TSD 2020

# Costra 1.1: An Inquiry into Geometric Properties of Sentence Spaces

Petra Barančíková, Ondřej Bojar

September, 2020

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

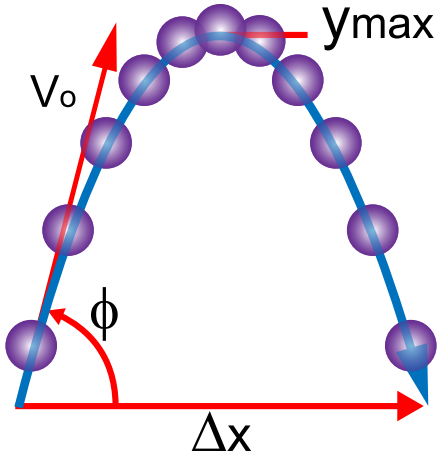
unless otherwise stated

# Outline

- The grand vision: Sentence meaning calculable.
- COSTRA 1.0
- COSTRA 1.1
  - Getting more sentences
  - Organizing sentences in space
- Examined sentence embeddings
- Results
- Conclusion

# The Grand Vision

# The Grand Vision: Calculability

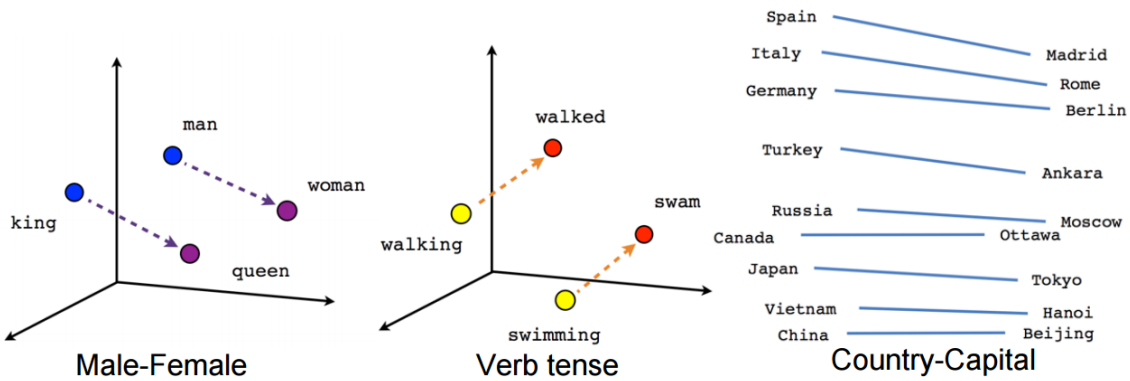



From: https://en.wikipedia.org/wiki/Classical_mechanics

# The Grand Vision: Calculability


From: `CatPhysicsApp`

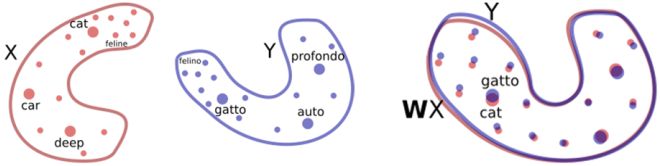
# Calculability for Words: Word Embeddings

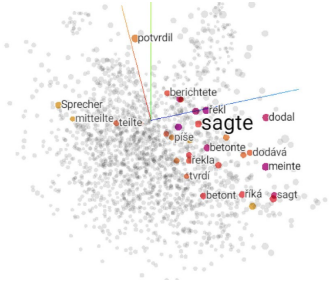



$$v_{\text{walking}} + (v_{\text{swam}} - v_{\text{swimming}}) \approx v_{\text{walked}}$$

From: http://hunterheidenreich.com/blog/intro-to-word-embeddings/

# Word Embeddings Cross-Lingually




A certain level of isomorphism across languages can be assumed.

