# Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering

Clayton Greenberg, Asad Sayeed, and Vera Demberg
Saarland University / M$^2$CI Cluster of Excellence

June 1, 2015

# Thematic fit

| | |
|---|---|
| Homer ate the donut with | his fingers<br>pliers<br>sprinkles<br>a friend |

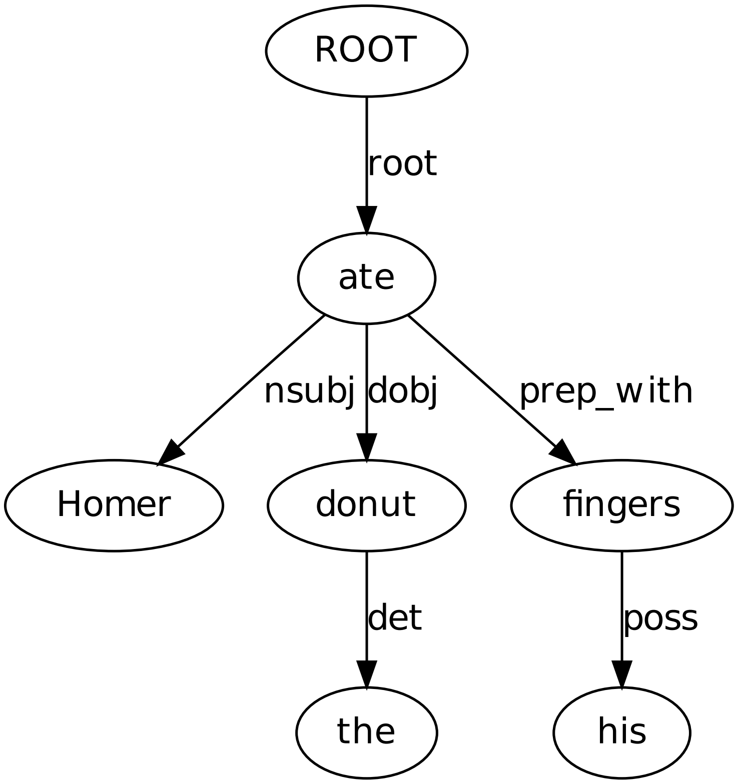
# Thematic fit judgements

Ferretti et al. (2001): "[On a scale from 1 to 7, h]ow common is it to use each of the following to perform the action of eating?"

| | |
|---|---|
| cup | 3.3 |
| fork | 6.7 |
| knife | 6.3 |
| napkin | 3.8 |
| pliers | 1.0 |
| spoon | 6.3 |
| toothpick | 2.1 |

# Estimating thematic fit (1/3)

Baroni & Lenci (2010): Distributional Memory (DM) approach:

