Thematic Roles and Semantic Space
Insights from Distributional Semantic Models

Gabriella Lapesa[1] & Stefan Evert[2]

[1]Institute of Cognitive Science, University of Osnabrück
[2]Corpus Linguistics Group, FAU Erlangen-Nürnberg

Quantitative Investigations in Theoretical Linguistics
12-14 September 2013

UNIVERSITÄT OSNABRÜCK

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

Outline

Distributional Semantic Models

- Distributional semantic models (DSM) implement the Distributional Hypothesis (Harris 1954): difference in meaning $\rightarrow$ difference in distribution
- Distributional meaning of a word is usually operationalized in terms of its co-occurrence patterns with other words

	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
pig	12	17	3	2	9	27

- Distance between word vectors $\Longleftrightarrow$ semantic similarity
  - empirical correlate of the amount of shared meaning

Two important (and open) research questions

1. Lots of tasks, lots of parameters: how do different parameters affect DSM performance in a particular task?
2. Are distributional meaning representations comparable to those of human speakers?

The Generalized Event Knowledge Framework

"Speakers use their knowledge of common events to understand language, and they do so as quickly as possible" (McRae and Matsuki 2009)

- Event Knowledge includes: actions, primary participants (agents, patients), instruments, locations, time course of events
- Event Knowledge is not separated from linguistic knowledge (no time delay, no modularity)

Generalized Event Knowledge

Priming Datasets

D	Relation	N	Prime_c	Prime_i	Target	Fac
V-N	AGENT	28	Pay	Govern	Customer	27*
	PATIENT	18	Invite	Arrest	Guest	32*
	PATIENT FEATURE	20	Comfort	Hire	Upset	33*
	INSTRUMENT	26	Cut	Dust	Rag	32*
	LOCATION	24	Confess	Dance	Court	- 5
N-V	AGENT	30	Reporter	Carpenter	Interview	18*
	PATIENT	30	Bottle	Ball	Recycle	22*
	INSTRUMENT	32	Chainsaw	Detergent	Cut	16*
	LOCATION	24	Beach	Pub	Tan	18*

Priming datasets overview: V-N (Ferretti *et al.* 2001), N-V (McRae *et al.* 2005)

Task: Identification of Consistent Primes

Verb-Noun

Congr. Prime: Adopt
Incongr Prime: Investigate
Target: Baby
Relation: Verb-Patient

Noun-Verb

Congr. Prime: Keyboard
Incongr. Prime: Knife
Target: Type
Relation: Instrument-Verb

If DSMs representation is sensitive to typicality effects based on event knowledge, we expect that:

$$\text{Distance(target, congruent prime)} < \text{Distance(target, incongruent prime)}$$

Why modeling these experiments?

- Contribute to a wider debate concerning the way human semantic representations are built and handled through the integration of experiential and language-based distributional data
- A practical reason: a (quite) large amount of data is available, from different experimental paradigms and with different types of information for each experimental item (norming, reaction times, etc.).

Why do we expect DSMs to be successful?

Distributional Similarity as relatedness

- Verbs and prototypical fillers co-occur, therefore they tend to occur in the same contexts
- Shared meaning relevant to these experiments is understood in terms of **shared topic** (the event), rather than in terms of synonymy

Overview of the models

- Term-term distributional semantic models (bag-of-words)
  - no syntax
  - no word order
- Target terms (rows)
  - vocabulary from Baroni and Lenci (2010) plus GEK tasks
  - 27,688 lemmas / 31,713 tagged lemmas
- Feature terms (columns)
  - filtered by part-of-speech (nouns, verbs, adjectives, adverbs)
  - filtered by frequency thresholds (same for all corpora)

Distributional models were compiled and evaluated using the IMS Corpus Workbench[a], the UCS toolkit[b] and the `wordspace` package[c] for R.

