Fitting shoes to polysemous feet:

multi-prototype vector-space thematic fit modeling

Clayton Greenberg

Department of Computational Linguistics and Phonetics, Saarland University

July 21, 2015

Outline

1. How do words fit?
   - Motivation
   - Polysemy hypotheses
   - Experimental design
   - Experiment results
2. A new modeling framework
   - Existing techniques
   - Framework architecture
   - Modeling results
   - Conclusions

Thematic fit

Alice played soccer / croquet / the harpsichord / the cheese in the garden with a flamingo.

Thematic roles

agent patient location instrument

Alice played croquet in the garden with a flamingo.

McRae et al. (1998) procedure for agents

How common is it for a

- snake
- nurse
- monster
- baby
- cat

to **frighten** someone/something?

McRae et al. (1998) procedure for patients

How common is it for a

- snake
- nurse
- monster
- baby
- cat

to **be frightened by** someone/something?

Datasets of human judgements

verbal	role-filler	thematic role	score
advise	doctor	`Arg0`	6.8
advise	doctor	`Arg1`	4.0
caution	friend	`Arg0`	5.6
caution	friend	`Arg2`	5.0
confuse	baby	`Arg0`	3.7
confuse	baby	`Arg1`	6.0
eat	lunch	`Arg0`	1.1
eat	lunch	`Arg1`	6.9
kill	lion	`Arg0`	2.7
kill	lion	`Arg1`	4.9
kill	man	`Arg0`	3.4
kill	man	`Arg1`	5.4

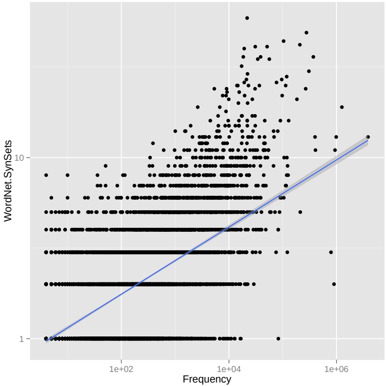
Sample of judgements from Padó (2007).

Polysemy

First pass: meanings per verb
"play": "croquet", "harpsichord"

Also track *Fit* variable: how the role-filler fits (e.g. SENSE1, SENSE2, BAD)

Role-fillers are shoes.
What happens with POLYSEMOUS feet?



Polysemy versus frequency of the most frequent verbs in COCA. Corpus obtained from Davies (2008).

Sense frequency

How common is it for croquet/the harpsichord to be played?

WordNet (Fellbaum, 1998) orders SynSets based on their frequencies.

`play_1`: participate in games or sport.
“We played hockey all afternoon”; “play cards”; “Pele played for the Brazilian teams in many important matches”

`play_7`: perform music on (a musical instrument).
“He plays the flute”; “Can you play on this old recorder?”

The Equal Sense Hypothesis

Definition

The thematic fit value for a POLYSEMOUS verb is the arithmetic mean of the thematic fit values for each individual sense.

$$\begin{aligned} &\textit{thematicFit}(\texttt{patient}(\text{“play”}\ \ (\text{“croquet”}))) = \\ 0.5\times &\textit{thematicFit}(\texttt{patient}(\textit{PLAY}_1(\text{“croquet”}))) + \\ 0.5\times &\textit{thematicFit}(\texttt{patient}(\textit{PLAY}_2(\text{“croquet”}))) \end{aligned}$$

Predictions

- POLYSEMOUS $\rightarrow$ ratings towards the middle of the scale
- Symmetrical ratings $\rightarrow$ no main effect of *Polysemy*
- No difference between more frequent and less frequent senses

The Autonomous Sense Hypothesis

Definition

The thematic fit value for a POLYSEMOUS verb is inherited from the thematic fit value for the most appropriate sense given the role-filler, irrespective of the number or distribution of verb senses.

$$\textit{thematicFit}(\texttt{patient}(\text{“play”}\ \ (\text{“croquet”}))) =$$
$$\textit{thematicFit}(\texttt{patient}(\textit{PLAY}_2(\text{“croquet”})))$$

Predictions

- More POLYSEMOUS $\rightarrow$ higher ratings
- Main effect of *Polysemy* does not change over the scale
- No difference between more frequent and less frequent senses

The Sense Frequency Hypothesis

Definition

Each sense of the verb contributes a share of the thematic fit value, weighted by its relative frequency, not conditioned by the role-filler.

