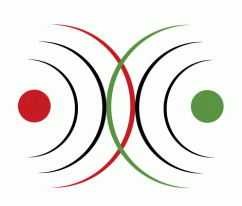
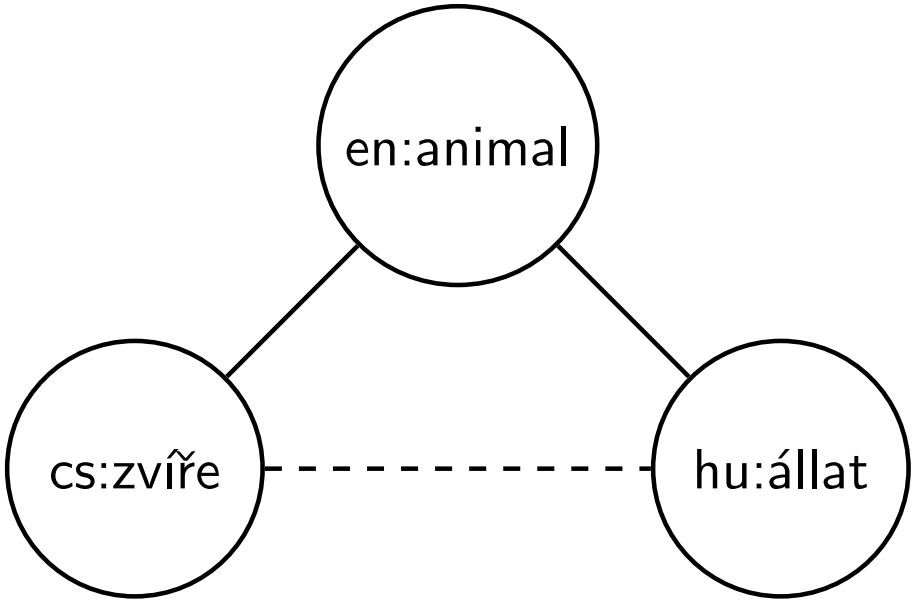
Filtering Wiktionary triangles by linear mapping between distributed word models