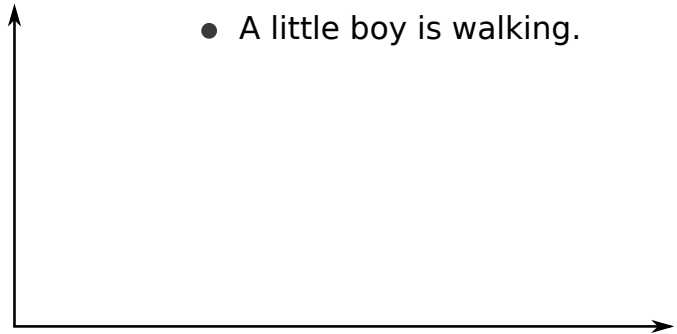
- Successfully used to seed unsupervised MT (Lample et al., 2018).
- Isomorphism not always realistic. (Søgaard et al., 2018; Patra et al., 2019)
- Different construction of the space can improve that. (Ormazabal et al., 2019)

$\Rightarrow$ Language-agnostic continuous word representations are conceivable.

# How about sentences?

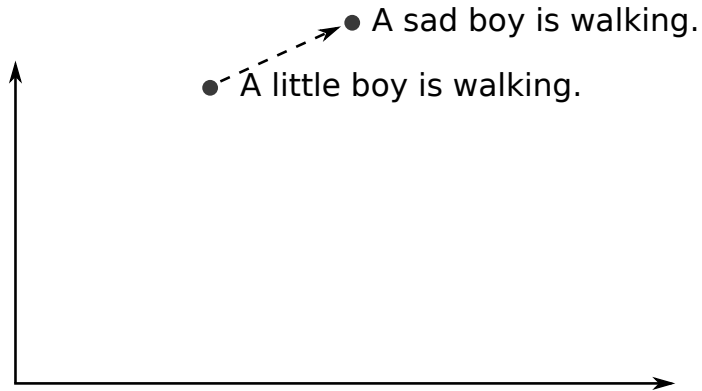
# Space of Sentences

# Space of Sentences

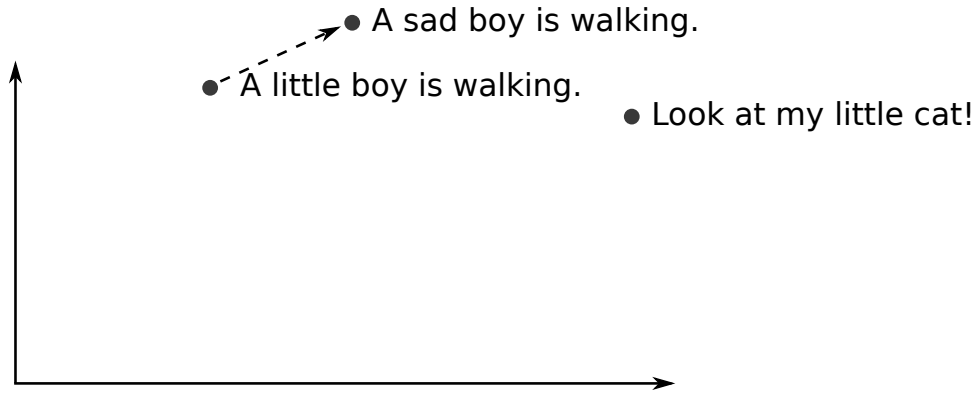


- A little boy is walking.

