# Computational approaches to semantic change detection
# Day 4
# Back to linguistics: grammatical profiling for semantic change detection

Andrey Kutuzov, Lidia Pivovarova

University of Oslo, University of Helsinki

ESSLLI'2023

# Contents

# Setting the stage

## Lexical semantic change detection

- Well-represented in NLP (mostly diachronic changes)
- Shared tasks:
  - English, German, Latin and Swedish [Schlechtweg et al., 2020]
  - Italian [Basile et al., 2020]
  - Russian [Kutuzov and Pivovarova, 2021]
  - Spanish [Zamora-Reina et al., 2022]
- The systems are to rank a set of words according to the degree of their semantic change between two or more given time bins.

The dominant approaches use distributional word embeddings, encoding semantics via language modeling pre-training. Difference between word embeddings = difference between meanings.

# Setting the stage

## Lexical semantic change detection

- Well-represented in NLP (mostly diachronic changes)
- Shared tasks:
  - English, German, Latin and Swedish [Schlechtweg et al., 2020]
  - Italian [Basile et al., 2020]
  - Russian [Kutuzov and Pivovarova, 2021]
  - Spanish [Zamora-Reina et al., 2022]
- The systems are to rank a set of words according to the degree of their semantic change between two or more given time bins.

The dominant approaches use distributional word embeddings, encoding semantics via language modeling pre-training. Difference between word embeddings = difference between meanings.

But what about changes in the morphosyntactic behaviour of words?

# Setting the stage

- Semantics, morphology and syntax are strongly interdependent
- Semantic change <–> changes in the distribution of grammatical features (no matter the causal direction)

# Setting the stage

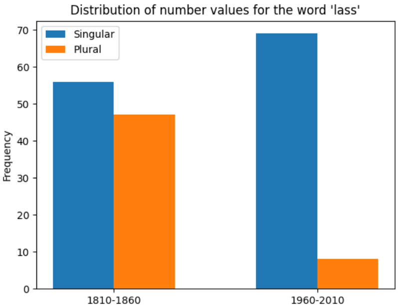
- Semantics, morphology and syntax are strongly interdependent
- Semantic change <–> changes in the distribution of grammatical features (no matter the causal direction)

English noun '*lass*':

1. 19th century: 'YOUNG WOMAN' sense more dominant ('*lasses are dancing*')
2. 20th century: 'SWEETHEART' sense more dominant ('*the young hero and his lass*')

A sharp decrease in plural usages!

**Are there systematic correlations between diachronic semantic change and morphosyntactic changes?**




*(English corpora of SemEval 2020)*

# Setting the stage

- Semantics, morphology and syntax are strongly interdependent
- Semantic change <–> changes in the distribution of grammatical features (no matter the causal direction)

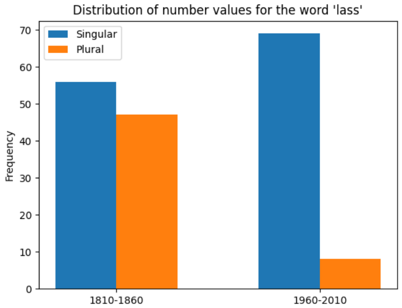
English noun ‘*lass*’:

1. $19^{th}$ century: ‘YOUNG WOMAN’ sense more dominant (‘*lasses are dancing*’)
2. $20^{th}$ century: ‘SWEETHEART’ sense more dominant (‘*the young hero and his lass*’)

A sharp decrease in plural usages!

**Are there systematic correlations between diachronic semantic change and morphosyntactic changes?**

The following is a retelling of [Kutuzov et al., 2021] (`https://aclanthology.org/2021.conll-1.33/`).




*(English corpora of SemEval 2020)*

# Setting the stage

## Grammatical profiling

Grammatical profile: the set of morphosyntactic properties preferred by a given word

# Setting the stage

## Grammatical profiling

Grammatical profile: the set of morphosyntactic properties preferred by a given word

- At the *RuShiftEval* shared task, [Ryzhova et al., 2021] already used grammatical profiles.
- Case and number frequency distributions.
- The system did not win, but an interesting approach.
- Semantic change detection without access to semantics?

# Setting the stage

## Grammatical profiling

Grammatical profile: the set of morphosyntactic properties preferred by a given word

- At the *RuShiftEval* shared task, [Ryzhova et al., 2021] already used grammatical profiles.
- Case and number frequency distributions.
- The system did not win, but an interesting approach.
- Semantic change detection without access to semantics?

## What we found in short

- Tracing changes in the morphosyntactic categories does outperform count-based static word embeddings.
- Morphological and syntactic categories are complementary.
- Useful categories are language-dependent.
- Predictions are interpretable (unlike embedding-based methods).

