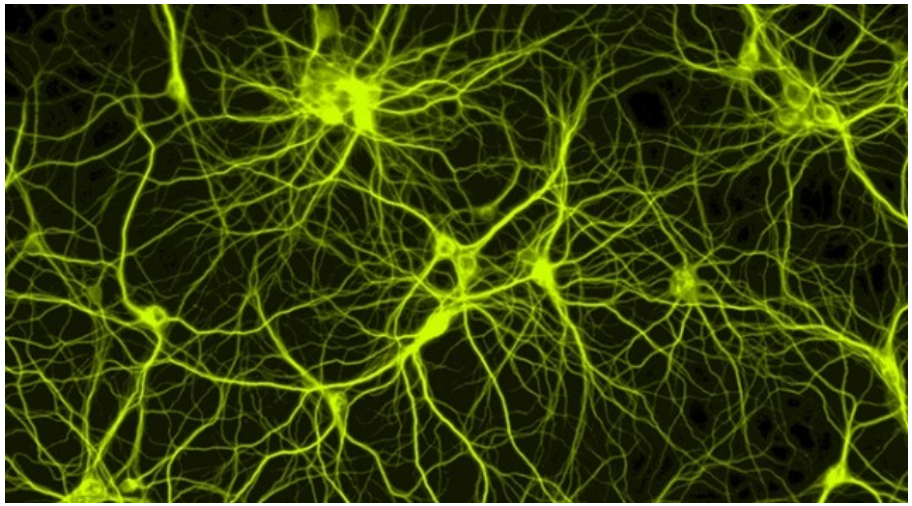
# Disambiguated linear word translation in medium European languages

