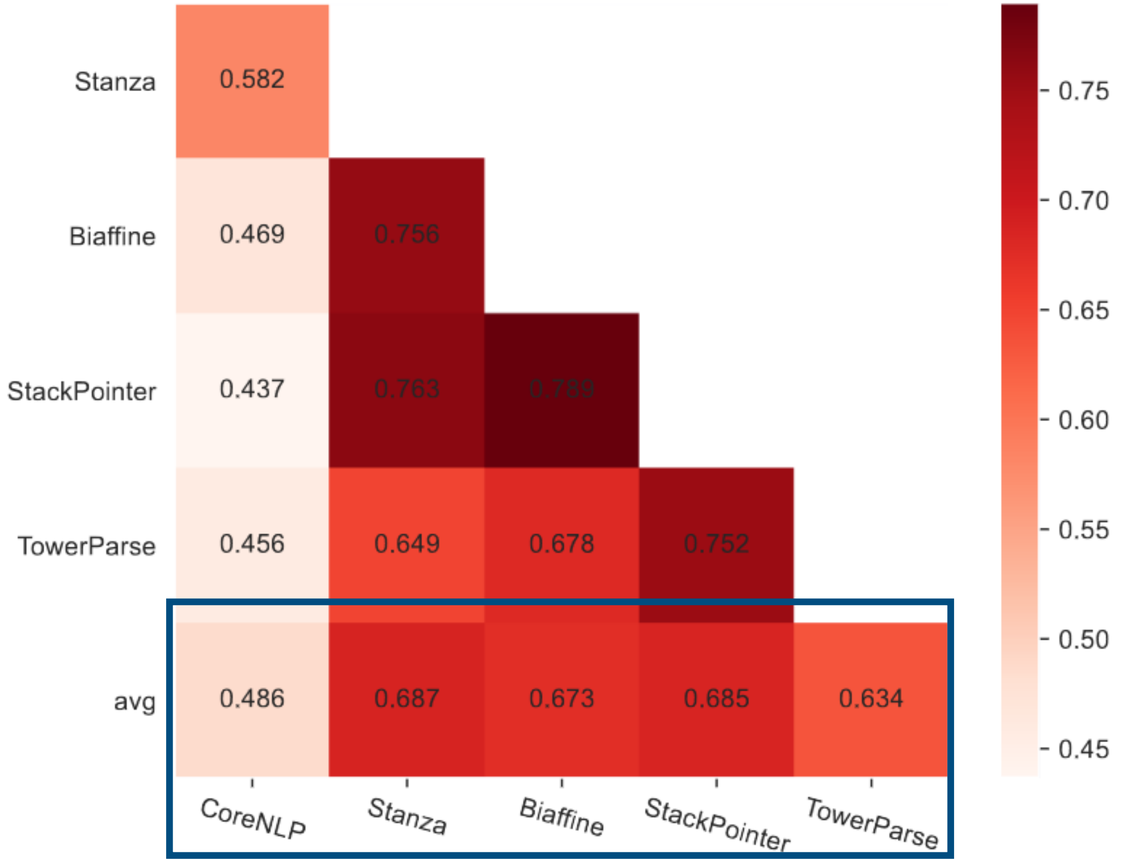
Datasets for analysis

- Balance: anchor length [5,10,15,20,30,40,50,60,70]
  - anchor length <= len < anchor length + 3 → one length group
  - We draw 450 sentences per each decade/length group
    - 450 x 8 (decade groups) x 9 (length groups) = 32,400 sentences for each language
- Mann Kendall trend test (MK)
  - Three outputs: increasing, decreasing, no trend

RQ: Are trend predictions of syntactic change from different parsers consistent?