# Setting the stage

**Standard subtasks**

- Subtask 2: ranking (graded change detection)
- Subtask 1: binary classification (any senses gained or lost?)

Standard datasets to evaluate against:

1. *SemEval* dataset: both subtasks, English, German, Latin, Swedish [Schlechtweg et al., 2020]
2. *EvalIta* dataset: Subtask 1 only, Italian [Basile et al., 2020]
3. *RuShiftEval* dataset: Subtask 2 only, Russian [Kutuzov and Pivovarova, 2021]

# Setting the stage

## Standard subtasks

- Subtask 2: ranking (graded change detection)
- Subtask 1: binary classification (any senses gained or lost?)

Standard datasets to evaluate against:

1. *SemEval* dataset: both subtasks, English, German, Latin, Swedish [Schlechtweg et al., 2020]
2. *EvalIta* dataset: Subtask 1 only, Italian [Basile et al., 2020]
3. *RuShiftEval* dataset: Subtask 2 only, Russian [Kutuzov and Pivovarova, 2021]

In total: 274 manually annotated words from 3 Indo-European language groups: Italic, Germanic and Slavic.

# Contents

# Morphosyntax methods for semantic change detection

## Basic procedure 1/2

1. Target historical corpora are tagged and parsed with *UDPipe* [Straka and Straková, 2017]
2. Dictionary of frequencies of morphosyntactic categories (keys) for each target word in both corpora: `'lass': {'Number=Sing': 338, 'Number=Plur': 114}`
   - (in real data, keys are combinations of categories)
3. For syntax, keys are labels of the dependency arc between the target word and its head (`DEPREL` in the CONLLU format).
4. Feature list: union of all keys for a target word
5. Feature vectors $\vec{x}_1$ and $\vec{x}_2$ for two time bins (if a feature does not occur, its value is 0).

# Morphosyntax methods for semantic change detection

## Basic procedure 1/2

1. Target historical corpora are tagged and parsed with *UDPipe* [Straka and Straková, 2017]
2. Dictionary of frequencies of morphosyntactic categories (keys) for each target word in both corpora: `'lass': {'Number=Sing': 338, 'Number=Plur': 114}`
   - (in real data, keys are combinations of categories)
3. For syntax, keys are labels of the dependency arc between the target word and its head (`DEPREL` in the CONLLU format).
4. Feature list: union of all keys for a target word
5. Feature vectors $\vec{x}_1$ and $\vec{x}_2$ for two time bins (if a feature does not occur, its value is 0).

Feature vectors are actually time-dependent grammatical profiles.

# Morphosyntax methods for semantic change detection

## Basic procedure 2/2

1. Cosine distance $cos(\vec{x}_1, \vec{x}_2)$ is the change in the grammatical profiles of the target word
2. Separate distance scores for morphology and syntax: $d_{morph}$ and $d_{synt}$
3. Subtask 2: distances map directly to semantic change
4. Subtask 1: top $n$ target words by distance score are labeled as ‘changed’ (1) and the rest as ’stable’ (0)
5. $d_{morph}$ and $d_{synt}$ can be averaged to combine morphology and syntax signals.

# Morphosyntax methods for semantic change detection

## Basic procedure 2/2

1. Cosine distance $cos(\vec{x}_1, \vec{x}_2)$ is the change in the grammatical profiles of the target word
2. Separate distance scores for morphology and syntax: $d_{morph}$ and $d_{synt}$
3. Subtask 2: distances map directly to semantic change
4. Subtask 1: top $n$ target words by distance score are labeled as ‘changed’ (1) and the rest as ’stable’ (0)
5. $d_{morph}$ and $d_{synt}$ can be averaged to combine morphology and syntax signals.

In the end, 3 solutions for each task:
‘morphology’, ‘syntax’, ‘averaged’

# Some small improvements help a lot

## 1. Filtering

- Exclude rare grammatical categories
- Sum of feature occurrences < 5% of total word usages? Discard.
- NB: we did not tune this threshold to avoid over-fitting.

# Some small improvements help a lot

## 1. Filtering

- Exclude rare grammatical categories
- Sum of feature occurrences < 5% of total word usages? Discard.
- NB: we did not tune this threshold to avoid over-fitting.

## 2. Category separation (example for the verb '*to circle*')

Basic procedure: features comprise all categories of a word occurrence (`FEATS`):

```
Tense=Pres|VerbForm=Part:50
Mood=Ind|Tense=Past|VerbForm=Fin:24
Tense=Past|VerbForm=Part|Voice=Pass:17
VerbForm=Inf:9
Mood=Ind|Tense=Pres|VerbForm=Fin:1
Tense=Past|VerbForm=Part:1
```

# Some small improvements help a lot

## 1. Filtering

- Exclude rare grammatical categories
- Sum of feature occurrences < 5% of total word usages? Discard.
- NB: we did not tune this threshold to avoid over-fitting.

