Distributional, yes—but *semantics*?

Timothee Mickus

November 2nd, 2023
Language Technology Research Group
Research seminar

What is a distributional semantics model?

How it started:

- "You shall know a word by the company it keeps" (Firth, 1957)

What is a distributional semantics model?

How it started:

- "You shall know a word by the company it keeps" (Firth, 1957)
- Something that models

  Pr(word | context)

What is a distributional semantics model?

How it started:

- "You shall know a word by the company it keeps" (Firth, 1957)
- Something that models

  $$\Pr(\text{word} \mid \text{context})$$

- Matches a wide array of actual training objectives for static and contextualized embeddings (CBOW, MLM, ...)

What is a distributional semantics model?

How it started:

- "You shall know a word by the company it keeps" (Firth, 1957)
- Something that models

$$\Pr(\text{word} \mid \text{context})$$

- Matches a wide array of actual training objectives for static and contextualized embeddings (CBOW, MLM, ...)
- Matches theoretical expectations (Sahlgren, 2008)

What is a distributional semantics model?

How it started:

- "You shall know a word by the company it keeps" (Firth, 1957)
- Something that models

  Pr(word | context)

- Matches a wide array of actual training objectives for static and contextualized embeddings (CBOW, MLM, ...)
- Matches theoretical expectations (Sahlgren, 2008)

How it's going:

- BayesOpt for HPO, looking for wordvec hyperparameters that increase the mass on attested words in Pr(word | context)

What is a distributional semantics model?

How it started:

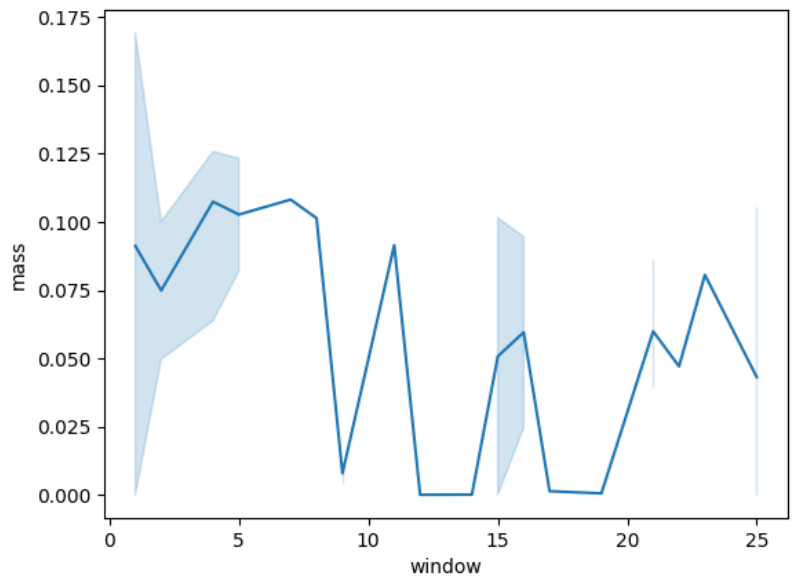
- "You shall know a word by the company it keeps" (Firth, 1957)
- Something that models

  Pr(word | context)

- Matches a wide array of actual training objectives for static and contextualized embeddings (CBOW, MLM, ...)
- Matches theoretical expectations (Sahlgren, 2008)

How it's going:

- BayesOpt for HPO, looking for wordvec hyperparameters that increase the mass on attested words in Pr(word | context)



A thought experiment

A thought experiment

A thought experiment

A thought experiment

A thought experiment

A thought experiment

Do you really need semantics to model
Pr(word | context)**?**

A thought experiment

Do you really need semantics to model
Pr(word | context)**?**

For today

- high-level talk
- linguistic focus
- borrowing results from recent research
- focusing on models that are easy to interpret

Outline

1. Segonne and Mickus (2023)
2. Mickus and Copot (In prep.)
3. Mickus and Bernard (2023)

Outline

1. Segonne and Mickus (2023)

2. Mickus and Copot (In prep.)

3. Mickus and Bernard (2023)

"*Definition Modeling* : `To model definitions.`"
Generating Definitions With Little to No Semantics

Vincent Segonne*
Université Grenoble Alpes
vincent.segonne
@univ-grenoble-alpes.fr

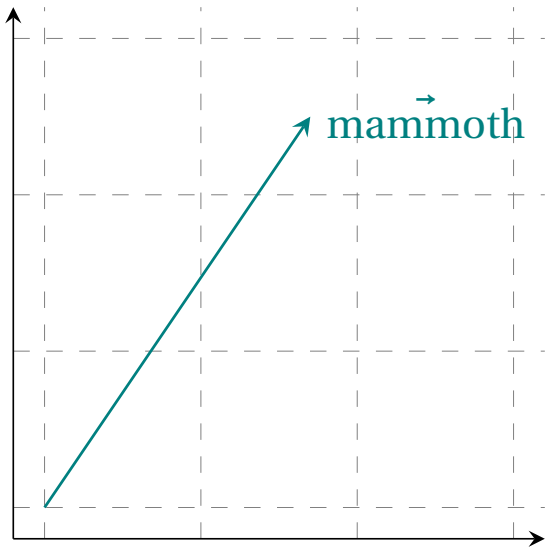
Timothee Mickus*
Helsinki University
timothee.mickus
@helsinki.fi

Definition Modeling

- Noraset et al. (2017): Well-trained distributional representations should capture enough semantics to derive definitions

Definition Modeling

- Noraset et al. (2017): Well-trained distributional representations should capture enough semantics to derive definitions

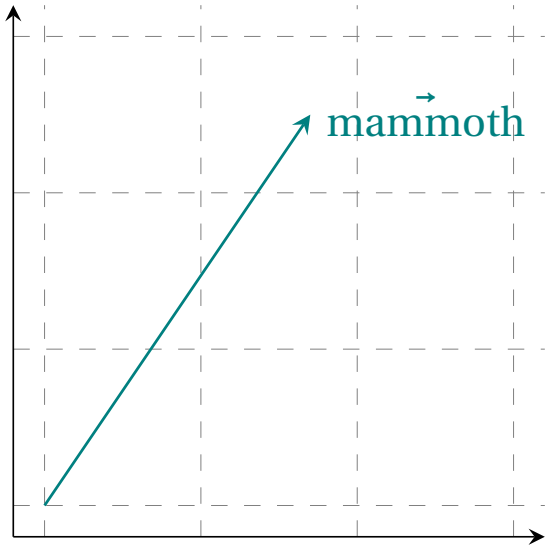


Definition Modeling

Any of a genus (Mammuthus) of extinct Pleistocene mammals of the elephant family distinguished from recent elephants by highly ridged molars, usually large size, very long tusks that curve upward, and well-developed body hair.

Definition Modeling

- Noraset et al. (2017): Well-trained distributional representations should capture enough semantics to derive definitions



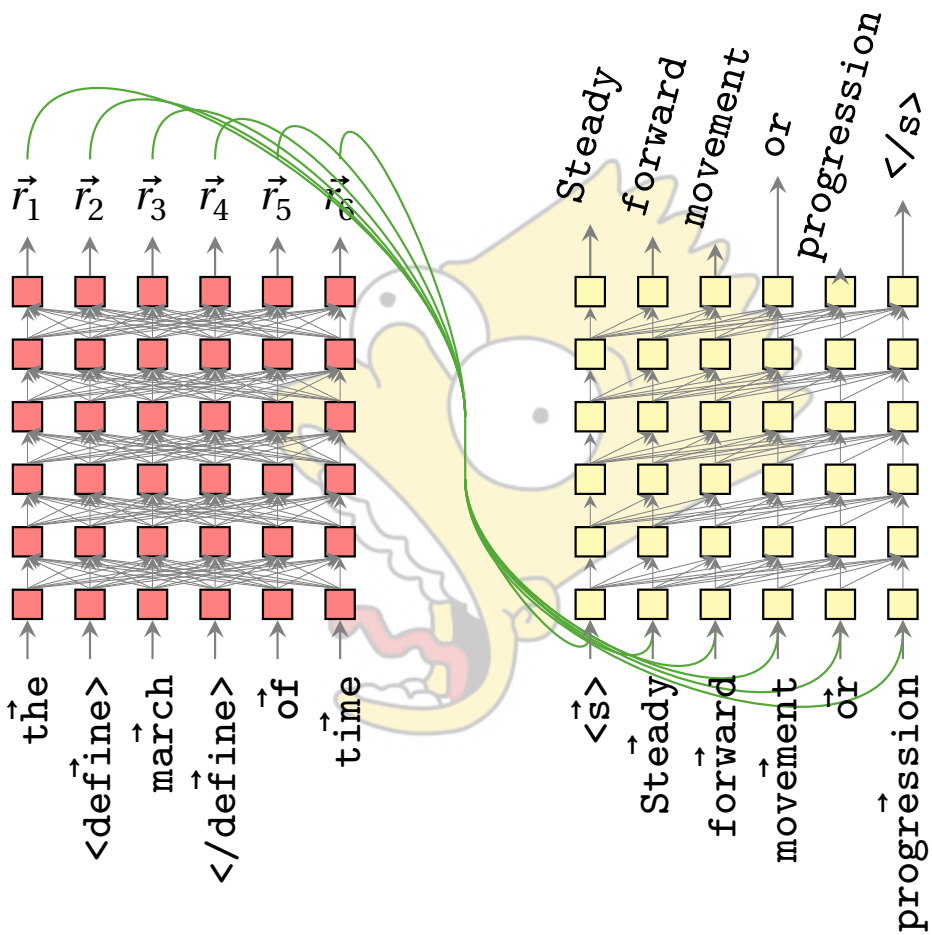
Definition Modeling

Any of a genus (Mammuthus) of extinct Pleistocene mammals of the elephant family distinguished from recent elephants by highly ridged molars, usually large size, very long tusks that curve upward, and well-developed body hair.

- Do related factors like polysemy and frequency impact the ability to generate definitions?