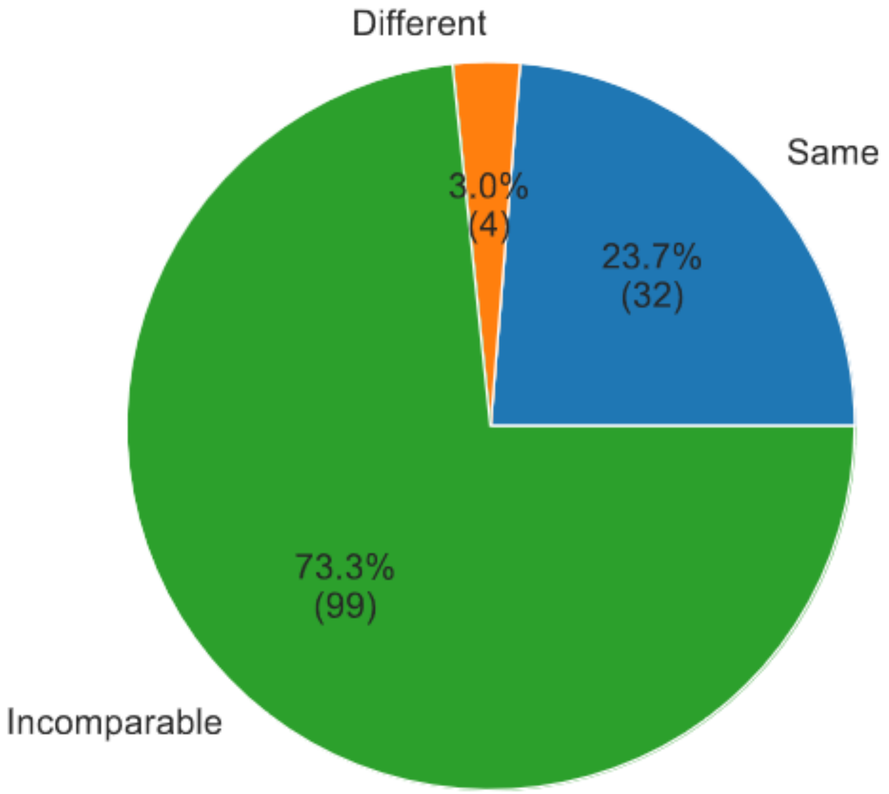


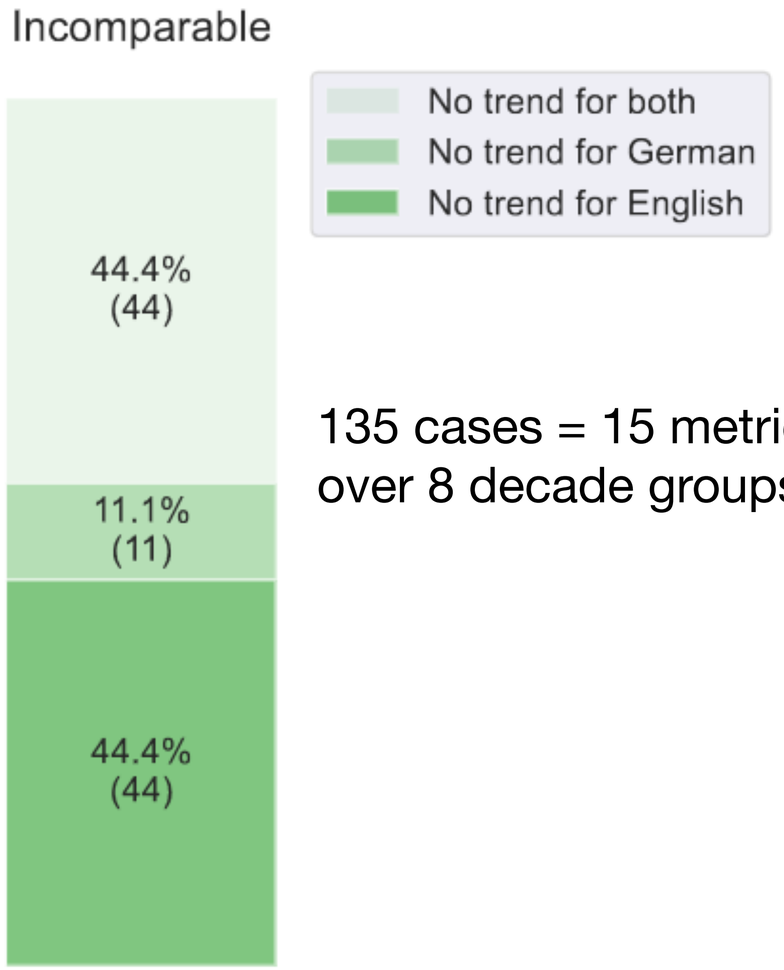
(a) English



(b) German

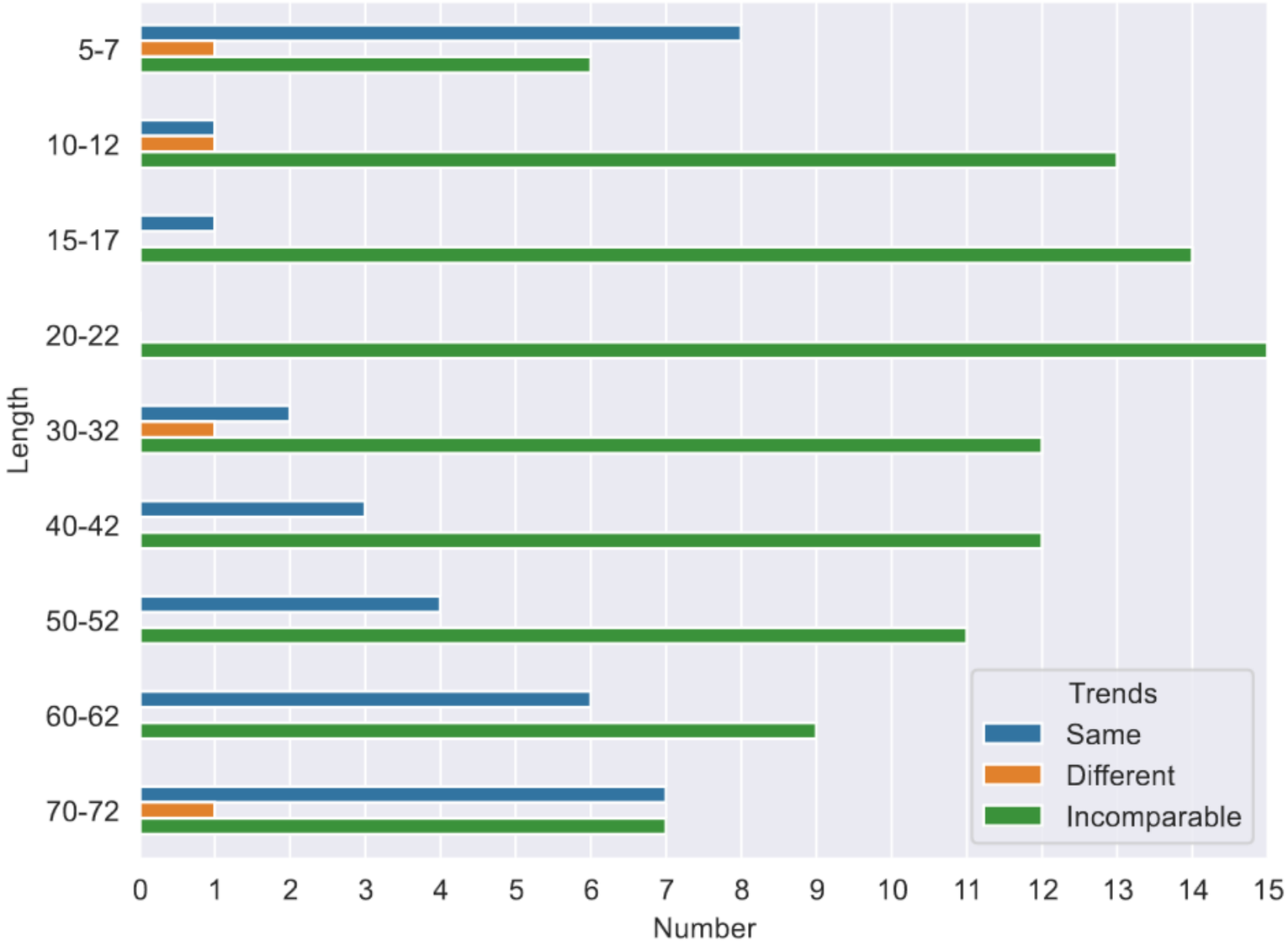
RQ: Are syntactic changes in English and German similar?



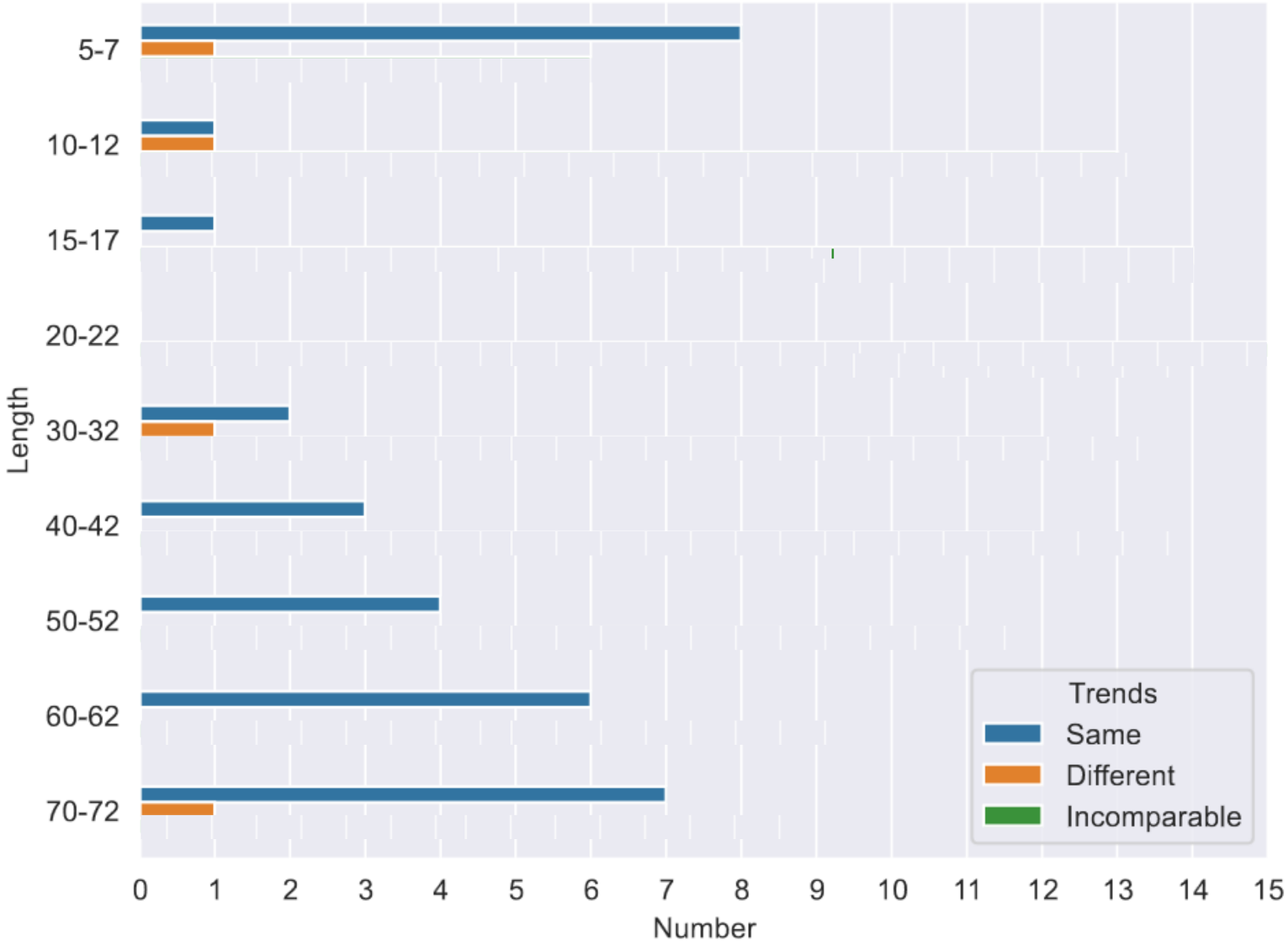


135 cases = 15 metrics x 9 length groups over 8 decade groups

RQ: Are syntactic changes in English and German similar?