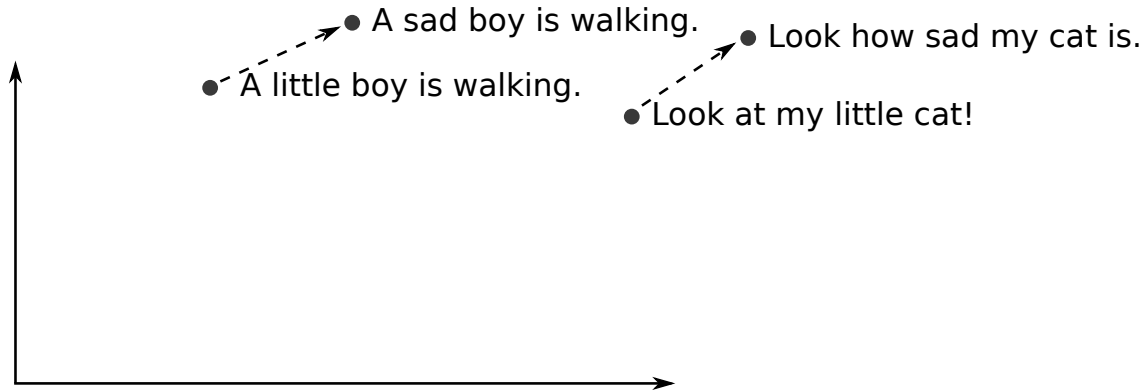
# Space of Sentences



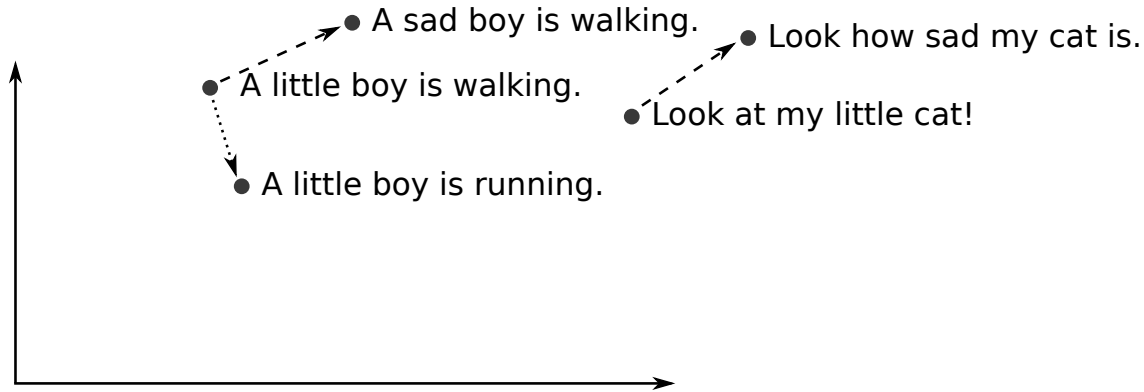
# Space of Sentences



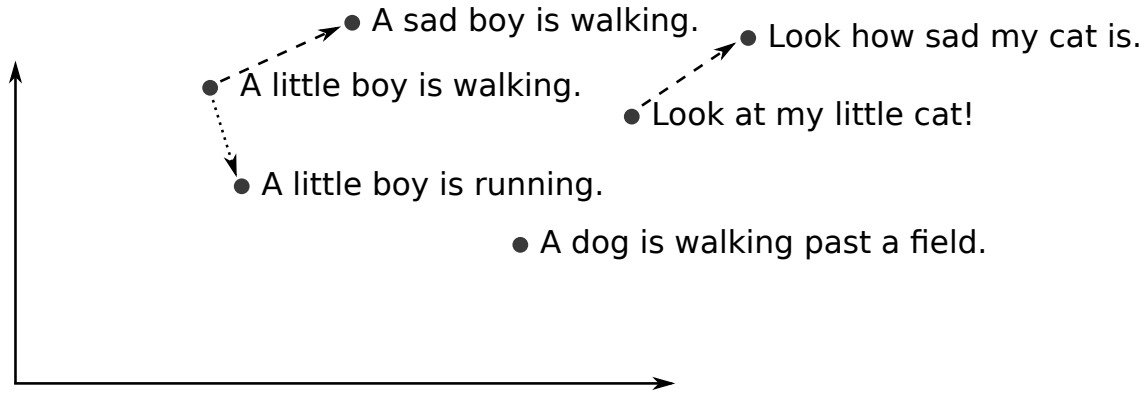
# Space of Sentences



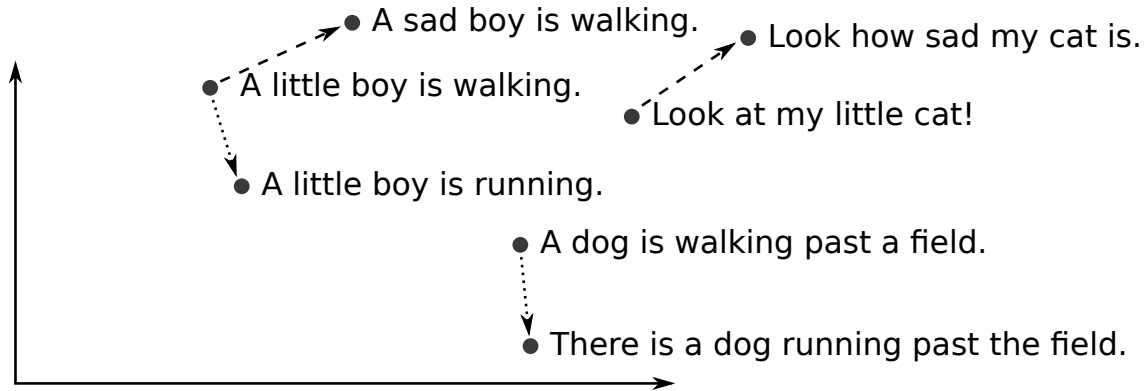
# Space of Sentences



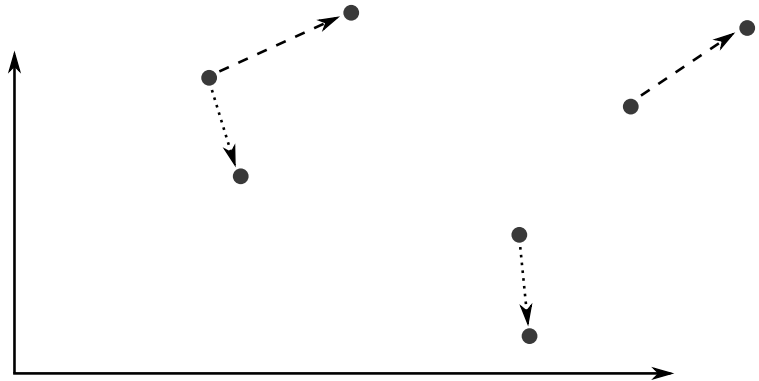
# Space of Sentences



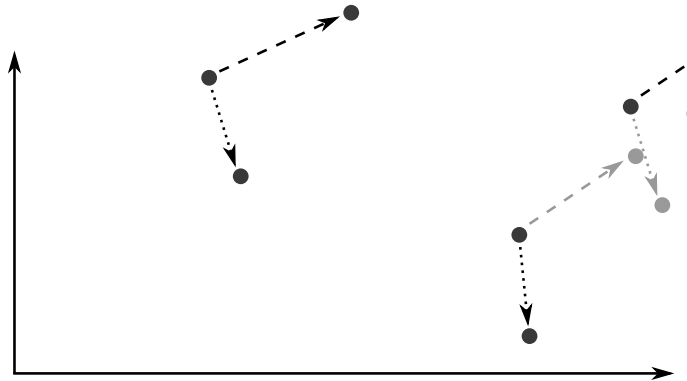