## 2. Category separation (example for the verb ‘*to circle*’)

Basic procedure: features comprise all categories of a word occurrence (`FEATS`):

```
Tense=Pres|VerbForm=Part:50
Mood=Ind|Tense=Past|VerbForm=Fin:24
Tense=Past|VerbForm=Part|Voice=Pass:17
VerbForm=Inf:9
Mood=Ind|Tense=Pres|VerbForm=Fin:1
Tense=Past|VerbForm=Part:1
```

Too sparse, conflates different phenomena. Convert to separate feature vector per category:

# Some small improvements help a lot

## 1. Filtering

- Exclude rare grammatical categories
- Sum of feature occurrences < 5% of total word usages? Discard.
- NB: we did not tune this threshold to avoid over-fitting.

## 2. Category separation (example for the verb '*to circle*')

Basic procedure: features comprise all categories of a word occurrence (`FEATS`):

```
Tense=Pres|VerbForm=Part:50
Mood=Ind|Tense=Past|VerbForm=Fin:24
Tense=Past|VerbForm=Part|Voice=Pass:17
VerbForm=Inf:9
Mood=Ind|Tense=Pres|VerbForm=Fin:1
Tense=Past|VerbForm=Part:1
```

Too sparse, conflates different phenomena. Convert to separate feature vector per category:

```
Tense:{Past 42, Pres 51}
VerbForm:{Part 68, Fin 25, Inf 9}
Mood:{Ind 25}
Voice:{Pass 17}
```

# Some small improvements help a lot

## 1. Filtering

- Exclude rare grammatical categories
- Sum of feature occurrences < 5% of total word usages? Discard.
- NB: we did not tune this threshold to avoid over-fitting.

## 2. Category separation (example for the verb '*to circle*')

Basic procedure: features comprise all categories of a word occurrence (`FEATS`):

```
Tense=Pres|VerbForm=Part:50
Mood=Ind|Tense=Past|VerbForm=Fin:24
Tense=Past|VerbForm=Part|Voice=Pass:17
VerbForm=Inf:9
Mood=Ind|Tense=Pres|VerbForm=Fin:1
Tense=Past|VerbForm=Part:1
```

Too sparse, conflates different phenomena. Convert to separate feature vector per category:

```
Tense:{Past 42, Pres 51}
VerbForm:{Part 68, Fin 25, Inf 9}
Mood:{Ind 25}
Voice:{Pass 17}
```

Distances computed separately for each category. Change score is max-pooled.

# We also tested some improvements

## 3. Combination of morphology and syntax

- With category separation, we have an array of morphological distances.
- It is now possible to treat syntax more flexibly:
  1. average morphological and syntactic distances
     - morphology and syntax are weighted equally
  2. append $d_{synt}$ to morphological distances, choose the maximum value (max pooling )
     - syntax is weighted down depending on the richness of the morphological profile (number of categories) for a particular word.

# We also tested some improvements

## 3. Combination of morphology and syntax

- With category separation, we have an array of morphological distances.
- It is now possible to treat syntax more flexibly:
  1. average morphological and syntactic distances
     - morphology and syntax are weighted equally
  2. append $d_{synt}$ to morphological distances, choose the maximum value (max pooling )
     - syntax is weighted down depending on the richness of the morphological profile (number of categories) for a particular word.

All these ‘boosters’ do improve the results of our semantic change detection system without semantics.
Let’s see.