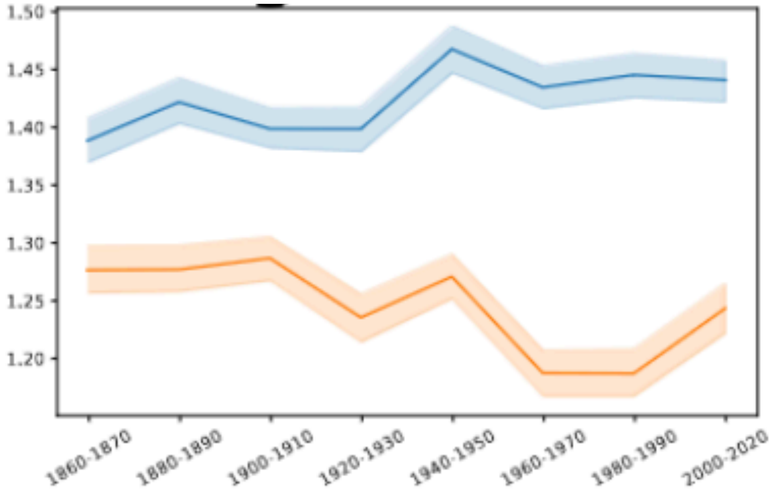
RQ: Are syntactic changes in English and German similar?



RQ: Are syntactic changes in English and German similar?

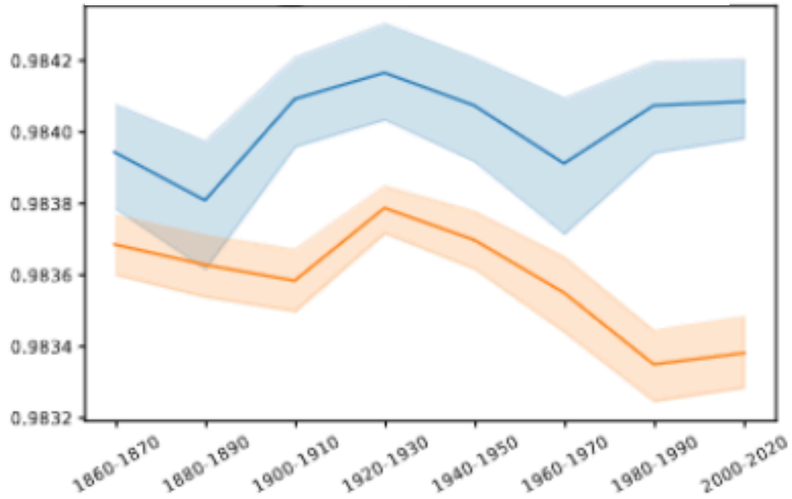
English
German

length: 30-32



nDD

length: 70-72



degreeMean

- Trends are not different at every point in time
  - Over some decades groups trends are similar but not captured by MK trend test.
- The degree of changes is marginal

Summary

- Are trend predictions of syntactic change from different parsers consistent? **No, they are not.**
- Are parsers trained on modern treebanks reliable to parse historical data? **Yes, but..**
- Are syntactic changes in English and German similar? **Yes, though many cases show incomparable trends.**
- Can we capture syntactic change based on graph properties instead of dependency distance? **Yes, but..**

Limitations

- MK trend test is not smart enough
  - Trends are partly similar
  - The degree of changes is marginal
- Data corpus over the past hundred years - not long enough for significant syntactic changes (word order) to happen
- Unknown quality of human annotations on historical treebanks

Graph-based Clustering for Detecting Semantic Change Across Time and Languages (EACL-24)

Xianghe Ma, Michael Strube, Wei Zhao

Motivation

- Discovery of lexical semantic change is important for HL, CL and NLP researchers
  - (HL) Inform a theory of semantic change
  - (CL) Automate the human process of detection
  - (NLP) maybe useful for downstream NLP tasks such as historical MT
- Potentially useful for the lexicography industry

Problem Definition

Target word: arm

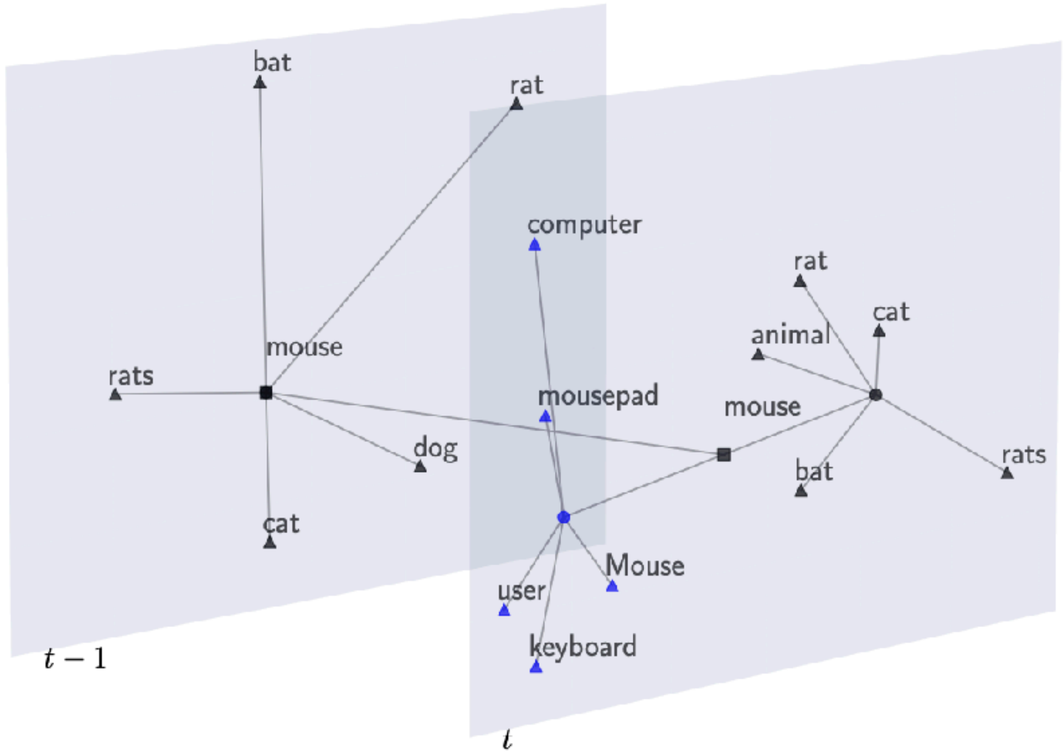
- A: ..taking a knife from her pocket, she opened a vein in her little **arm**..
  - Hobomok, A Tale of Early Times, published before 1850
- B: Dear Grace, " said Henry, passing his **arm** round her neck, " I have something to say..
  - The Rebels: Boston Before the Revolution, published before 1850
- C: …near a point where a long **arm** of land thrust out into the sea and shut off the wind..
  - Blix by Norris, Frank, published after 1850

Tasks

- **Binary classification:** whether ‘arm’ changes its usage across two time periods.
- **Ranking:** score the degree of semantic change if any.