[a] `http://cwb.sf.net/`
[b] `http://www.collocations.de/software.html`
[c] `https://r-forge.r-project.org/projects/wordspace/`

DSM parameters

1. Source corpus (*corpus*)
2. Size of the context window (*window*)
3. Use of part-of-speech information (*pos*)
4. Association score for feature weighting (*score*)
5. Transformation function (*transformation*)
6. Distance measure (*distance*)
7. Dimensionality reduction (*dim.reduction*)
8. Index of semantic relatedness (*relatedness index*)

DSM parameters

Source corpus

- British National Corpus (100 million words)
- Wackypedia: 2009 dump of the english Wikipedia (820 million words)
- WP500: subset of Wackypedia, reduced to initial 500 words of each article (200 million words)
- ukWaC: Web corpus from `.uk` domain (1.9 billion words)
- Joint corpus: BNC + Wackypedia + UkWaC

Small and balanced (BNC) or big and messy (ukWaC)?

DSM parameters

Size of the context window

Symmetric, undirected, flat window of size:

- 2 words left/right
- 5 words left/right
- 15 words left/right

Small windows are expected to find paradigmatically related words, large windows are expected to find topically related words (Sahlgren 2006).

DSM parameters

Use of part of speech information

- targets: lemma / features: lemma (*no_pos*)
- targets: tagged lemma / features: lemma (*pos_t*)
- targets: tagged lemma / features: tagged lemma (*pos_t+f*)

Part-of-speech information reduces ambiguity (*light/A* vs. *light/N*) but results in sparser representation.

DSM parameters

Association score for feature weighting

- co-occurrence frequency (*freq*)
- Dice coefficient (*Dice*)
- Mutual Information (*MI*)
- simple log-likelihood (*s-ll*)
- t-score (*t-sc*)
- z-score (*z-sc*)

Q: Are association measures cognitively plausible? E.g., do they allow for incremental updates?

DSM parameters
Transformation function

- no transformation
- logarithmic
- square root
- sigmoid

Transformations reduce Zipfian skew of co-occurrence frequencies.

DSM parameters
Dimensionality Reduction

- no dimensionality reduction
- Random Indexing to 1000 dimensions (*ri*)
- (randomized) Singular Value Decomposition to 300 dimensions (*rsvd*)

Dimensionality reduction expected to improve semantic representation (SVD) and/or make computations more efficient (SVD, RI), but some researchers also report detrimental effect (e.g. for composition by pointwise multiplication).

DSM parameters
Distance measure

- cosine similarity $\rightarrow$ angular distance
- Euclidean distance
- Manhattan distance

Problem: all these distance measures are symmetric, while cognitive processes (among which priming - Hare et al.2009) are often asymmetric

DSM parameters
Relatedness index

- distance between prime and target (*dist*)
- rank of prime among nearest neighbors of target (*back_rank*)
- rank of target among nearest neighbors of prime (*forw_rank*)
- average rank = mean of *back_rank* and *forw_rank* (*rank_avg*)

Michelbacher *et al.* (2011) use rank-based measures to predict asymmetric syntagmatic association. Hare *et al.* (2009) apply them to their noun-noun priming data.

Step 1: Range and Mean Accuracy

Dataset	Relation	Distance		Forward rank	
		Range	M	Range	M
Verb-Noun	AGENT	43-100	79.3	39-100	85.6
	PATIENT	44-100	83.4	50-100	87.8
	INSTRUMENT	42-100	80.2	38-100	82.6
	LOCATION	30-96	73.6	42-100	82.9
Noun-Verb	AGENT	40-100	77.1	47-100	87.5
	PATIENT	47-100	85.6	60-100	93.6
	INSTRUMENT	40-100	75.4	47-100	87.6
	LOCATION	42-96	79.4	46-96	85.2

Range and Mean Accuracy over Thematic Relations

Step 1: Range and Mean Accuracy

Mean Accuracy: Ranking of Thematic Roles

- Verb-Noun
  - Distance: PATIENT>INSTRUMENT>AGENT>LOCATION
  - Forward rank: PATIENT>AGENT>LOCATION>INSTRUMENT
- Noun-Verb
  - Distance: PATIENT>LOCATION>AGENT>INSTRUMENT
  - Forward rank: PATIENT>INSTRUMENT>AGENT>LOCATION

Step 2: Evaluating DSM parameters

Making sense of 38800 results: a proposal

We use linear models to analyze the influence of parameters and their interactions on performance