# Contents

# Results on Subtask 2 (graded change detection, Spearman $\rho$)

| Categories | SemEval 2020 languages | | | | | Russian | | | |
|---|---|---|---|---|---|---|---|---|---|
| | English | German | Latin | Swedish | Mean | Russian1 | Russian2 | Russian3 | Mean |
| | | | | | Basic procedure | | | | |
| **Morphology** | 0.234 | 0.043 | 0.241 | 0.207 | 0.181 | **0.137** | 0.210 | **0.327** | **0.225** |
| **Syntax** | 0.319 | 0.163 | 0.328 | -0.017 | 0.198 | 0.060 | 0.101 | 0.269 | 0.143 |
| **Average** | 0.293 | 0.147 | 0.304 | 0.088 | 0.208 | 0.101 | 0.191 | 0.294 | 0.195 |
| | | | | | 5% filtering | | | | |
| **Morphology** | 0.211 | 0.080 | 0.285 | 0.191 | 0.192 | 0.127 | 0.185 | 0.264 | 0.192 |
| **Syntax** | **0.331** | 0.146 | 0.265 | 0.184 | 0.231 | 0.056 | 0.111 | 0.279 | 0.149 |
| **Average** | 0.315 | 0.171 | 0.345 | 0.263 | 0.273 | 0.094 | 0.183 | 0.278 | 0.185 |
| | | | | | Category separation and 5% filtering | | | | |
| **Morphology** | 0.218 | 0.074 | 0.519 | 0.303 | 0.278 | 0.028 | **0.241** | 0.293 | 0.187 |
| **Average** | 0.321 | 0.227 | 0.523 | **0.381** | 0.363 | 0.002 | 0.179 | 0.278 | 0.153 |
| **Combination / max pool** | 0.320 | **0.298** | **0.525** | 0.334 | **0.369** | 0.000 | 0.149 | 0.242 | 0.130 |
| | **Prior SemEval results** | | | | | **Prior RuShiftEval results** | | | |
| Count baseline | 0.022 | 0.216 | 0.359 | -0.022 | 0.144 | 0.314 | 0.302 | 0.381 | 0.332 |
| Best shared task system | 0.422 | 0.725 | 0.412 | 0.547 | 0.527 | 0.798 | 0.803 | 0.822 | 0.807 |
| [Ryzhova et al., 2021] | - | - | - | - | - | 0.157 | 0.199 | 0.343 | 0.233 |

# Results on Subtask 1 (binary change detection, accuracy)

| Categories | English | German | Latin | Swedish | Mean | Italian |
|---|---|---|---|---|---|---|
| | Basic procedure with $n = 43\%$ threshold | | | | | |
| **Morphology** | 0.595 | 0.521 | 0.525 | 0.581 | 0.555 | 0.722 |
| **Syntax** | 0.541 | **0.646** | 0.575 | 0.645 | 0.602 | 0.611 |
| **Average** | 0.568 | 0.583 | 0.475 | **0.710** | 0.584 | 0.722 |
| | Automatic change point detection with dynamic programming [Truong et al., 2020] | | | | | |
| **Morphology** | **0.622** | 0.479 | **0.625** | 0.548 | 0.569 | 0.722 |
| **Syntax** | 0.514 | 0.625 | 0.500 | 0.677 | 0.579 | 0.611 |
| **Average** | 0.595 | 0.542 | 0.525 | 0.677 | 0.585 | **0.778** |
| | Category separation, change point detection and 5% filtering | | | | | |
| **Morphology** | **0.622** | 0.583 | **0.625** | 0.581 | **0.603** | 0.500 |
| **Average** | 0.595 | 0.625 | 0.450 | **0.710** | 0.595 | 0.667 |
| **Combination / max pool** | 0.541 | 0.583 | 0.575 | 0.645 | 0.586 | 0.500 |
| | **Prior SemEval results** | | | | | **Prior Evallita results** |
| Baseline | 0.595 | 0.688 | 0.525 | 0.645 | 0.613 | 0.611 |
| Best shared task system | 0.622 | 0.750 | 0.700 | 0.677 | 0.687 | 0.944 |

*NB: All binary change detection methods are fundamentally based on the scores produced by ranking (Subtask 2)*

# Contents

# When it works?

Many cases of broadening and narrowing are captured.

## Just one English example

- The noun '*stab*' is ranked 4th/37 target words.
- Syntactic changes
- The word used as oblique argument:
  - 19th century: 13% of all occurrences
  - 20th century: 27% of all occurrences
- Why?
- Emergent sense of 'SUDDEN SHARP FEELING':
- '*...left me with a sharp stab of sadness*', etc.

# When it works?

Many cases of broadening and narrowing are captured.

## Just one English example

- The noun '*stab*' is ranked 4th/37 target words.
- Syntactic changes
- The word used as oblique argument:
  - 19th century: 13% of all occurrences
  - 20th century: 27% of all occurrences
- Why?
- Emergent sense of 'SUDDEN SHARP FEELING':
- '*...left me with a sharp stab of sadness*', etc.


More examples for other languages in [Kutuzov et al., 2021].