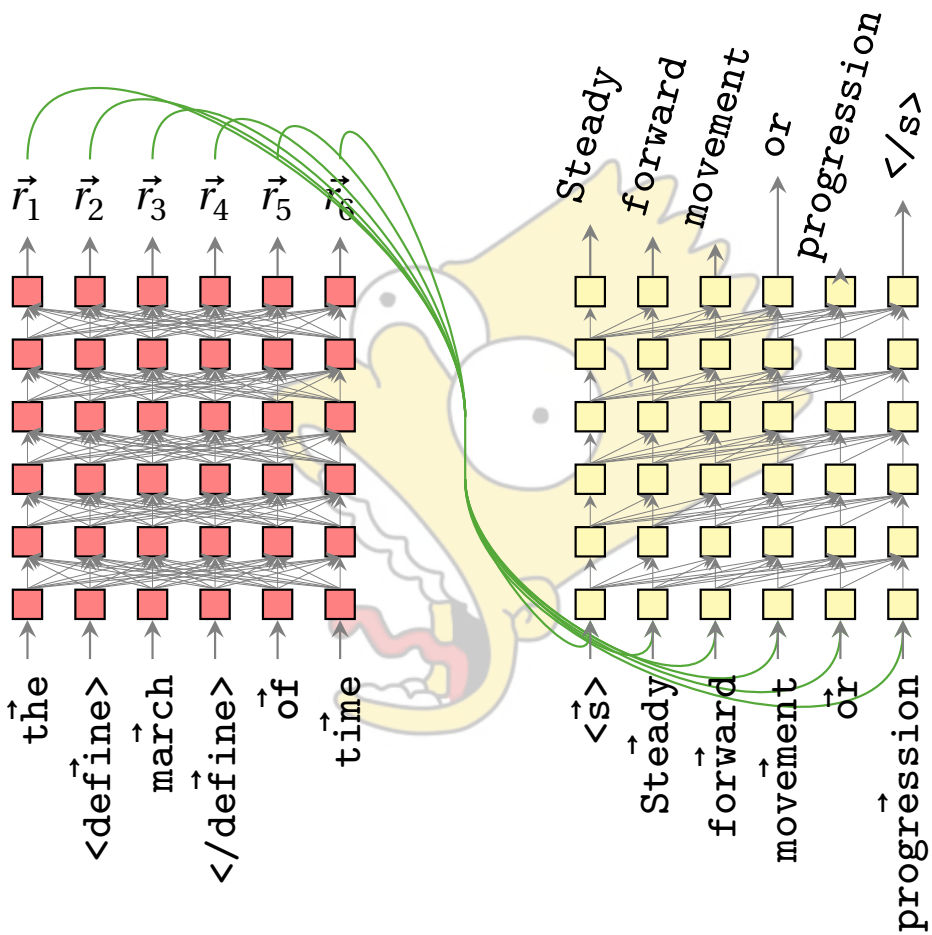
Setup

- Setup borrowed from Bevilacqua, Maru, and Navigli (2020)



Setup

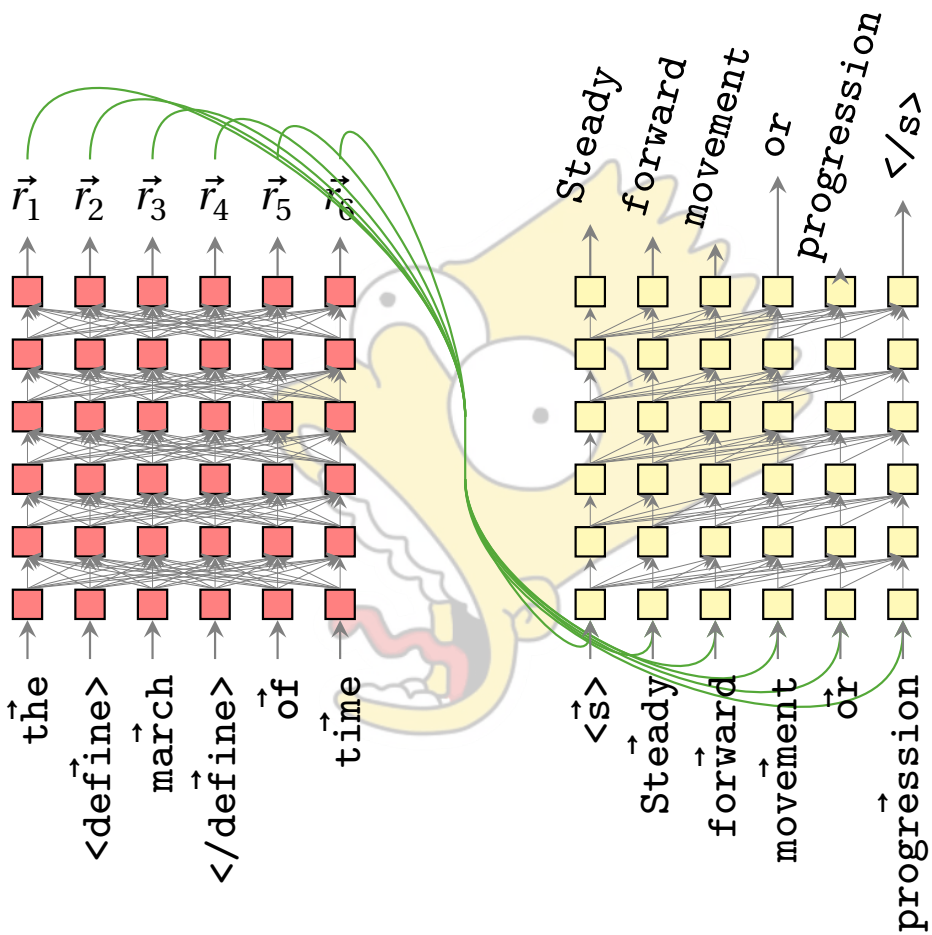
- Setup borrowed from Bevilacqua, Maru, and Navigli (2020)



- Training models with or without explicit polysemy (train set ablation)

Setup

- Setup borrowed from Bevilacqua, Maru, and Navigli (2020)



- Training models with or without explicit polysemy (train set ablation)
- Training models on frequent words, testing on rare words

Results

Polysemy	Val.	Test Splits		
		iid.	rare	0-freq
with	9.07	9.13	11.15	10.85
without	8.49	8.53	11.06	10.87

Average BLEU performances on held-out sets. Averaged on 5 runs; std. dev. $< \pm 0.001$ always.

Results

Polysemy	Val.	Test Splits		
		iid.	rare	0-freq
with	9.07	9.13	11.15	10.85
without	8.49	8.53	11.06	10.87

Average BLEU performances on held-out sets. Averaged on 5 runs; std. dev. $< \pm 0.001$ always.

- Performances are comparable across all setups

Results

Polysemy	Val.	Test Splits		
		iid.	rare	0-freq
with	9.07	9.13	11.15	10.85
without	8.49	8.53	11.06	10.87

Average BLEU performances on held-out sets. Averaged on 5 runs; std. dev. $< \pm 0.001$ always.

- Performances are comparable across all setups
- **Polysemy and frequency do not appear to play a major role**

What? Why?

- Manual annotation of a subset of 800 productions in four traits:
  - **Fluency (FL)**: if the output is free of grammar or commonsense mistakes
    ✓"`(architecture) A belfry`"
    ✗"`(intransitive) To go too far; to go too far.`"
  - **Factuality (FA)**: if the output contains only & all facts relevant to the target sense
    ✓"**flaglet:** `A small flag.`"
    ✗"**unsatined:** `Not stained.`"
  - **Pos-appropriateness (PA)**: if the generated gloss matches the headword's POS
    ✓"**unsubstantiate:** `(intransitive) To make unsubstantiated claims.`"
    ✗"**fried:** `(transitive) To cook (something) in a frying pan.`"
  - **Pattern-based (PB)**: if the generated gloss relies on morphological relatedness
    ✓"**clacky:** `Resembling or characteristic of clacking.`"
    ✗"**fare:** `(intransitive) To do well or poorly.`"

What? Why?

- Manual annotation of a subset of 800 productions in four traits:
  - **Fluency (FL)**: if the output is free of grammar or commonsense mistakes
    ✓"`(architecture) A belfry`"
    ✗"`(intransitive) To go too far; to go too far.`"
  - **Factuality (FA)**: if the output contains only & all facts relevant to the target sense
    ✓"**flaglet:** `A small flag.`"
    ✗"**unsatined:** `Not stained.`"
  - **Pos-appropriateness (PA)**: if the generated gloss matches the headword's POS
    ✓"**unsubstantiate:** `(intransitive) To make unsubstantiated claims.`"
    ✗"**fried:** `(transitive) To cook (something) in a frying pan.`"
  - **Pattern-based (PB)**: if the generated gloss relies on morphological relatedness
    ✓"**clacky:** `Resembling or characteristic of clacking.`"
    ✗"**fare:** `(intransitive) To do well or poorly.`"

- 36.5% of productions are PBs; 10% involve a straight copy of the headword

What? Why?

- Manual annotation of a subset of 800 productions in four traits:
  - **Fluency (FL)**: if the output is free of grammar or commonsense mistakes
    ✓"`(architecture) A belfry`"
    ✗"`(intransitive) To go too far; to go too far.`"
  - **Factuality (FA)**: if the output contains only & all facts relevant to the target sense
    ✓"**flaglet:** `A small flag.`"
    ✗"**unsatined:** `Not stained.`"
  - **Pos-appropriateness (PA)**: if the generated gloss matches the headword's POS
    ✓"**unsubstantiate:** `(intransitive) To make unsubstantiated claims.`"
    ✗"**fried:** `(transitive) To cook (something) in a frying pan.`"
  - **Pattern-based (PB)**: if the generated gloss relies on morphological relatedness
    ✓"**clacky:** `Resembling or characteristic of clacking.`"
    ✗"**fare:** `(intransitive) To do well or poorly.`"

- 36.5% of productions are PBs; 10% involve a straight copy of the headword
- Non-PB outputs have lower FL ($p < 3 \cdot 10^{-6}$, $f = 42.3\%$)

What? Why?

- Manual annotation of a subset of 800 productions in four traits:
  - **Fluency (FL)**: if the output is free of grammar or commonsense mistakes
    ✓"`(architecture) A belfry`"
    ✗"`(intransitive) To go too far; to go too far.`"
  - **Factuality (FA)**: if the output contains only & all facts relevant to the target sense
    ✓"**flaglet:** `A small flag.`"
    ✗"**unsatined:** `Not stained.`"
  - **Pos-appropriateness (PA)**: if the generated gloss matches the headword's POS
    ✓"**unsubstantiate:** `(intransitive) To make unsubstantiated claims.`"
    ✗"**fried:** `(transitive) To cook (something) in a frying pan.`"
  - **Pattern-based (PB)**: if the generated gloss relies on morphological relatedness
    ✓"**clacky:** `Resembling or characteristic of clacking.`"
    ✗"**fare:** `(intransitive) To do well or poorly.`"

- 36.5% of productions are PBs; 10% involve a straight copy of the headword
- Non-PB outputs have lower FL ($p < 3 \cdot 10^{-6}$, $f = 42.3\%$)
- Non-PB outputs have lower FA ($p < 2 \cdot 10^{-9}$, $f = 37.7\%$)

What? Why?

- Manual annotation of a subset of 800 productions in four traits:
  - **Fluency (FL)**: if the output is free of grammar or commonsense mistakes
    ✓"`(architecture) A belfry`"
    ✗"`(intransitive) To go too far; to go too far.`"
  - **Factuality (FA)**: if the output contains only & all facts relevant to the target sense
    ✓"**flaglet:** `A small flag.`"
    ✗"**unsatined:** `Not stained.`"
  - **Pos-appropriateness (PA)**: if the generated gloss matches the headword's POS
    ✓"**unsubstantiate:** `(intransitive) To make unsubstantiated claims.`"
    ✗"**fried:** `(transitive) To cook (something) in a frying pan.`"
  - **Pattern-based (PB)**: if the generated gloss relies on morphological relatedness
    ✓"**clacky:** `Resembling or characteristic of clacking.`"
    ✗"**fare:** `(intransitive) To do well or poorly.`"

- 36.5% of productions are PBs; 10% involve a straight copy of the headword
- Non-PB outputs have lower FL ($p < 3 \cdot 10^{-6}$, $f = 42.3\%$)
- Non-PB outputs have lower FA ($p < 2 \cdot 10^{-9}$, $f = 37.7\%$)
- PB and non-PB outputs have similar BLEU scores ($p = 0.262$)

What? Why?