Márton Makrai
makrai@nytud.hu



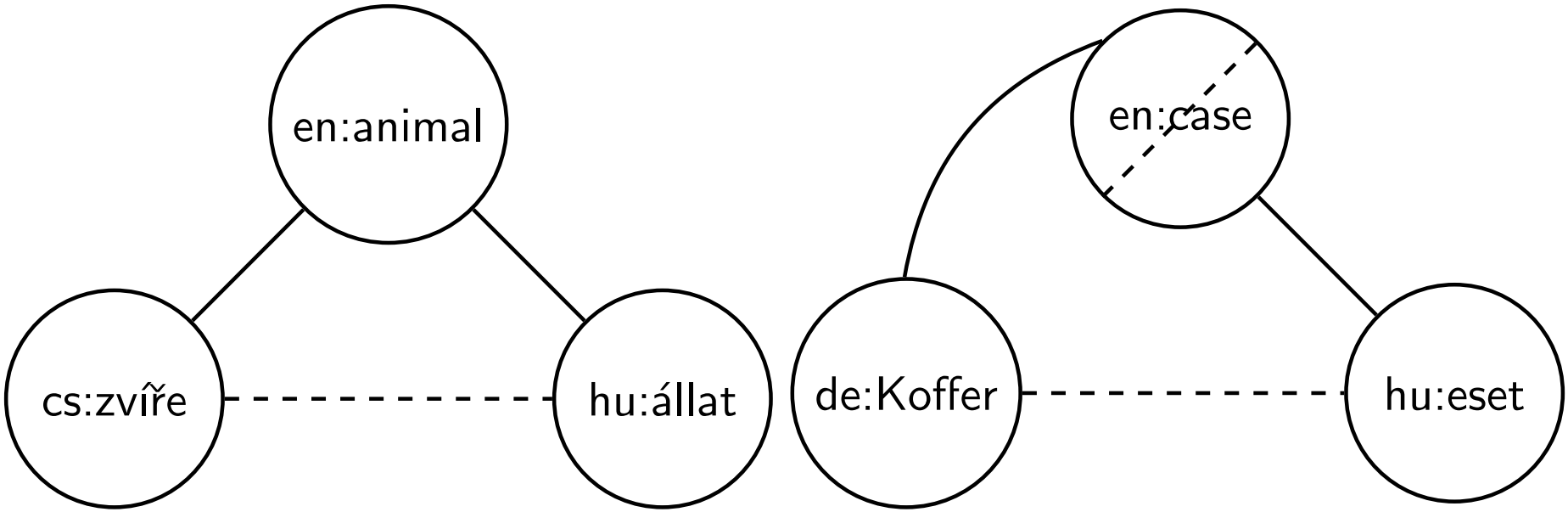
LREC 2016

Triangulation



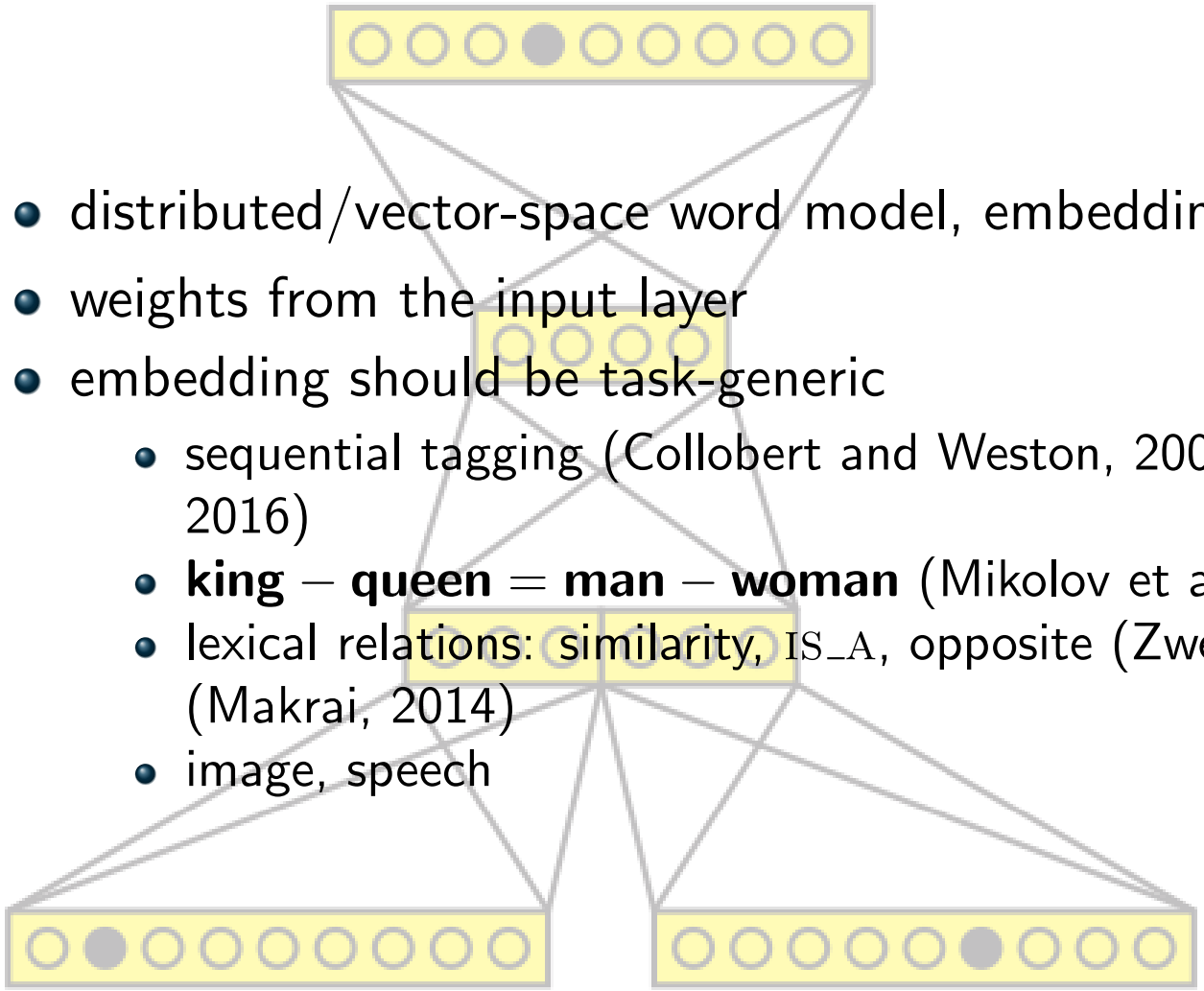
- pruning triangles
  - number of pivots (Tanaka and Umemura, 1994)
  - based on distributional similarity
    - comparable corpora (Saralegi et al., 2011)
    - now: with word embeddings, smoother