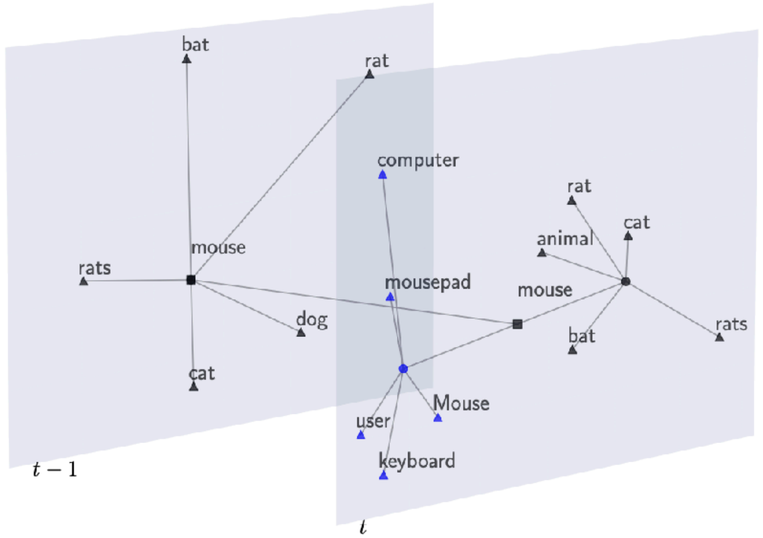
Schematic Overview

1. Produce usage embeddings for each target word
2. Partition usages within each time period into clusters based on their embeddings
3. Detect semantic change by computing similarities between clusters.
4. Represent clusters in a temporal graph.



Our Contributions

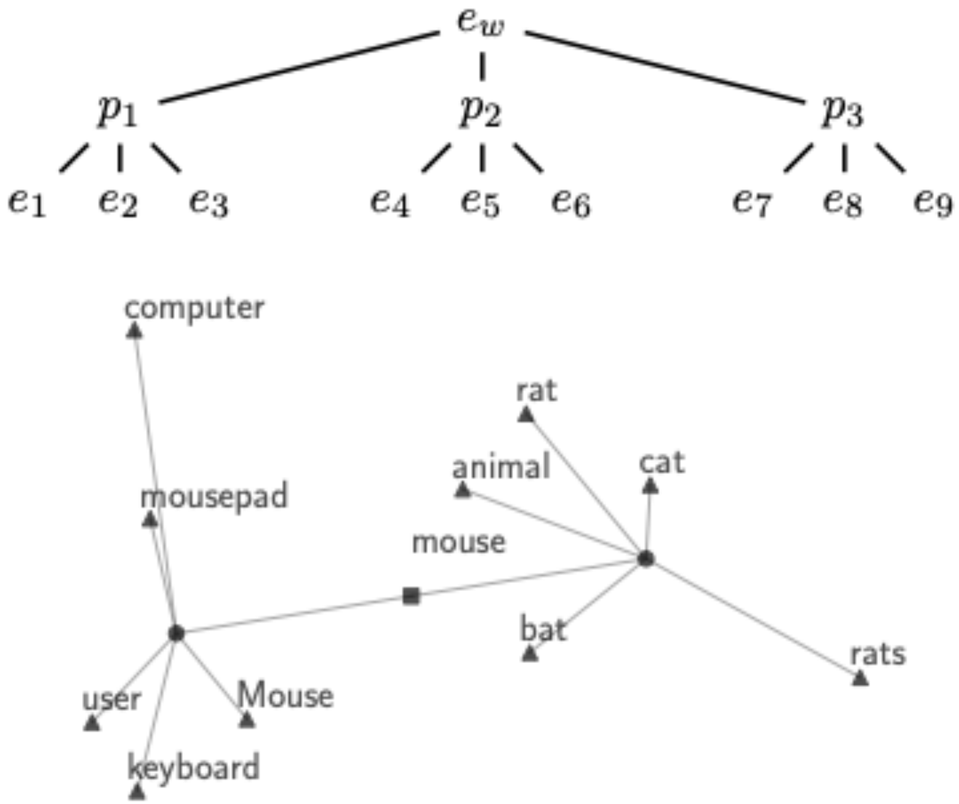
- Detect low-frequency sense clusters
  - An agglomerative-like clustering process
  - A neighbor-based distance metric
- Visualization tool (temporal graph)
  - Show semantic change over time
  - Compare cross-language semantic change



Graph Construction

For a target word **w**:

- **Root node:** the average word usage embedding
- **Nodes on the second layer:** the centroid (average usage embedding) of each sense cluster
- **Nodes on the third layer:** k-nearest neighbors to the centroid of each sense cluster

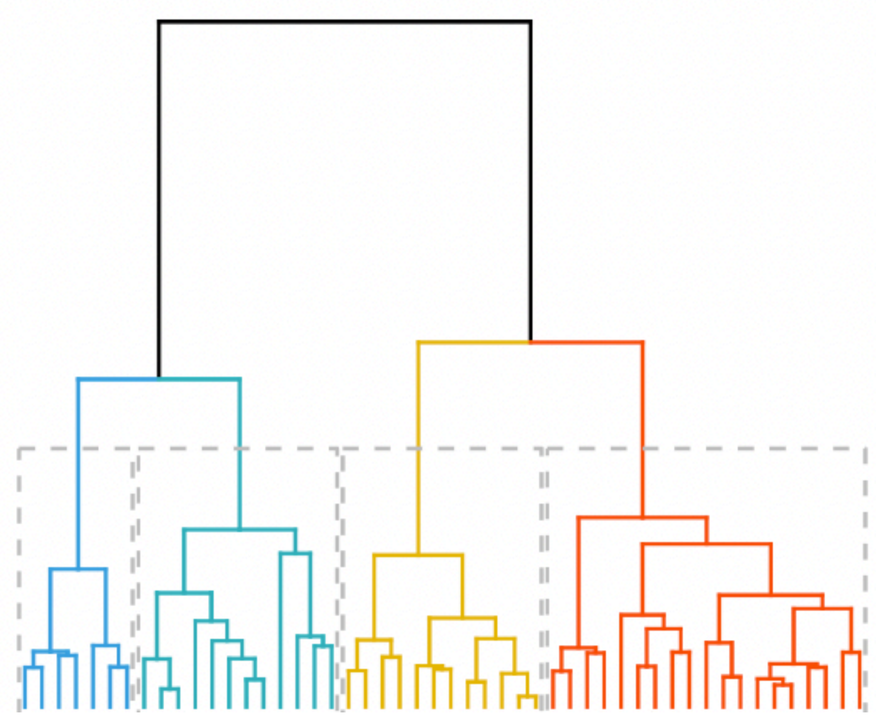


at a point in time

Our Clustering Process

Require: $\mathcal{C}_w = \{c_i\}_{i=1}^{n}$ as a set of contextualized embeddings of each word w, t_{sc} as the maximum distance between similar clusters, t_{low} as the minimum cluster size for a low-frequency sense cluster.

Initial centroids of clusters: $\mathcal{P}_w = \{p_i | p_i = c_i\}_{i=1}^{n}$
while $\min_{p_i \in \mathcal{P}_w, p_j \in \mathcal{P}_w, i \neq j} d(p_i, p_j) < t_{sc}$ **do**
    $\mathcal{P}_w = \mathcal{P}_w \setminus \{p_i, p_j\} \cup \{\frac{p_i + p_j}{2}\}$
end while
for $p_i \in \mathcal{P}_w$ **do**
    if $|\mathcal{C}_w(p_i)| < t_{low}$ **then**
        $\mathcal{P}_w = \mathcal{P}_w \setminus \{p_i\}$
    end if
end for
return $\mathcal{P}_w$