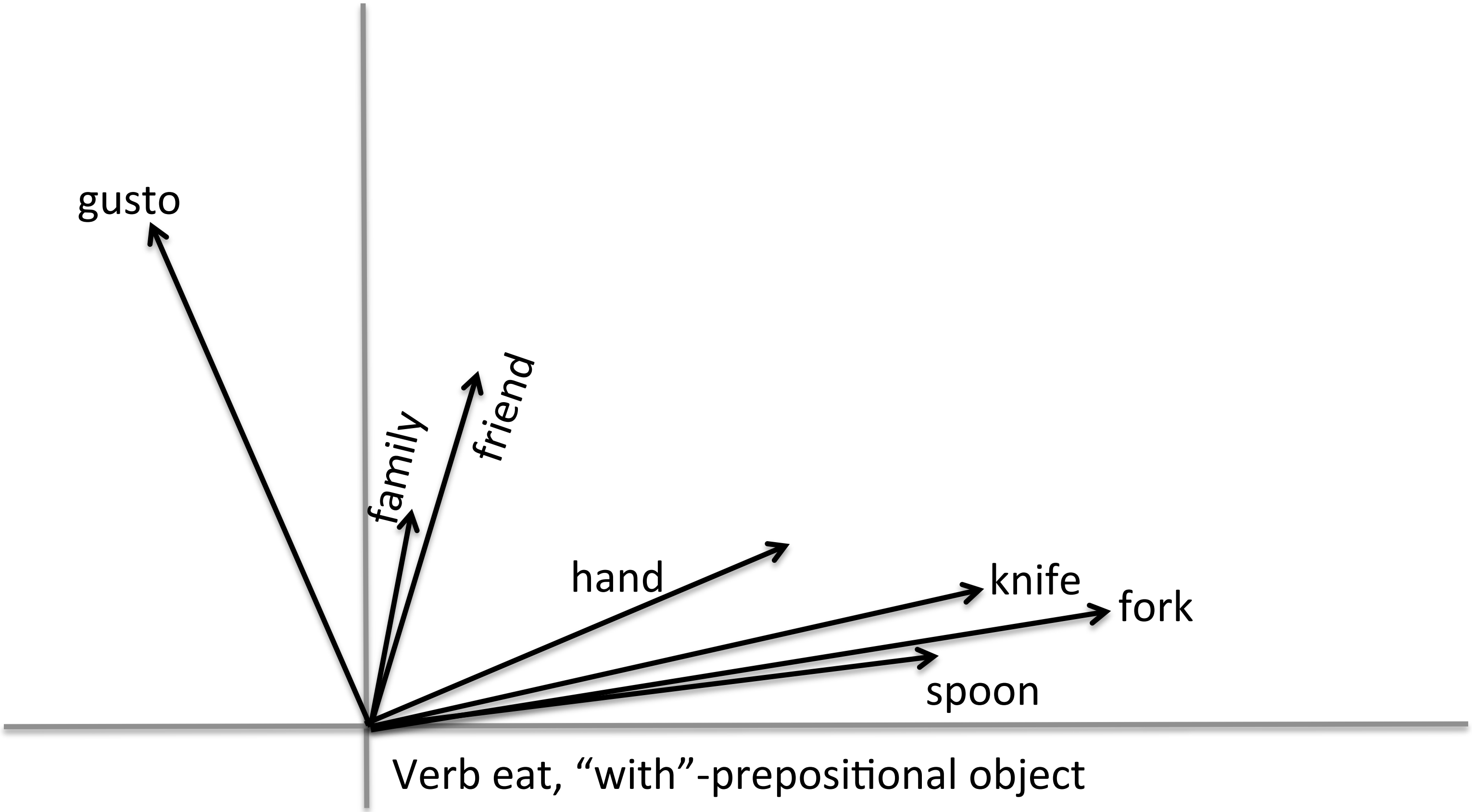



Tree generated at http://eztreesee.coli.uni-saarland.de/ which uses the Stanford Dependency Parser (de Marneffe et al., 2006)

(1) Count verb-role-filler triples and adjust counts by local mutual information (LMI)

# Estimating thematic fit (2/3)

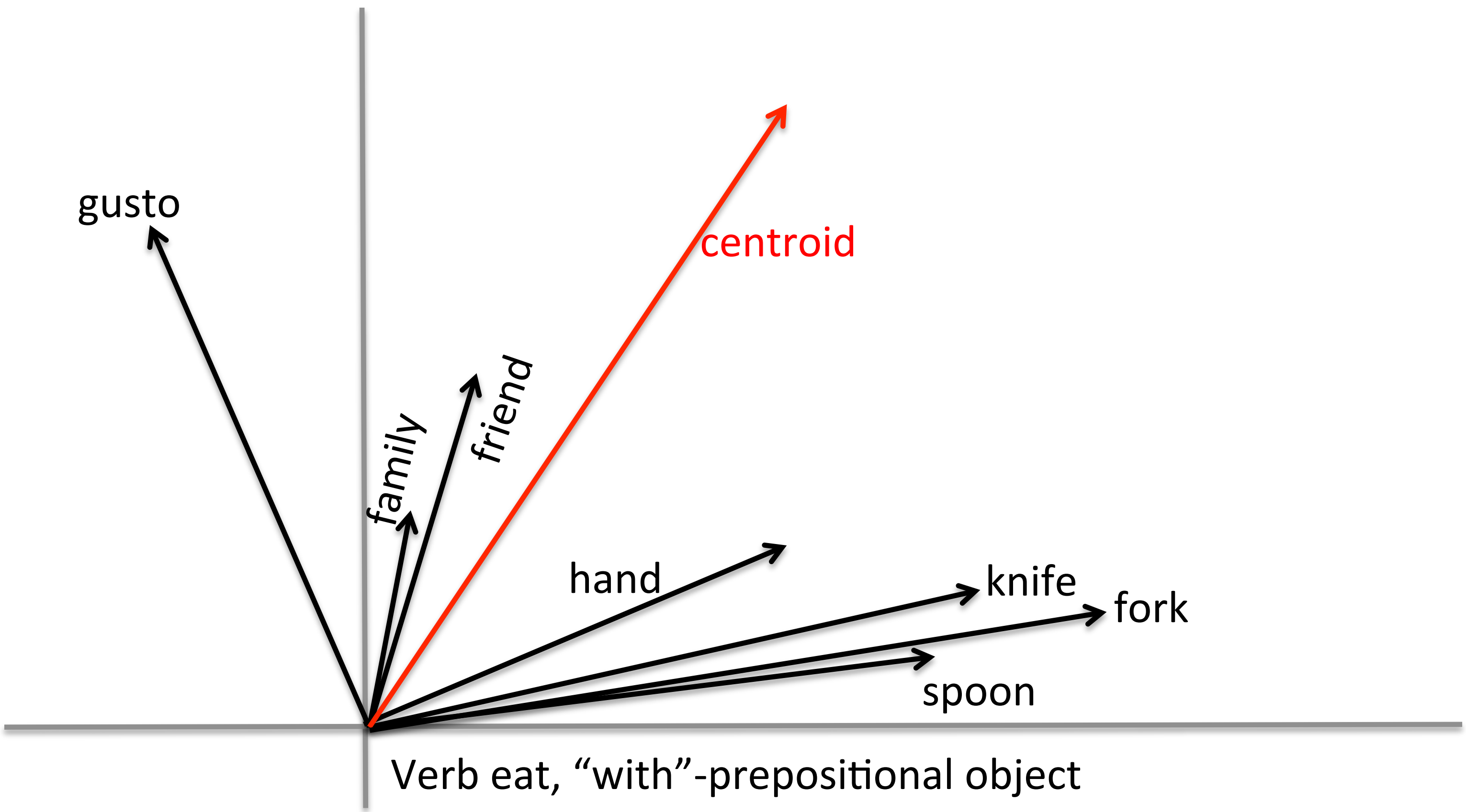
Baroni & Lenci (2010): Distributional Memory (DM) approach:




(2) Query the top 20 highest scoring fillers and compute the centroid

# Estimating thematic fit (2/3)

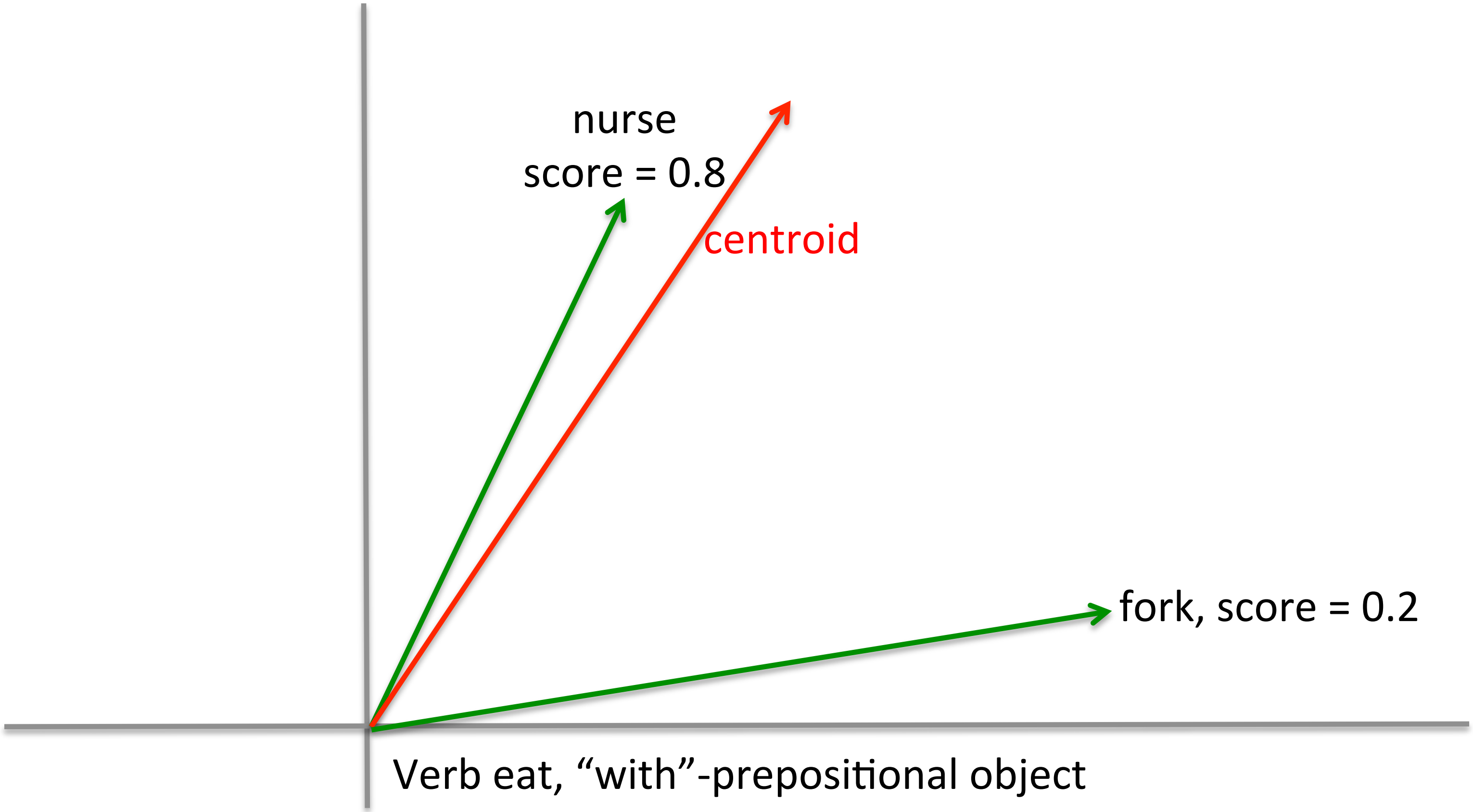
Baroni & Lenci (2010): Distributional Memory (DM) approach:




(2) Query the top 20 highest scoring fillers and compute the centroid

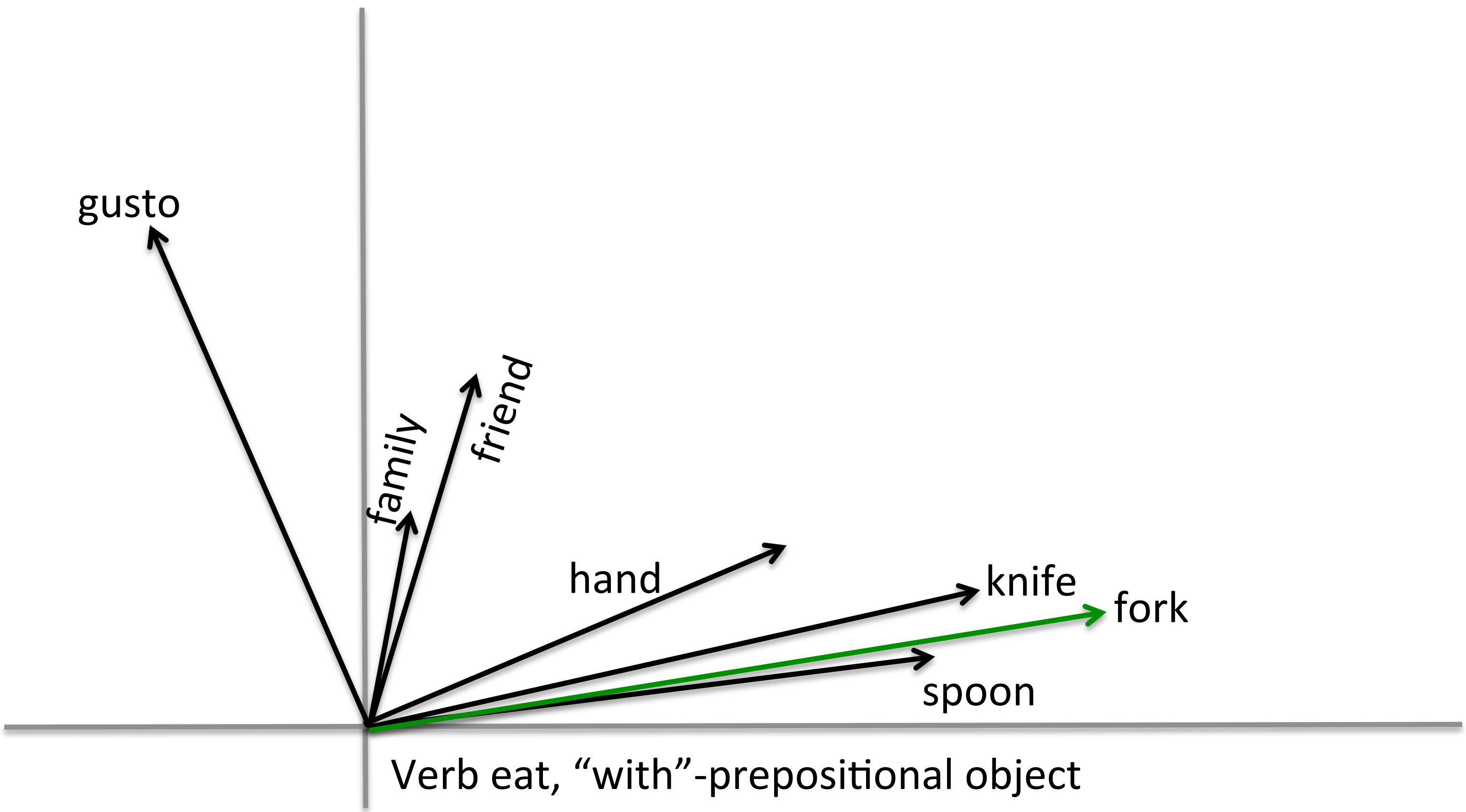
# Estimating thematic fit (3/3)

Baroni & Lenci (2010): Distributional Memory (DM) approach:

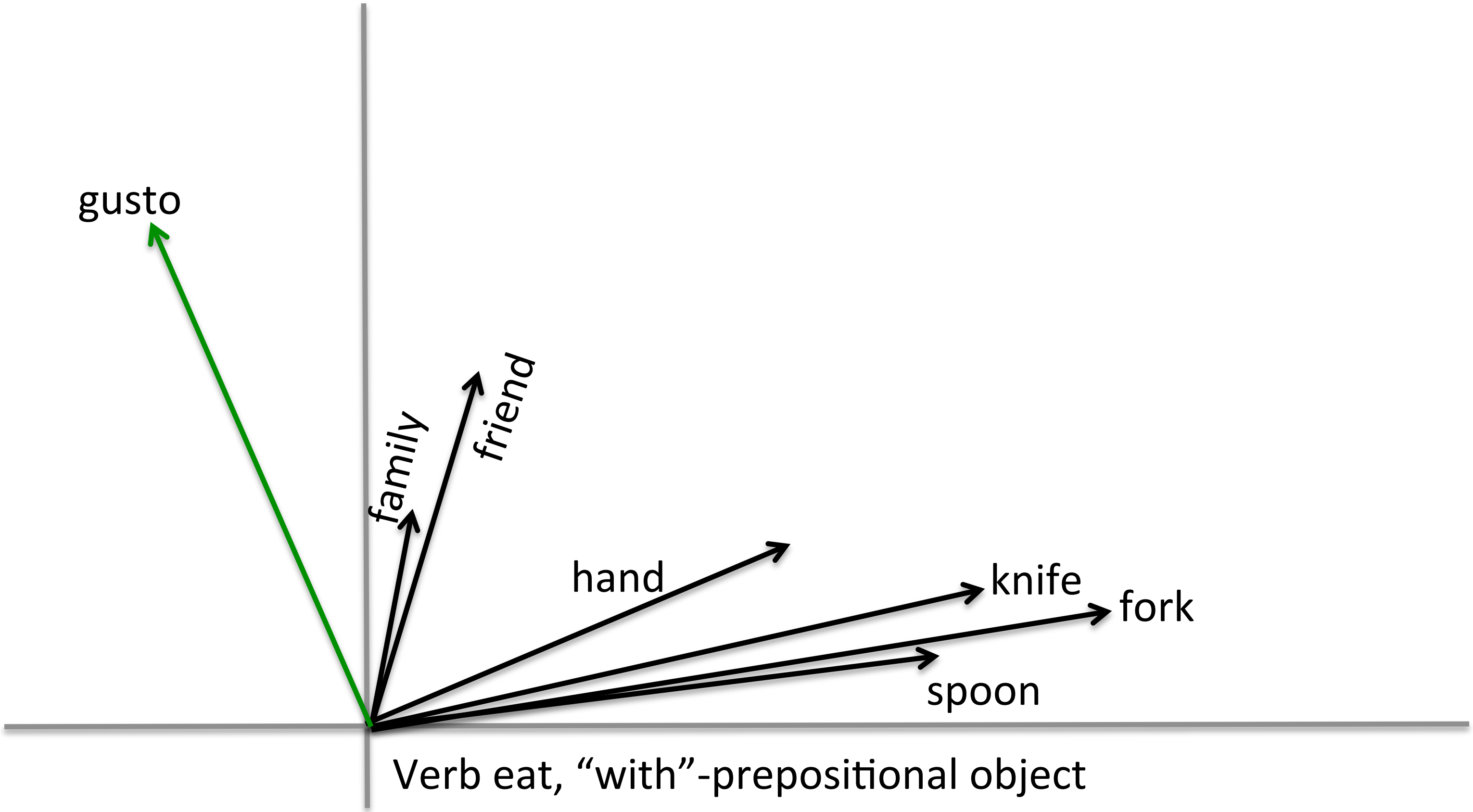



(3) Return cosine similarity of test role-filler and centroid

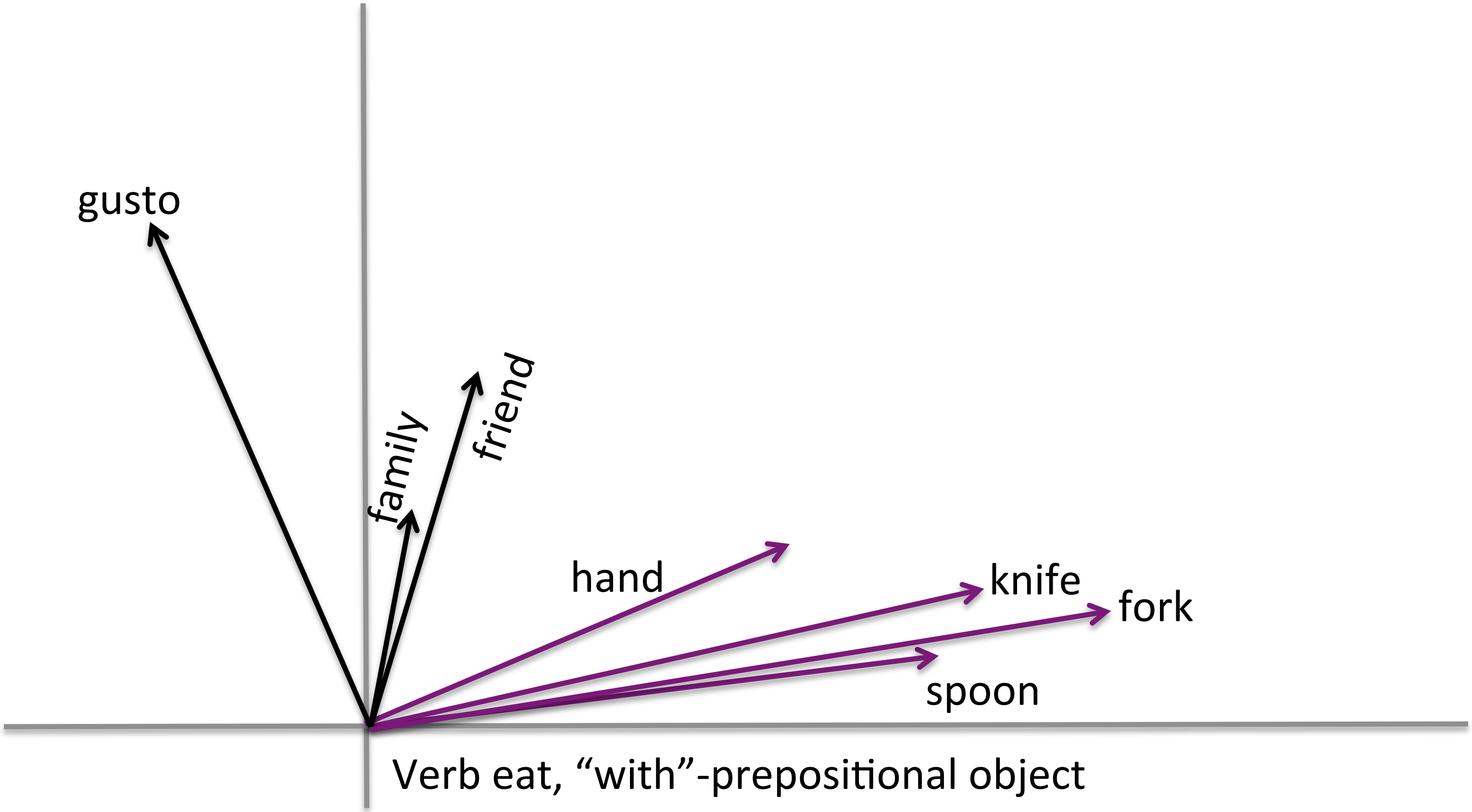
# The OneBest method (1/2)