Triangulation



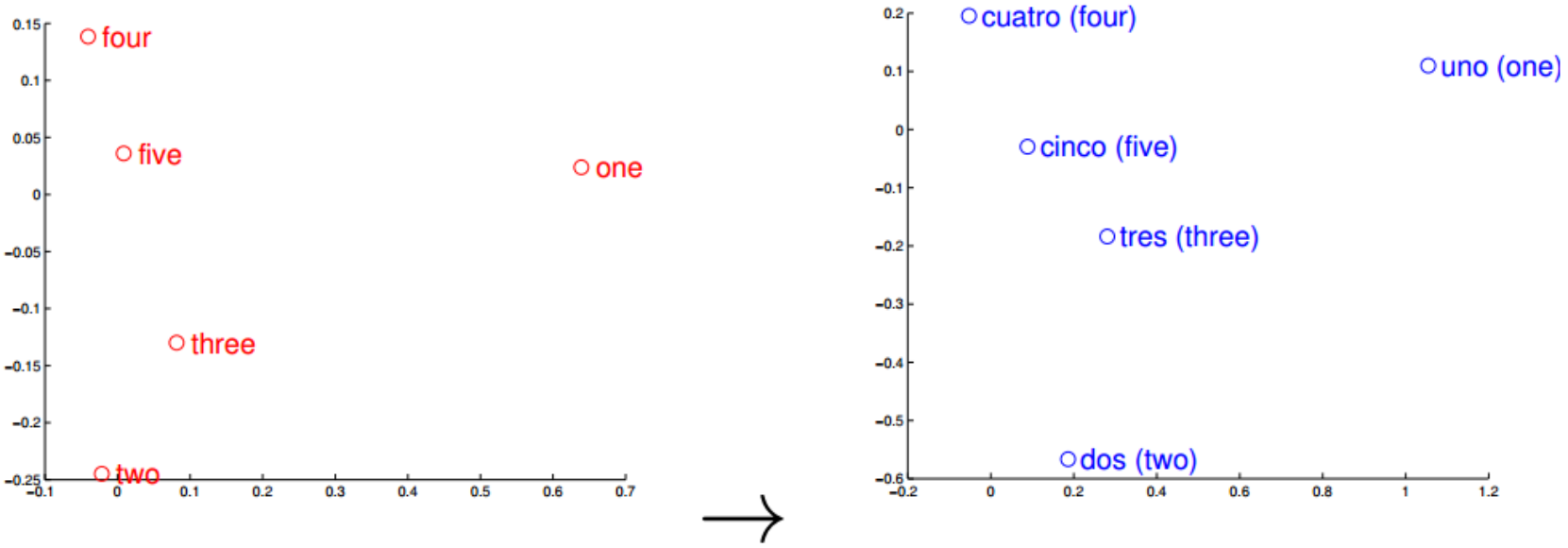
- pruning triangles
  - number of pivots (Tanaka and Umemura, 1994)
  - based on distributional similarity
    - comparable corpora (Saralegi et al., 2011)
    - now: with word embeddings, smoother

Embeddings



- distributed/vector-space word model, embedding
- weights from the input layer
- embedding should be task-generic
  - sequential tagging (Collobert and Weston, 2008; Pajkossy and Zséder, 2016)
  - **king** − **queen** = **man** − **woman** (Mikolov et al., 2013d; Makrai, 2015)
  - lexical relations: similarity, IS_A, opposite (Zweig, 2014), cause (Makrai, 2014)
  - image, speech