- dependent variable = performance (accuracy)
- independent variables = model parameters

$$\begin{aligned}\text{accuracy} = \beta_0 &+ \beta_1(\text{corpus}) + \beta_2(\text{window}) + \beta_3(\text{pos}) + \beta_4(\text{score}) \\ &+ \beta_5(\text{trans}) + \beta_6(\text{dist}) + \beta_7(\text{dim.red}) + \beta_8(\text{rel.index}) + \epsilon\end{aligned}$$

ANOVA:

- shows effect of each parameter and its significance, as well as interactions between parameters
- interpretation based on partial effects plots

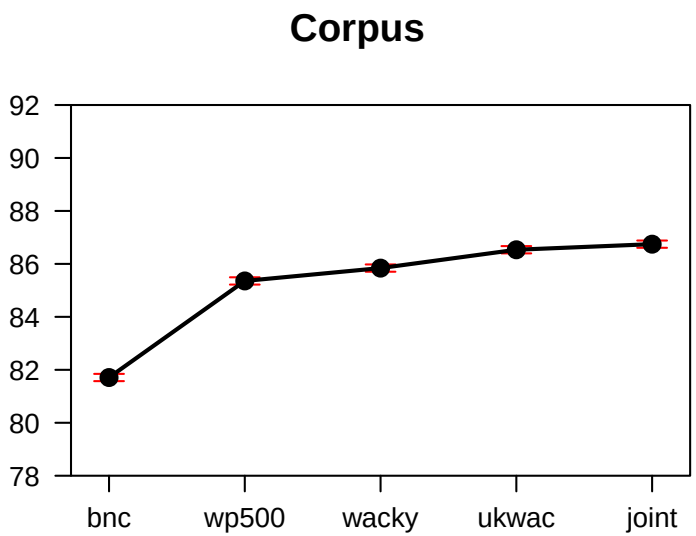
Step 2: Verb-Noun, Patient

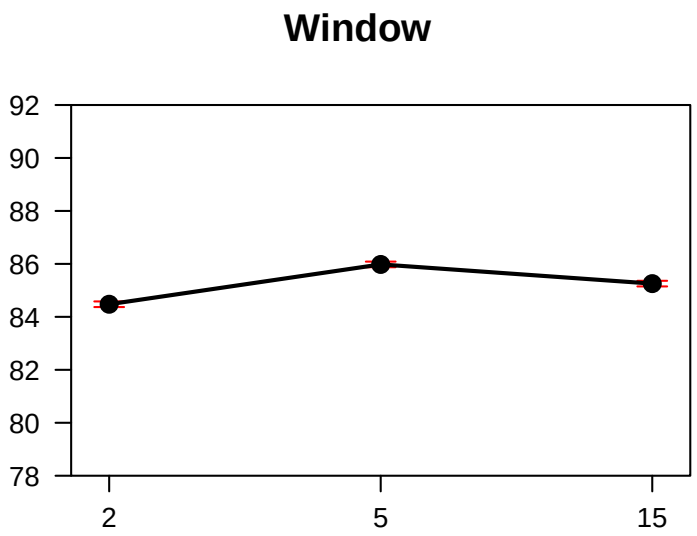
Model Parameters and Interactions (R^2:0.61)

Parameter	df	R^2	signif
corpus	4	3.47	***
window	2	0.39	***
pos	2	2.09	***
score	5	3.95	***
transformation	3	2.01	***
distance	2	11.79	***
dimensionality reduction	2	7.90	***
relatedness index	3	3.73	***
score:transformation	15	6.53	***
window:dim.reduction	4	1.80	***
distance:dim.reduction	4	1.71	***
pos:dimensionality reduction	4	1.71	***
pos:distance	4	1.51	***
window:transformation	6	1.18	***
distance:relatedness index	12	1.03	***