# When it does not work?

## False positives

- Erroneously high semantic change score
- Can be caused by sharp increase in word frequency
- Grammatical profile becomes more diverse:
  - German ‘*Lyzeum*’ (‘LYCEUM’), $19^{th}$ vs $20^{th}$ century
  - Latin ‘*jus*’ (a ‘RIGHT’, the ‘LAW’), BCE vs CE

# When it does not work?

## False positives

- Erroneously high semantic change score
- Can be caused by sharp increase in word frequency
- Grammatical profile becomes more diverse:
  - German '*Lyzeum*' ('LYCEUM'), $19^{th}$ vs $20^{th}$ century
  - Latin '*jus*' (a 'RIGHT', the 'LAW'), BCE vs CE

## False negatives

- Erroneously low semantic change score
- Semantic shift is not reflected (enough) in morphosyntax
- German '*Ohrwurm*' ('EARWORM' →'CATCHY SONG')
  - no significant changes in the grammatical profile
- Latin '*pontifex*' (a 'BISHOP' →the 'POPE')
  - singular usages increased from 63% to 83%, but still low change score
  - ranked $22^{nd}$ out of 40 target words

# Contents

# Category importance

Which categories are most related to semantic change?

# Category importance

Which categories are most related to semantic change?

| Language | Important categories | Accuracy | F1 |
|---|---|---|---|
| **English nouns** | *number* | 0.576 | 0.523 |
| **English verbs** | *verb form, syntax* | 0.750 | 0.733 |
| **German** | *number, syntax, gender* | 0.542 | 0.541 |
| **Swedish** | *syntax, mood, voice, definiteness, number* | 0.839 | 0.797 |
| **Latin** | *voice, number, degree, case, gender, mood, aspect, person, tense* | 0.650 | 0.649 |
| **Italian** | *number, tense, syntax* | 0.778 | 0.723 |

Categories with positive weights in binary logistic regression classifiers of semantic change.
Features are cosine distances between frequency vectors of categories from different time bins.

# Category importance

Spearman $\rho$ between per-category diachronic grammatical profile distances and manually annotated semantic change estimations:

| | Number | Mood | Degree | Gender | Case | Syntax |
|---|---|---|---|---|---|---|
| **English** | - | - | - | - | - | 0.331 |
| **German** | - | - | - | - | - | - |
| **Latin** | 0.304 | - | 0.301 | - | - | - |
| **Swedish** | 0.402 | 0.397 | - | - | - | - |
| **Russian 1** | - | - | - | 0.218 | 0.196 | - |
| **Russian 2** | - | - | - | 0.231 | 0.324 | - |
| **Russian 3** | 0.246 | - | - | 0.218 | 0.327 | 0.279 |

Only significant correlations shown ($p < 0.05$). Note no correlations for German, a fusional language - why so? Correlations with gender in Russian are also unexpected.

# Contents

# Grammatical profiles ensembled with LMs?

## Are LMs and explicit morphosyntax complementary?

- LMs capture approximations of grammatical information in their deep representations [Warstadt et al., 2020].
- They are widely used in LSCD.
- Can they reliably detect meaning shifts accompanied by morphosyntactic changes in word usage?
- Does adding grammatical profiles to LMs result in improved performance? .
- The following is a retelling of [Giulianelli et al., 2022] (`https://aclanthology.org/2022.lchange-1.6/`)

# Grammatical profiles ensembled with LMs?

## Let's evaluate on an enriched company of datasets

| | EN | DE | IT | LA | NO-1 | NO-2 | RU-1 | RU-2 | RU-3 | SW |
|---|---|---|---|---|---|---|---|---|---|---|
| **Period 1** | 1810-1860 | 1800-1899 | 1945-1970 | -200-0 | 1929-1965 | 1980-1990 | 1700-1916 | 1918-1990 | 1700-1916 | 1790-1830 |
| **Period 2** | 1960-2010 | 1946-1990 | 1990-2014 | 0-2000 | 1970-2013 | 2012-2019 | 1918-1990 | 1992-2016 | 1992-2016 | 1895-1903 |
| **Tokens** (mln) | 7+7 | 70+72 | 52+197 | 2+9 | 57+175 | 43+649 | 93+122 | 122+107 | 93+107 | 71+110 |
| **Targets** | 37 | 48 | 18 | 40 | 80 | 80 | 99 | 99 | 99 | 32 |
| **Ranking** | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Classification** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ |

# Grammatical profiling

To recall:

- Target historical corpora are tagged and parsed with *UDPipe*
- **MORPH**: profile is the set of frequency values for each morphological category

| | | | | |
|---|---|---|---|---|
| Tense | Past | 42/70 | $\longrightarrow$ cosine | |
| | Pres | 51/55 | | |
| | Part | 68/40 | | $\longrightarrow$ max |
| VerbForm | Fin | 25/9 | $\longrightarrow$ cosine | |
| | Inf | 9/25 | | |

- Additional filtering: discarding rare features (<5% in both corpora)
- **SYNT**: profile is the set of frequency values for the dependency labels that govern the target word
- **MORPHSYNT**: the syntactic profile is appended to the morphological profile
- Semantic change scores computed by measuring distance between two profiles