# Space of Sentences



# Space of Sentences

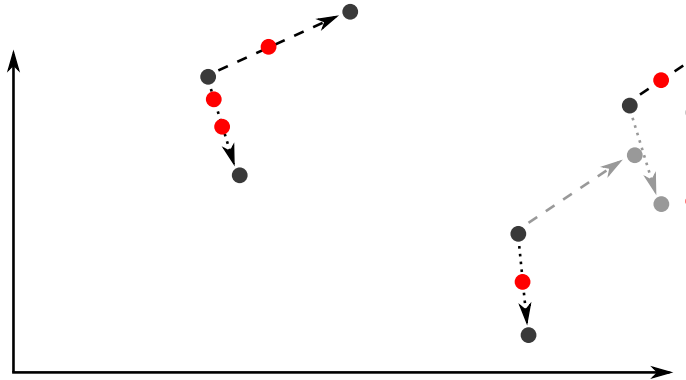


# Space of Sentences



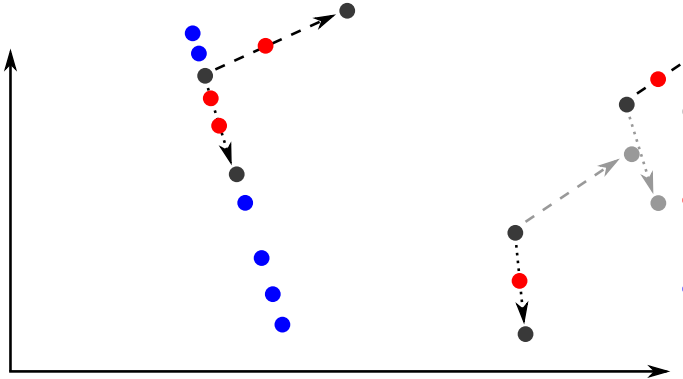
- Some transformations may or may not make sense.

# Space of Sentences



- Some transformations may or may not make sense.
- Sentences ‘in between’ may exist.