Step 2: Verb-Noun, Patient

Best Parameter Values: Corpus and Window

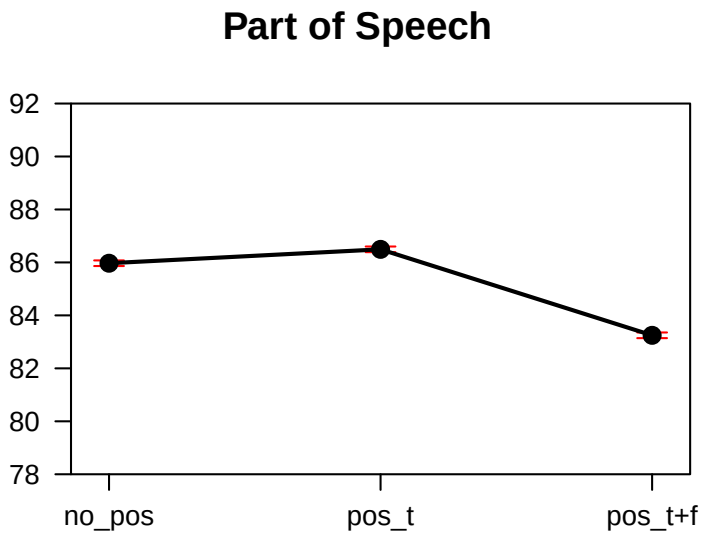


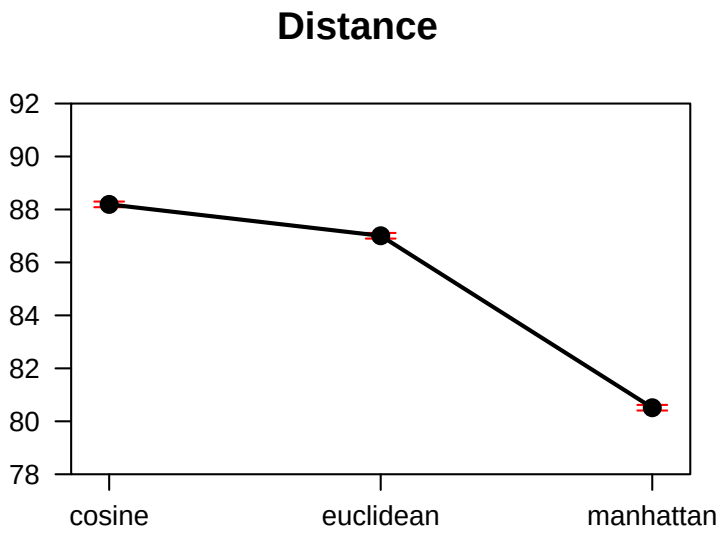


Higher accuracy for models trained on bigger corpora and with medium context windows