$$\begin{aligned} \mathit{thematicFit}(\texttt{patient}(\text{“play”}\ \ (\text{“croquet”}))) &= \\ 0.8 \times \mathit{thematicFit}(\texttt{patient}(\mathit{PLAY}_1(\text{“croquet”}))) &+ \\ 0.2 \times \mathit{thematicFit}(\texttt{patient}(\mathit{PLAY}_2(\text{“croquet”}))) & \end{aligned}$$

$$\mathit{senseEntropy}(\mathit{verb}) = -\sum_{s \in \mathit{Senses}} p(s) \log_2 p(s)$$

Predictions

- High sense entropy $\rightarrow$ Sense Frequency H. $\approx$ Equal Sense H.
- Large effect of *Sense*, small effect of *Polysemy*
- *Polysemy* should interact with *Fit*

The Conditioned Sense Hypothesis

Definition

Create custom sense distributions conditioned on the sense frequencies and the plausibilities of the role-filler in each sense.

$$\begin{aligned} \textit{thematicFit}(\texttt{patient}(\text{``play''}\ (\text{``croquet''}))) &= \\ 0.3\times\textit{thematicFit}(\texttt{patient}(\textit{PLAY}_1(\text{``croquet''}))) &+ \\ 0.7\times\textit{thematicFit}(\texttt{patient}(\textit{PLAY}_2(\text{``croquet''}))) & \end{aligned}$$

Predictions

- High sense entropy $\rightarrow$ Conditioned Sense H. $\approx$ Equal Sense H.
- Small effect of *Sense*, small effect of *Polysemy*
- *Polysemy* should interact with *Fit*

Existing dataset analysis

Predictor	**Est.**	**Std. Err.**	$t(1439)$	**Sig. level**
LOGVERBPOLYSEMY	-0.15	0.08	-1.89	.
LOGVERBFREQUENCY	0.13	0.04	3.12	**
LOGNOUNPOLYSEMY	-0.09	0.08	-1.08	
LOGNOUNFREQUENCY	0.12	0.03	3.84	***

A linear model of McRaeNN thematic fit ratings based on polysemy and frequency of both verbs and nouns, $\Delta r^2 = 0.01846$.

Existing datasets: stimuli selection

McRaeNN	Padó (2007)
Many purposes	One purpose
Many verbs have “well-defined” roles	Verbs are most frequent in Penn Treebank and FrameNet
Many role-fillers selected to fit their roles well	Role-fillers selected to have a wide range of fit ratings
Animate role-fillers preferred	Fully mixed animacy
146 verbs	18 verbs
1,444 (F,R,V) triples	414 (F,R,V) triples

New formulation of the task

How common is it for croquet/soccer to be played?

The relative unigram frequencies of “croquet”, “soccer”, and “harpsichord” over the years 1820 to 2000 in the Google Books corpus (Michel et al., 2011).

New formulation of the task

Agreement scale: croquet is *something* that is played.

The relative unigram frequencies of “croquet”, “soccer”, and “harpsichord” over the years 1820 to 2000 in the Google Books corpus (Michel et al., 2011).

Verb selection

- Start with 500,000 most common word forms in COCA.
- Filter for verbs.
- Lemmatize using the WordNet lemmatizer in NLTK (Bird et al., 2009).
- Filter for only those that retrieve exactly one SynSet.
- Sort by frequency.
- Choose the first 48 that fit the paradigm (transitive, etc...).

For each MONOSEMOUS verb

Find a POLYSEMOUS verb with similar unigram frequency.
(at least 2 salient senses, $\approx$ 7 SynSets)

Stimuli examples

Filler type	**Freq.**	*whip* (1686, 6 SynSets)	*punish* (2908, 1 SynSet)
SENSE1	HIGH	horse (32384)	criminal (9271)
	LOW	stallion (818)	outlaw (1487)
SENSE2	HIGH	cream (19727)	-
	LOW	frosting (905)	-
BAD	HIGH	party (118292)	criminal (9271)
	LOW	gathering (7025)	outlaw (1487)

- To find a good patient-filler, query COCA for: `VERB [at*] [nn*]`.
- Find a much higher or lower ($\approx 10\times$) frequency synonym.
- For POLYSEMOUS verbs, repeat for second sense.
- Randomly shuffle good patient-fillers to assign poor ones.
- Reshuffle all of the ones that are too good.