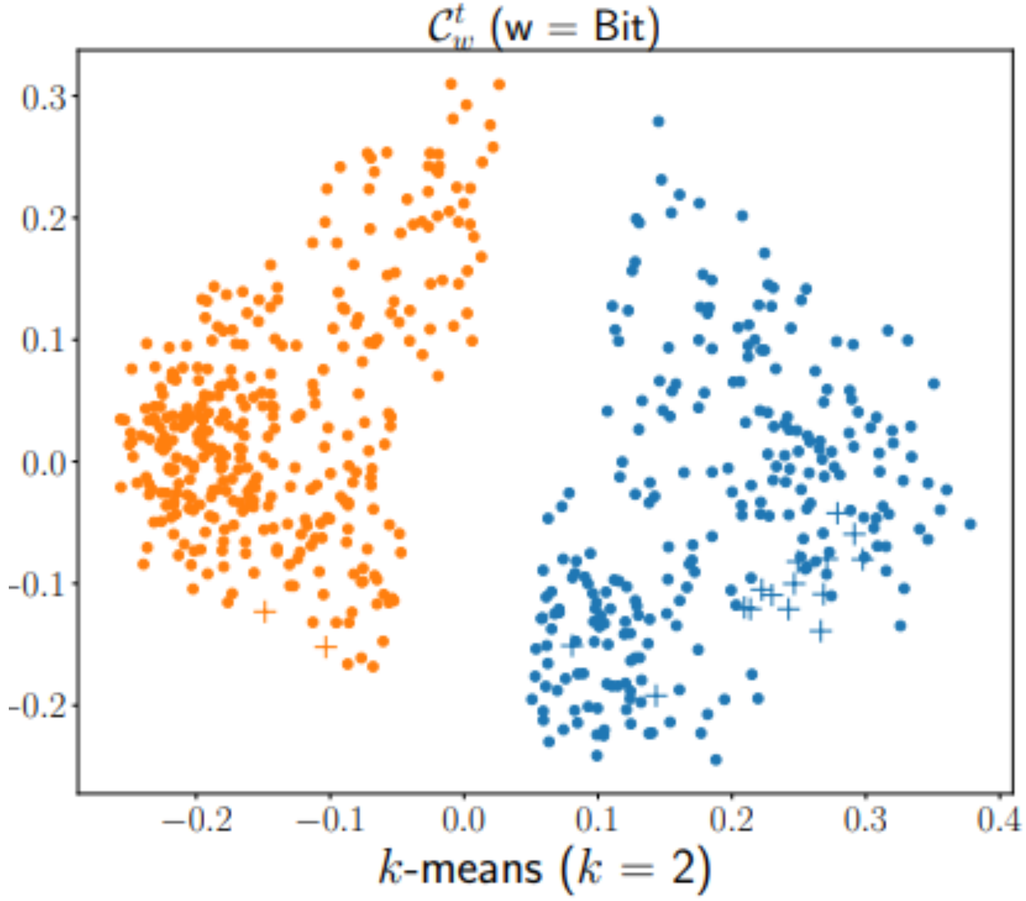
- Treat each usage embedding as a separate cluster and iteratively merge two clusters when similar.
- Neighbor-based metric: for a low-frequency sense with very few usages, replying on them to decide a standalone sense cluster is not reliable
  - ask nearest neighbors of these usages to participate in decision making
- Remove noisy clusters vs. wrongly remove low-frequency sense clusters.

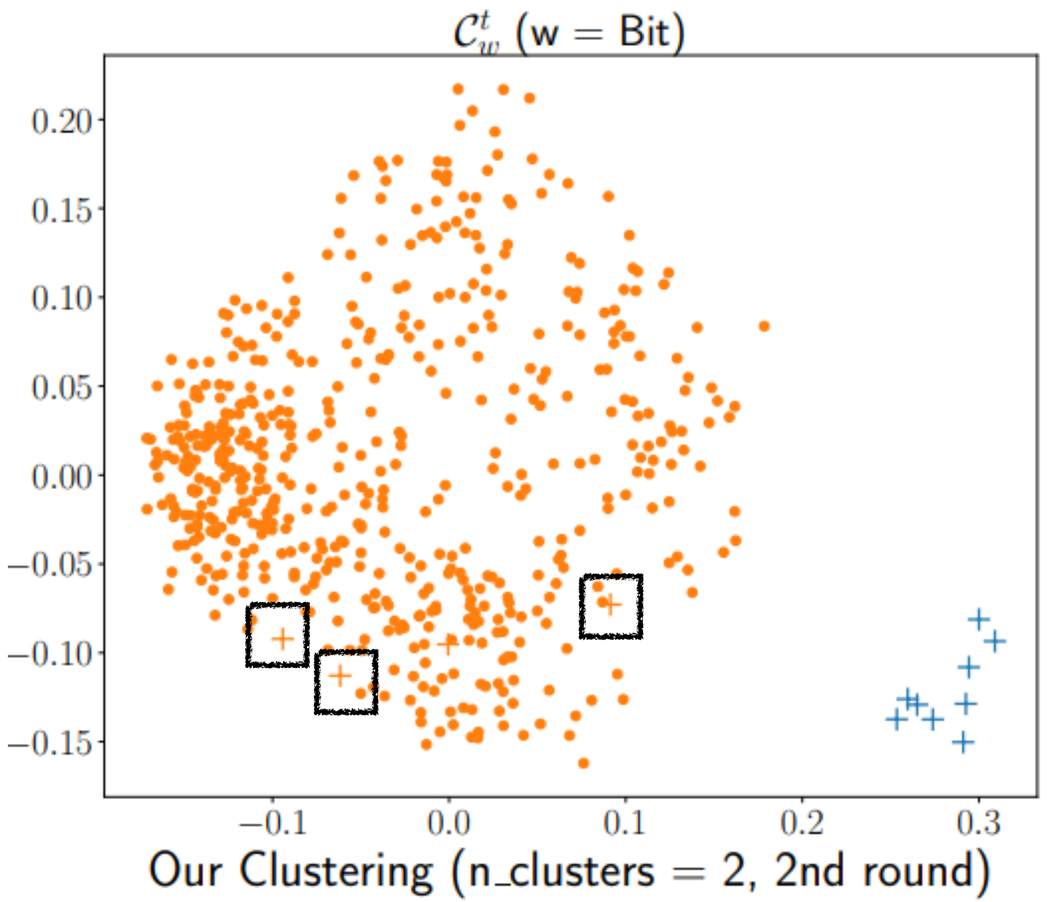
Data Sources

SemEval2020 corpora:

- English: Corpus of Historical American English - spelling normalization
  - 1810s–2000s
- German: the DTA corpus (newspaper) - OCR errors, spelling normalization
  - 16th–20th centuries
- Latin: the LatinISE corpus (literature, history)
  - 2nd century B.C. to the 21st century A.D
- Sweden: the Kubhist corpus (newspaper) - OCR errors
  - 18th–20th century

Analysis: Detecting low-frequency sense clusters





. High-frequency sense (a small piece)

+ Low-frequency sense (binary digit)

Analysis: Detecting low-frequency sense clusters

Algorithms	EN	DE	LA	SV
K-means	0.975	0.778	0.664	0.775
Gaussian Mixture	0.939	0.775	0.670	0.754
Affinity Propagation	0.891	0.741	0.686	0.662
Our Clustering	**0.994**	**0.879**	**0.877**	**0.909**

- Report in purity score
- 8 target words per language. Each word has a high-freq sense (100 usages) and a low-freq sense (20 usages)
- K as a hyperparameter: k-means and GMM
- Adaptive clustering: AP and our method