Step 2: Verb-Noun, Patient

Best Parameter Values: Part of Speech and Distance

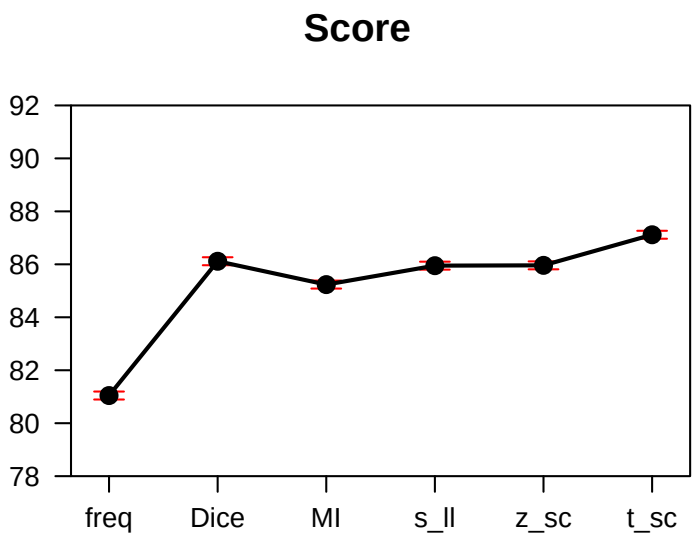


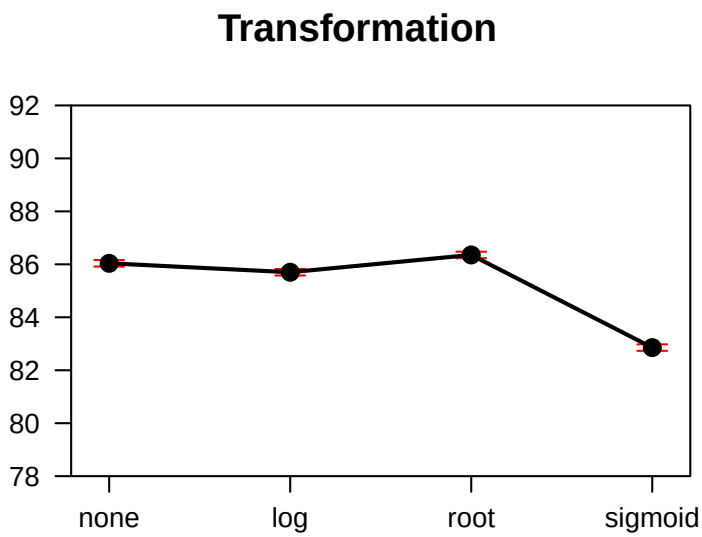


Models with no part of speech info or with pos info only on the target perform better (trade off between disambiguating effect and sparseness?). Cosine and euclidean distance are the best value for the *distance measure* parameter.

Step 2: Verb-Noun, Patient

Best Parameter Values: Score and Transformation

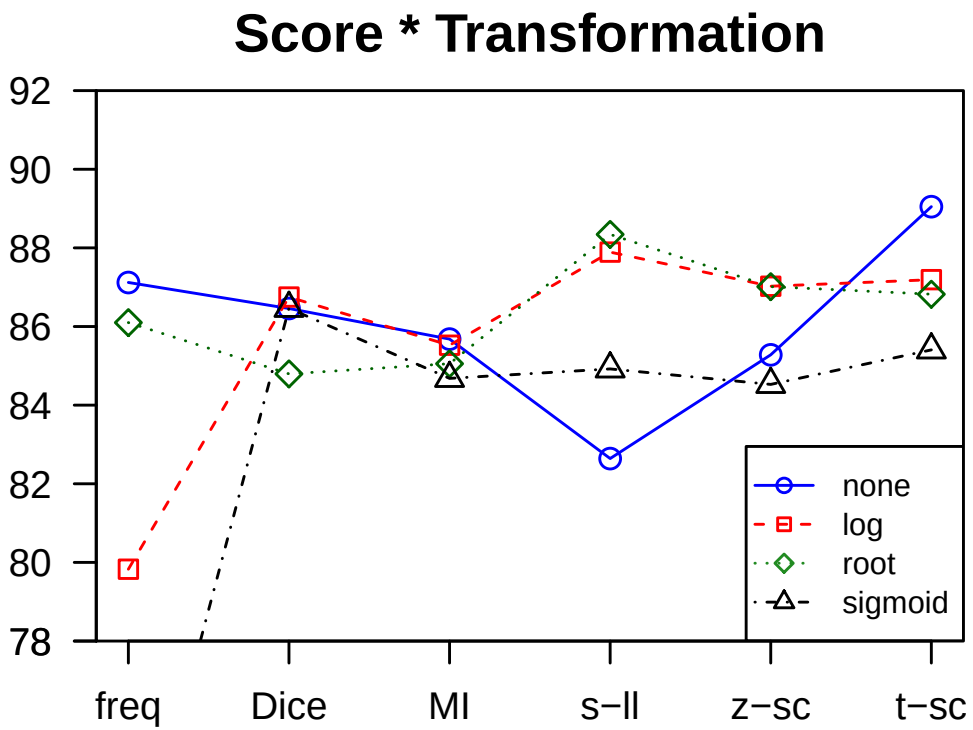




Best performances for association measures vs frequency, worse performances for sigmoid vs other values of the transformation parameter