Stimuli examples

Filler type	Freq.	*whip* (1686, 6 SynSets)	*punish* (2908, 1 SynSet)
SENSE1	HIGH	horse (32384)	criminal (9271)
	LOW	stallion (818)	outlaw (1487)
SENSE2	HIGH	cream (19727)	-
	LOW	frosting (905)	-
BAD	HIGH	party (118292)	baby (70498)
	LOW	gathering (7025)	fetus (2329)

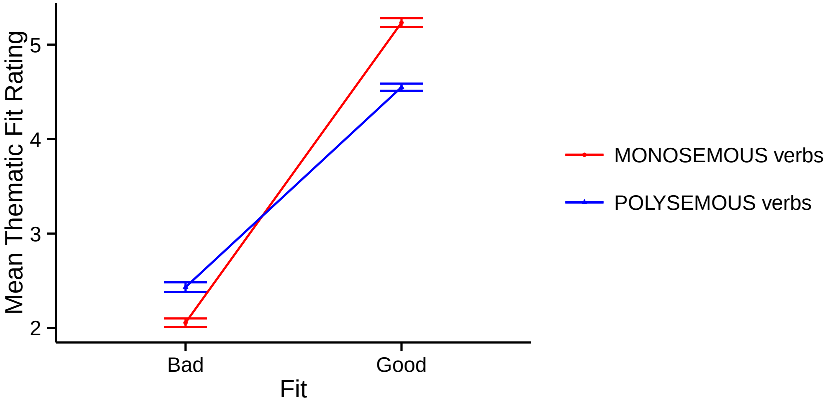
- To find a good patient-filler, query COCA for: `VERB [at*] [nn*]`.
- Find a much higher or lower ($\approx 10\times$) frequency synonym.
- For POLYSEMOUS verbs, repeat for second sense.
- Randomly shuffle good patient-fillers to assign poor ones.
- Reshuffle all of the ones that are too good.

Procedure

- Rewrite each verb in its past-participle form.
- Normalize each role-filler to singular with appropriate determiner.
- Choose either the `+human` or the `−human` template:
  - `+human: ___ is someone who is ___`
  - `−human: ___ is something that is ___`

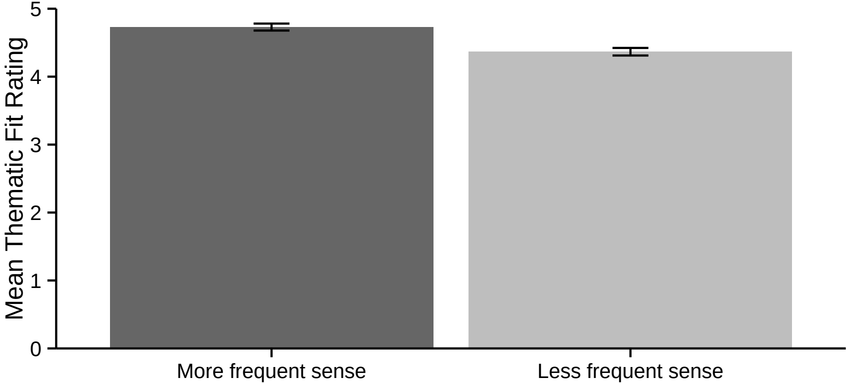
- One survey
  - 6 POLYSEMOUS, 4 MONOSEMOUS, 5 fillers
  - Filler items: the 240 most frequent triples from McRaeNN.
  - Workers do not see an experimental verb in more than one condition.
  - Compensation: $0.15
  - 159 workers participated, 10 ratings per item.

ANOVA results: *Polysemy-Fit* interaction



Interaction is inconsistent with the Autonomous Sense Hypothesis.

Comparing senses



Effect is probably too small for the Sense Frequency Hypothesis.

Effect is probably too large for the Equal Sense Hypothesis.

This just leaves the Conditioned Sense Hypothesis!

Linear modeling results

Predictor	Est.	Std. Err.	$t(1439)$	Sig. level
LOGVERBPOLYSEMY	0.003	0.08	0.04	
LOGVERBFREQUENCY	0.253	0.09	2.74	**
LOGNOUNPOLYSEMY	0.069	0.12	0.55	
LOGNOUNFREQUENCY	0.001	0.06	0.02	

A linear model of Greenberg et al. (2015a) thematic fit ratings based on polysemy and frequency of both verbs and nouns, $\Delta r^2 = 0.01911$. Ignoring the other three predictors, there is a positive correlation between rating and LOGVERBFREQUENCY, Pearson's $r(478) = 0.134, p = 0.003$.

Conclusions

- This is the first thematic fit dataset to vary unigram frequency and verb polysemy systematically.
- POLYSEMOUS: good role-fillers not as good, bad role-fillers not as bad.
- The good role-fillers of a *more frequent sense* get higher ratings.
- Verb frequency positively correlates with ratings.
- Noun frequency does not show a correlation with ratings.
- The Conditioned Sense Hypothesis is the most supported “linear” model.