Results on SemEval-2020

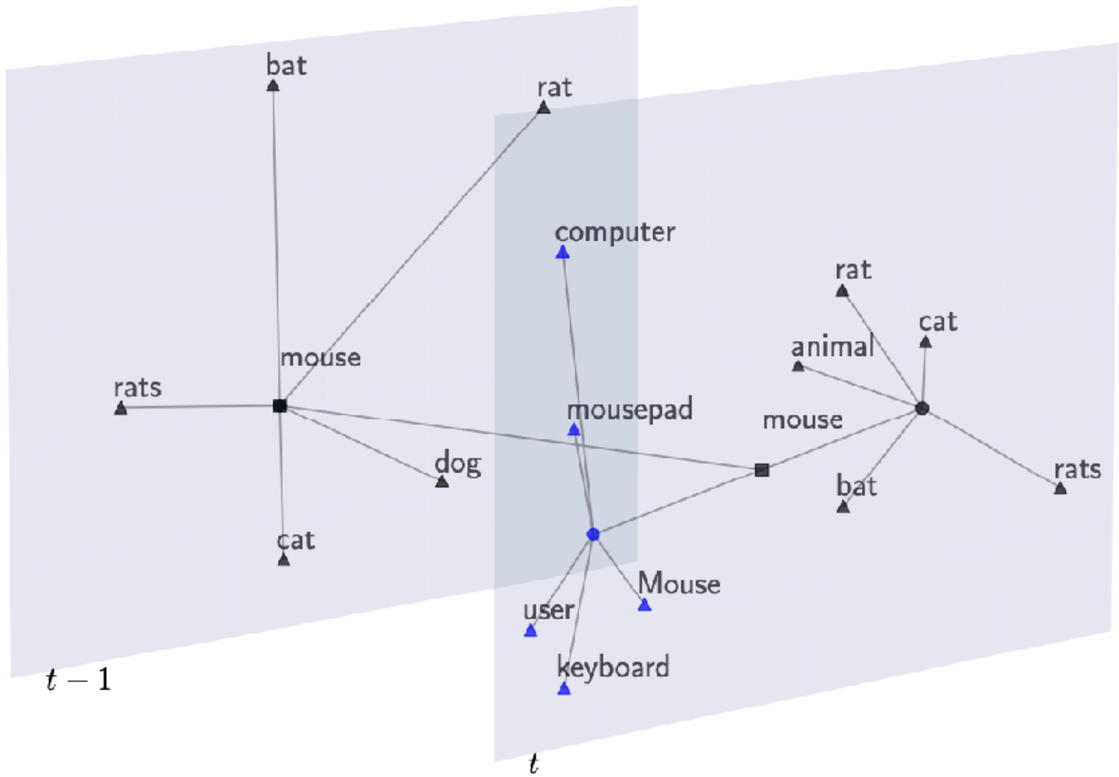
Binary classification

Approaches	Avg	EN	DE	LA	SV
Static Word Embeddings					
UWB (Pražák et al., 2020)	.687	.622	.750	**.700**	.677
Life-Language (Asgari et al., 2020)	.686	.703	.750	.550	.742
Contextualized Word Embeddings					
NLP@IDSIA (Kanjirangat et al., 2020)	.637	.622	.625	.625	.677
Skurt (Cuba Gyllensten et al., 2020)	.629	.568	.562	.675	.710
Our Approach	**.776**	**.784**	**.813**	**.700**	**.806**

- German and Swedish > English, perhaps German and Swedish corpora got fewer low-frequency senses than English.
- Our results in the ranking task are not so good. It is more challenging than binary classification.

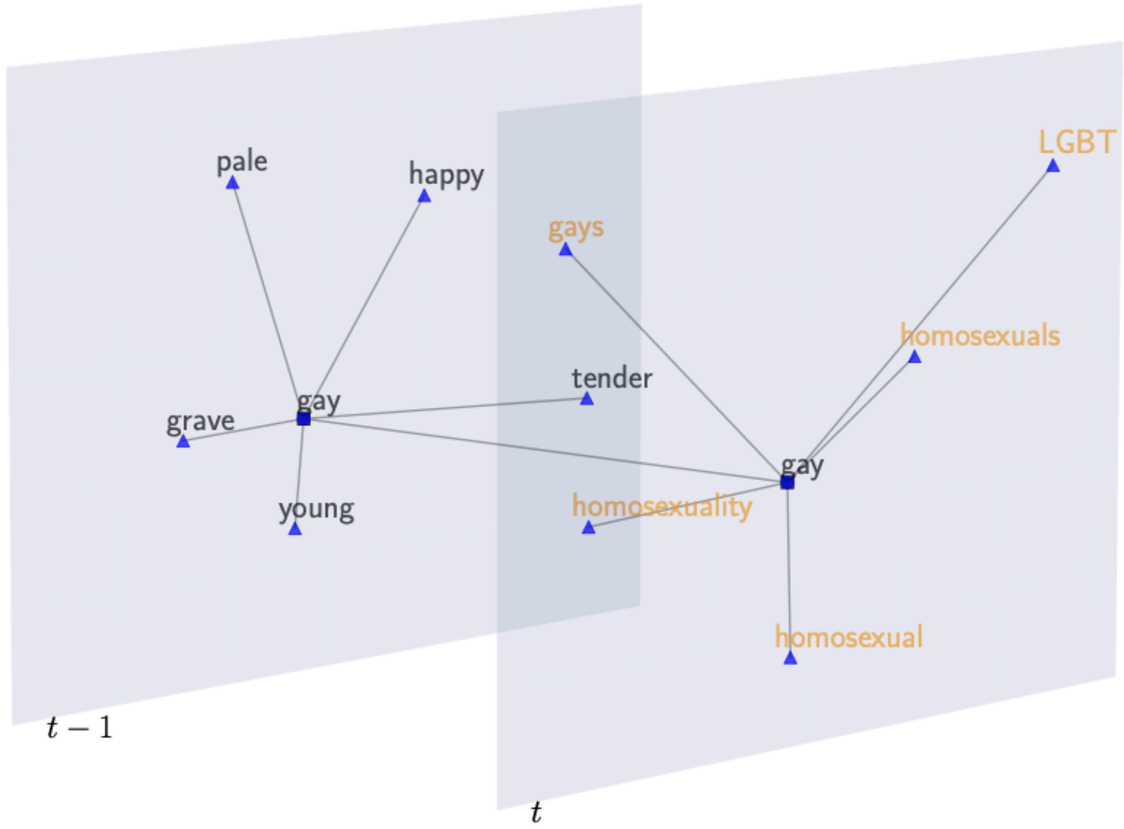
Use Case: Broadening

- Word sense becomes more general than it used to be.
- A word gains a new sense while retaining the original sense: A => AB