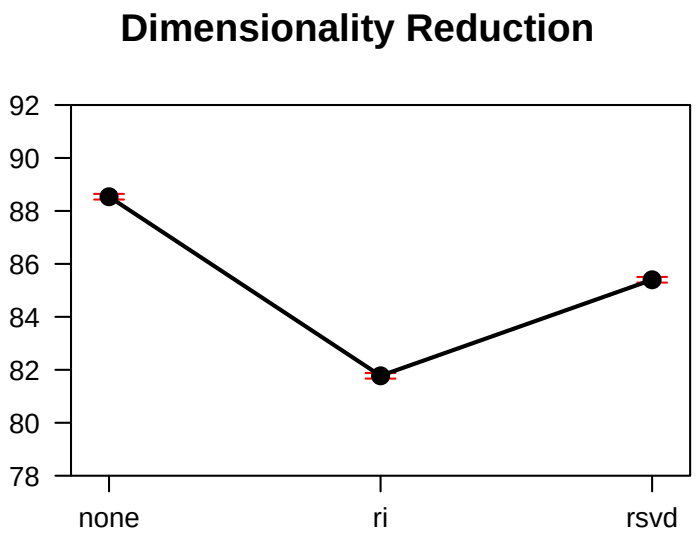
Step 2: Verb-Noun, Patient

Best Parameter Values: Interaction Score and Transformation



Step 2: Verb-Noun, Patient

Best Parameter Values: Relatedness Index and Dimensionality Reduction



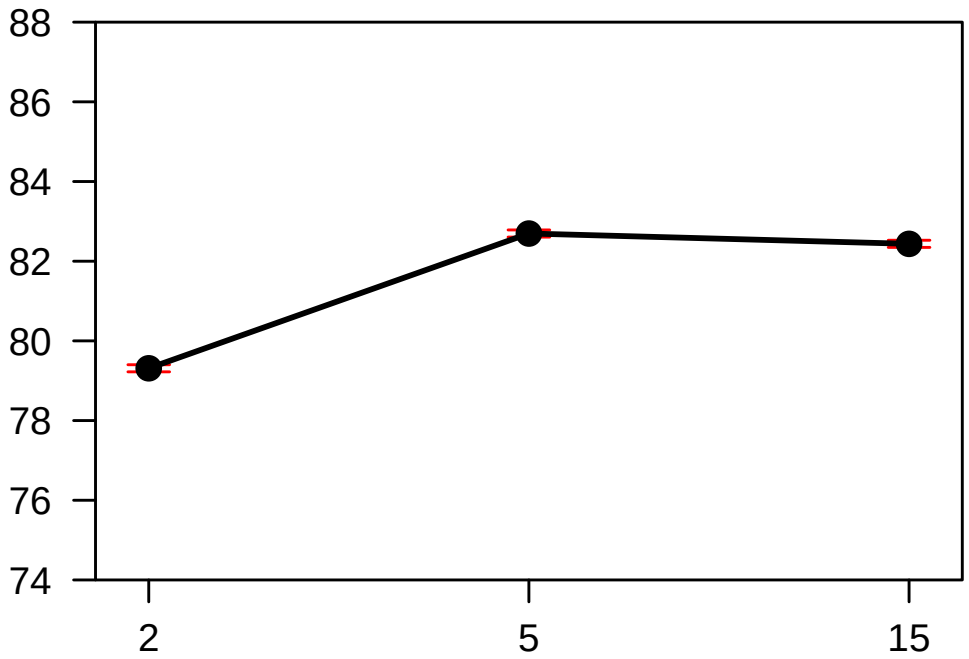
Non reduced models perform significantly better than the reduced ones. Forward rank is the best relatedness index.

Step 2: How about other relations?

- Same analysis on all relations: no significant difference in terms of both explained variance and best parameter values
- Most explanatory parameters:
  - *Distance* and *dimensionality reduction* (followed by *relatedness index* and by *corpus* - with more fluctuations. Score:transformation is always very explanatory)
- Best parameter values:
  - bigger corpora
  - medium-big context windows
  - no part of speech or part of speech only on target
  - association measures better than frequency
  - better accuracy without vector transformation (or with log and root)
  - negative effect of dimensionality reduction
  - cosine as best distance measure
  - forward rank as best relatedness index