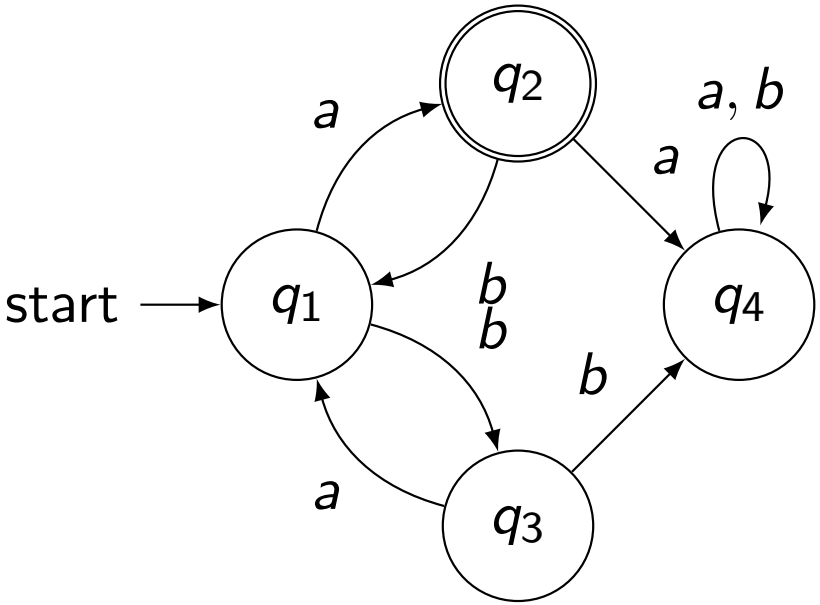
Márton Makrai
makrai@nytud.hu

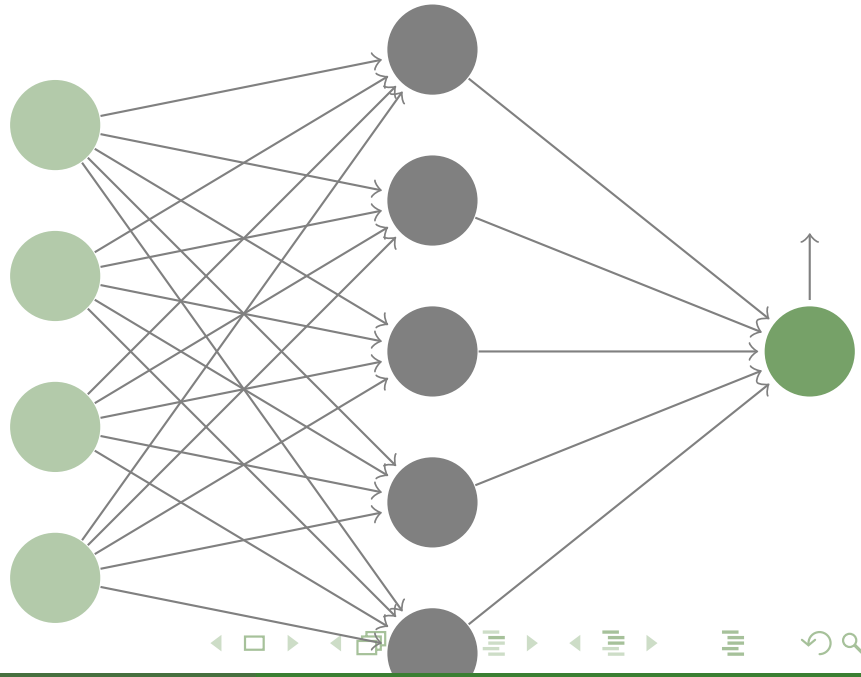
CogInfoCom 2015

# Overview

# The cognitive inspiration





# Overview

1. Neural language models
   - Distributed representations applied
   - Counts and synapses
2. Translation
   - neural machine translation
   - pivot-based dictionary induction
3. Experiments

# Modalities

- image recognition (Krizhevsky and Sutskever, 2012)
- speech recognition (Hinton et al., 2012)
- natural language processing
  - topic classification
  - sentiment analysis (Socher et al., 2011)
  - question answering and
  - machine translation (Sutskever et al., 2014)