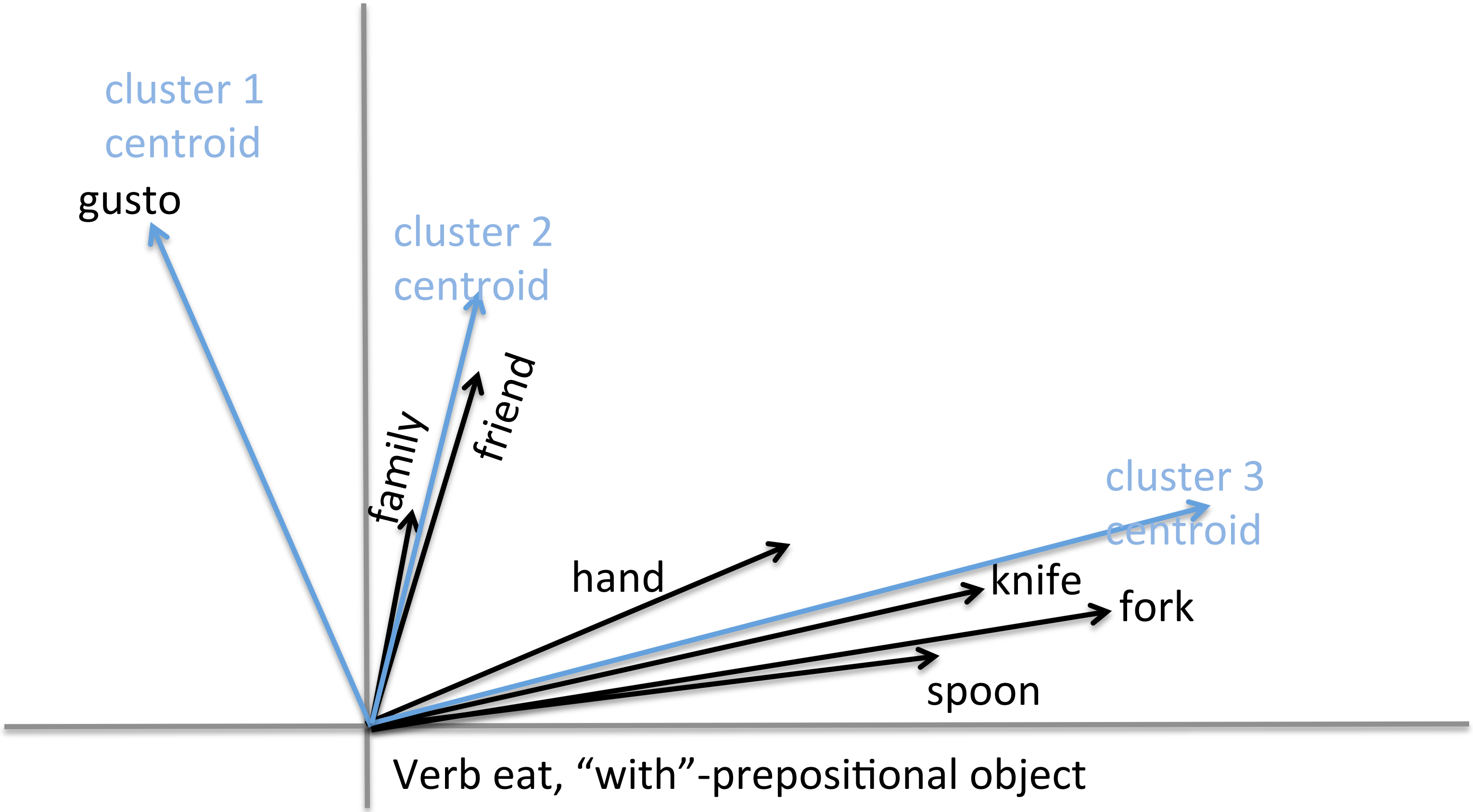
# The OneBest method (2/2)



# Prototype clustering (1/2)



# Prototype clustering (2/2)



# kClusters

- Use hierarchical agglomerative clustering package from NLTK (Bird et al., 2009).
- To set number of clusters, use variance ratio criterion (VRC) (Calinski and Harabasz, 1974).

$$VRC(k) = \frac{SS_B}{k-1} / \frac{SS_W}{n-k}$$

$$\omega_k = (VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1})$$

- VRC cannot evaluate fewer than 3 clusters, capped at 10 clusters.
- 2Clusters method forces 2 clusters for all verb-roles.