# Contextualized embeddings (LMs)

- XLM-R [Conneau et al., 2020]: a pre-trained multilingual language model, can be applied to all languages under analysis
- We fine-tune XLM-R on each monolingual diachronic corpus, separately
- Embeddings are extracted for all occurrences of the target words in both time periods

# Contextualized embeddings (LMs)

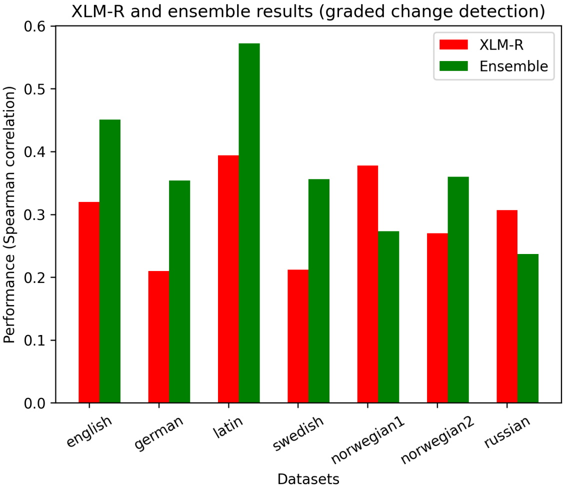
- XLM-R [Conneau et al., 2020]: a pre-trained multilingual language model, can be applied to all languages under analysis
- We fine-tune XLM-R on each monolingual diachronic corpus, separately
- Embeddings are extracted for all occurrences of the target words in both time periods
- Semantic change metrics:
  - **APD**: averaged pairwise cosine distance [Giulianelli et al., 2020]
  - **PRT**: cosine distance between *prototypes*, i.e. averaged contextualised embeddings [Kutuzov and Giulianelli, 2020]
  - **APD-PRT**: averaging of the two previous methods [Kutuzov et al., 2022]
  - **JSD**: clustering contextualised embeddings and then computing the Jensen-Shannon divergence between cluster distributions [Martinc et al., 2020, Giulianelli et al., 2020]

# Contextualized embeddings (LMs)

- XLM-R [Conneau et al., 2020]: a pre-trained multilingual language model, can be applied to all languages under analysis
- We fine-tune XLM-R on each monolingual diachronic corpus, separately
- Embeddings are extracted for all occurrences of the target words in both time periods
- Semantic change metrics:
  - **APD**: averaged pairwise cosine distance [Giulianelli et al., 2020]
  - **PRT**: cosine distance between *prototypes*, i.e. averaged contextualised embeddings [Kutuzov and Giulianelli, 2020]
  - **APD-PRT**: averaging of the two previous methods [Kutuzov et al., 2022]
  - **JSD**: clustering contextualised embeddings and then computing the Jensen-Shannon divergence between cluster distributions [Martinc et al., 2020, Giulianelli et al., 2020]
- Ensembling with profiling: the geometric mean $\sqrt{c_g c_e}$ between the change score $c_g$ obtained using grammatical profiles and the score $c_e$ output by an embedding-based metric (e.g., **PRT-MORPHSYNT**)

# Contents

# Grammatical profiles improve the performance of LMs




*Performance of XLM-R (PRT) and an ensemble (PRT-MORPHSYNT) on the ranking task.*