An “instrument” example

Homer ate the donut with

- pliers
- his fingers
- sprinkles
- a friend

Instrument thematic fit judgements

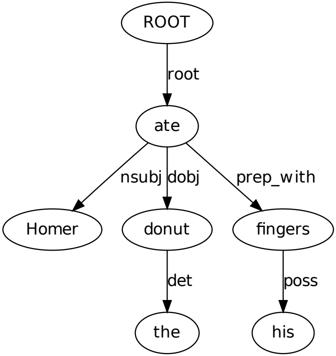
Ferretti et al. (2001): “[On a scale from 1 to 7, h]ow common is it to use each of the following to perform the action of eating?”

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

Step 1 of 3 (Baroni and Lenci, 2010)

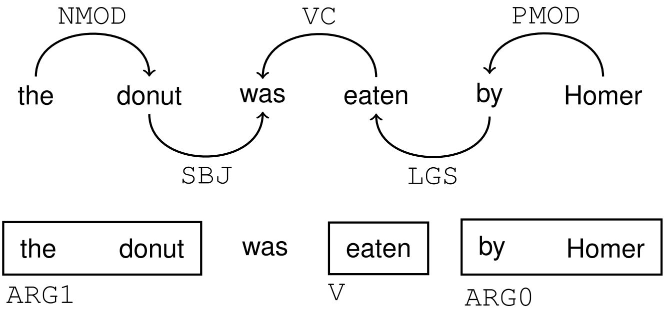
Count verb-role-filler triples & adjust counts by local mutual information (LMI).

$LMI(V, R, F) = O_{VRF} \log \frac{O_{VRF}}{E_{VRF}}$



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

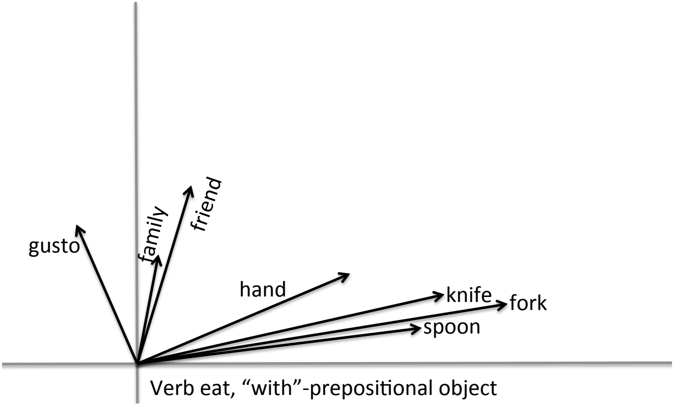
Syntactic or semantic links (Sayeed and Demberg, 2014)



The same sentence with MaltParser (above) and SENNA (below) labels. Sayeed and Demberg (2014) used a simplified approach similar to the head percolation table of Magerman (1994) to find head nouns from SENNA annotation.

Step 2 of 3 (Baroni and Lenci, 2010)

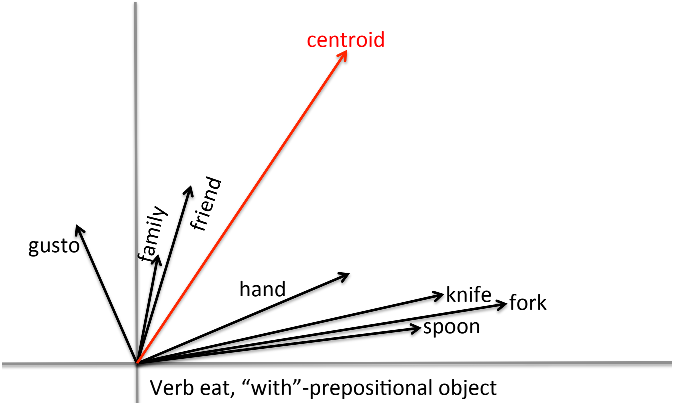
Query the top 20 highest scoring fillers and compute the centroid.



The most typical with-PP arguments of the verb “eat” according to *TypeDM*.

Step 2 of 3 (Baroni and Lenci, 2010)

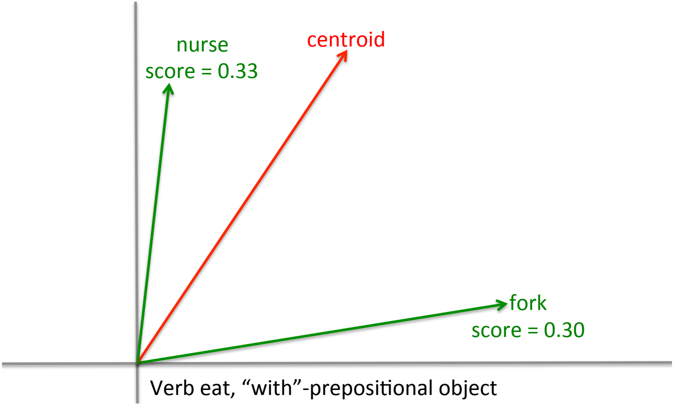
Query the top 20 highest scoring fillers and compute the centroid.