# Overall results

| Method | Spearman's rho (TypeDM), range = [-1,1] |
|---|---|
| Centroid | 0.35 |
| OneBest | 0.36 |
| 2Clusters | 0.37 |
| kClusters | 0.39 |

Correlation between human judgements from the McRae et al. (1998), Ferretti et al. (2001), and Padó (2007) datasets and automatic scores using LMIs from TypeDM, by prototype generation method.

# Agents and patients results

| Method | Padó (2007) agents | Padó (2007) patients |
|---|---|---|
| Centroid | 0.54 | 0.53 |
| kClusters | 0.46 | 0.56 |

Correlation between human judgements from the Padó (2007) dataset, with agents and patients separated, and automatic scores using LMIs from TypeDM, by prototype generation method.

# Instruments results (1/2)

| Method | Spearman's rho (TypeDM) |
|---|---|
| Centroid | 0.36 |
| OneBest | 0.39 |
| 2Clusters | 0.39 |
| kClusters | 0.42 |

Correlation between human judgements on instruments from the Ferretti et al. (2001) dataset and automatic scores using LMIs from TypeDM, by prototype generation method.

# Instruments results (2/2)

| Method | SENNA-DepDM | TypeDM |
|---|---|---|
| Centroid | 0.19 | 0.36 |
| OneBest | 0.27 | 0.39 |
| kClusters | 0.34 | 0.42 |

Correlation between human judgements on instruments from the Ferretti et al. (2001) dataset and automatic scores using LMIs from SENNA-DepDM (Sayeed and Demberg, 2014) and TypeDM, by prototype generation method.

# Locations results

| Method | SDDMX | TypeDM |
|---|---|---|
| Centroid | 0.25 | 0.23 |
| OneBest | 0.28 | 0.24 |
| kClusters | 0.33 | 0.29 |

Correlation between human judgements on locations from the Ferretti et al. (2001) dataset and automatic scores using LMIs from TypeDM and a novel extension of SENNA-DepDM (SDDMX), by prototype generation method.

