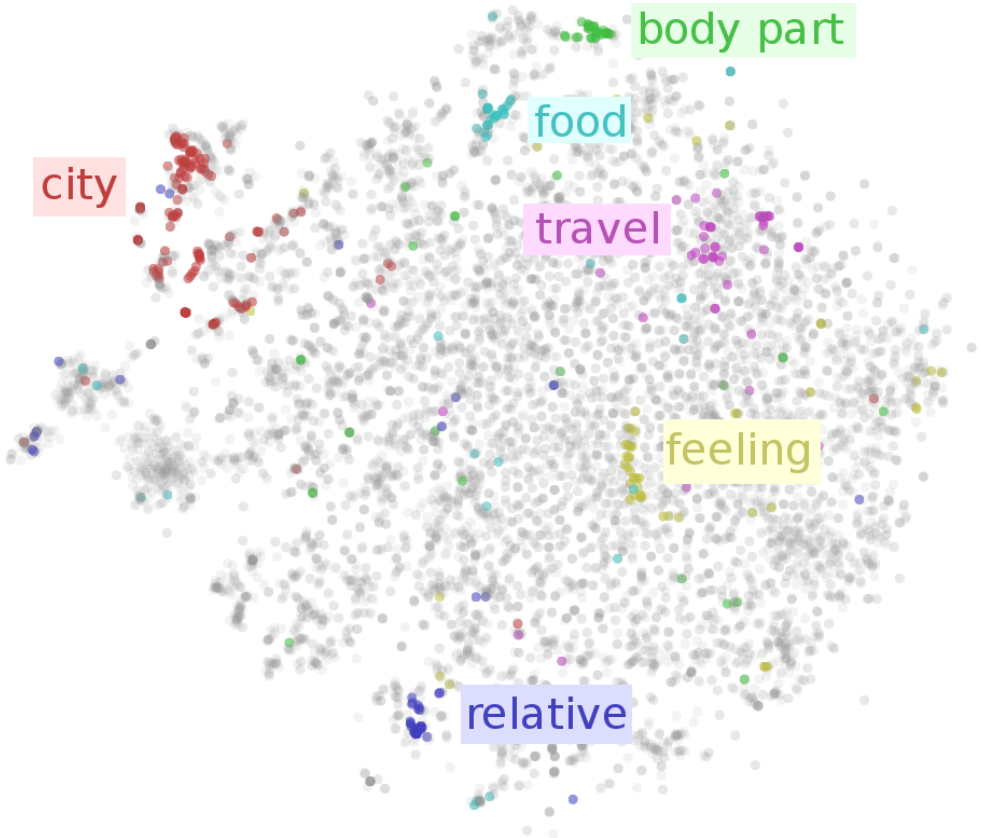
UNIVERSITY OF TARTU

# Linear Ensembles of Word Embeddings

Avo Muromägi, **Kairit Sirts**, Sven Laur

**Nodalida 2017**

# Dense word embeddings are very popular in NLP




- Distributional word representations
- Semantically and/or morphologically similar words are located together
- Provide useful features for many NLP tasks

# Training high quality word embeddings requires lots of data

- To reliably estimate the word's distributional representation the word must be observed in many contexts
- Most research on word embeddings has been done on English
- There are several large corpora available for English:
  - Wikipedia – 2B words
  - Gigaword – 4B words
  - Common Crawl – 840B words

# Training word embeddings for Estonian

- Estonian Reference Corpus – 250M words
- Wikipedia (as of 2013) – 23M words
- Morphologically rich language:
  - Type-token ratio much higher than in English
  - Thus, the training corpus would need to be even larger than in English

# How to improve word embeddings?

- Collect more data
- Develop better methods for training word embeddings
- Construct an ensemble of several embedding models
  - Cancel out noise
  - Reinforce useful regularities

# Ensemble of word embeddings

- The useful relations in the embedding space are linear
- Use linear transformations to align word embedding models into a common embedding space
- Combine the aligned embedding vectors via averaging