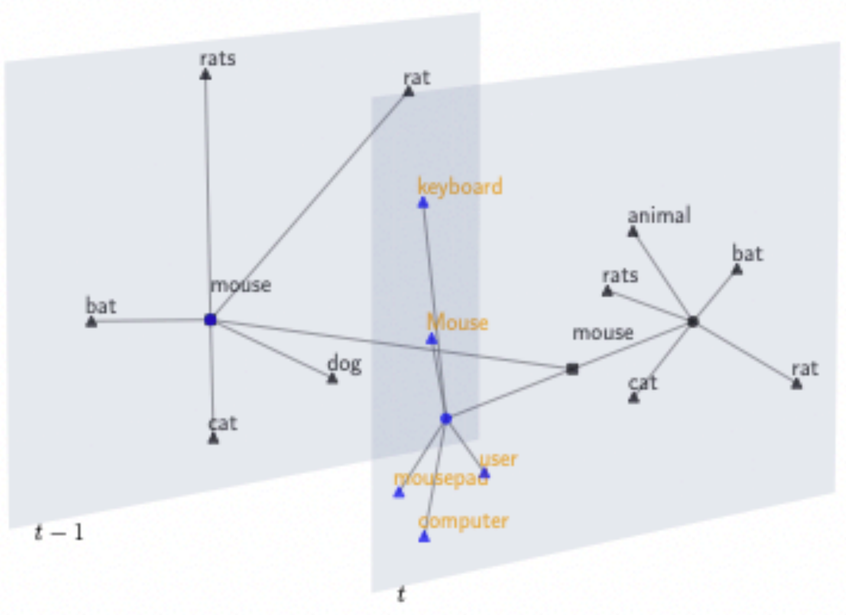
Use Case: Shifts

- A word loses a sense that it had previously, and gains a new sense: A=>B



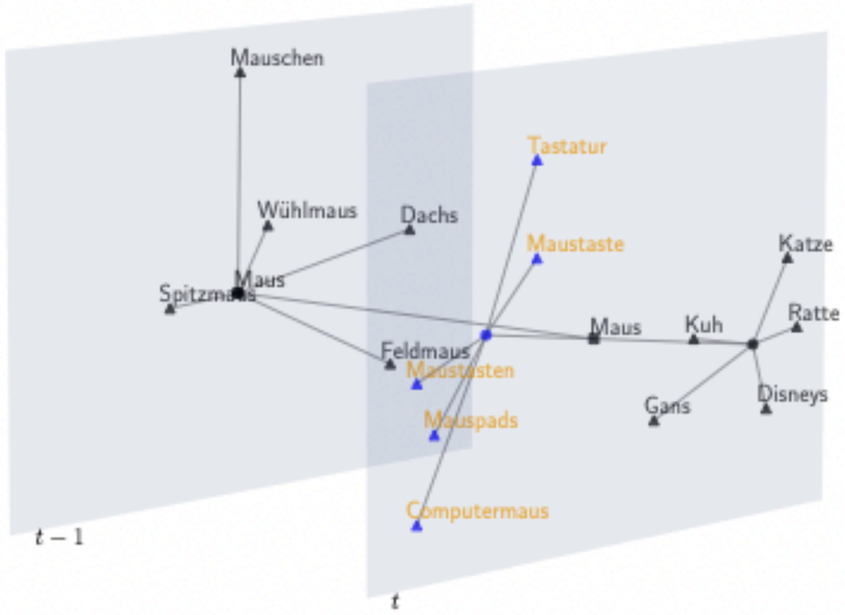
Use Case: Cross-language Comparison

English



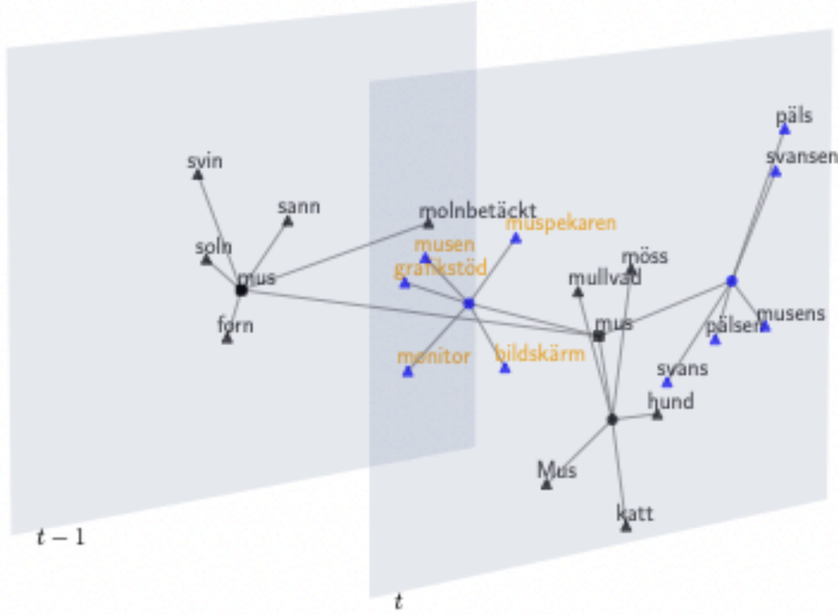
A => AB

German



A => AB

Swedish



A => ABC

C = {tail, fur, …}

Summary

- Our approach could detect low-frequency sense clusters, but..
- Good results in binary classification, but..
- Visualization tool
  - to track senses gained or lost over time in a temporal graph
  - to compare cross-language semantic change

Presence or Absence: Are Unknown Word Usages in Dictionaries? (Arxiv-24)

Xianghe Ma, Dominik Schlechtweg, Wei Zhao

Motivation

- Gap between lexical semantic change detection and lexicography
  - Detect word senses gained or lost over time
  - Unclear whether they are covered by dictionaries.
- Bridge between LSCD and lexicography
  - Discover new senses
  - Profile these senses lexicographically: generate sense definition, collect word usages for each sense, separate usages with different sense IDs.
  - AXOLOTL-24 shared task

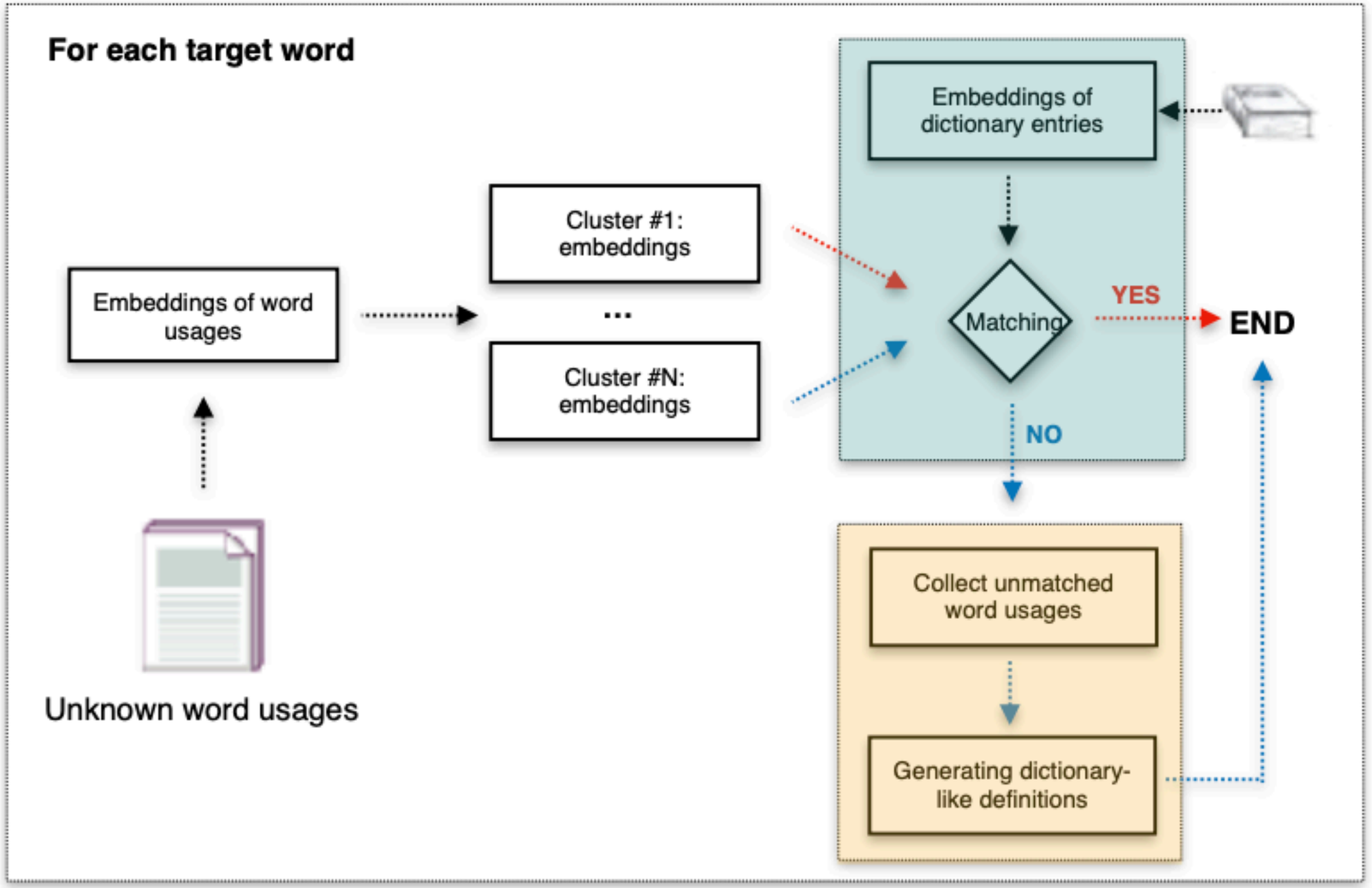
Research Question

Can we adapt LSDC to the lexicography problem - dictionary updating?

Schematic Overview

- Our lexicography system
  - **Subtask 1:** apply the semantic change detector to detect senses gained over time
  - **Subtask 1:** compare detected senses with dictionary entries to see if they are unrecorded in a dictionary
  - **Subtask 2:** if they are not, use LLMs to generate their sense definitions
- Evaluate our system in the AXOLOTL-24 shared task.