The most typical with-PP arguments of the verb “eat” according to *TypeDM*.

Step 3 of 3 (Baroni and Lenci, 2010)

Return cosine similarity of test role-filler and centroid.



Sample thematic fit scores using the Baroni and Lenci (2010) method.

A new vector-space framework for thematic fit modeling

Key idea

Each verb-role has multiple prototypes (vectors).
Use only the closest prototype to determine the thematic fit score.

The *Centroid* method

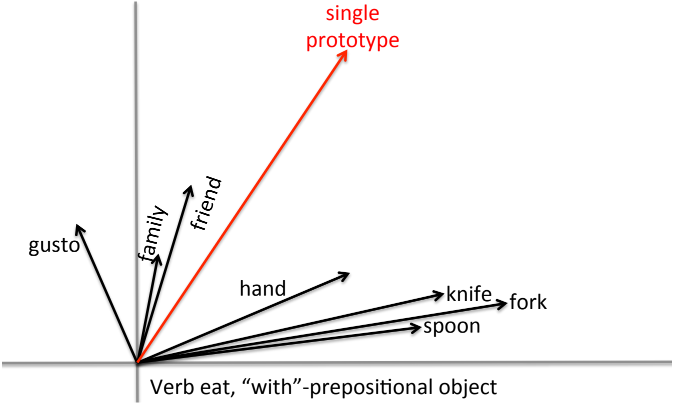


Illustration of the *Centroid* method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to *TypeDM*.

The *OneBest* method

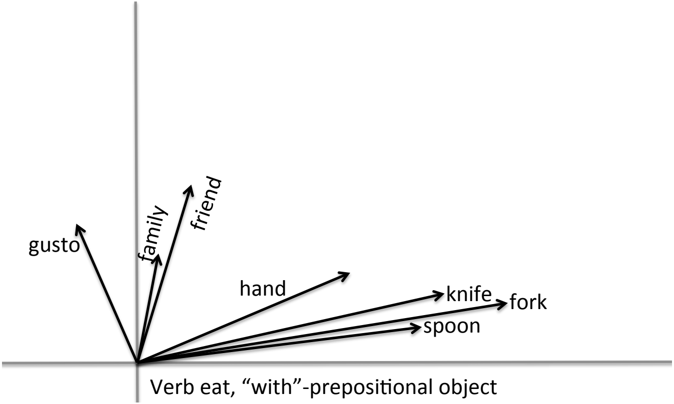


Illustration of the *OneBest* method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to *TypeDM*.

The *OneBest* method

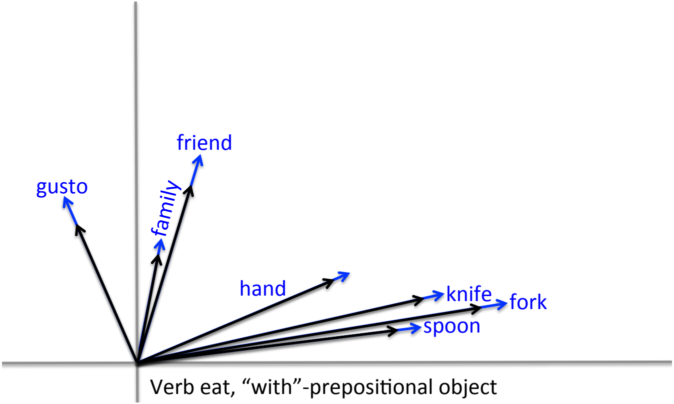


Illustration of the *OneBest* method for prototype generation, using the most typical with-PP arguments of the verb "eat" according to *TypeDM*.

The *OneBest* method

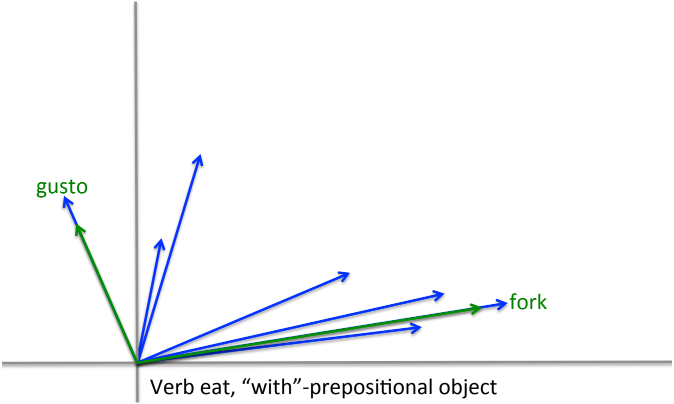


Illustration of the *OneBest* method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to *TypeDM*.