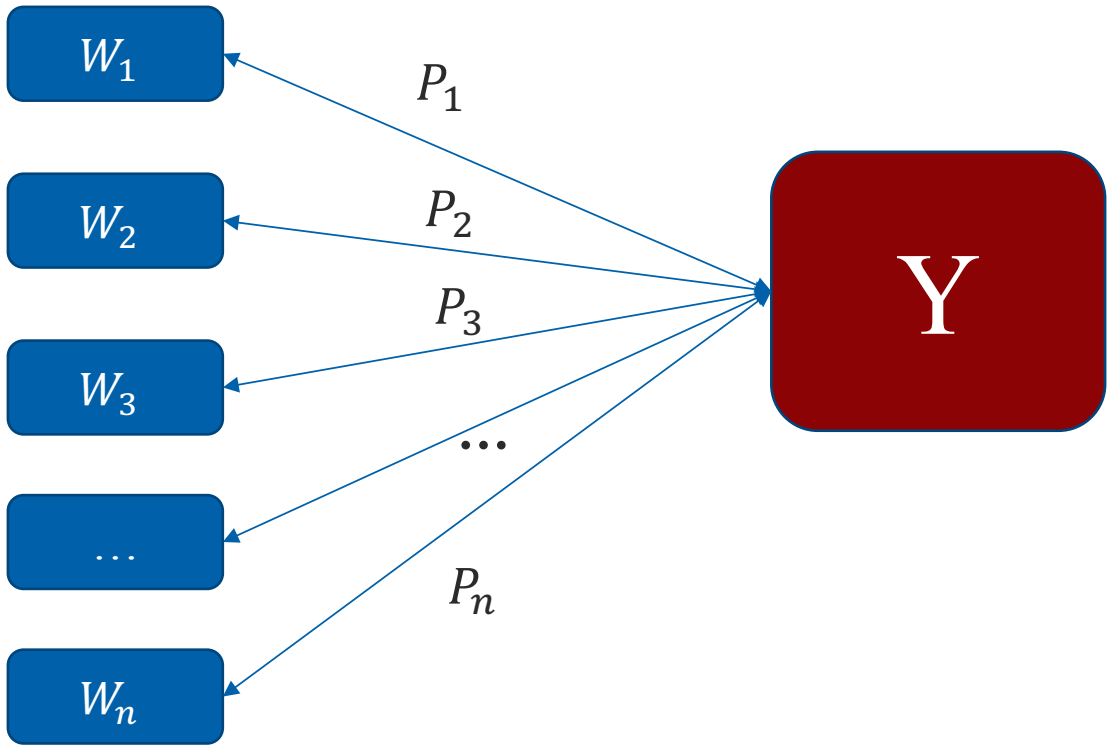
# Previous and related work

- Learning Word Meta-Embeddings (Yin and Schütze, ACL 2016)
  - Combined the embeddings pre-trained with 5 different embedding learning systems
  - Experiments on English only
  - The ensemble barely outperforms the Glove embeddings (42B words)
- Align series of word embeddings to detect the semantic changes over time
  - Kulkarni et al., 2015 (WWW)
  - Hamilton et al., 2016 (ACL)
- Align the embedding spaces of two languages
  - Mikolov et al., 2013 (arxiv)
  - Mogdala and Rettinger, 2016 (NAACL)

# This work

- Combine several word embedding models into an ensemble
- All models are trained on the same dataset and with the same method (word2vec)
- Due to stochastic training all embedding models are different
- Experiment with two different linear methods for combining:
  - Linear regression
  - Orthogonal Procrustes (Schönemann, 1966)
- Experiments on Estonian

# Iterative solution

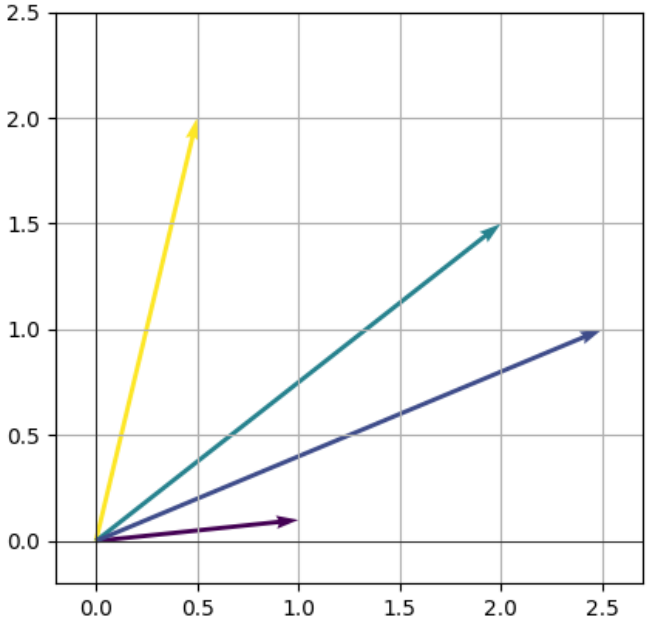



- $W_1, \ldots, W_n$: initial embeddings
- Y: combined embeddings (ensemble)
- $P_1, \ldots, P_n$: linear transformation matrices