Linear 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)

Data

- seed dictionary extracted from Wiktionary with `https://github.com/juditacs/wikt2dict` handles 43 editions
- embeddings
  - German
    - SdeWaC (Baroni et al., 2009)
  - Hungarian
    - Webkorpusz (Halácsy et al., 2004) $\oplus$ Hungarian National Corpus (Oravecz et al., 2014)

Data

- seed dictionary extracted from Wiktionary with `https://github.com/juditacs/wikt2dict` handles 43 editions
- embeddings
  - German
    - SdeWaC (Baroni et al., 2009)
    - continuous bag of words (cbow) , 300 dim, cut-off 100
  - Hungarian
    - Webkorpusz (Halácsy et al., 2004) $\oplus$ Hungarian National Corpus (Oravecz et al., 2014)
    - cbow, 600 dim, cut-off 10
    - embeddings trained with word2vec (Mikolov et al., 2013a,c)
- translation matrix
  - tool: we forked Dinu et al. (2015) to `https://github.com/makrai/dinu15/`
  - trained on the 5 K direct word pairs that are supported by the most pivots in Wiktionary

Evaluation

- two rankings
- gold dictionary from the OPUS project extracted by (Tiedemann, 2012) from the OpenSubtitles2013 parallel corpus,
  - a collection of translated movie subtitles `http://www.opensubtitles.org/` in 59 languages

documents	3208
sentences	3.2 M
German tokens	23.3 M
Hungarian tokens	19.7 M
extracted word pairs	29.1 K