The *OneBest* method

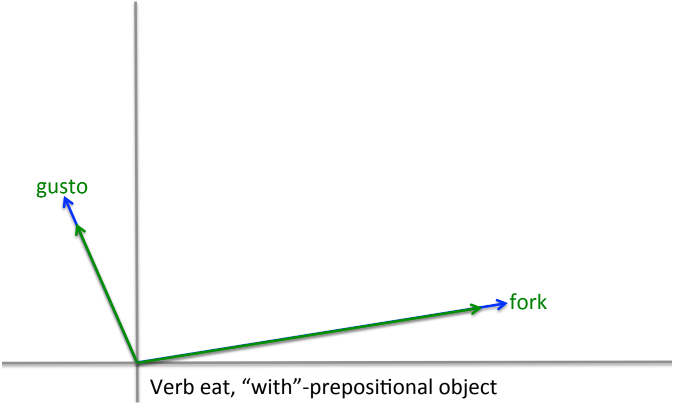


Illustration of the *OneBest* method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to *TypeDM*.

The *OneBest* method

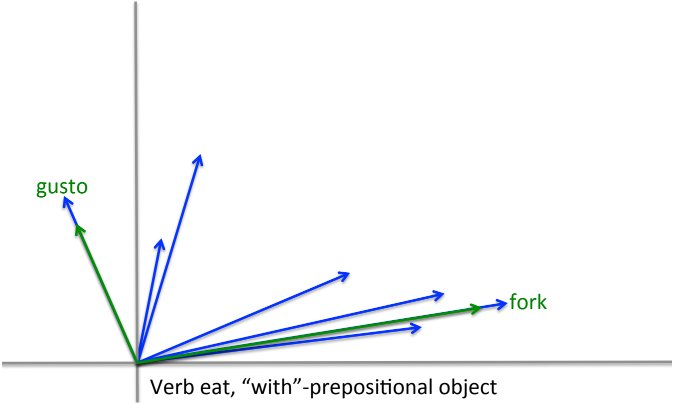


Illustration of the *OneBest* method for prototype generation, using the most typical with-PP arguments of the verb "eat" according to *TypeDM*.

The *OneBest* method

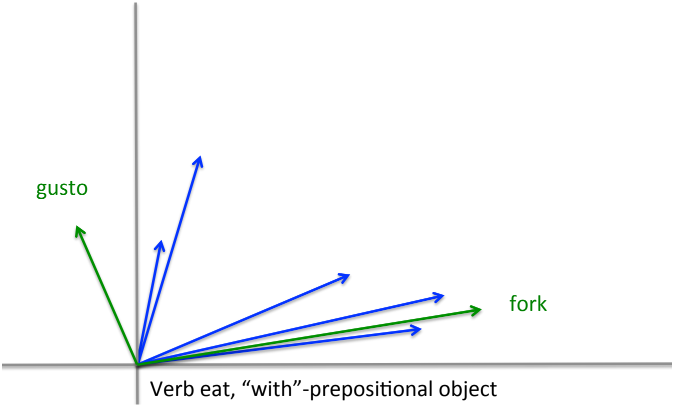


Illustration of the *OneBest* method for prototype generation, using the most typical with-PP arguments of the verb "eat" according to *TypeDM*.

The *OneBest* method

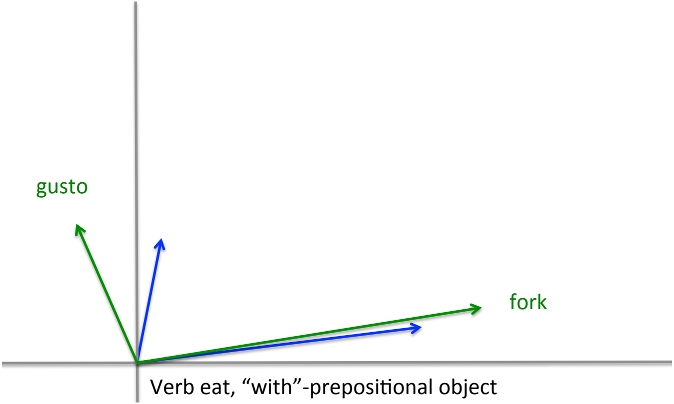


Illustration of the *OneBest* method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to *TypeDM*.