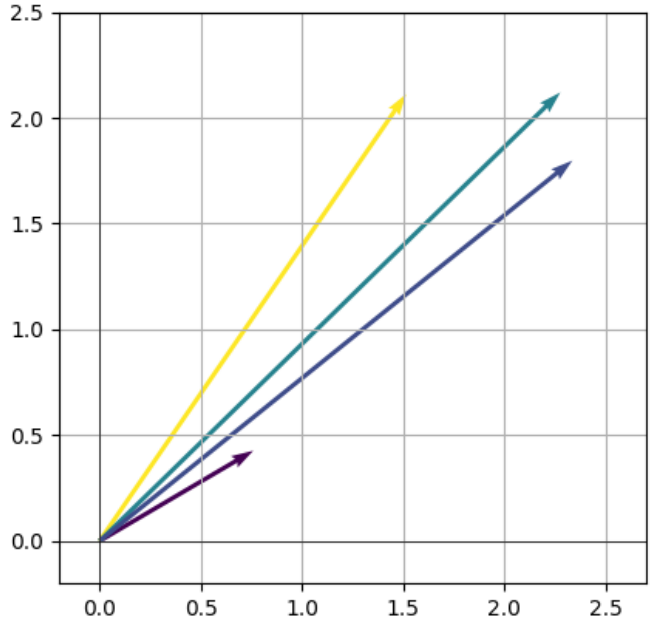
- Estimate Y and $P_1, \ldots, P_n$ iteratively