# similarity and relatedness

- similarity

$$\textbf{Lake Baikal} \approx \textbf{Aral Sea}$$

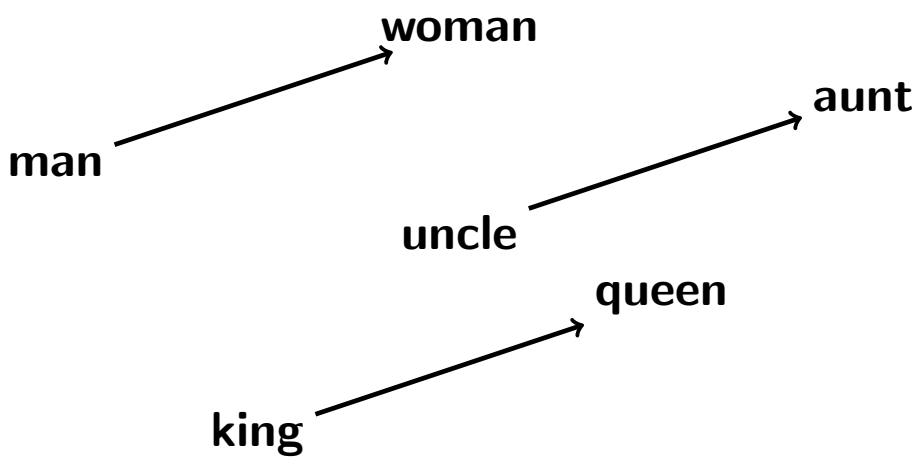
  - synonymy

$$\textbf{graffiti} \approx \textbf{spray paint}$$

  - relatedness

$$\textbf{apple} \approx \textbf{pear}$$

- relational similarity



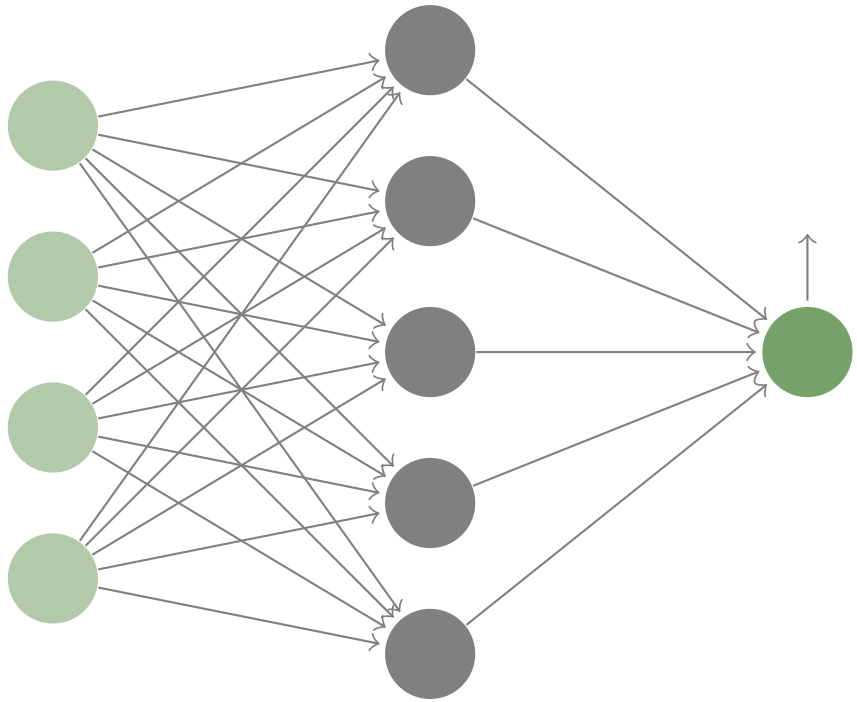
# Overview

1. Neural language models
   - Distributed representations applied
   - Counts and synapses
2. Translation
   - neural machine translation
   - pivot-based dictionary induction
3. Experiments

# Vector space language models

| | the | , | ... | table | ... | dog | ... |
|---|---|---|---|---|---|---|---|
| the | 30507 | 126192 | ... | 99067 | ... | 16785 | ... |
| , | 10488729 | 3462 | ... | 5164 | ... | 1645 | ... |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ | ⋱ |
| table | 1307 | 39189 | ... | 36 | ... | 62 | ... |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ | ⋱ |
| dog | 257 | 10268 | ... | 91 | ... | 46 | ... |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋱ | ⋮ | ⋱ |

- + feature engineering (domain specific)



- unsupervised learning
- representation sharing among
  - words
  - NLP tasks (Collobert et al., 2011)
  - modalities