# Spearman rank-correlation scores in the ranking task

| Method | EN | DE | LA | SW | NO-1 | NO-2 | RU-1 | RU-2 | RU-3 | AVG |
|---|---|---|---|---|---|---|---|---|---|---|
| PROFILES | | | | | | | | | | |
| MORPH | 0.218 | 0.120 | 0.519 | 0.303 | 0.106 | *0.409* | 0.028 | *0.241* | *0.293* | *0.248* |
| SYNT | *0.331* | 0.146 | 0.265 | 0.184 | *0.179* | 0.006 | *0.056* | 0.111 | 0.279 | 0.173 |
| MORPHSYNT | 0.320 | *0.298* | *0.525* | *0.334* | 0.064 | 0.265 | 0.000 | 0.149 | 0.242 | 0.244 |
| CONTEXTUALISED (XLM-R) | | | | | | | | | | |
| APD | ***0.514*** | 0.073 | 0.162 | *0.310* | 0.389 | *0.387* | 0.372 | ***0.480*** | ***0.457*** | *0.349* |
| PRT | 0.320 | 0.210 | *0.394* | 0.212 | 0.378 | 0.270 | 0.294 | 0.313 | 0.313 | 0.300 |
| APD-PRT | 0.457 | 0.202 | 0.370 | 0.220 | ***0.394*** | 0.325 | ***0.376*** | 0.374 | 0.384 | **0.345** |
| Clustering/JSD | 0.127 | *0.287* | 0.318 | -0.108 | 0.160 | -0.137 | 0.247 | 0.267 | 0.362 | 0.169 |
| ENSEMBLES | | | | | | | | | | |
| APD-MORPH | 0.262 | 0.140 | 0.506 | 0.350 | 0.151 | ***0.503*** | 0.062 | *0.288* | 0.340 | 0.289 |
| APD-SYNT | 0.384 | 0.159 | 0.264 | 0.255 | *0.262* | 0.119 | *0.093* | 0.181 | *0.354* | 0.230 |
| APD-MORPHSYNT | *0.390* | *0.290* | *0.513* | ***0.397*** | 0.180 | 0.364 | 0.036 | 0.216 | 0.299 | *0.298* |
| PRT-MORPH | 0.278 | 0.204 | 0.528 | 0.305 | 0.236 | *0.478* | 0.112 | *0.309* | 0.336 | 0.309 |
| PRT-SYNT | 0.448 | 0.213 | 0.401 | 0.280 | *0.351* | 0.146 | *0.186* | 0.246 | *0.351* | 0.291 |
| PRT-MORPHSYNT | *0.451* | ***0.354*** | ***0.572*** | *0.356* | 0.273 | 0.360 | 0.117 | 0.269 | 0.326 | *0.342* |
| APD-PRT-MORPH | 0.277 | 0.188 | 0.518 | 0.338 | 0.189 | *0.497* | 0.092 | *0.310* | 0.340 | 0.305 |
| APD-PRT-SYNT | 0.405 | 0.189 | 0.376 | 0.295 | *0.330* | 0.121 | *0.147* | 0.235 | *0.367* | 0.274 |
| APD-PRT-MORPHSYNT | *0.418* | *0.337* | *0.554* | *0.377* | 0.236 | 0.359 | 0.092 | 0.255 | 0.328 | *0.328* |

# Contents

# When does ensembling fail?

- Profiling works best for Latin and German, languages with rich morphology

# When does ensembling fail?

- Profiling works best for Latin and German, languages with rich morphology
- Grammatical profiles are consistently worse than XLM-R on Russian datasets, Norwegian-1 and English
- English has quite poor morphology
- However, Russian and Norwegian are also morphologically rich

# When does ensembling fail?

- Profiling works best for Latin and German, languages with rich morphology
- Grammatical profiles are consistently worse than XLM-R on Russian datasets, Norwegian-1 and English
- English has quite poor morphology
- However, Russian and Norwegian are also morphologically rich

*Possible explanation has to do with the time gap between the two periods:*

| | EN | DE | IT | LA | NO-1 | NO-2 | RU-1 | RU-2 | RU-3 | SW |
|---|---|---|---|---|---|---|---|---|---|---|
| **Period 1** | 1810-1860 | 1800-1899 | 1945-1970 | -200-0 | *1929-1965* | 1980-1990 | *1700-1916* | 1918-1990 | 1700-1916 | 1790-1830 |
| **Period 2** | 1960-2010 | 1946-1990 | 1990-2014 | 0-2000 | *1970-2013* | 2012-2019 | *1918-1990* | *1992-2016* | 1992-2016 | 1895-1903 |
| **Tokens** (mln) | 7+7 | 70+72 | 52+197 | 2+9 | 57+175 | 43+649 | 93+122 | 122+107 | 93+107 | 71+110 |
| **Targets** | 37 | 48 | 18 | 40 | 80 | 80 | 99 | 99 | 99 | 32 |
| **Ranking** | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **Classification** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ |

# When does ensembling help?

- False positives and negatives of best XLM-R method corrected via ensembling

# When does ensembling help?

- False positives and negatives of best XLM-R method corrected via ensembling
- **False positives**: E.g., ranking of *'tree'* (English dataset) improves by 20 positions; predicted change score decreases thanks to small distance between grammatical profiles

```
Number=Plur:  56.27% -> 56.33%
Number=Sing:  43.73% -> 43.67%
```

```
nmod:   36.52% -> 30.04%
obl:    28.78% -> 33.87%
obj:    11.89% -> 13.28%
nsubj:  11.26% -> 10.92%
conj:   11.55% -> 11.88%
```

# When does ensembling help?

- False positives and negatives of best XLM-R method corrected via ensembling
- **False positives**: E.g., ranking of *'tree'* (English dataset) improves by 20 positions; predicted change score decreases thanks to small distance between grammatical profiles

```
Number=Plur:  56.27% -> 56.33%
Number=Sing:  43.73% -> 43.67%
```

```
nmod:   36.52% -> 30.04%
obl:    28.78% -> 33.87%
obj:    11.89% -> 13.28%
nsubj:  11.26% -> 10.92%
conj:   11.55% -> 11.88%
```