- Manual annotation of a subset of 800 productions in four traits:
  - **Fluency (FL)**: if the output is free of grammar or commonsense mistakes
    ✓"`(architecture) A belfry`"
    ✗"`(intransitive) To go too far; to go too far.`"
  - **Factuality (FA)**: if the output contains only & all facts relevant to the target sense
    ✓"**flaglet:** `A small flag.`"
    ✗"**unsatined:** `Not stained.`"
  - **Pos-appropriateness (PA)**: if the generated gloss matches the headword's POS
    ✓"**unsubstantiate:** `(intransitive) To make unsubstantiated claims.`"
    ✗"**fried:** `(transitive) To cook (something) in a frying pan.`"
  - **Pattern-based (PB)**: if the generated gloss relies on morphological relatedness
    ✓"**clacky:** `Resembling or characteristic of clacking.`"
    ✗"**fare:** `(intransitive) To do well or poorly.`"

- 36.5% of productions are PBs; 10% involve a straight copy of the headword
- Non-PB outputs have lower FL ($p < 3 \cdot 10^{-6}$, $f = 42.3\%$)
- Non-PB outputs have lower FA ($p < 2 \cdot 10^{-9}$, $f = 37.7\%$)
- PB and non-PB outputs have similar BLEU scores ($p = 0.262$)

- **Valid generated definitions often entail relying on morphological relatedness**

In short

- **Some semantic tasks can be (partially) solved without semantics**

Outline

1. Segonne and Mickus (2023)

2. Mickus and Copot (In prep.)

3. Mickus and Bernard (2023)

Stranger than Paradigms
Word Embedding Benchmarks Don't Align
With Morphology

Timothee Mickus[1] *and Maria Copot*[2]
[1] University of Helsinki
[2] LLF

So, is it morphology then?

- Going back to our definition:

  $$\Pr(\text{word} \mid \text{context})$$

  distributional models are models of the lexicon

So, is it morphology then?

- Going back to our definition:

  $$\Pr(\text{word} \mid \text{context})$$

  distributional models are models of the lexicon
- To what extent do they model morphological relations?

CBOW & Negative sampling crash course

- **CBOW**: predict a word given its context

CBOW & Negative sampling crash course

- **CBOW**: predict a word given its context
- context is modelled as a bag of words

CBOW & Negative sampling crash course

- **CBOW**: predict a word given its context
- context is modelled as a bag of words
- inefficient to train due to the softmax over the vocabulary

CBOW & Negative sampling crash course

- **CBOW**: predict a word given its context
- context is modelled as a bag of words
- inefficient to train due to the softmax over the vocabulary

- **Negative sampling**: replace softmax by binary classification task (attested or not)

CBOW & Negative sampling crash course

- **CBOW**: predict a word given its context
- context is modelled as a bag of words
- inefficient to train due to the softmax over the vocabulary

- **Negative sampling**: replace softmax by binary classification task (attested or not)
- negative examples are constructed by randomly picking words for the same context

CBOW & Negative sampling crash course

- **CBOW**: predict a word given its context
- context is modelled as a bag of words
- inefficient to train due to the softmax over the vocabulary

- **Negative sampling**: replace softmax by binary classification task (attested or not)
- negative examples are constructed by randomly picking words for the same context
- probability of sampling as negative:

$$q(W) \propto p(w)^{\alpha}$$

with $\alpha = 0.75$

It's grid search time

Tasks:

It's grid search time

Tasks:

- Simlex-999 (Barzegar et al., 2018)
- FEEL (Abdaoui et al., 2017)
- GATS (Grave et al., 2018)
- POS tagging using OMW (Bond and Paik, 2012)

It's grid search time

Tasks:

- Simlex-999 (Barzegar et al., 2018)
- FEEL (Abdaoui et al., 2017)
- GATS (Grave et al., 2018)
- POS tagging using OMW (Bond and Paik, 2012)

- One-cell and two-cell clustering scores for inflection (SCC, PCC)
- One-cell and two-cell prediction scores for inflection (SCP, PCP)
- Two-cell clustering scores for derivation, based on process semantics or form (DerCS, DerCF)
- Two-cell prediction scores for derivation, based on process semantics or form (DerPS, DerPF)

It's grid search time

Grid search over CBOW hyper-parameters:

Tasks:

- Simlex-999 (Barzegar et al., 2018)
- FEEL (Abdaoui et al., 2017)
- GATS (Grave et al., 2018)
- POS tagging using OMW (Bond and Paik, 2012)

- One-cell and two-cell clustering scores for inflection (SCC, PCC)
- One-cell and two-cell prediction scores for inflection (SCP, PCP)
- Two-cell clustering scores for derivation, based on process semantics or form (DerCS, DerCF)
- Two-cell prediction scores for derivation, based on process semantics or form (DerPS, DerPF)

It's grid search time

Tasks:

- Simlex-999 (Barzegar et al., 2018)
- FEEL (Abdaoui et al., 2017)
- GATS (Grave et al., 2018)
- POS tagging using OMW (Bond and Paik, 2012)

- One-cell and two-cell clustering scores for inflection (SCC, PCC)
- One-cell and two-cell prediction scores for inflection (SCP, PCP)
- Two-cell clustering scores for derivation, based on process semantics or form (DerCS, DerCF)
- Two-cell prediction scores for derivation, based on process semantics or form (DerPS, DerPF)

Grid search over CBOW hyper-parameters:

1. window size

$$w \in \{5, 10, 15, 20, 25\}$$

It's grid search time

Tasks:

- Simlex-999 (Barzegar et al., 2018)
- FEEL (Abdaoui et al., 2017)
- GATS (Grave et al., 2018)
- POS tagging using OMW (Bond and Paik, 2012)

- One-cell and two-cell clustering scores for inflection (SCC, PCC)
- One-cell and two-cell prediction scores for inflection (SCP, PCP)
- Two-cell clustering scores for derivation, based on process semantics or form (DerCS, DerCF)
- Two-cell prediction scores for derivation, based on process semantics or form (DerPS, DerPF)

Grid search over CBOW hyper-parameters:

1. window size

$$w \in \{5, 10, 15, 20, 25\}$$

2. number of negative examples per positive example

$$\#N \in \{5, 10, 15, 20, 25\}$$

It's grid search time

Tasks:

- Simlex-999 (Barzegar et al., 2018)
- FEEL (Abdaoui et al., 2017)
- GATS (Grave et al., 2018)
- POS tagging using OMW (Bond and Paik, 2012)

- One-cell and two-cell clustering scores for inflection (SCC, PCC)
- One-cell and two-cell prediction scores for inflection (SCP, PCP)
- Two-cell clustering scores for derivation, based on process semantics or form (DerCS, DerCF)
- Two-cell prediction scores for derivation, based on process semantics or form (DerPS, DerPF)

Grid search over CBOW hyper-parameters:

1. window size

$$w \in \{5, 10, 15, 20, 25\}$$

2. number of negative examples per positive example

$$\#N \in \{5, 10, 15, 20, 25\}$$

3. number of epochs

$$e \in \{1, 3, 5\}$$

It's grid search time

Tasks:

- Simlex-999 (Barzegar et al., 2018)
- FEEL (Abdaoui et al., 2017)
- GATS (Grave et al., 2018)
- POS tagging using OMW (Bond and Paik, 2012)

- One-cell and two-cell clustering scores for inflection (SCC, PCC)
- One-cell and two-cell prediction scores for inflection (SCP, PCP)
- Two-cell clustering scores for derivation, based on process semantics or form (DerCS, DerCF)
- Two-cell prediction scores for derivation, based on process semantics or form (DerPS, DerPF)

Grid search over CBOW hyper-parameters:

1. window size

$$w \in \{5, 10, 15, 20, 25\}$$

2. number of negative examples per positive example

$$\#N \in \{5, 10, 15, 20, 25\}$$

3. number of epochs

$$e \in \{1, 3, 5\}$$

4. negative sampling distribution exponent

$$\alpha \in \{0.2, 0.6, 1.0, 1.4\}$$

It's grid search time

Tasks:

- Simlex-999 (Barzegar et al., 2018)
- FEEL (Abdaoui et al., 2017)
- GATS (Grave et al., 2018)
- POS tagging using OMW (Bond and Paik, 2012)

- One-cell and two-cell clustering scores for inflection (SCC, PCC)
- One-cell and two-cell prediction scores for inflection (SCP, PCP)
- Two-cell clustering scores for derivation, based on process semantics or form (DerCS, DerCF)
- Two-cell prediction scores for derivation, based on process semantics or form (DerPS, DerPF)

Grid search over CBOW hyper-parameters:

1. window size

$$w \in \{5, 10, 15, 20, 25\}$$

2. number of negative examples per positive example

$$\#N \in \{5, 10, 15, 20, 25\}$$

3. number of epochs

$$e \in \{1, 3, 5\}$$

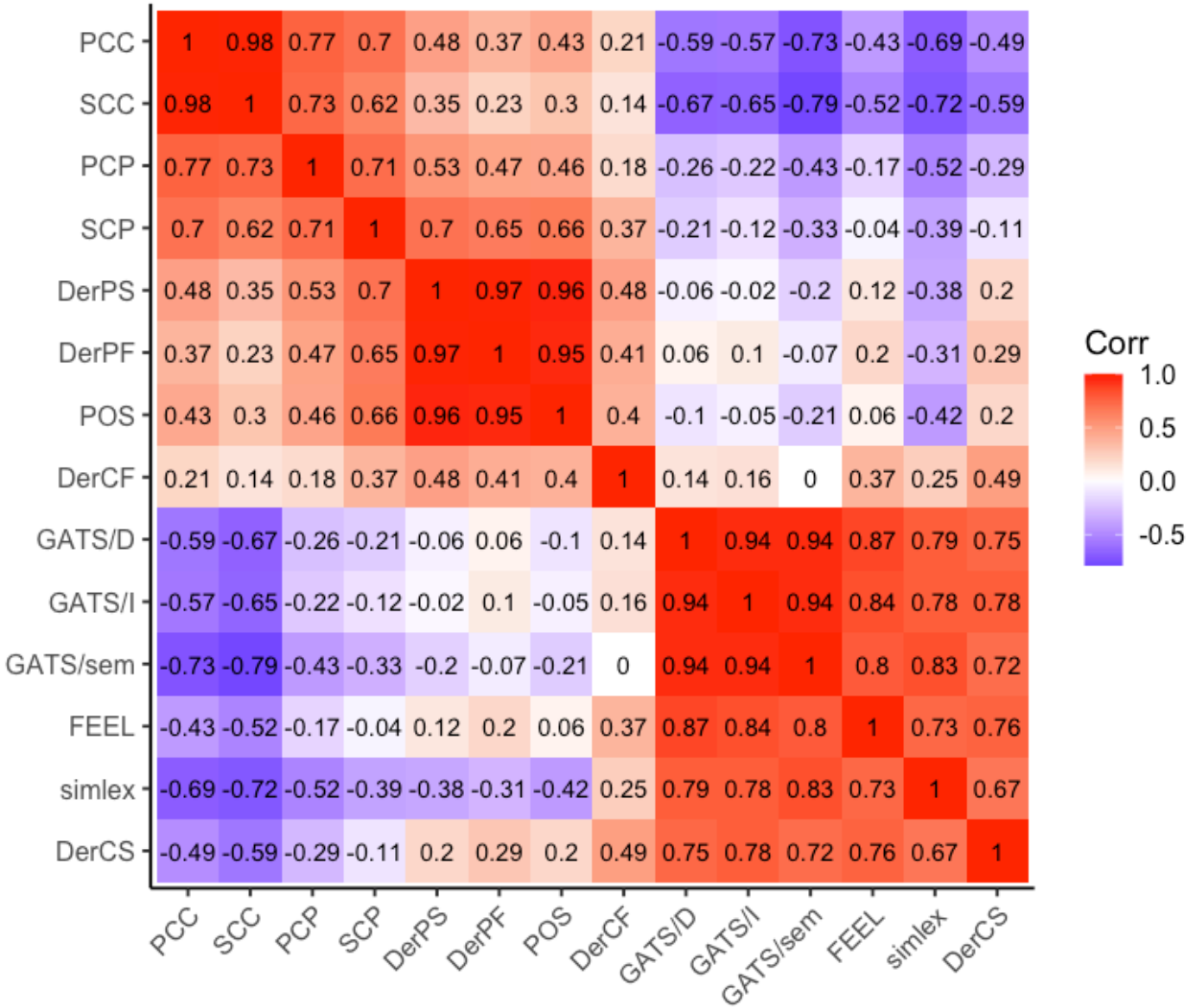
4. negative sampling distribution exponent