Our Lexicography System



Example

Target word: zersetzen

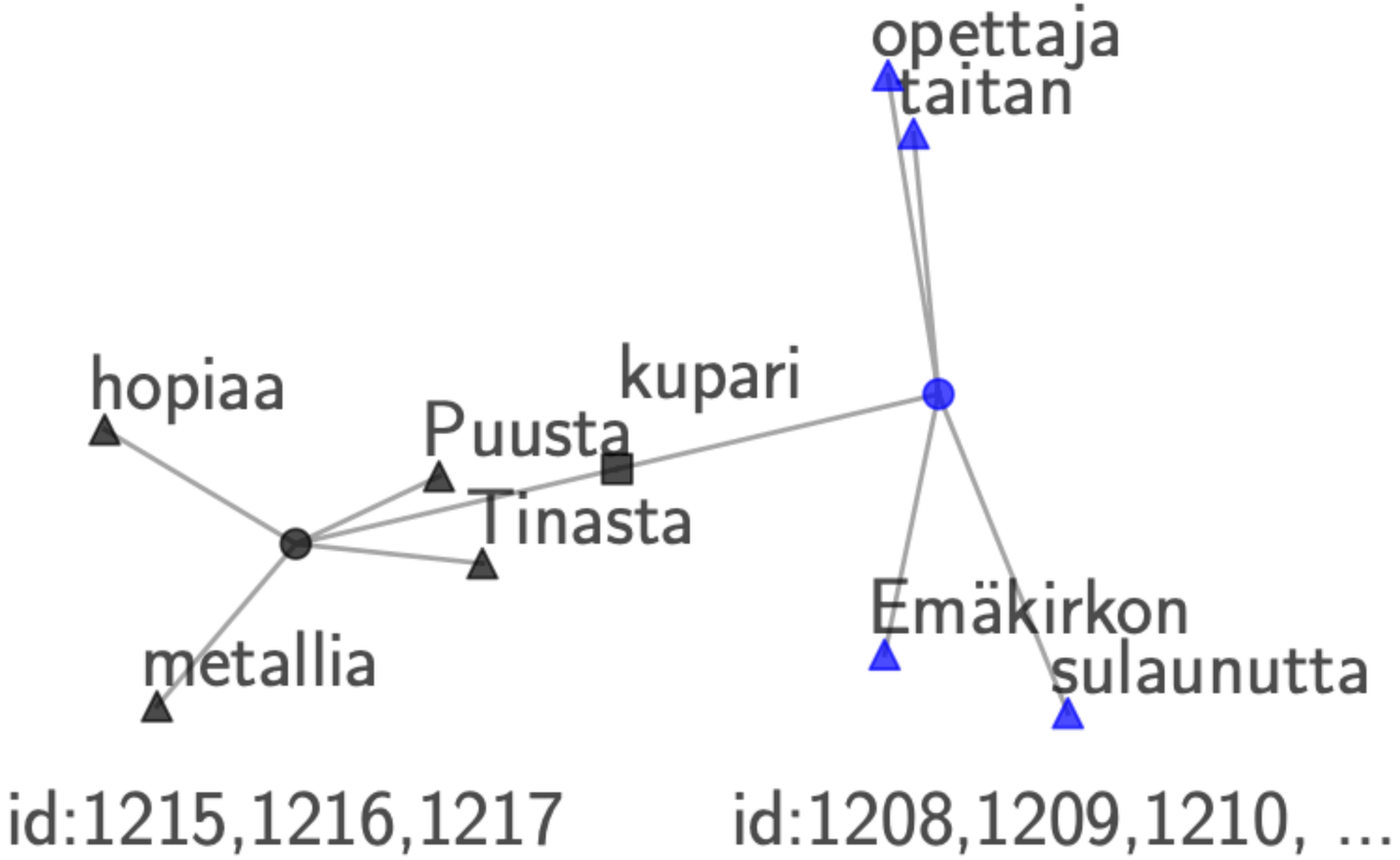
- SenseID-1: eine schädigende, zerstörende Wirkung auf den Bestand von etw. ausüben, etw. untergraben (damage)
- SenseID-2: etw., sich auflösen (dissolve/decompose)

Unknown word usage: …der unauflösliche Humus wird wahrscheinlich von ihm zersetzt.

Mappings: (Unknown usage, [senseID-1, senseID-2, new sense])

Our Lexicography System

Subtask 1 - mapping



Our Lexicography System

Subtask 2 - sense definition generation

[Instruction]:
Imagine that you are a lexicographer, given a headword {target_word} in {lang}, write the dictionary definition of its usage in the following quotations:

1. First quotation
2. Second quotation

[Requirements]:
The definition you create should be brief. A maximum of ten words is allowed. The definition ends at the first period.

[Response]:
Definition (string): {definition}

- Created from scratch
- Not optimal
- Start with English prompts
- Translate to other languages
- Definition length < 10
- Stop generating
- Number of word usages

Data Sources

AXOLOTL-24 corpora:

- Finnish: Dictionary of Old Literary Finnish
  - 1550s–1750s
- Russian: Explanatory Dictionary of the Living Great Russian Language
  - 1800s–now
- German: DWUG DE Sense
  - 1800s–1990s

Note: Word usages perhaps have been cleaned up as they are not collected from raw corpora but from dictionary entries, **but…**

Results on AXOLOTL-24

Subtask 1 - mapping

		Finnish		Russian		German	
Systems	#Entries	ARI	macro-F1	ARI	macro-F1	ARI	macro-F1
deep-change(1)	17	**0.649**	**0.760**	**0.247**	0.640	0.322	0.510
deep-change(2)	16	**0.649**	**0.760**	0.048	**0.750**	**0.521**	**0.740**
Holotniekat	4	0.596	0.630	0.043	0.660	0.298	0.610
ABDN-NLP (Ours)	2	0.553	0.590	0.009	0.570	0.102	0.300
Baseline	5	0.023	0.230	0.079	0.260	0.022	0.130

- Our system is unsupervised, not sure about other systems
- ARI vs. macro-FI in Russian and German

Analysis on ARI and F1

- Only unrecorded sense IDs are considered when computing F1
- Both recorded and unrecorded sense IDs are considered when computing ARI.
- Why F1 is sometimes higher than ARI?
  - Our system in Russian: 0.570 (F1) vs. 0.043 (ARI)
  - F1 will not penalize wrong predictions over recorded sense IDs
  - Russian: 47% unrecorded senses vs. German (10.2%) and Finnish (5.8%)

Metrics	Finnish	Russian	German
macro-F1 (recorded)	0.590	0.570	0.300
ARI_both	0.596	0.043	0.298
ARI_unrecorded	0.633	0.039	0.524
ARI_recorded	0.619	0.754	0.260

Results on AXOLOTL-24

Subtask 2 - sense definition generation

		Finnish		Russian		German	
Systems	#Entries	BLEU	BERTScore	BLEU	BERTScore	BLEU	BERTScore
ABDN-NLP (Ours)	3	**0.107**	**0.706**	0.027	0.677	0.000	**0.714**
TartuNLP	1	0.028	0.679	**0.587**	**0.869**	**0.010**	0.630
t-montes	7	0.023	0.675	0.027	0.656	**0.010**	0.650
Baseline	6	0.033	0.403	0.005	0.377	0.000	0.490

- Our system is unsupervised (GPT-3.5-turbo), not sure about other systems
- Bad in BLEU but good in BERTScore (our definitions are not lexically but semantically similar to gold standard)

Summary

- Can we adapt LSDC to the lexicography problem?
  - Yes, through two step: unrecorded sense detection and sense definition generation
- Good results in the AXOLOTL-24 shared task
- But results by LLMs may be misleading (data contamination)
  - The data sources of AXOLOTL test sets are publicly accessible.
    - target words, word usages, definitions
    - LLM’s training data?

Thank you