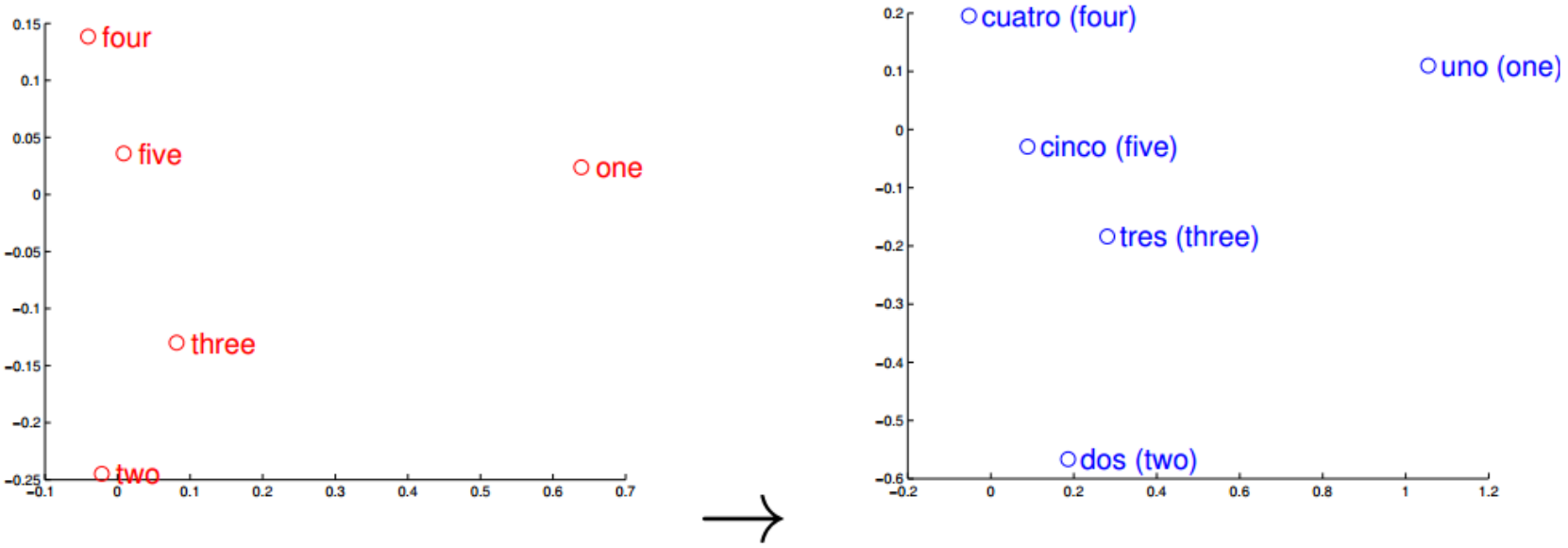
# Overview

1. Neural language models
   - Distributed representations applied
   - Counts and synapses
2. Translation
   - neural machine translation
   - pivot-based dictionary induction
3. Experiments

# Overview

1. Neural language models
   - Distributed representations applied
   - Counts and synapses
2. Translation
   - neural machine translation
   - pivot-based dictionary induction