$$\alpha \in \{0.2, 0.6, 1.0, 1.4\}$$

5. dynamic uniform sampling of window size

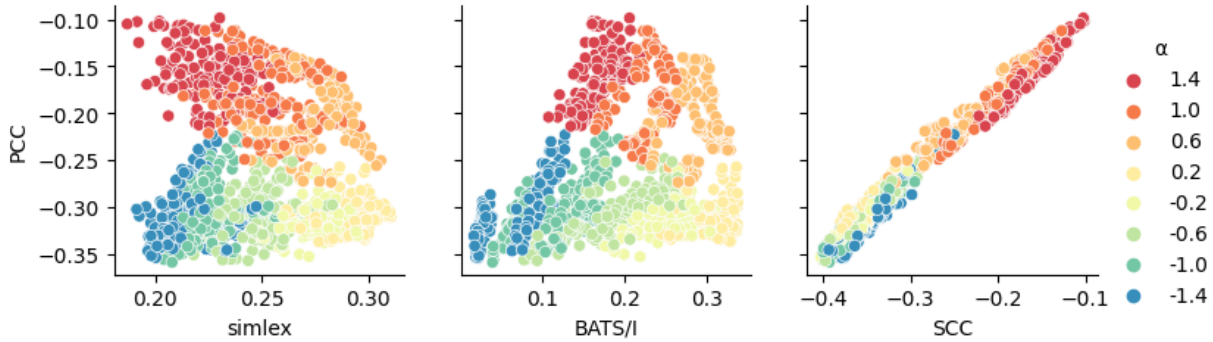
$$s \in \{\top, \bot\}$$

Results


PCC 1 0.98 0.77 0.7 0.48 0.37 0.43 0.21 -0.59 -0.57 -0.73 -0.43 -0.69 -0.49
SCC 0.98 1 0.73 0.62 0.35 0.23 0.3 0.14 -0.67 -0.65 -0.79 -0.52 -0.72 -0.59
PCP 0.77 0.73 1 0.71 0.53 0.47 0.46 0.18 -0.26 -0.22 -0.43 -0.17 -0.52 -0.29
SCP 0.7 0.62 0.71 1 0.7 0.65 0.66 0.37 -0.21 -0.12 -0.33 -0.04 -0.39 -0.11
DerPS 0.48 0.35 0.53 0.7 1 0.97 0.96 0.48 -0.06 -0.02 -0.2 0.12 -0.38 0.2
DerPF 0.37 0.23 0.47 0.65 0.97 1 0.95 0.41 0.06 0.1 -0.07 0.2 -0.31 0.29
POS 0.43 0.3 0.46 0.66 0.96 0.95 1 0.4 -0.1 -0.05 -0.21 0.06 -0.42 0.2
DerCF 0.21 0.14 0.18 0.37 0.48 0.41 0.4 1 0.14 0.16 0 0.37 0.25 0.49
GATS/D -0.59 -0.67 -0.26 -0.21 -0.06 0.06 -0.1 0.14 1 0.94 0.94 0.87 0.79 0.75
GATS/I -0.57 -0.65 -0.22 -0.12 -0.02 0.1 -0.05 0.16 0.94 1 0.94 0.84 0.78 0.78
GATS/sem -0.73 -0.79 -0.43 -0.33 -0.2 -0.07 -0.21 0 0.94 0.94 1 0.8 0.83 0.72
FEEL -0.43 -0.52 -0.17 -0.04 0.12 0.2 0.06 0.37 0.87 0.84 0.8 1 0.73 0.76
simlex -0.69 -0.72 -0.52 -0.39 -0.38 -0.31 -0.42 0.25 0.79 0.78 0.83 0.73 1 0.67
DerCS -0.49 -0.59 -0.29 -0.11 0.2 0.29 0.2 0.49 0.75 0.78 0.72 0.76 0.67 1
PCC SCC PCP SCP DerPS DerPF POS DerCF GATS/D GATS/I GATS/sem FEEL simlex DerCS
Corr
1.0
0.5
0.0
-0.5

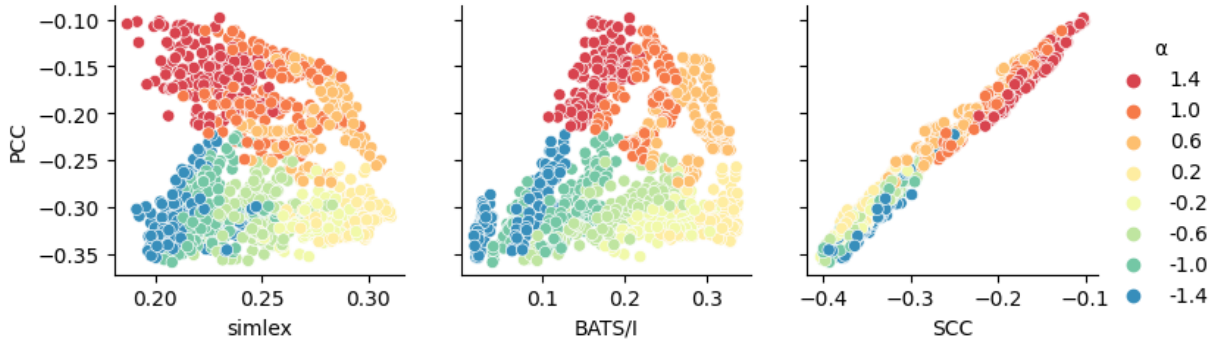
Why so different?

- The distribution is determined by the negative sampling hyperparameter:



Why so different?

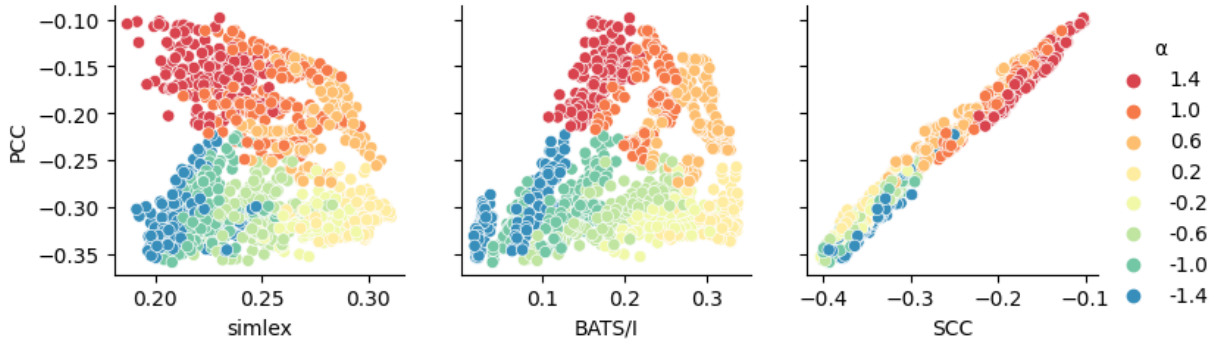
- The distribution is determined by the negative sampling hyperparameter:



Contexts constrain words in (at least) two different manners

Why so different?

- The distribution is determined by the negative sampling hyperparameter:



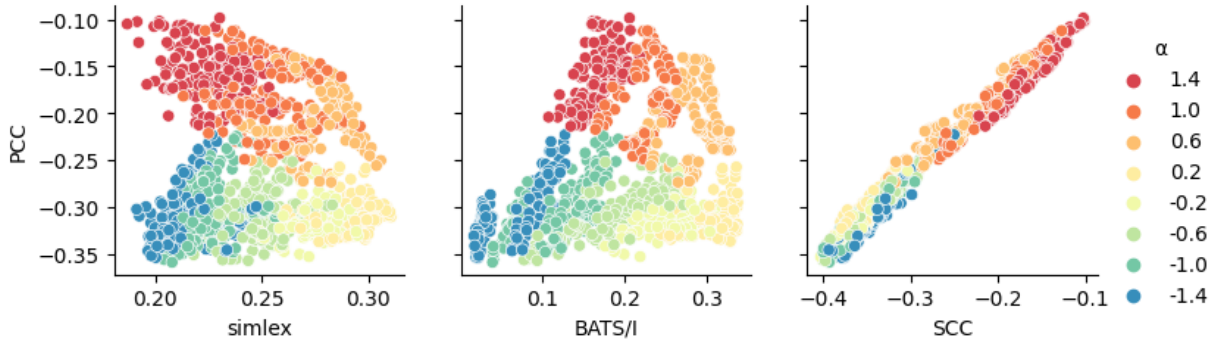
Contexts constrain words in (at least) two different manners

1. through lexical semantic requirements, e.g.,
   You know, this is the way we eat in _____.

2. through morphosyntactic dependencies, e.g.,
   I think this game is really _____.

Why so different?

- The distribution is determined by the negative sampling hyperparameter:



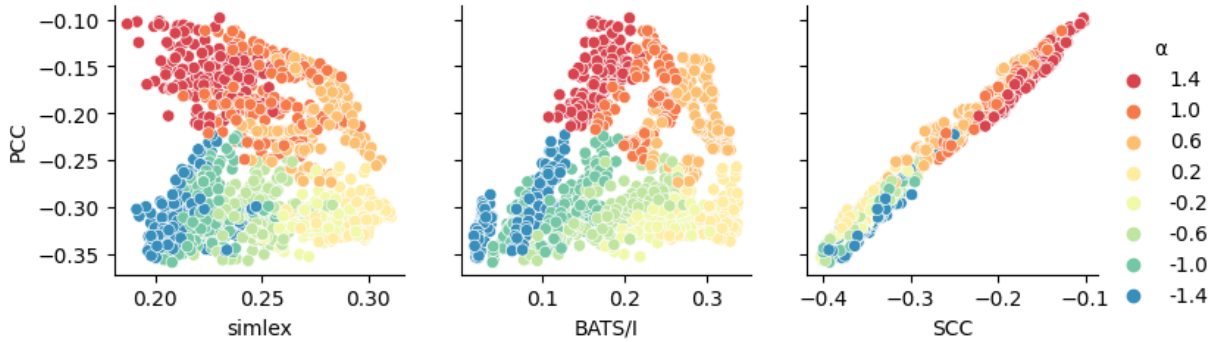
Contexts constrain words in (at least) two different manners

1. through lexical semantic requirements, e.g.,
   You know, this is the way we eat in _____.
   - Words that are frequent occur in many contexts

2. through morphosyntactic dependencies, e.g.,
   I think this game is really _____.

Why so different?