- **False negatives**: E.g., ranking of *'plane'* improves by 15 positions; predicted change score increases thanks to large distance between grammatical profiles

```
Number=Sing:  83.59% -> 72.48%
Number=Plur:  16.41% -> 27.52%
```

```
nmod:   35.34% -> 20.36%
nsubj:  12.85% -> 24.13%
obj:    13.25% -> 19.67%
conj:    9.24% ->  5.44%
```

# Contents

# Summary

## Part 1

- Grammatical profiling allows to build a successful semantic change detection system without access to lexical semantics at all.
- It consistently outperforms count-based word embeddings.
- For Latin, it outperforms all official results in Subtask 2 of SemEval'20 Task 1.

# Summary

## Part 1

- Grammatical profiling allows to build a successful semantic change detection system without access to lexical semantics at all.
- It consistently outperforms count-based word embeddings.
- For Latin, it outperforms all official results in Subtask 2 of SemEval'20 Task 1.
- Importantly, the predictions are highly interpretable.

# Summary

## Part 1

- Grammatical profiling allows to build a successful semantic change detection system without access to lexical semantics at all.
- It consistently outperforms count-based word embeddings.
- For Latin, it outperforms all official results in Subtask 2 of SemEval'20 Task 1.
- Importantly, the predictions are highly interpretable.

# Summary

## Part 2

- Providing large pre-trained language models with explicit morphosyntactic information (ensembling) can help detect and quantify lexical semantic change
- The ensemble predictions mostly outperform standalone grammatical profiles or contextualised embeddings in the ranking task
- Do not fire the linguist yet!

# Summary

## Part 2

- Providing large pre-trained language models with explicit morphosyntactic information (ensembling) can help detect and quantify lexical semantic change
- The ensemble predictions mostly outperform standalone grammatical profiles or contextualised embeddings in the ranking task
- Do not fire the linguist yet!
- Datasets where grammatical profiles fail to help are:
  1. languages with poor morphology
  2. long time periods separated by narrow time gaps: not enough for morphosyntactic changes to manifest themselves.
- Grammatical profiling should become one of the standard baselines for semantic change detection.

# References I


- Basile, P., Caputo, A., Caselli, T., Cassotti, P., and Varvara, R. (2020).
  Diacr-ita@ evalita2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task.
  *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020).
  Unsupervised cross-lingual representation learning at scale.
  In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

- Giulianelli, M., Del Tredici, M., and Fernández, R. (2020).
  Analysing lexical semantic change with contextualised word representations.
  In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

# References II


Giulianelli, M., Kutuzov, A., and Pivovarova, L. (2022).
Do not fire the linguist: Grammatical profiles help language models detect semantic change.
In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 54–67, Dublin, Ireland. Association for Computational Linguistics.

Kutuzov, A. and Giulianelli, M. (2020).
UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection.
In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

# References III


Kutuzov, A. and Pivovarova, L. (2021).
RuShiftEval: a shared task on semantic shift detection for Russian.
*Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Kutuzov, A., Pivovarova, L., and Giulianelli, M. (2021).
Grammatical profiling for semantic change detection.
In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 423–434, Online. Association for Computational Linguistics.

Kutuzov, A., Velldal, E., and Øvrelid, L. (2022).
Contextualized embeddings for semantic change detection: Lessons learned.
In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.

# References IV


Martinc, M., Montariol, S., Zosa, E., and Pivovarova, L. (2020).
Capturing evolution in word usage: just add more clusters?
In *Companion Proceedings of the Web Conference 2020*, pages 343–349.

Ryzhova, A., Ryzhova, D., and Sochenkov, I. (2021).
Detection of semantic changes in Russian nouns with distributional models and grammatical features.
*Computational linguistics and intellectual technologies: Papers from the annual conference Dialogue*.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., and Tahmasebi, N. (2020).
SemEval-2020 task 1: Unsupervised lexical semantic change detection.
In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

# References V


Straka, M. and Straková, J. (2017).
Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe.
In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Truong, C., Oudre, L., and Vayatis, N. (2020).
Selective review of offline change point detection methods.
*Signal Processing*, 167:107299.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020).
BLiMP: The benchmark of linguistic minimal pairs for English.
*Transactions of the Association for Computational Linguistics*, 8:377–392.

# References VI


Zamora-Reina, F. D., Bravo-Marquez, F., and Schlechtweg, D. (2022). LSCDiscovery: A shared task on semantic change discovery and detection in Spanish.
In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.