Table: The German Hungarian section of the OpenSubtitles2013

Filtering

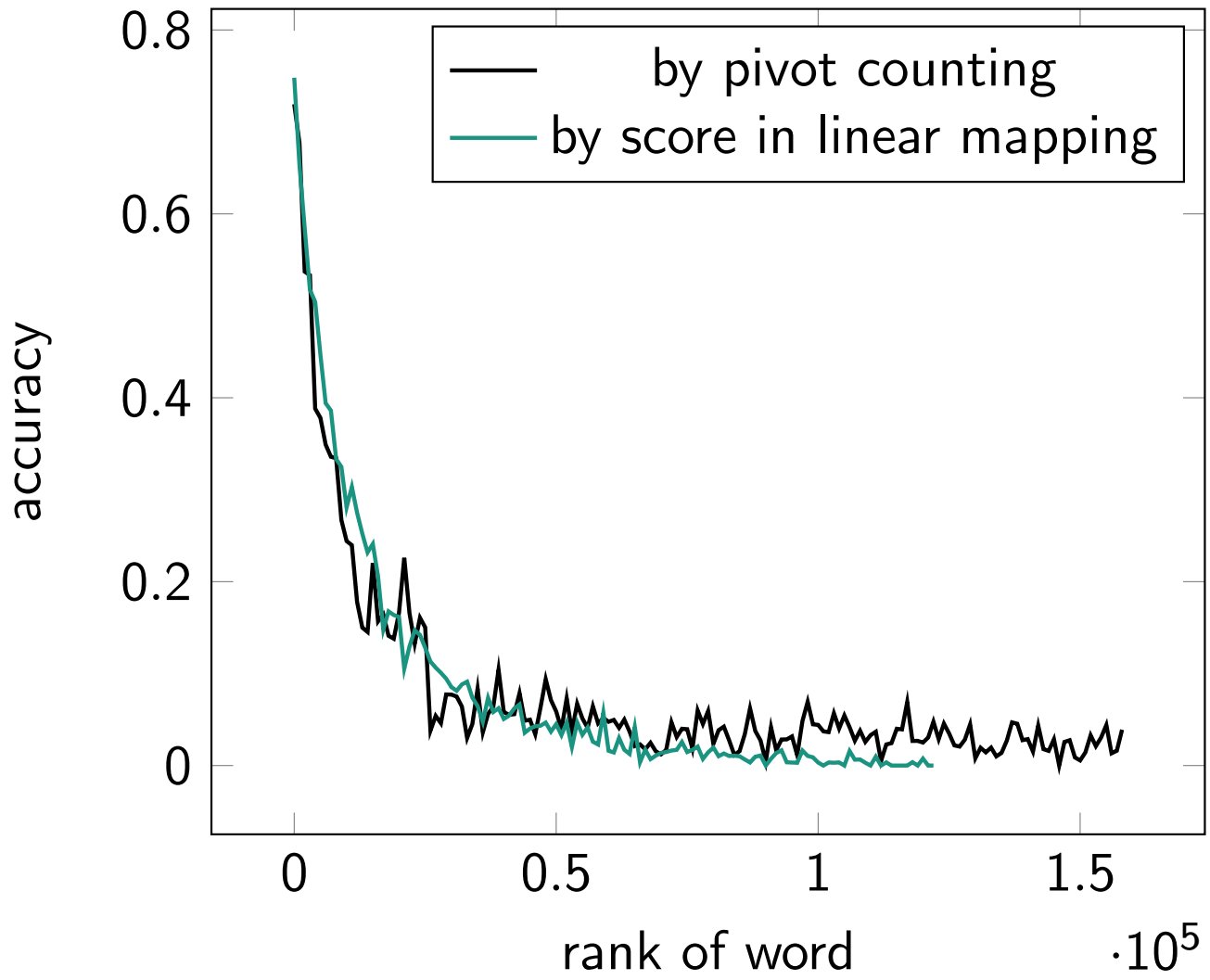