3. Experiments

# neural word translation (Mikolov et al., 2013b)




$$W : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2} \quad z \approx Wx$$

- learning the mapping: supervised by a seed dictionary

$$\min_W \sum_i ||Wx_i - z_i||^2$$

- 5 K train + 1 K test

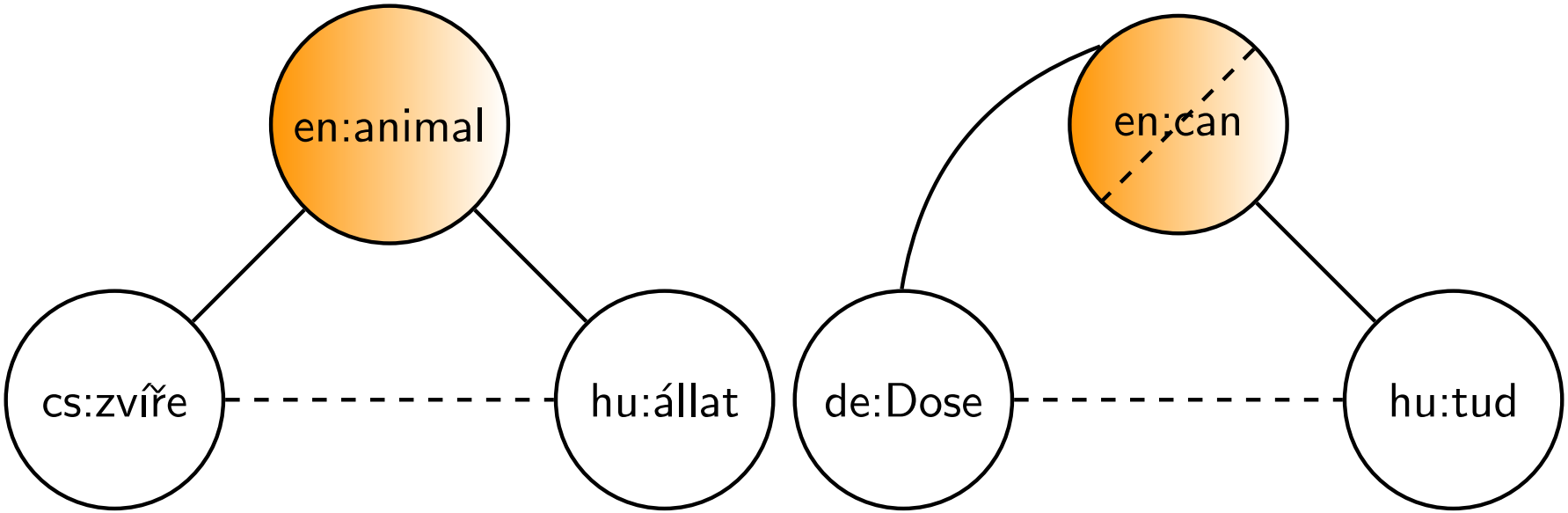
- generate or score translations
- hub: some targets are erroneous translations of many sources (Dinu et al., 2015)