# Multivariate linear regression

- $Y = WP$
- Can be solved analytically



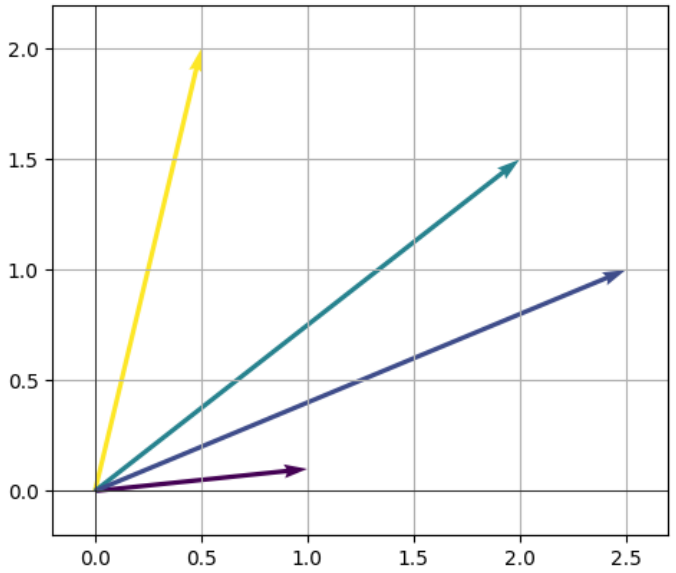

W



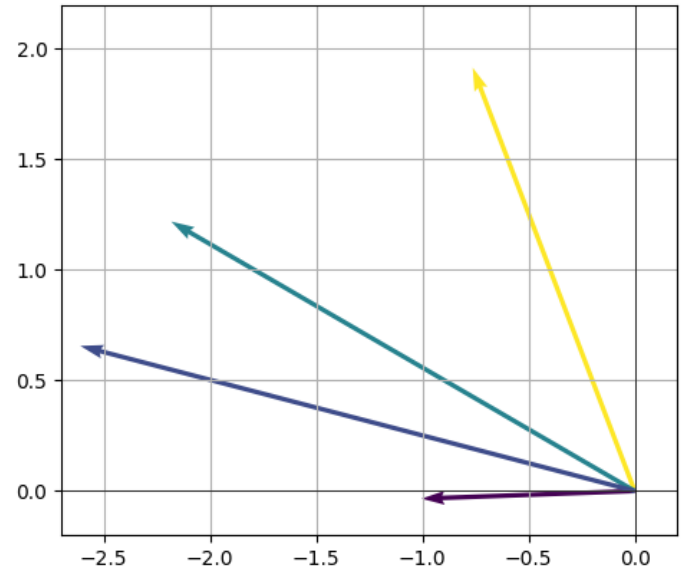

Y = WP

# Solution to Orthogonal Procrustes problem

- $Y = WP, \quad s.t. PP^T = P^T P = I$
- Transformation matrices are orthonormal
- The lengths and the angles between the word vectors are preserved
- Can be solved analytically using singular value decomposition




W




Y = WP

UNIVERSITY OF TARTU

# Data

- Embedding models trained on Estonian Reference Corpus using Word2Vec
  - 250M word tokens
  - 800K word types
- 10 embedding models combined into an ensemble
- Dimensionality of the embedding vectors: 50, 100, 150, 200, 250, 300

# Evaluation

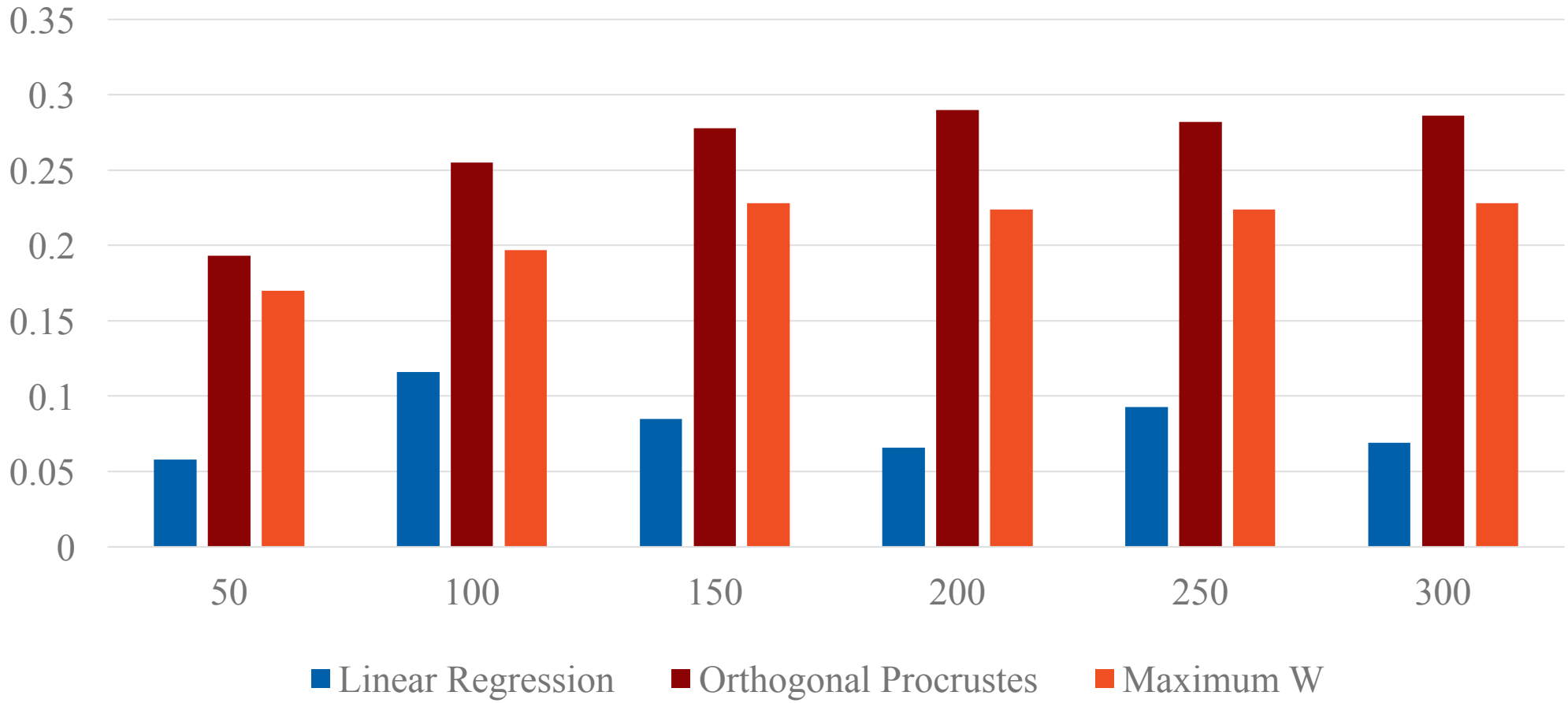
- Accuracy of analogy tests:
  - $w_{king} - w_{man} + w_{woman} \approx ?$
- Mean synonym ranks:
  - Replacement for a more popular word similarity test
  - Assumption: synonymous words are organized close in the embedding space
  - Find the similarity rank of  the synonym pairs
  - Compute the mean rank over all synonym pairs