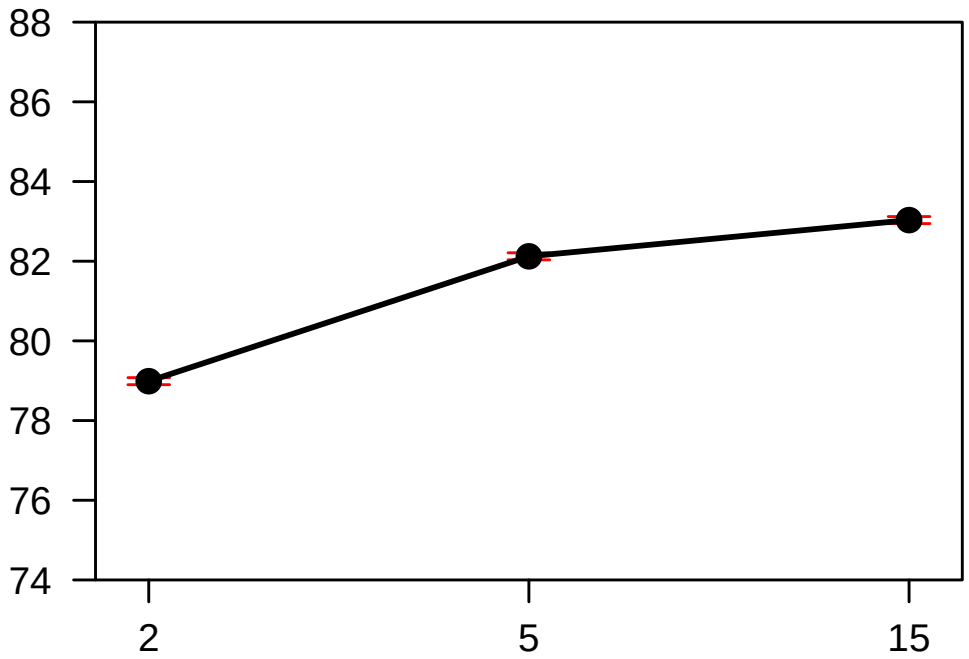
Step 2: Window

Verb-Noun, Agent



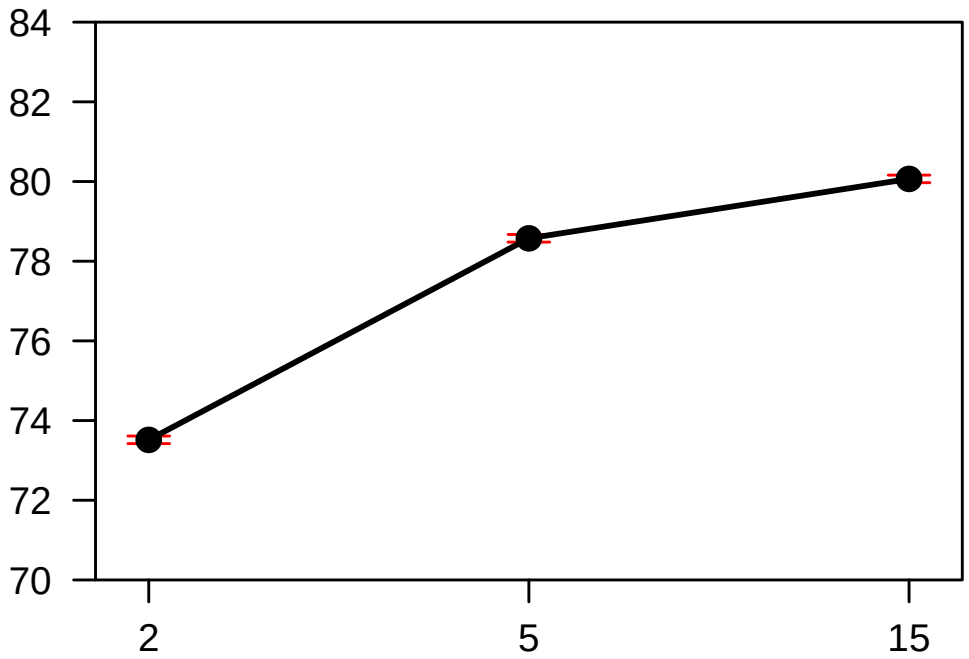
Step 2: Window

Verb-Noun, Instrument