# multi-prototype VSMs

- meaning depends on context
- prototype $\leftarrow$ psychological concept modeling
- Reisinger and Mooney (2010); Huang et al. (2012)
- problems
  - uniform number of senses
  - word sense induction proceeds VSM learning
  - efficiency
- solution?, free code?? (Neelakantan et al., 2014; Chen et al., 2014; Bartunov et al., 2015)

# Overview

1. Neural language models
   - Distributed representations applied
   - Counts and synapses
2. Translation
   - neural machine translation
   - pivot-based dictionary induction
3. Experiments




- pruning triangles
  - number of pivots (Tanaka and Umemura, 1994)
  - based on distributional similarity
    - comparable corpora (Saralegi et al., 2011)
    - now: with monolingual corpora

# Overview

1. Neural language models
   - Distributed representations applied
   - Counts and synapses
2. Translation
   - neural machine translation
   - pivot-based dictionary induction
3. Experiments

# Less-resourced languages

- project
  - EFNILEX
  - European Federation of National Institutions for Language
  - machine translation for lexicography in less-resourced Official EU languages
- scoring Wiktionary triangles
  - mapping trained with direct pairs
- linear mapping between Multi-prototype VSMs

# Data and tools

## Corpora

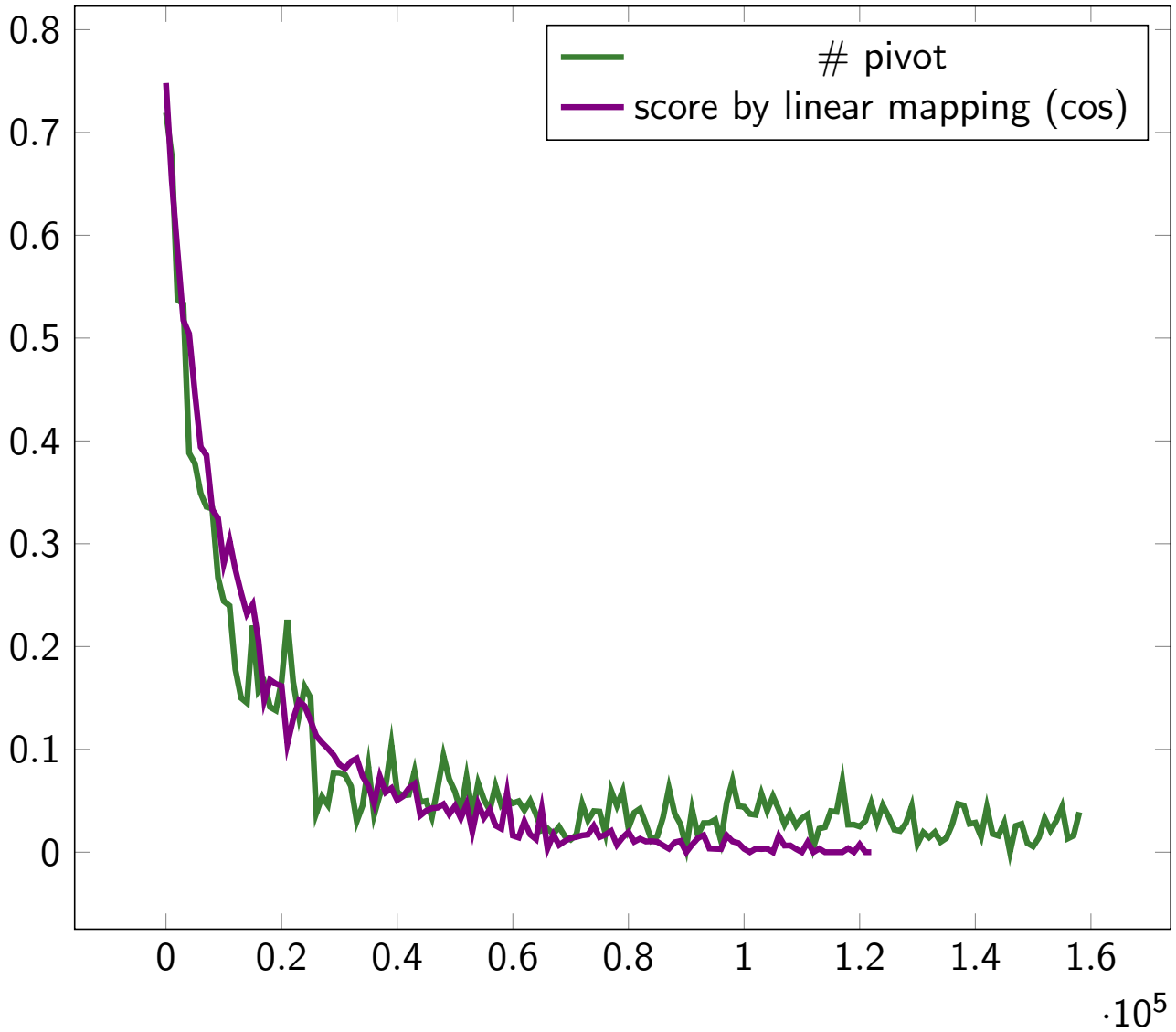
| | | # words |
|---|---|---|
| * Czech | CNK-SYN (Hnátková et al., 2014) | 2.2 B |
| Croatian | hrWaC2.0 Ljubešić and Klubička (2014) | 2.0 B |
| * Slovenian | slWaC (Ljubešić and Erjavec, 2011) | 1.6 B |
| Polish | Araneum Polonicum Maius (Benko, 2014) | 1.1 B |
| Serbian | srWaC (Ljubešić and Klubička, 2014) | 1.0 B |
| * German | SdeWac (Baroni et al., 2009) | 0.8 B |
| * Hungarian | HNC (Oravecz et al., 2014) | 0.8 B |
| * Hungarian | webcorpus (HW) (Halácsy et al., 2004) | 0.7 B |