The *2Clusters* method

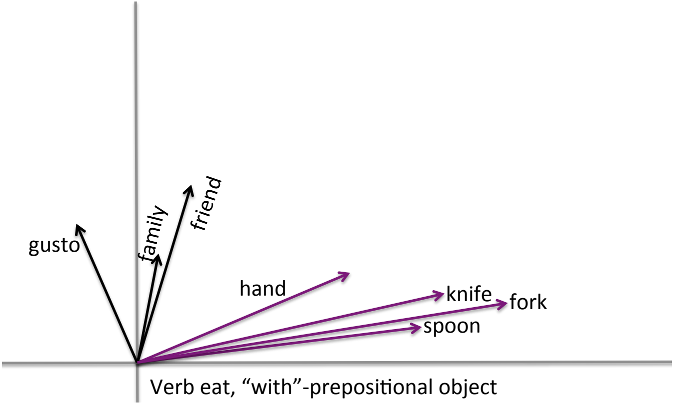


Illustration of the *2Clusters* method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to *TypeDM*.

The *2Clusters* method

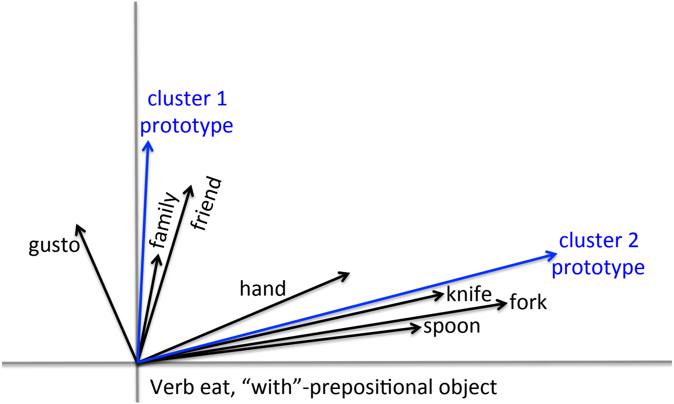


Illustration of the *2Clusters* method for prototype generation, using the most typical with-PP arguments of the verb “eat” according to *TypeDM*.

The *kClusters* method

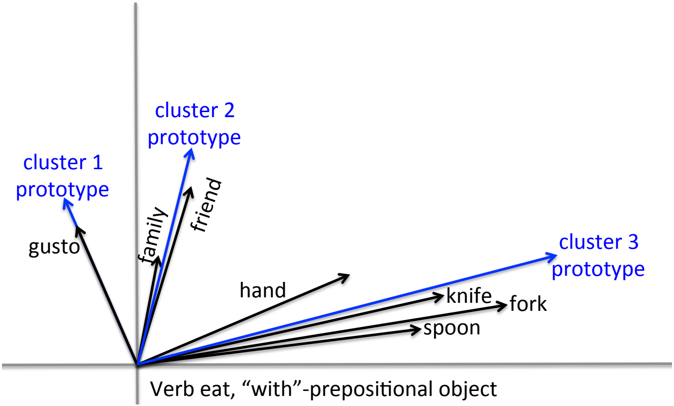


Illustration of the *kClusters* method for prototype generation, using the most typical with-PP arguments of the verb "eat" according to *TypeDM*.

Choosing the number of clusters (K)

Use hierarchical agglomerative clustering package from NLTK (Bird et al., 2009).

Use the Variance Ratio Criterion (VRC) (Caliński and Harabasz, 1974).

$$VRC_k = \frac{SS_B}{k-1} \Big/ \frac{SS_W}{n-k}$$

$$\hat{K} = \underset{k}{\operatorname{argmin}}(VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1})$$

VRC cannot evaluate fewer than 3 clusters, capped at 10 clusters.

Post-processing for thematic fit scores

Greenberg et al. (2015a) dataset:

- LOGVERBFREQUENCY matters!
- LOGNOUNFREQUENCY does not.

Scale each cosine by the log frequency of the verb.

Overall results

Method	Spearman's ρ, range $= [-1, 1]$
Centroid	$0.35 \rightarrow 0.37$
OneBest	$0.36 \rightarrow 0.37$
2Clusters	$0.37 \rightarrow 0.38$
kClusters	$0.39 \rightarrow 0.40$

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

Padó (2007) dataset: agents and patients results

Method	agents	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.

Greenberg et al. (2015a) dataset: overall results

Method	Spearman's ρ, range $= [-1, 1]$
Centroid	0.53
OneBest	0.54
kClusters	0.55

Correlation between human judgements from the Greenberg et al. (2015a) dataset (patients) and automatic scores using LMIs from *TypeDM*, by prototype generation method.