UNIVERSITY OF TARTU

# Analogy tests

- 259 analogy questions
- Positive and comparative adjectives
  - pime : pimedam, jõukas : jõukam (dark : darker, wealthy : wealthier)
- Nominative singular and plural nouns
  - vajadus : vajadused, võistlus : võistlused (need : needs, competition : competitions)
- The lemma and the 3rd person past form of the verbs
  - aitama : aitas, katsuma : katsus (help : helped, touch : touched)

# Accuracy of analogy tests

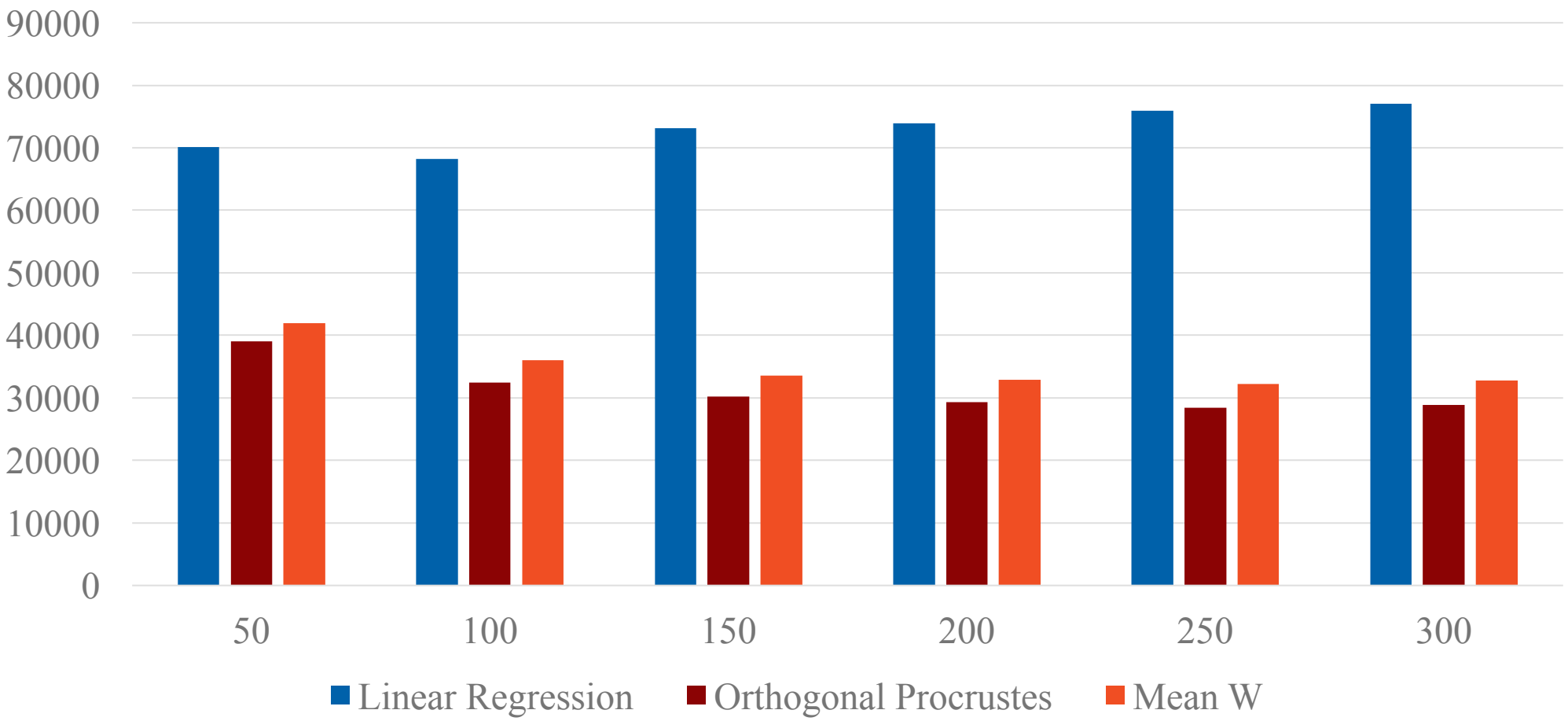


UNIVERSITY OF TARTU

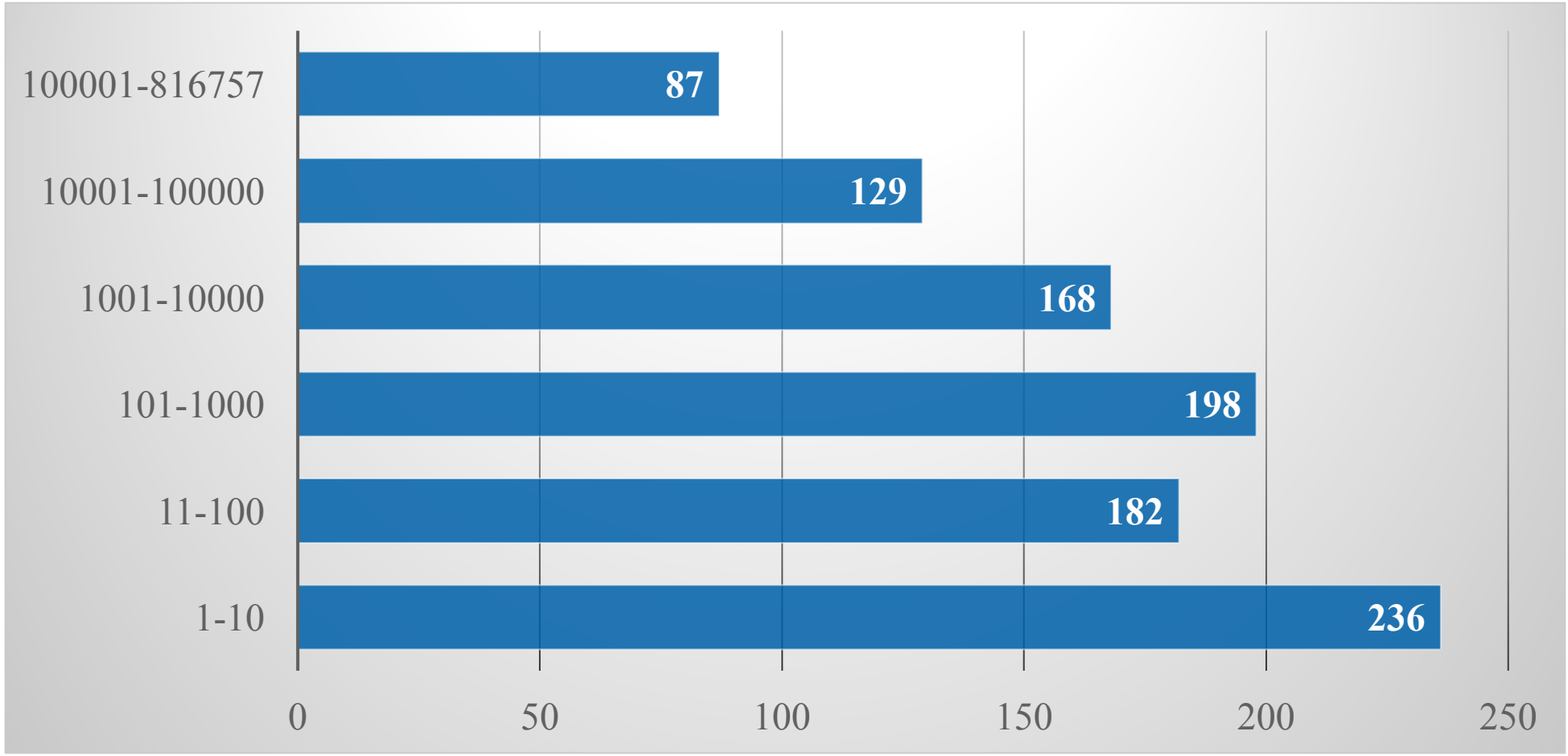
# Mean synonym ranks

- 1000 synonym pairs
- Extracted from Estonian synonym dictionary
- The first word in the pair is chosen based on frequency
- The second word is the first synonym offered by the dictionary