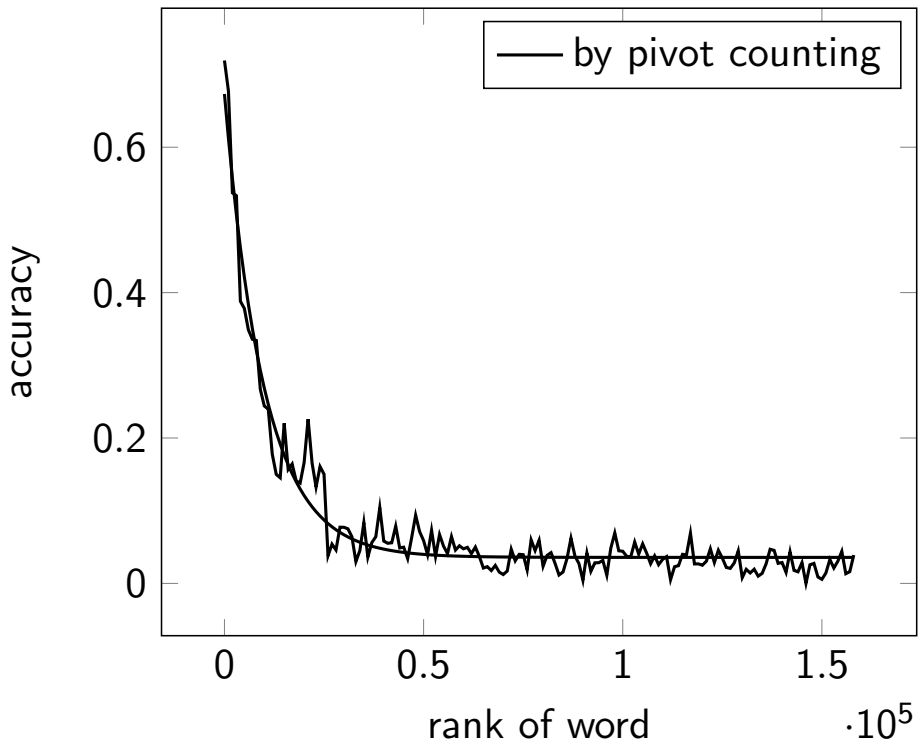
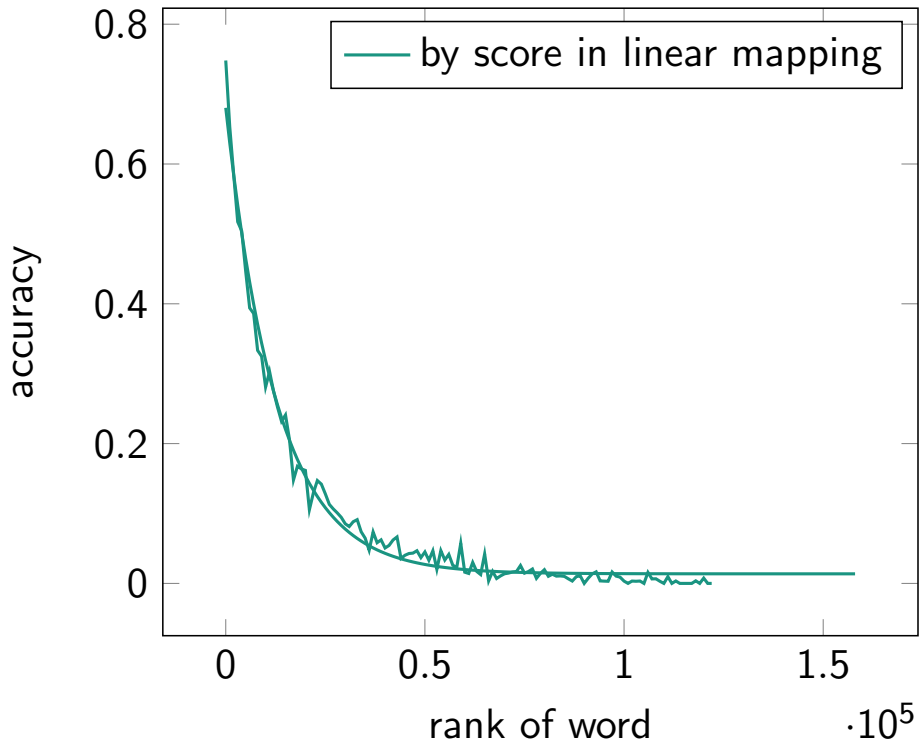