Greenberg et al. (2015a) dataset: results by verb type

Method	POLYSEMOUS	MONOSEMOUS
Centroid	0.41	0.66
OneBest	0.45	0.64
kClusters	0.43	0.67

Correlation between human judgements from the Greenberg et al. (2015a) dataset (patients) and automatic scores using LMIs from *TypeDM*, by prototype generation method and verb type.

Ferretti et al. (2001) dataset: instruments results (1/2)

Method	Spearman's ρ, range $= [-1, 1]$
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.

Ferretti et al. (2001) dataset: instruments results (2/2)

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

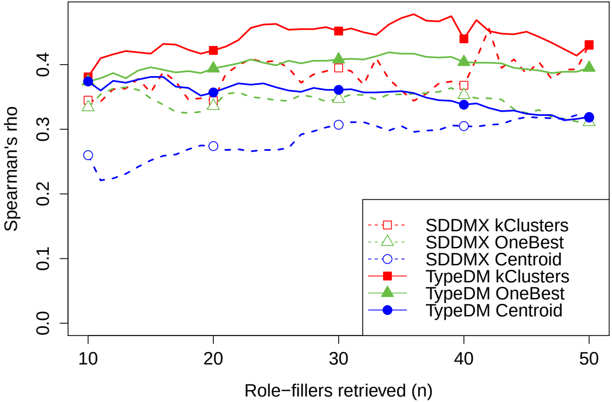
Correlation between human judgements on instruments from the Ferretti et al. (2001) dataset and automatic scores using LMIs from *SENNA-DepDM* (Sayeed and Demberg, 2014) and *TypeDM*, by prototype generation method.

Ferretti et al. (2001) dataset: 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 *SDDMX* (Greenberg et al., 2015b) and *TypeDM*, by prototype generation method.

Deep parameter tuning



Spearman's ρ values for the Ferretti et al. (2001) instruments dataset versus the number of vectors retrieved.

The MONOSEMOUS verb “obey”

1. *injunction*
2. *will*
3. *wish*
4. *limit*
5. *equation*
6. *master*
7. *law, rule, commandment, principle, regulation, teaching, convention*
8. *voice, word*
9. *order, command, instruction, call, summons*

The POLYSEMOUS verb "observe"

1. *day* (`observe_5`)
2. *silence* (`observe_8`)
3. *difference, change* (`observe_1`)
4. *object, star, bird* (`observe_7`)
5. *effect, phenomenon, pattern, behaviour, practice, behavior, reaction, movement, trend*
6. *rule, custom, law, condition* (`observe_9`)

Unsuccessful extensions

- Density peaks clustering (Rodriguez and Laio, 2014)
- Non-negative matrix factorization (Xu et al., 2003)
- Scale cosines by LMI-mass of cluster
- Scale cosines by LMIs
- Use LMIs alone
- Scale centroids by LMI
- Separating PropBank roles for “objects”

Future work

- Knowledge-based number of senses (implemented)
- Using an unlabelled vector-space for cosines
- Examining verb predictability instead of verb frequency
- More detailed modeling of predictions for method comparison
- More sophisticated clustering
  - Expectation-maximization (generalize to weighted centroid)
  - Revisit non-negative matrix factorization

Conclusions

- Thematic fit judgements are sensitive to verb polysemy and frequency.
- Judgements are not sensitive to noun polysemy and frequency.
- Having multiple prototypes improves correlation with humans.
- Prototype clustering navigates a trade-off between polysemy and noise.
- Plausibility is important for psycholinguistic modeling and statistical NLP.

References I

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.

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

Davies, M. (2008). The corpus of contemporary american english: 450 million words, 1990-present. Available online at `http://corpus.byu.edu/coca/`.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454.

References II

Fellbaum, C. (1998). *WordNet: an electronic lexical database*. Wiley Online Library.

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.

Greenberg, C., Demberg, V., and Sayeed, A. (2015a). Verb polysemy and frequency effects in thematic fit modeling. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–57, Denver, Colorado. Association for Computational Linguistics.

Greenberg, C., Sayeed, A., and Demberg, V. (2015b). Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 21–31, Denver, Colorado. Association for Computational Linguistics.

References III

Magerman, D. M. (1994). *Natural Lagnuage Parsing as Statistical Pattern Recognition*. PhD thesis, Stanford University.

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.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

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

Rodriguez, A. and Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496.

References IV

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)*.

Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM.

Follow-up ANOVAs

GOOD:	*Polysemy*	(***)	*NounFrequency*	(**)
BAD:	*Polysemy*	(***)	*NounFrequency*	( )
POLYSEMOUS:	*Fit*	(***)	*NounFrequency*	(.)
MONOSEMOUS:	*Fit*	(***)	*NounFrequency*	(***)