# Mean synonym ranks



# Orthogonal Procrustes 100-dimensional

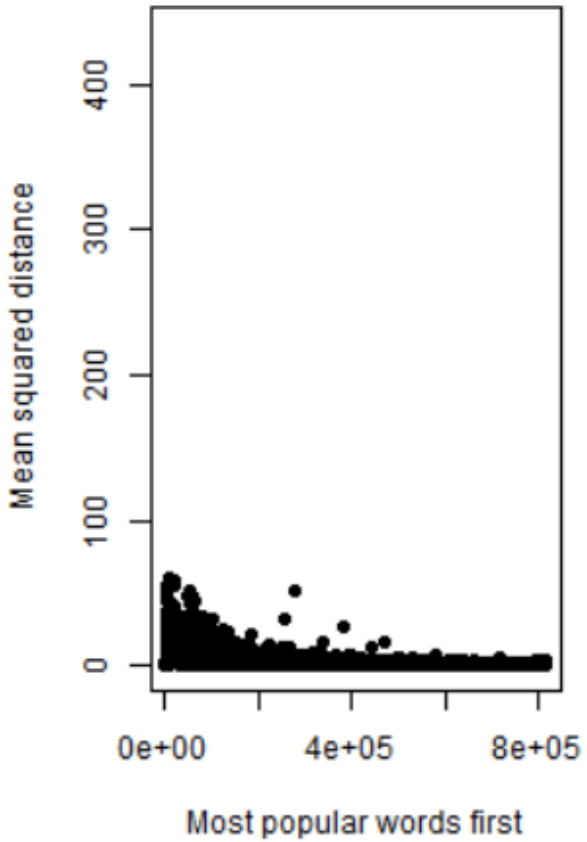


# Dissimilar synonyms

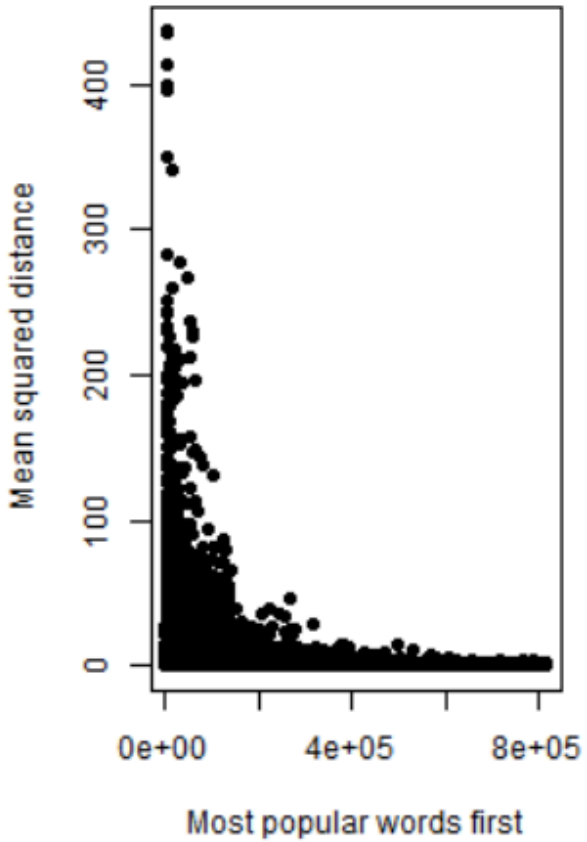
- Kaks – puudulik (two – insufficient)
- Ida – ost (east – ost)
- Rubla – kull (rouble – bank note in slang)

# Scatteredness of the embedding vectors

Linear regression




Orthogonal Procrustes



# Future work

- Test Orthogonal Procrustes method on more languages
- Test in Estonian on semantic analogy questions
- Study the relations between the training corpus size and the effectiveness of the ensembling

# Conclusion

- Two linear methods for combining word embedding models into ensemble
- Evaluated on synonymy and analogy tests on Estonian
- Orthogonal Procrustes performed well, improving significantly over the best initial model
- Linear regression was worse than the initial models
- Embeddings with more dimensions (200+) perform better on Estonian