# Space of Sentences



- Some transformations may or may not make sense.
- Sentences ‘in between’ may exist.
- Extrapolation can be possible.

# COSTRA 1.0

# COSTRA 1.0 (Barančíková and Bojar, 2020)

- Paraphrases and inference were not diverse enough for us.
- COSTRA 1.0: 4k Czech sents across 15 complex transformations.

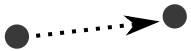
| Change | Instructions |
|---|---|
| paraphrase 1 | Reformulate the sentence using different words |
| paraphrase 2 | Reformulate the sentence using other different words |
| different meaning | Shuffle words in the sentence in order to get different meaning |
| opposite meaning | Reformulate the sentence to get a sentence with opposite meaning |
| nonsense | Shuffle words in sentence to make grammatical sentence with no sense.<br>E.g. *A hen pecked grain.* → *Grain pecked a hen.* |
| minimal change | Significantly change the meaning of the sentence using only a minimal alternation. |
| generalization | Make the sentence more general. |
| gossip | Rewrite the sentence in a gossip style – strongly exaggerated meaning on the sentence. |
| formal sentence | Rewrite the sentence in a more formal style. |
| non-standard | Rewrite the sentence in non-standard, colloquial style. |
| simple sentence | Rewrite the sentence in a simplistic style, with a limited vocabulary. |
| possibility | Change the modality of the sentence into a possibility. |
| ban | Change the modality of the sentence into a ban. |
| future | Move the sentence into the future. |
| past | Move the sentence into the past. |

# COSTRA 1.0 (Barančíková and Bojar, 2020)

- Paraphrases and inference were not diverse enough for us.
- COSTRA 1.0: 4k Czech sents across 15 complex transformations.

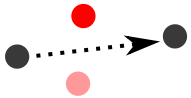
| Change | Instructions |
|---|---|
| paraphrase 1 | Reformulate the sentence using different words |
| paraphrase 2 | Reformulate the sentence using other different words |
| different meaning | Shuffle words in the sentence in order to get different meaning |
| opposite meaning | Reformulate the sentence to get a sentence with opposite meaning |
| nonsense | Shuffle words in sentence to make grammatical sentence with no sense. E.g. *A hen pecked grain.* → *Grain pecked a hen.* |
| minimal change | Significantly change the meaning of the sentence using only a minimal alternation. |
| generalization | Make the sentence more general. |
| gossip | Rewrite the sentence in a gossip style – strongly exaggerated meaning on the sentence. |
| formal sentence | Rewrite the sentence in a more formal style. |
| non-standard | Rewrite the sentence in non-standard, colloquial style. |
| simple sentence | Rewrite the sentence in a simplistic style, with a limited vocabulary. |
| possibility | Change the modality of the sentence into a possibility. |
| ban | Change the modality of the sentence into a ban. |
| future | Move the sentence into the future. |
| past | Move the sentence into the past. |

# COSTRA 1.1 (this work)



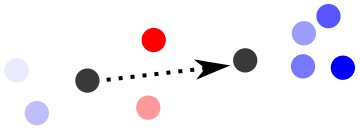
- Adding interpolations and extrapolations given a seed and transformed sentence.

# COSTRA 1.1 (this work)



- Adding interpolations and extrapolations given a seed and transformed sentence.

# COSTRA 1.1 (this work)



- Adding interpolations and extrapolations given a seed and transformed sentence.

# COSTRA 1.1 (this work)

- Adding interpolations and extrapolations
  given a seed and transformed sentence.
- Comparing sentences
  aiming to organize them to linear scales in each examined dimension.

# Phase 1: Interpolations and Extrapolations

- Data collected for: *formal sentence*, *future*, *generalization*, *nonstandard sentence*, *opposite meaning* and *past*.