Figure: Accuracy curves in two ranking

Smoothness



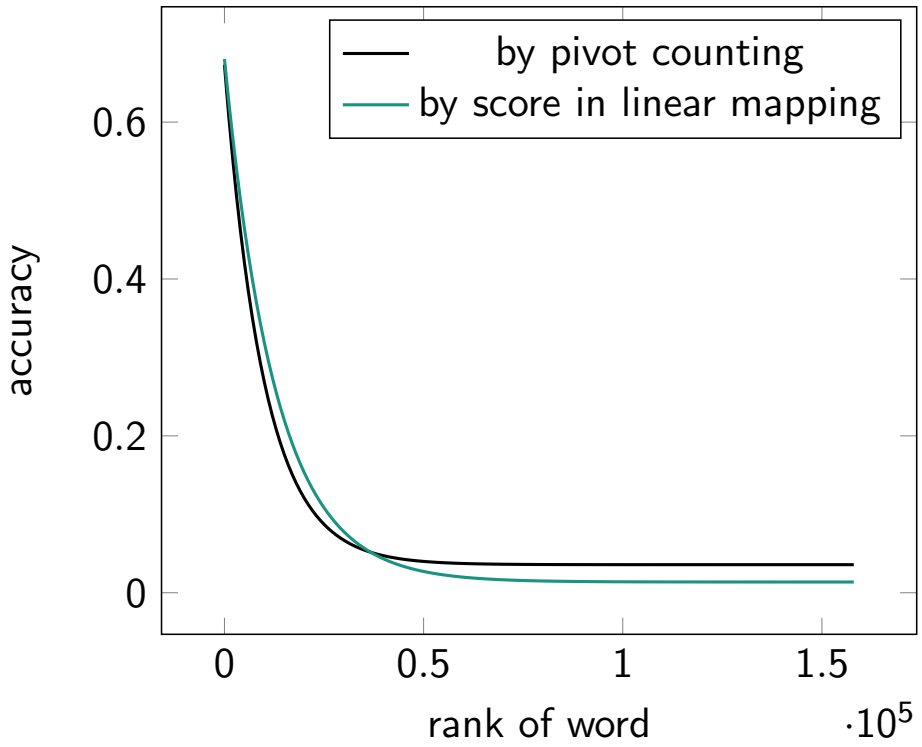


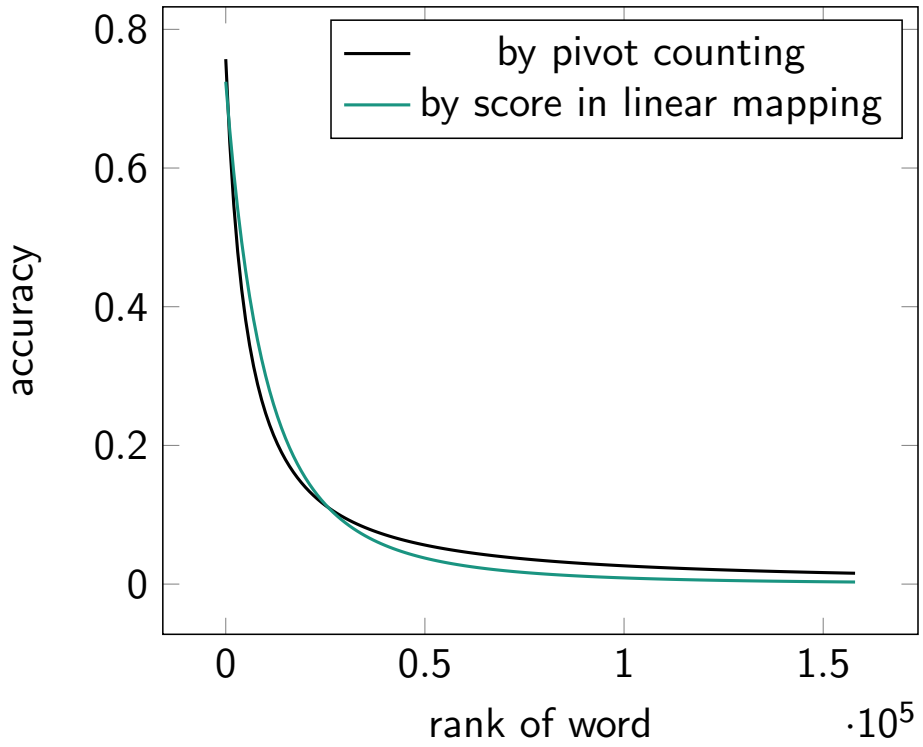
- mean squared error from
- exponential, $a \cdot \exp(-bx) + c$
- power law, $a \cdot (bx + c)^k$

scoring method	exp	power law
pivot counting	6.1859e-04	5.2182e-04
linear mapping	**2.4574e-04**	**1.1789e-04**
ratio	2.51	4.42