- The distribution is determined by the negative sampling hyperparameter:



Contexts constrain words in (at least) two different manners

1. through lexical semantic requirements, e.g.,

   You know, this is the way we eat in _____.

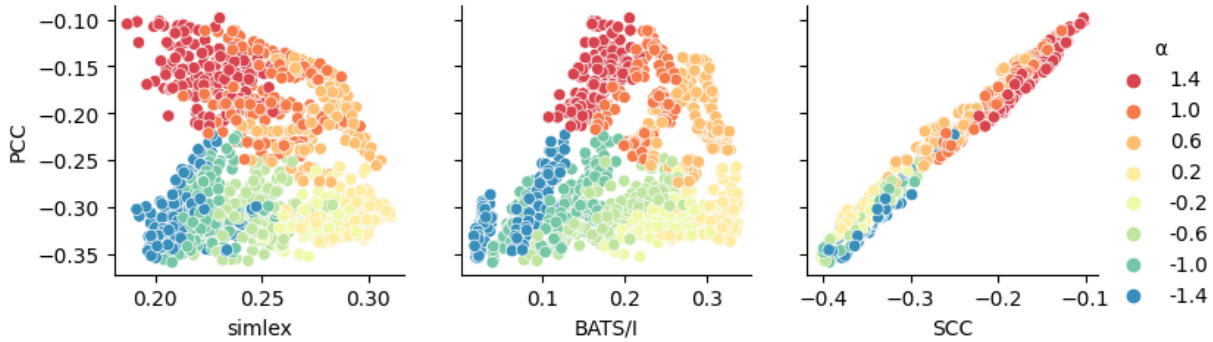
   - Words that are frequent occur in many contexts
   - They are not useful for capturing the specific semantics of a given context

2. through morphosyntactic dependencies, e.g.,

   I think this game is really _____.

Why so different?

- The distribution is determined by the negative sampling hyperparameter:



Contexts constrain words in (at least) two different manners

1. through lexical semantic requirements, e.g.,

   You know, this is the way we eat in _____.

   - Words that are frequent occur in many contexts
   - They are not useful for capturing the specific semantics of a given context

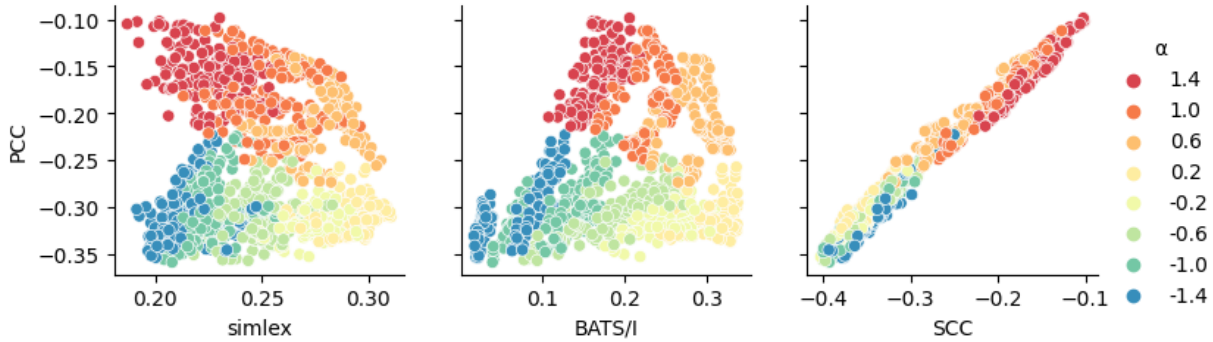
2. through morphosyntactic dependencies, e.g.,

   I think this game is really _____.

   - Frequency and morphological regularity are inversely correlated (Wu, Cotterell, and O'Donnell, 2019)

Why so different?

- The distribution is determined by the negative sampling hyperparameter:



Contexts constrain words in (at least) two different manners

1. through lexical semantic requirements, e.g.,

   You know, this is the way we eat in _____.

   - Words that are frequent occur in many contexts
   - They are not useful for capturing the specific semantics of a given context

2. through morphosyntactic dependencies, e.g.,

   I think this game is really _____.

   - Frequency and morphological regularity are inversely correlated (Wu, Cotterell, and O'Donnell, 2019)
   - To model morphology, one should focus on frequent (= irregular) words

In short

- **Not every distributional constraint is semantics**

Outline

1. Segonne and Mickus (2023)

2. Mickus and Copot (In prep.)

3. Mickus and Bernard (2023)

Distributional, yes—but *semantics*?
Comparing distributional representations, semantics and syntax

Timothee Mickus
University of Helsinki, Finland

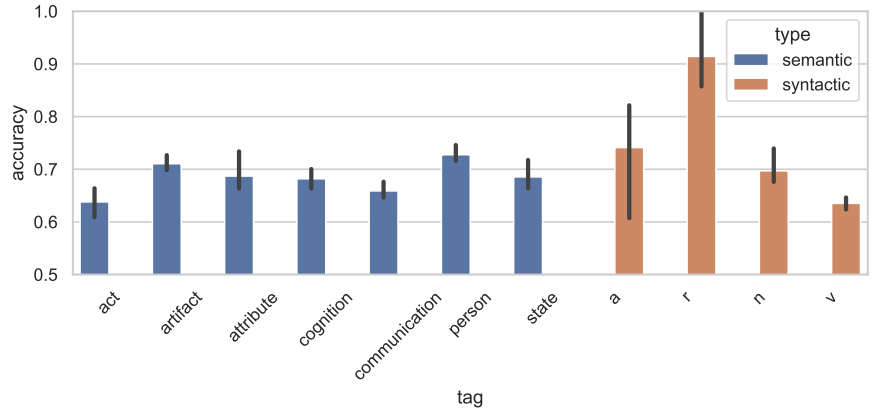
Timothée Bernard
LLF, Université Paris Cité, France

timothee.lastname@{helsinki.fi,u-paris.fr}

What about syntax?

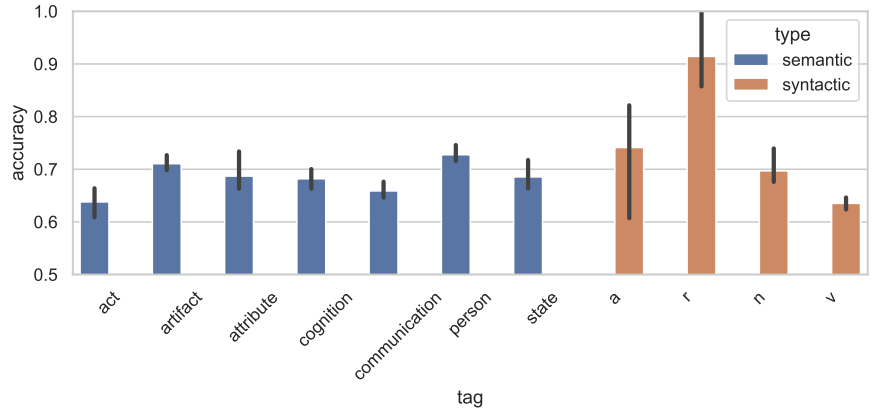
- Simple tagging experiment using decision trees, comparing POS tags and supersense tags

What about syntax?



- Simple tagging experiment using decision trees, comparing POS tags and supersense tags

What about syntax?



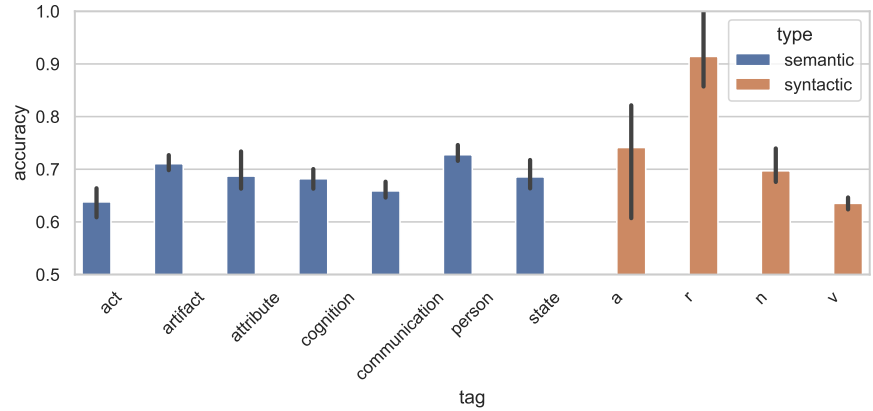
- Simple tagging experiment using decision trees, comparing POS tags and supersense tags

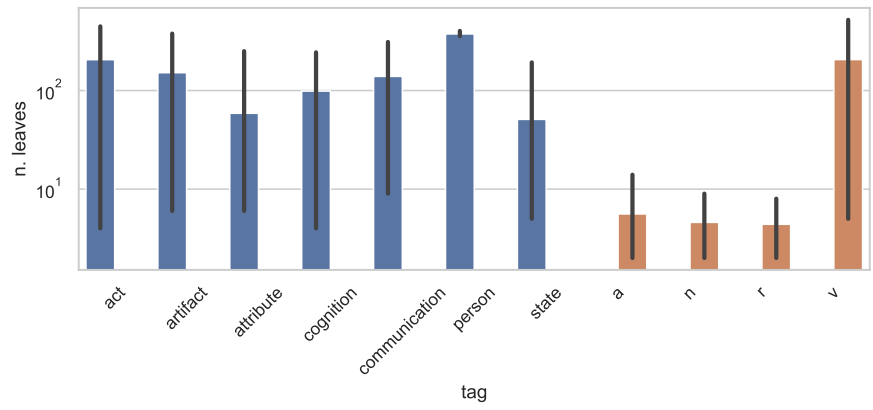
- **Syntax generally yields classifier trees that are more accurate**

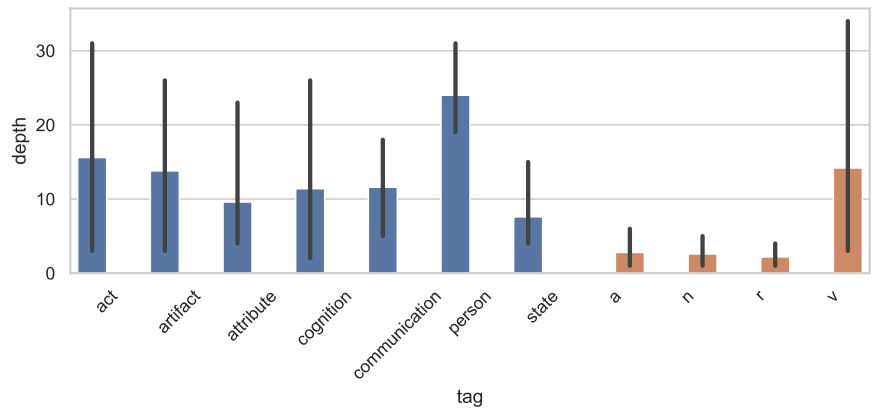
What about syntax?