# Phase 1: Interpolations and Extrapolations

- Data collected for: *formal sentence*, *future*, *generalization*, *nonstandard sentence*, *opposite meaning* and *past*.

| | |
|---|---|
| seed | "*Neudržel jsem hlavu vzhůru ani otevřené oči.*"<br>I couldn't keep my head up or my eyes open. |
| transformation | "*Hlavu jsem měl vzpřímenou a oči otevřené.*"<br>My head was upright and my eyes were open. |

# Phase 1: Interpolations and Extrapolations

- Data collected for: *formal sentence*, *future*, *generalization*, *nonstandard sentence*, *opposite meaning* and *past*.

| | |
|---|---|
| seed | "*Neudržel jsem hlavu vzhůru ani otevřené oči.*" |
| | I couldn't keep my head up or my eyes open. |
| interpolation | "*Neudržel jsem hlavu vzhůru, ale otevřené oči ano.*" |
| | I couldn't keep my head up, but my eyes were open. |
| transformation | "*Hlavu jsem měl vzpřímenou a oči otevřené.*" |
| | My head was upright and my eyes were open. |

# Phase 1: Interpolations and Extrapolations

- Data collected for: *formal sentence*, *future*, *generalization*, *nonstandard sentence*, *opposite meaning* and *past*.

| | |
|---|---|
| seed | "*Neudržel jsem hlavu vzhůru ani otevřené oči.*" |
| | I couldn't keep my head up or my eyes open. |
| interpolation | "*Neudržel jsem hlavu vzhůru, ale otevřené oči ano.*" |
| | I couldn't keep my head up, but my eyes were open. |
| transformation | "*Hlavu jsem měl vzpřímenou a oči otevřené.*" |
| | My head was upright and my eyes were open. |
| extrapolation | "*Ležel jsem v posteli a byl zcela čilý.*" |
| | I was lying in bed and I was completely awake. |

# Phase 1: Interpolations and Extrapolations

- Data collected for: *formal sentence*, *future*, *generalization*, *nonstandard sentence*, *opposite meaning* and *past*.

| | |
|---|---|
| seed | "*Neudržel jsem hlavu vzhůru ani otevřené oči.*" |
| | I couldn't keep my head up or my eyes open. |
| interpolation | "*Neudržel jsem hlavu vzhůru, ale otevřené oči ano.*" |
| | I couldn't keep my head up, but my eyes were open. |
| transformation | "*Hlavu jsem měl vzpřímenou a oči otevřené.*" |
| | My head was upright and my eyes were open. |
| extrapolation | "*Ležel jsem v posteli a byl zcela čilý.*" |
| | I was lying in bed and I was completely awake. |

- 7 annotators, almost 1,500 annotations $\Rightarrow$ 2,749 new unique sents.
- With Costra 1.0, the total volume of Costra 1.1 is 6,968 sentences.

# Phase 2: Comparing Sentences

- Again, only linearly scalable axes considered:
  - changes in tense (*future*, *past*),
  - changes in style (*formal sentence*, *simple sentence*, and the merge of *gossip* and *nonstandard sentence* into one axis),
  - significant changes in meaning (*generalization*, *opposite meaning*).
- Possible answers:
  - $S_1$ is more general/formal/in the past/non-standard/... than $S_2$.
  - $S_2$ is more general/formal/in the past/non-standard/... than $S_1$.
  - $S_1$ and $S_2$ are **too similar** to be clearly ordered.
  - $S_1$ and $S_2$ **too dissimilar**, they became incomparable.

# Phase 2: Comparing Sentences

- Again, only linearly scalable axes considered:
  - changes in tense (*future*, *past*),
  - changes in style (*formal sentence*, *simple sentence*, and the merge of *gossip* and *nonstandard sentence* into one axis),
  - significant changes in meaning (*generalization*, *opposite meaning*).
- Possible answers:
  - $S_1$ is more general/formal/in the past/non-standard/... than $S_2$.
  - $S_2$ is more general/formal/in the past/non-standard/... than $S_1$.
  - $S_1$ and $S_2$ are **too similar** to be clearly ordered.
  - $S_1$ and $S_2$ **too dissimilar**, they became incomparable.
- Collected 25k pairwise comparisons.
  - Inter-annotator agreement: 0.62 ($\kappa = 0.49$).
  - Intra-annotator agreement: 0.77 ($\kappa = 0.70$).
  - 16,385 sent pairs with clear result; 1,620 pairs disregarded (disagreement).

# Evaluation of Sentence Embedding Methods

# Available Sentence Embedding Methods for Czech

- Not many off-the-shelf embedding methods support Czech.
- We found the following:

| Method | Level | Description |
|---|---|---|
| LASER | whole sentence | multi-lingual biLSTM across words |
| Flair | aggregated | contextualized char-level embs. of words |
| mBERT | aggregated | contextualized word-level embs. from Transformer |
| SentBERT | aggregated | contextualized word embs. from sentence encoder initialized with multilingual BERT and finetuned on inference tasks |

For BERTs, the embedding of the special token "CLS" is also considered as sentence representation.