# Deep parameter tuning

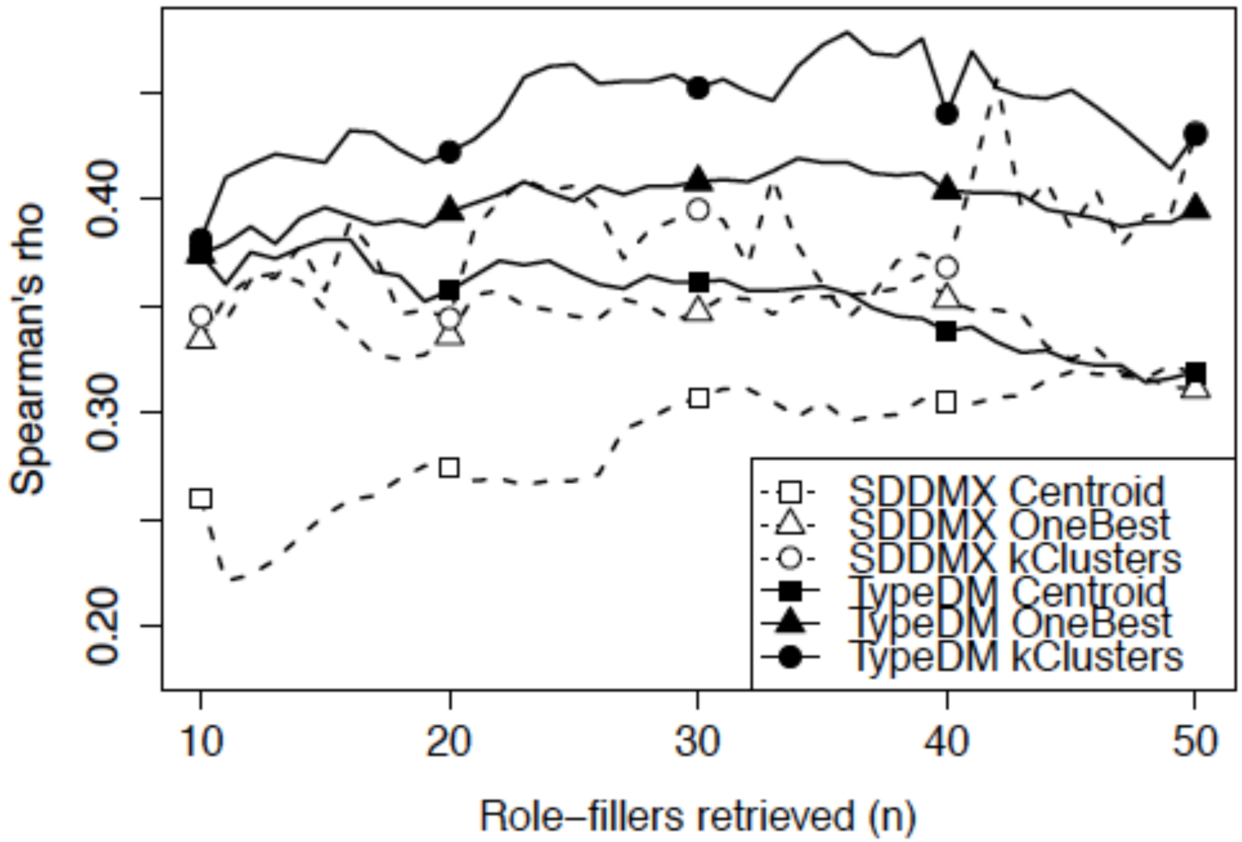



Figure 3: Spearman's $\rho$ values for Ferretti et al. (2001) instruments vs. the number of vectors retrieved.

# Future directions

- More sophisticated clustering
  - Expectation-maximization (generalize to weighted centroid)
  - Non-negative matrix factorization
  - Density peaks
- Knowledge-based number of senses
- More detailed modeling of predictions for method comparison
- Evaluation on a dataset that systematically varies polysemy and frequency

# Thank you!

Data from this project available at http://rollen.mmci.uni-saarland.de/

# All correlation results

| | Padó (2007) agents | | | McRae et al. (1998) agents | | | Ferretti et al. (2001) instruments | | |
|---|---|---|---|---|---|---|---|---|---|
| | *SDDM* | *SDDMX* | *TypeDM* | *SDDM* | *SDDMX* | *TypeDM* | *SDDM* | *SDDMX* | *TypeDM* |
| *Centroid* | 0.515 | 0.528 | **0.535** | 0.371 | 0.394 | 0.359 | 0.193 | 0.274 | 0.357 |
| *OneBest* | 0.321 | 0.324 | 0.464 | 0.375 | 0.376 | **0.431** | 0.274 | 0.336 | 0.394 |
| *2Clusters* | 0.489 | 0.412 | 0.522 | 0.367 | 0.373 | 0.370 | 0.252 | 0.331 | 0.388 |
| *kClusters* | 0.281 | 0.322 | 0.460 | 0.396 | 0.394 | 0.416 | 0.335 | 0.344 | **0.422** |
| | **Padó (2007) patients** | | | **McRae et al. (1998) patients** | | | **Ferretti et al. (2001) locations** | | |
| | *SDDM* | *SDDMX* | *TypeDM* | *SDDM* | *SDDMX* | *TypeDM* | *SDDM* | *SDDMX* | *TypeDM* |
| *Centroid* | 0.511 | 0.505 | 0.525 | 0.133 | 0.131 | 0.343 | 0.187 | 0.248 | 0.230 |
| *OneBest* | 0.447 | 0.467 | 0.509 | 0.214 | 0.233 | 0.307 | 0.234 | 0.276 | 0.244 |
| *2Clusters* | 0.526 | 0.498 | 0.551 | 0.175 | 0.166 | **0.353** | 0.294 | 0.249 | 0.235 |
| *kClusters* | 0.401 | 0.428 | **0.555** | 0.212 | 0.227 | 0.350 | 0.293 | **0.326** | 0.289 |
| | **All from Padó (2007)** | | | **All from McRae et al. (1998)** | | | **All datasets** | | |
| | *SDDM* | *SDDMX* | *TypeDM* | *SDDM* | *SDDMX* | *TypeDM* | *SDDM* | *SDDMX* | *TypeDM* |
| *Centroid* | 0.512 | 0.521 | 0.530 | 0.237 | 0.251 | 0.325 | 0.258 | 0.296 | 0.354 |
| *OneBest* | 0.385 | 0.395 | 0.482 | 0.273 | 0.287 | 0.345 | 0.275 | 0.304 | 0.359 |
| *2Clusters* | 0.508 | 0.458 | **0.532** | 0.252 | 0.256 | 0.336 | 0.287 | 0.289 | 0.366 |
| *kClusters* | 0.343 | 0.375 | 0.503 | 0.287 | 0.294 | **0.359** | 0.294 | 0.317 | **0.385** |

Table 3: Spearman's $\rho$ for each method on each dataset and on all datasets together, using the 20 highest ranked words per verb-role.

# References


Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus based semantics. *Computational Linguistics*, 36(4):673-721.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1):1-27.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493-2537.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516-547.

Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006*.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283-312.

Padó, U. (2007). The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing. PhD thesis, Saarland University.

Sayeed, A. and Demberg, V. (2014). Combining unsupervised syntactic and semantic models of thematic fit. In *Proceedings of the first Italian Conference on Computational Linguistics* (CLiC-it 2014).