- Simple tagging experiment using decision trees, comparing POS tags and supersense tags

- **Syntax generally yields classifier trees that are more accurate**

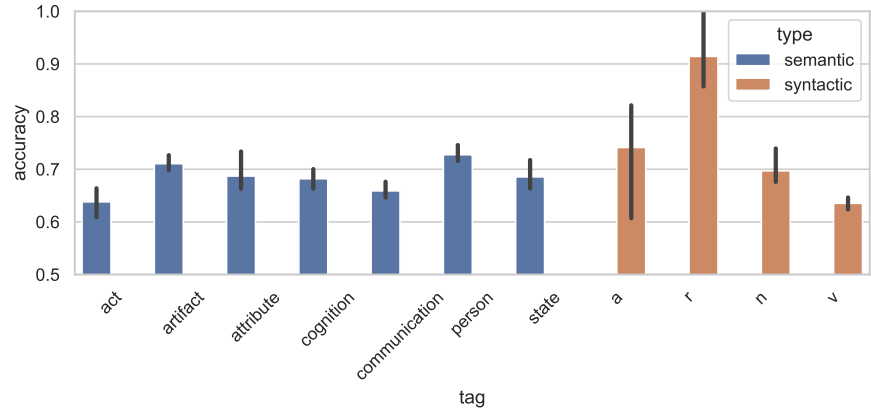


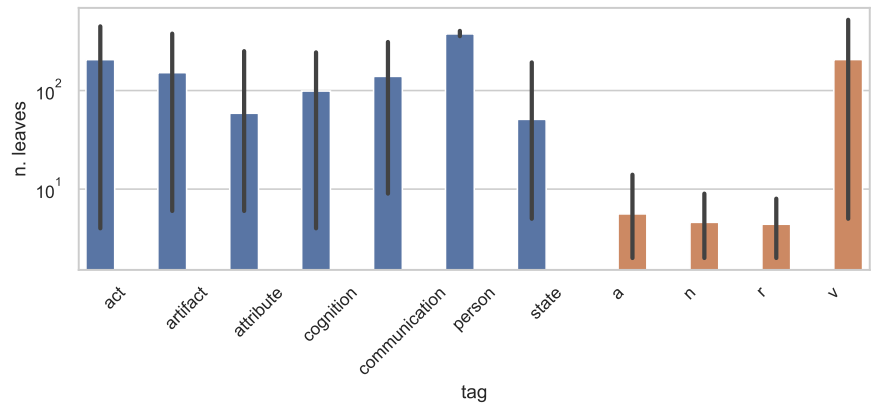


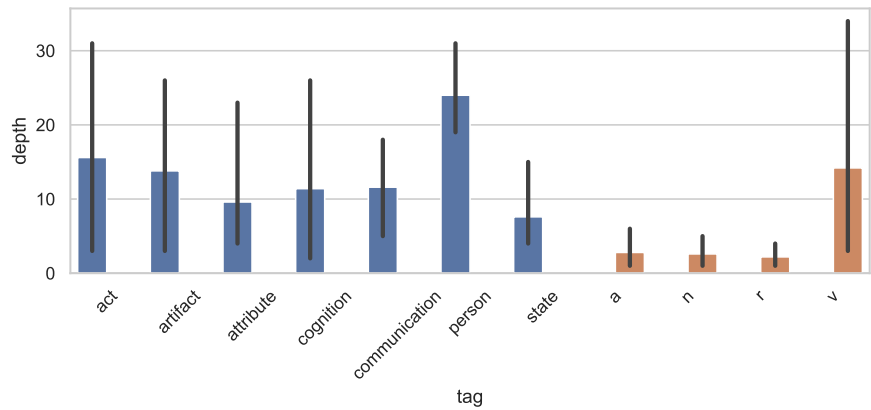


What about syntax?

- Simple tagging experiment using decision trees, comparing POS tags and supersense tags
- **Syntax generally yields classifier trees that are more accurate**
- **Syntax generally yields classifier trees that are structurally simpler**







What about sentence-level syntax?

Should we factor sentence-level structure?

What about sentence-level syntax?

Should we factor sentence-level structure?

- Given a directed labeled graph G with edges $\langle n_{\text{in}}, n_{\text{out}}, \ell \rangle$ and a length k, get the multiset of (possibly indirect) dependencies of length k:

$$\mathrm{v}_k(G) = \left\{ (\ell_1, \ldots, \ell_k) \mid \exists n_1, \ldots, n_{k+1}, \langle n_1, n_2, \ell_1 \rangle, \ldots, \langle n_k, n_{k+1}, \ell_k \rangle \in G \right\}$$

What about sentence-level syntax?

Should we factor sentence-level structure?

- Given a directed labeled graph G with edges $\langle n_{\text{in}}, n_{\text{out}}, \ell \rangle$ and a length k, get the multiset of (possibly indirect) dependencies of length k:

$$\mathrm{v}_k(G) = \left\{ (\ell_1, \ldots, \ell_k) \mid \exists n_1, \ldots, n_{k+1}, \langle n_1, n_2, \ell_1 \rangle, \ldots, \langle n_k, n_{k+1}, \ell_k \rangle \in G \right\}$$

- Combine all such dependencies up to some maximum length $\hat{k}$ as

$$\mathrm{v}_{\leq \hat{k}}(G) = \bigcup_{k=1}^{\hat{k}} \mathrm{v}_k(G)$$

What about sentence-level syntax?

Should we factor sentence-level structure?

- Given a directed labeled graph G with edges $\langle n_{\text{in}}, n_{\text{out}}, \ell \rangle$ and a length k, get the multiset of (possibly indirect) dependencies of length k:

$$\mathrm{v}_k(G) = \left\{ (\ell_1, \ldots, \ell_k) \mid \exists n_1, \ldots, n_{k+1}, \langle n_1, n_2, \ell_1 \rangle, \ldots, \langle n_k, n_{k+1}, \ell_k \rangle \in G \right\}$$

- Combine all such dependencies up to some maximum length $\hat{k}$ as

$$\mathrm{v}_{\leq \hat{k}}(G) = \bigcup_{k=1}^{\hat{k}} \mathrm{v}_k(G)$$

- Compare two graphs through the combined dependencies multisets:

$$\text{similarity}(G_a, G_b) = \frac{\left| \mathrm{v}_{\leq \hat{k}}(G_a) \cap \mathrm{v}_{\leq \hat{k}}(G_b) \right|}{\left| \mathrm{v}_{\leq \hat{k}}(G_a) \cup \mathrm{v}_{\leq \hat{k}}(G_b) \right|}$$

What about sentence-level syntax?

Should we factor sentence-level structure?

- Given a directed labeled graph G with edges $\langle n_{\text{in}}, n_{\text{out}}, \ell \rangle$ and a length k, get the multiset of (possibly indirect) dependencies of length k:

$$\mathrm{v}_k(G) = \left\{ (\ell_1, \ldots, \ell_k) \mid \exists n_1, \ldots, n_{k+1}, \langle n_1, n_2, \ell_1 \rangle, \ldots, \langle n_k, n_{k+1}, \ell_k \rangle \in G \right\}$$

- Combine all such dependencies up to some maximum length $\hat{k}$ as

$$\mathrm{v}_{\leq \hat{k}}(G) = \bigcup_{k=1}^{\hat{k}} \mathrm{v}_k(G)$$

- Compare two graphs through the combined dependencies multisets:

$$\text{similarity}(G_a, G_b) = \frac{\left| \mathrm{v}_{\leq \hat{k}}(G_a) \cap \mathrm{v}_{\leq \hat{k}}(G_b) \right|}{\left| \mathrm{v}_{\leq \hat{k}}(G_a) \cup \mathrm{v}_{\leq \hat{k}}(G_b) \right|}$$

- equally applicable to syntactic trees and semantic DAGs

What about sentence-level syntax?

Should we factor sentence-level structure?

- Given a directed labeled graph G with edges $\langle n_{\text{in}}, n_{\text{out}}, \ell \rangle$ and a length k, get the multiset of (possibly indirect) dependencies of length k:

$$\mathrm{v}_k(G) = \left\{ (\ell_1, \ldots, \ell_k) \mid \exists n_1, \ldots, n_{k+1}, \langle n_1, n_2, \ell_1 \rangle, \ldots, \langle n_k, n_{k+1}, \ell_k \rangle \in G \right\}$$

- Combine all such dependencies up to some maximum length $\hat{k}$ as

$$\mathrm{v}_{\leq \hat{k}}(G) = \bigcup_{k=1}^{\hat{k}} \mathrm{v}_k(G)$$

- Compare two graphs through the combined dependencies multisets:

$$\text{similarity}(G_a, G_b) = \frac{\left| \mathrm{v}_{\leq \hat{k}}(G_a) \cap \mathrm{v}_{\leq \hat{k}}(G_b) \right|}{\left| \mathrm{v}_{\leq \hat{k}}(G_a) \cup \mathrm{v}_{\leq \hat{k}}(G_b) \right|}$$

- equally applicable to syntactic trees and semantic DAGs
- can be compared to distribution-based similarity, e.g., BertScore or WMD, using RSA

What about sentence-level syntax?

Should we factor sentence-level structure?

- Given a directed labeled graph G with edges $\langle n_{\text{in}}, n_{\text{out}}, \ell \rangle$ and a length k, get the multiset of (possibly indirect) dependencies of length k:

$$\text{v}_k(G) = \left\{ (\ell_1, \ldots, \ell_k) \mid \exists n_1, \ldots, n_{k+1}, \langle n_1, n_2, \ell_1 \rangle, \ldots, \langle n_k, n_{k+1}, \ell_k \rangle \in G \right\}$$

- Combine all such dependencies up to some maximum length $\hat{k}$ as

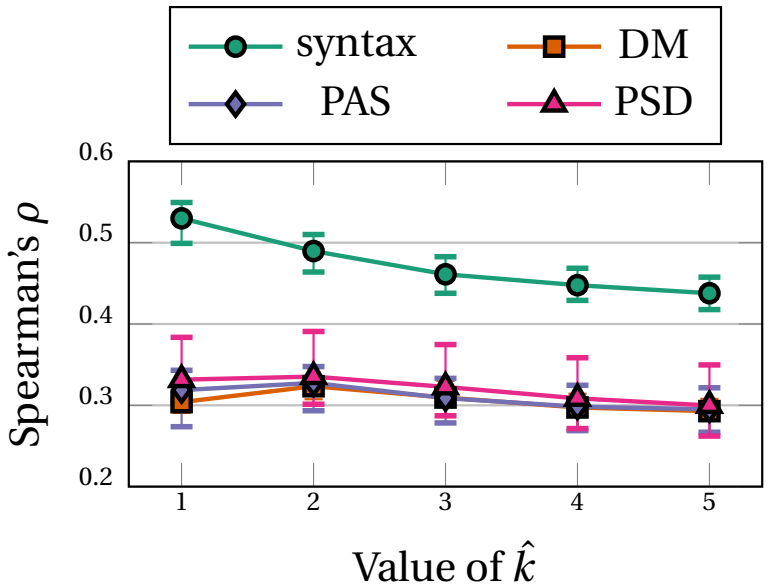
$$\text{v}_{\leq \hat{k}}(G) = \bigcup_{k=1}^{\hat{k}} \text{v}_k(G)$$

- Compare two graphs through the combined dependencies multisets:

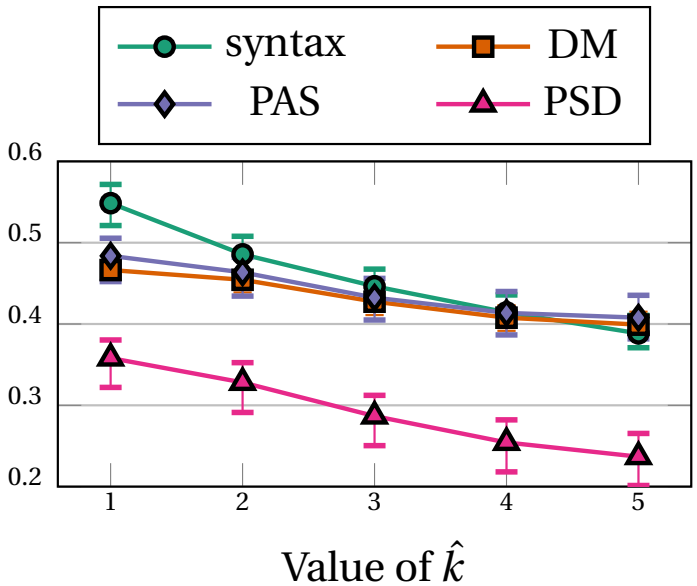
$$\text{similarity}(G_a, G_b) = \frac{\left| \text{v}_{\leq \hat{k}}(G_a) \cap \text{v}_{\leq \hat{k}}(G_b) \right|}{\left| \text{v}_{\leq \hat{k}}(G_a) \cup \text{v}_{\leq \hat{k}}(G_b) \right|}$$

- equally applicable to syntactic trees and semantic DAGs
- can be compared to distribution-based similarity, e.g., BertScore or WMD, using RSA
- using the data from SemEval 2015 shared-task 18 (Oepen et al., 2015)

Results

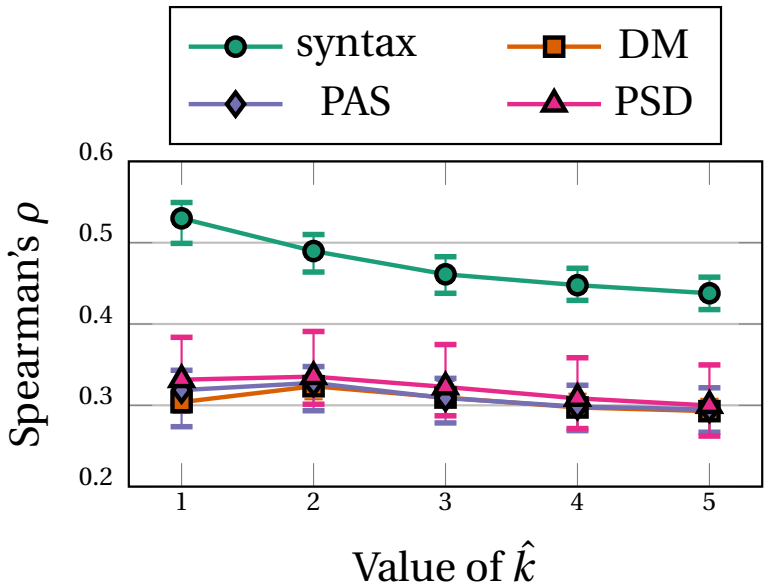


Using BertScore as distributional similarity

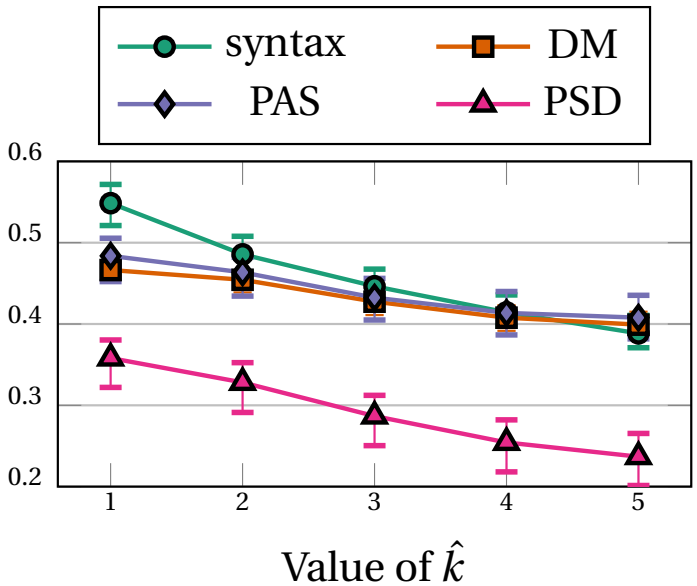


Using negative WMD between word2vec vectors as distributional similarity

Results



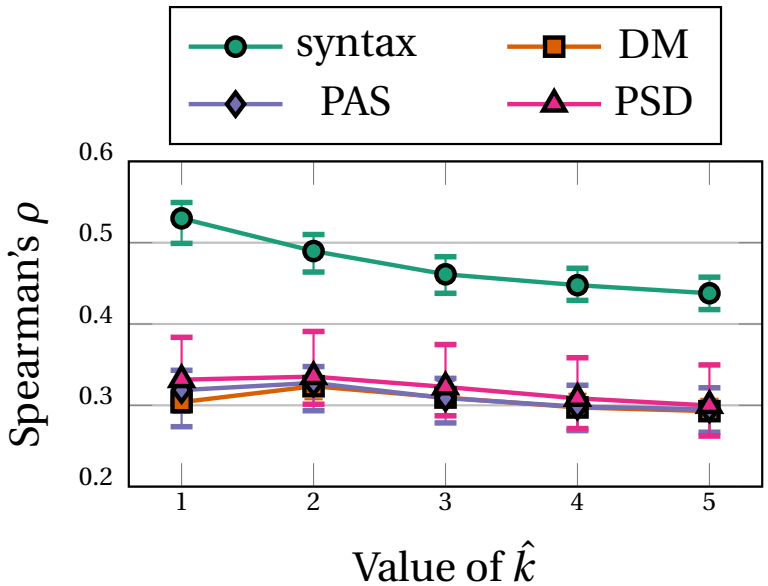
Using BertScore as distributional similarity



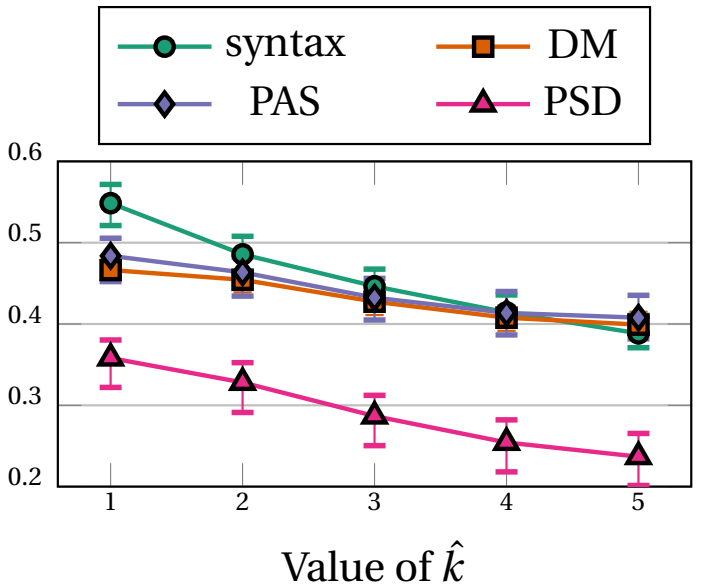
Using negative WMD between word2vec vectors as distributional similarity

- **In both cases, best results are achieved with syntax**

Results



Using BertScore as distributional similarity



Using negative WMD between word2vec vectors as distributional similarity

- **In both cases, best results are achieved with syntax**
- **Results deteriorate when factoring in more indirect dependencies**

In short

- **Off-the-shelf embeddings align more with (shallow) syntax than with semantics**

To recap

Do you really need semantics to model
Pr(word | context)**?**

To recap

Do you really need semantics to model Pr(word | context)?

We saw:

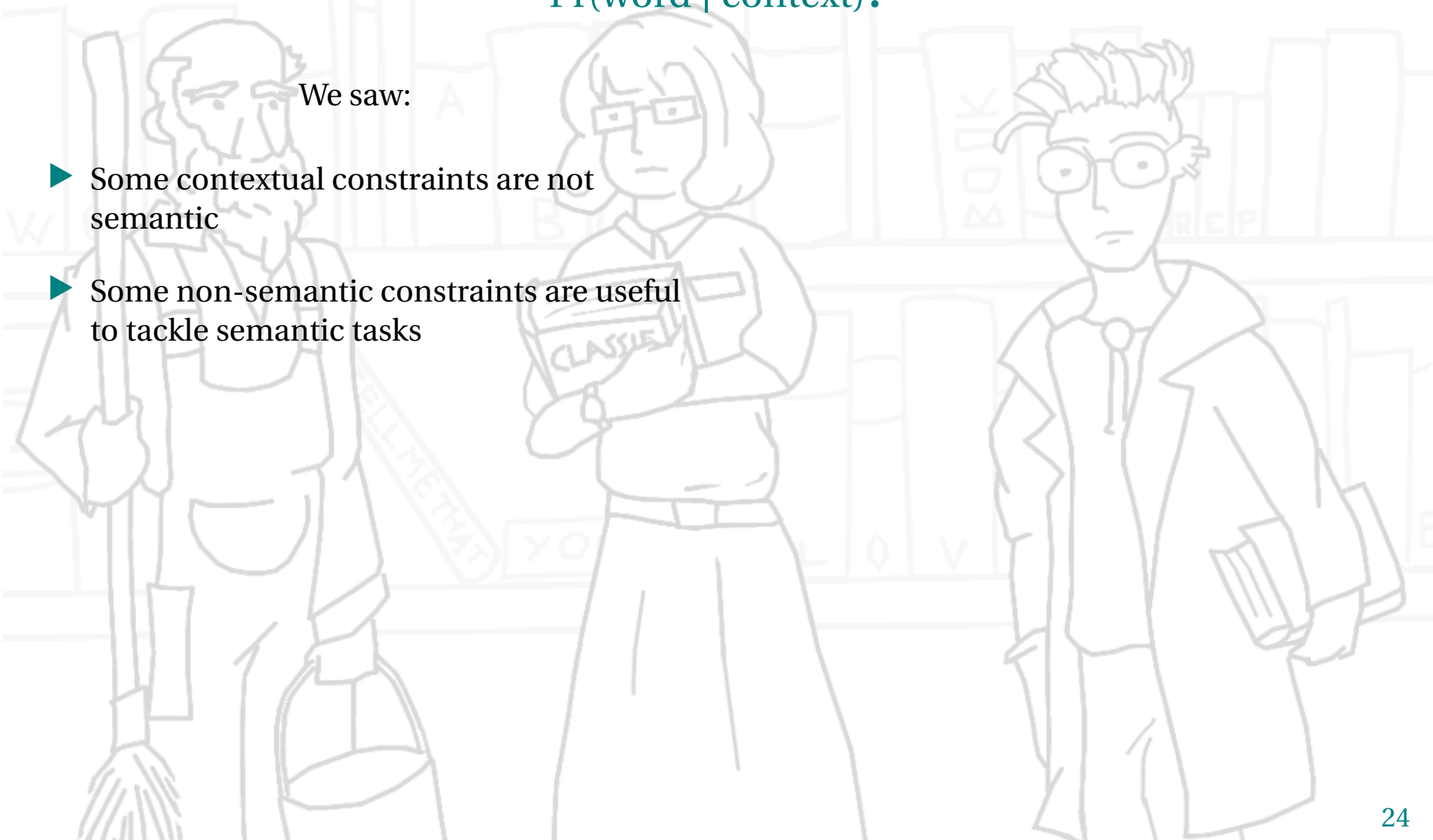
- Some contextual constraints are not semantic