# Results (1/2)

Results are collected for 12 scales, grouped to 6 classes for conciseness:

basic Paraphrases should be closer to their seed than any transformation that significantly changes the meaning of the seed (*different meaning*, *nonsense*, *minimal change*).

modality Paraphrases should be closer to their seed than any transformation which changes modality of the seed (*possibility*, *ban*).

time, style, generalization, opposite should reflect the manual linear ordering.

- **basic** and **modality** evaluated with precision (how many pairwise comparisons satisfy the expectation).
- Others evaluated by checking how often the order $A < B < C$ declared by humans is confirmed by cosine similarity in vector space.

# Results (2/2)

| | basic | modality | time | style | gener. | opposite | avg |
|---|---|---|---|---|---|---|---|
| SentBERT - mean | 0.150 | 0.251 | 0.667 | 0.588 | **0.718** | 0.685 | 0.510 |
| SentBERT - CLS | 0.172 | **0.303** | 0.654 | 0.577 | 0.690 | 0.654 | 0.508 |
| Flair - mean | 0.145 | 0.157 | **0.682** | **0.627** | 0.695 | **0.728** | 0.506 |
| mBERT - CLS | **0.262** | 0.274 | 0.616 | 0.579 | 0.603 | 0.640 | 0.496 |
| mBERT - mean | 0.103 | 0.115 | 0.674 | 0.621 | 0.691 | 0.727 | 0.489 |
| LASER | 0.255 | 0.244 | 0.583 | 0.533 | 0.667 | 0.636 | 0.486 |

# Results (2/2)

All methods fooled by superficial sentence similarity.

| | basic | modality | time | style | gener. | opposite | avg |
|---|---|---|---|---|---|---|---|
| SentBERT - mean | 0.150 | 0.251 | 0.667 | 0.588 | **0.718** | 0.685 | 0.510 |
| SentBERT - CLS | 0.172 | **0.303** | 0.654 | 0.577 | 0.690 | 0.654 | 0.508 |
| Flair - mean | 0.145 | 0.157 | **0.682** | **0.627** | 0.695 | **0.728** | 0.506 |
| mBERT - CLS | **0.262** | 0.274 | 0.616 | 0.579 | 0.603 | 0.640 | 0.496 |
| mBERT - mean | 0.103 | 0.115 | 0.674 | 0.621 | 0.691 | 0.727 | 0.489 |
| LASER | 0.255 | 0.244 | 0.583 | 0.533 | 0.667 | 0.636 | 0.486 |

# Results (2/2)

All methods fooled by superficial sentence similarity.

Linear orderings much better reflected.

| | basic | modality | time | style | gener. | opposite | avg |
|---|---|---|---|---|---|---|---|
| SentBERT - mean | 0.150 | 0.251 | 0.667 | 0.588 | **0.718** | 0.685 | 0.510 |
| SentBERT - CLS | 0.172 | **0.303** | 0.654 | 0.577 | 0.690 | 0.654 | 0.508 |
| Flair - mean | 0.145 | 0.157 | **0.682** | **0.627** | 0.695 | **0.728** | 0.506 |
| mBERT - CLS | **0.262** | 0.274 | 0.616 | 0.579 | 0.603 | 0.640 | 0.496 |
| mBERT - mean | 0.103 | 0.115 | 0.674 | 0.621 | 0.691 | 0.727 | 0.489 |
| LASER | 0.255 | 0.244 | 0.583 | 0.533 | 0.667 | 0.636 | 0.486 |

# Results (2/2)

All methods fooled by superficial sentence similarity.

Linear orderings much better reflected.

| | basic | modality | time | style | gener. | opposite | avg |
|---|---|---|---|---|---|---|---|
| SentBERT - mean | 0.150 | 0.251 | 0.667 | 0.588 | **0.718** | 0.685 | 0.510 |
| SentBERT - CLS | 0.172 | **0.303** | 0.654 | 0.577 | 0.690 | 0.654 | 0.508 |
| Flair - mean | 0.145 | 0.157 | **0.682** | **0.627** | 0.695 | **0.728** | 0.506 |
| mBERT - CLS | **0.262** | 0.274 | 0.616 | 0.579 | 0.603 | 0.640 | 0.496 |
| mBERT - mean | 0.103 | 0.115 | 0.674 | 0.621 | 0.691 | 0.727 | 0.489 |
| LASER | 0.255 | 0.244 | 0.583 | 0.533 | 0.667 | 0.636 | 0.486 |

LASER and CLS-based worse than mean-aggregated embs. for linear.