Table: Mean squared error

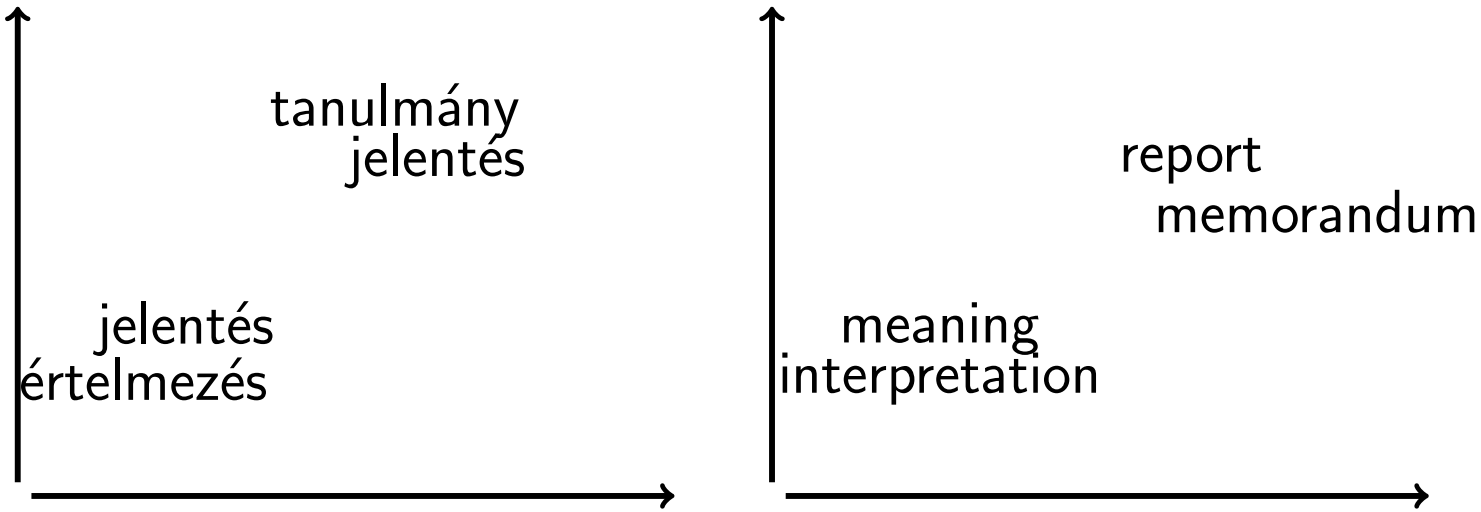
Top 20–30 000 pairs





Perspectives

- multi-sense embedding (Reisinger and Mooney, 2010; Huang et al., 2012)
- non-uniform number of senses (Neelakantan et al., 2014; Bartunov et al., 2015)
- in some NLP tasks (Li and Jurafsky, 2015)
- granularity (Borbély, Makrai, Nemeskey, and Kornai, submitted to repevalacl16)
  - number of senses correlated traditional lexicons
  - different sense $\Leftrightarrow$ different translation



URL

- `http://corpus.nytud.hu/efnilex-vect/`
- makrai@nytud.hu
- Work supported by the EFNILEX project of the European Federation of National Institutions for Language.

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.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA, 2008. ACM.

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.

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.

Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? In *EMNLP*, 2015.

Márton Makrai. Comparison of distributed language models on medium-resourced languages. In Attila Tanács, Viktor Varga, and Veronika Vincze, editors, *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015)*, 2015. ISBN 978-963-306-359-0.

Márton Makrai. Causality in vectors space language models. In *Spring Wind*, volume 6, pages 192–200. Association of Hungarian PhD and DLA Students (DOSZ), 2014.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013c. URL `http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-c` `pdf`.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 746–751, Atlanta, Georgia, 2013d. Association for Computational Linguistics.

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.

Katalin Pajkossy and Attila Zséder. The hunvec framework for nn-crf-based sequential tagging. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

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.

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.

Geoffrey Zweig. Explicit representation of antonymy in language modeling. Technical report, Microsoft Research, 2014.

German	OpenSubtitles2013	triangulated
Ernährung ‘nutrition’	táplálkozás ‘nutrition’	diéta ‘diet’
ekelhaft ‘disgusting’	gusztustalan ‘disgusting’	koszos ‘dirty’
Aufnahme	felvételt, felvétel	átvétel
Folge (noun)	kövesd (verb)	eredmény (noun)
Terror	terror	rettegés ‘dread’
strikt	szigorú**ak**	merev
dunkel (adj)	sötét (adj)	sötétedés (noun)
dünn	vékony, sovány	flamingó ‘flamenco’
Demonstration	bemutató	bemutatás (action noun)
Ablenkung ‘relief’	elterelés ‘distraction’	szünet ‘pause’
Rüssel	ormány ‘trunk’	szaglás ‘smelling (sense)’
Geruchssinn ‘smelling’	szaglás**od**	kürt ‘horn’
Koffer	bőröndöt, bőrönd	**eset** ‘de:Fall’
Verbindung ‘connecion’	kapcsolat ‘connecion’	kapcsolattartó ‘contact (person)’
Uhr ‘clock’	kor, óra ‘clock’	karóra ‘watch’
absorbieren ‘absorbieren’	nyelni ‘absorb’	furcsa ‘strange’
Schwule ‘gay’	meleg**ek** ‘gays’	sor ‘line, row, queue’
Hubschrauber ‘helicopter’	helikoptert, helikopter	húsbárd ‘chopper (for meat)’

Domain

- web corpora (SdeWaC, the Hungarian Webcorpus)
- "curated" corpus (the Hungarian National Corpus, 754 million words)
- Wiktionary: crowd-sourced but otherwise causal dictionary
- reference dictionary extracted from movie subtitles
- domain mismatch is negligible