To recap

Do you really need semantics to model
Pr(word | context)**?**

We saw:

- Some contextual constraints are not semantic
- Some non-semantic constraints are useful to tackle semantic tasks

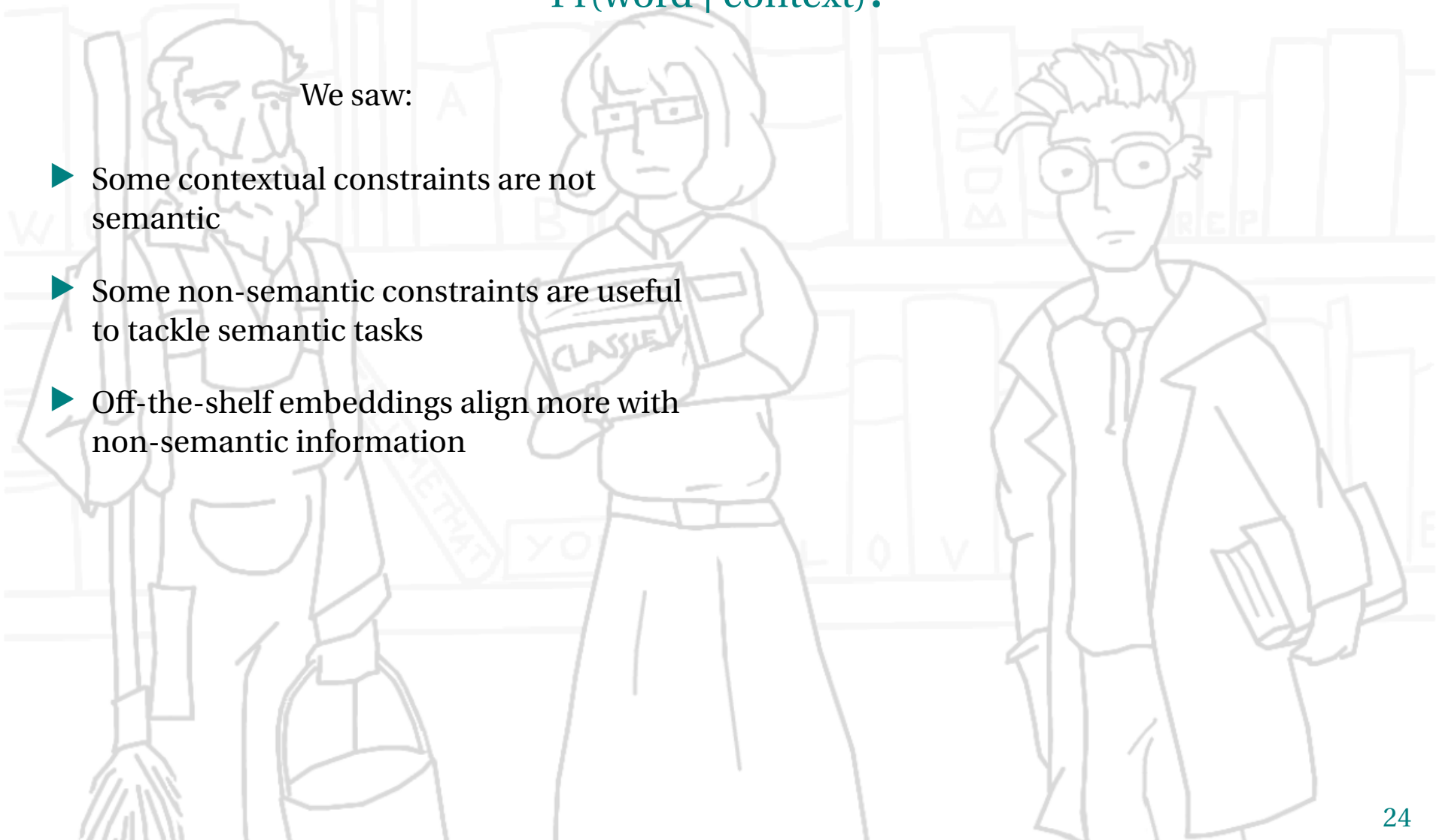


To recap

Do you really need semantics to model Pr(word | context)?

We saw:

- Some contextual constraints are not semantic
- Some non-semantic constraints are useful to tackle semantic tasks
- Off-the-shelf embeddings align more with non-semantic information



To recap

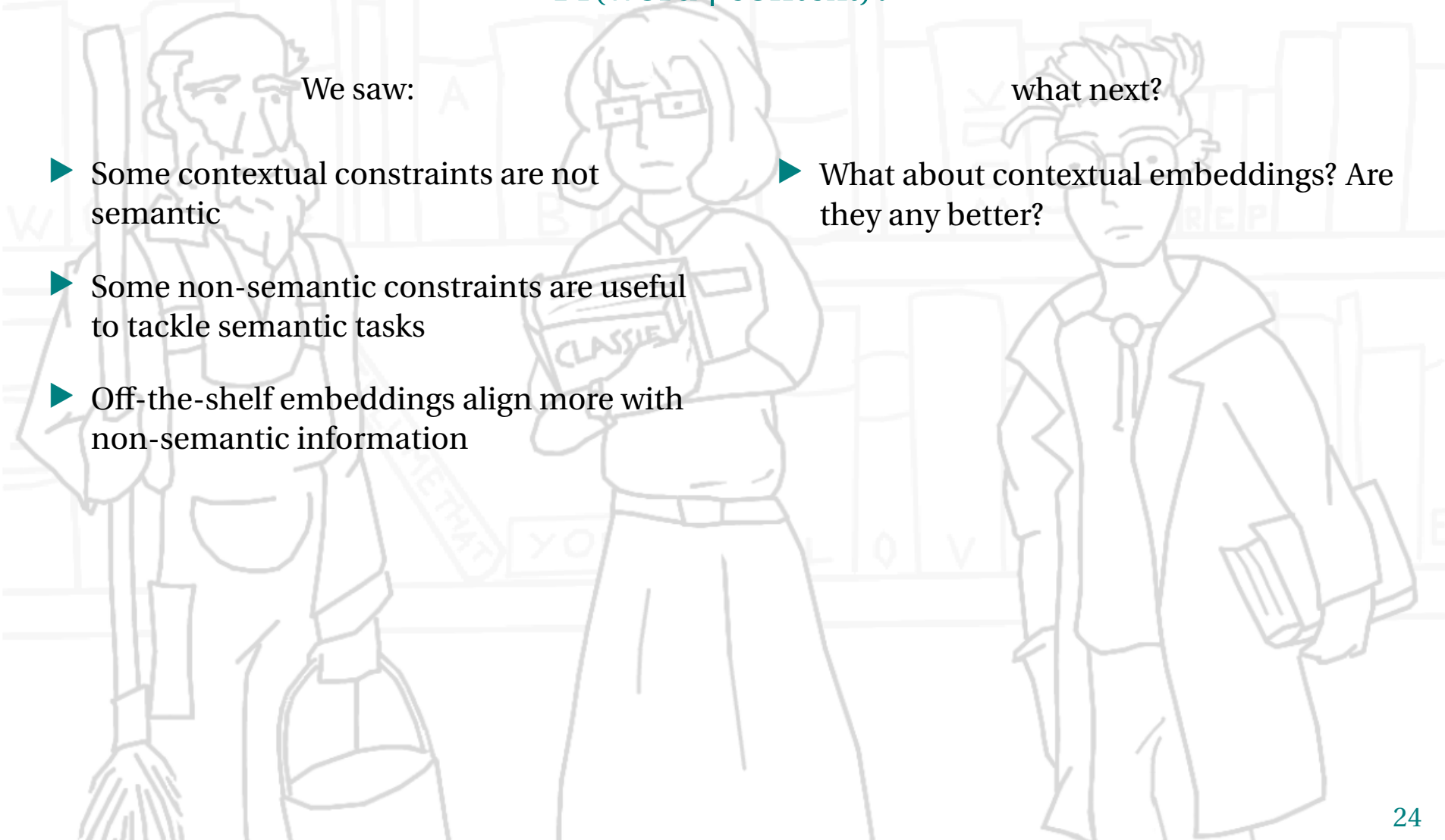
Do you really need semantics to model Pr(word | context)?

We saw:

- Some contextual constraints are not semantic
- Some non-semantic constraints are useful to tackle semantic tasks
- Off-the-shelf embeddings align more with non-semantic information

what next?

- What about contextual embeddings? Are they any better?



To recap

Do you really need semantics to model
Pr(word | context)**?**

We saw:

- Some contextual constraints are not semantic
- Some non-semantic constraints are useful to tackle semantic tasks
- Off-the-shelf embeddings align more with non-semantic information

what next?

- What about contextual embeddings? Are they any better?
- What about other aspects of semantics, e.g., grounding and interaction?

To recap

Do you really need semantics to model
Pr(word | context)**?**

We saw:

- Some contextual constraints are not semantic
- Some non-semantic constraints are useful to tackle semantic tasks
- Off-the-shelf embeddings align more with non-semantic information

what next?

- What about contextual embeddings? Are they any better?
- What about other aspects of semantics, e.g., grounding and interaction?
- What's the evidence for distributional *semantics*?

To recap

Do you really need semantics to model
Pr(word | context)**?**

We saw:

- Some contextual constraints are not semantic
- Some non-semantic constraints are useful to tackle semantic tasks
- Off-the-shelf embeddings align more with non-semantic information

what next?

- What about contextual embeddings? Are they any better?
- What about other aspects of semantics, e.g., grounding and interaction?
- What's the evidence for distributional *semantics*?

Thanks! any questions?

References

Abdaoui, Amine et al. (2017). "FEEL: a French Expanded Emotion Lexicon". In: *Language Resources and Evaluation* 51.3.

Barzegar, Siamak et al. (2018). "SemR-11: A Multi-Lingual Gold-Standard for Semantic Similarity and Relatedness for Eleven Languages". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bevilacqua, Michele, Marco Maru, and Roberto Navigli (2020). "Generationary or "How We Went beyond Word Sense Inventories and Learned to Gloss"". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bond, Francis and Kyonghee Paik (2012). "A Survey of WordNets and their Licenses". In:

Firth, John Rupert (1957). "A synopsis of linguistic theory 1930-55.". In: *Studies in Linguistic Analysis (special volume of the Philological Society)* 1952-59.

Grave, Edouard et al. (2018). "Learning Word Vectors for 157 Languages". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mickus, Timothee and Timothée Bernard (2023). *Distributional, yes—but semantics? Comparing distributional representations, semantics and syntax.*

Mickus, Timothee and Maria Copot (In prep.). *Stranger than Paradigm: word embedding benchmarks don't align with morphology.*

Noraset, Thanapon et al. (2017). "Definition Modeling: Learning to define word embeddings in natural language". In: *AAAI*.

Oepen, Stephan et al. (2015). "SemEval 2015 Task 18: Broad-Coverage Semantic Dependency Parsing". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Sahlgren, Magnus (2008). "The Distributional Hypothesis". In: *The Italian Journal of Linguistics* 20.

Segonne, Vincent and Timothee Mickus (2023). "Definition Modeling: To model definitionsGenerating Definitions With Little to No Semantics.". In: *Proceedings of the 15th International Conference on Computational Semantics (IWCS)*.

Wu, Shijie, Ryan Cotterell, and Timothy J. O'Donnell (2019). "Morphological Irregularity Correlates with Frequency". In: *CoRR* abs/1906.11483.