# Data and tools

## Tools

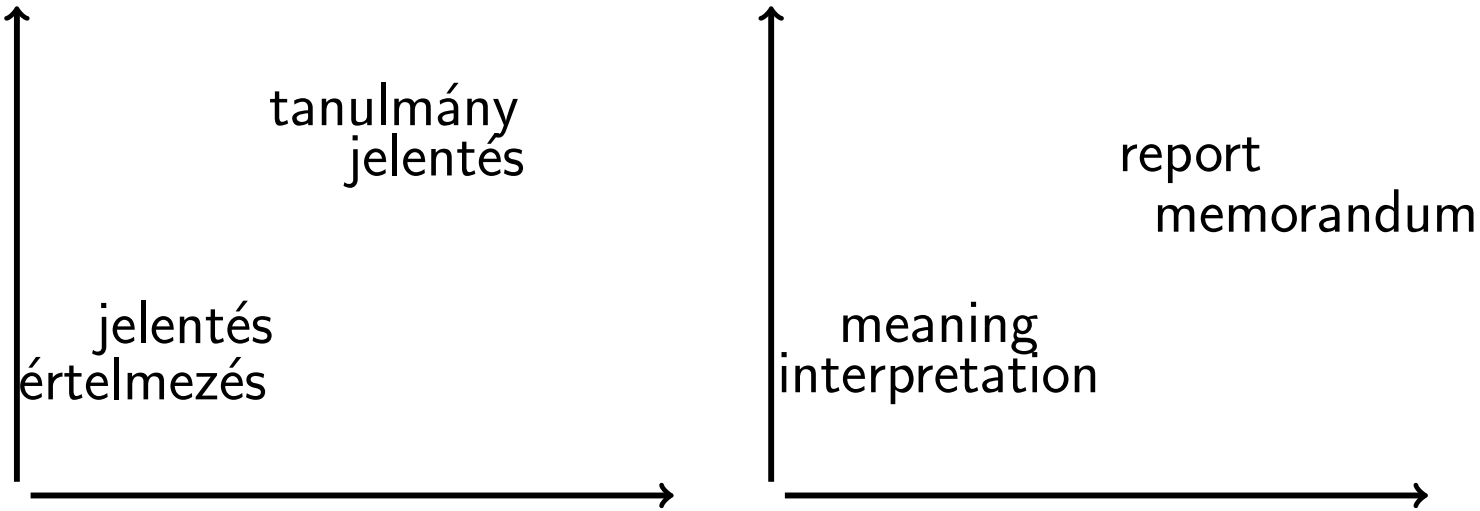
- VSM tools and pre-trained English models:
  - word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014), gensim (Řehůřek and Sojka, 2010)
- triangle scoring
  - seed dictionary extracted from Wiktionary by wikt2dict (Ács et al., 2013)
  - translational mapping: Dinu et al. (2015) forked `https://github.com/makrai/dinu15/`
  - evaluated against dictionaries extracted from parallel corpora (Tiedemann, 2012)
- MPVSM: AdaGram (Bartunov et al., 2015)

# exper 1: triangle scoring



# exper 2: linear mapping between MPVSMs

- idea




- preliminary results are poor $\Leftarrow$ prototypes don't match intuition

# url

http://corpus.nytud.hu/efnilex-vect/

makrai@nytud.hu

Judit Ács, Katalin Pajkossy, and András Kornai. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. In *LREC 2009*, volume 3, pages 209–226, 2009.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. Breaking sticks and ambiguities with adaptive skip-gram. *ArXiv preprint*, 2015.

Vladimír Benko. Aranea: Yet another family of (comparable) web corpora. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text and Speech and Dialogue. 17th International Conference, TSD 2014*, pages 257–264. Springer International Publishing Switzerland, 2014. ISBN 978-3-319-10815-2.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, 2014.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 2011.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *ICLR 2015, Workshop Track*, 2015.

Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. Creating open language resources for Hungarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 203–210. ELRA, 2004.

G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29: 82–97, 2012.

M. Hnátková, M. Křen, P. Procházka, and H. Skoumalová. The syn-series corpora of written czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 160–164. ELRA, 2014. ISBN 978-2-9517408-8-4.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 873–882, Jeju Island, Korea, 2012. Association for Computational Linguistics.

A. Krizhevsky and G. Sutskever, I.and Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS'2012*, 2012.

Nikola Ljubešić and Tomaž Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer, 2011.

Nikola Ljubešić and Filip Klubička. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *Proceedings of the ICLR 2013*, 2013a.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. Xiv preprint arXiv:1309.4168, 2013b.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*, 2014.

Csaba Oravecz, Tamás Váradi, and Bálint Sass. The Hungarian Gigaword Corpus. In *Proceedings of LREC 2014*, 2014.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. URL `http://is.muni.cz/publication/884893/en`.

Joseph Reisinger and Raymond J Mooney. Multi-prototype vector-space models of word meaning. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics, 2010.

Xabier Saralegi, Iker Manterola, and Iñaki San Vicente. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics, 2011.

Richard Socher, Eric H Huang, Jeffrey Pennington, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, pages 801–809, 2011.

I. Sutskever, O. Vinyals, and Le. Q. V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

Kumiko Tanaka and Kyoji Umemura. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 297–303. Association for Computational Linguistics, 1994.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, editor, *LREC*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.