# Results (2/2)

All methods fooled by superficial sentence similarity.

Linear orderings much better reflected.

| | basic | modality | time | style | gener. | opposite | avg |
|---|---|---|---|---|---|---|---|
| SentBERT - mean | 0.150 | 0.251 | 0.667 | 0.588 | **0.718** | 0.685 | 0.510 |
| SentBERT - CLS | 0.172 | **0.303** | 0.654 | 0.577 | 0.690 | 0.654 | 0.508 |
| Flair - mean | 0.145 | 0.157 | **0.682** | **0.627** | 0.695 | **0.728** | 0.506 |
| mBERT - CLS | **0.262** | 0.274 | 0.616 | 0.579 | 0.603 | 0.640 | 0.496 |
| mBERT - mean | 0.103 | 0.115 | 0.674 | 0.621 | 0.691 | 0.727 | 0.489 |
| LASER | 0.255 | 0.244 | 0.583 | 0.533 | 0.667 | 0.636 | 0.486 |

LASER and CLS-based worse th
mean-aggregated embs. for line

No big differences when averaged across the classes.

# Conclusion

# Summary

- COSTRA 1.1 released: `http://hdl.handle.net/11234/1-3248`
  - ∼7k Czech sents, 15 complex transformations, 25k pairwise comparisons.
  - IAA for pairwise comparisons reasonable, >90% of sent. pairs ordered unanimously.
- Available Czech sentence embeddings tested:
  - SentBERT, Flair, mBERT, LASER.
  - All fail at spotting meaning differences (accuracies under 30%).
  - Linear orderings better, 63–74% of tested triples correct.
  - Sentence-level representations (LASER and CLS-token) better in meaning changes but worse in linear relations.

# References


Petra Barančíková and Ondřej Bojar. 2020. COSTRA 1.0: A Dataset of Complex Sentence Transformations. In *Proceedings of the LREC 2020*. ELRA.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October-November. Association for Computational Linguistics.

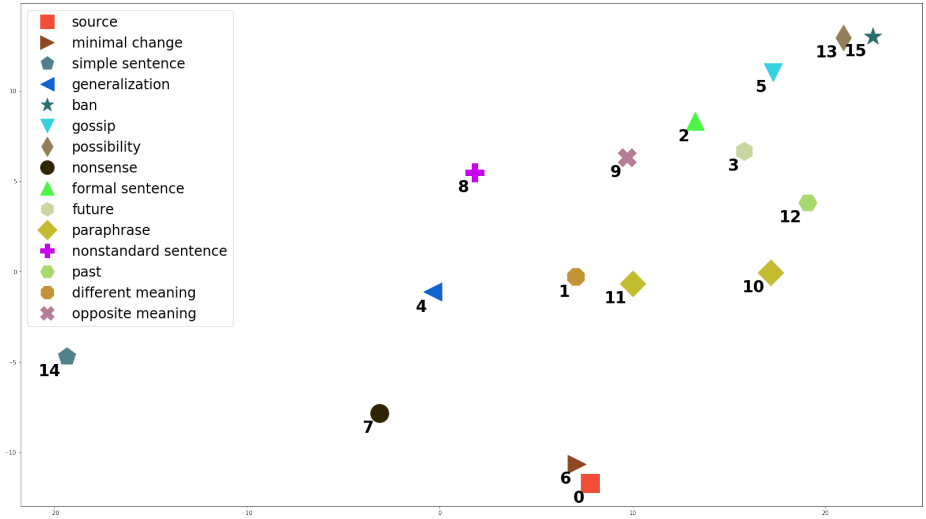
Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995, Florence, Italy, July. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. 2019. BLISS in non-isometric embedding spaces.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July. Association for Computational Linguistics.

# Extra Slides

# XXX: co je to?PB: to je PCA z Costry 1.0



# PCA Projection