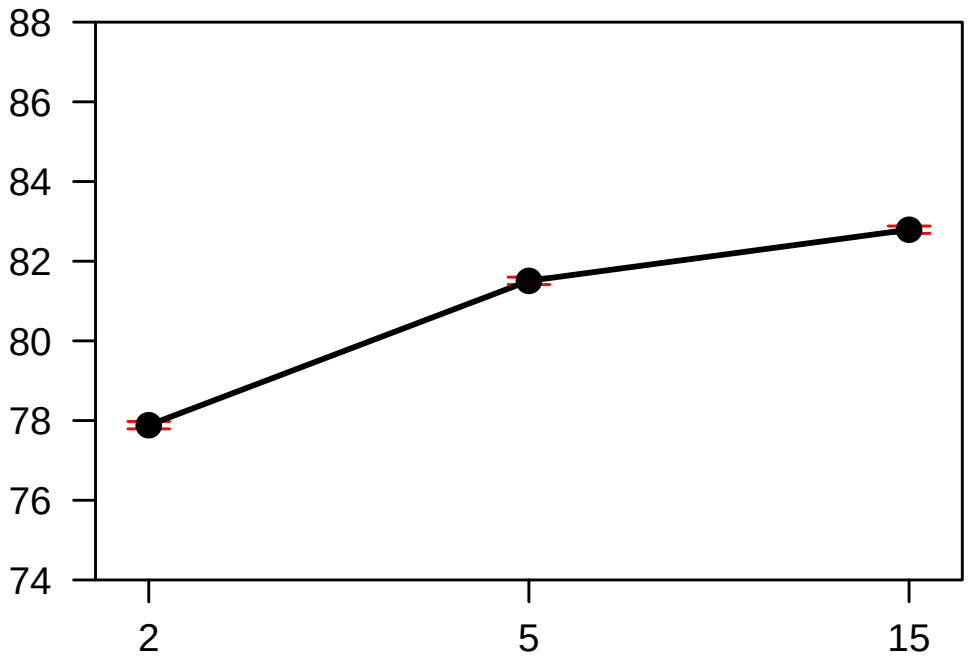
Step 2: Window

Verb-Noun, Location



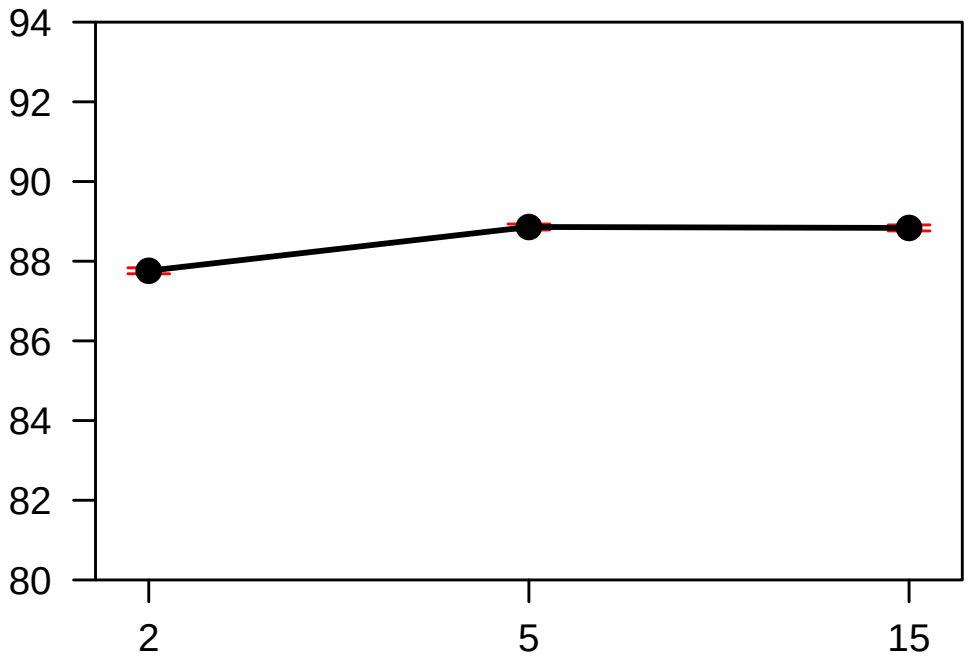
Step 2: Window

Noun-Verb, Agent



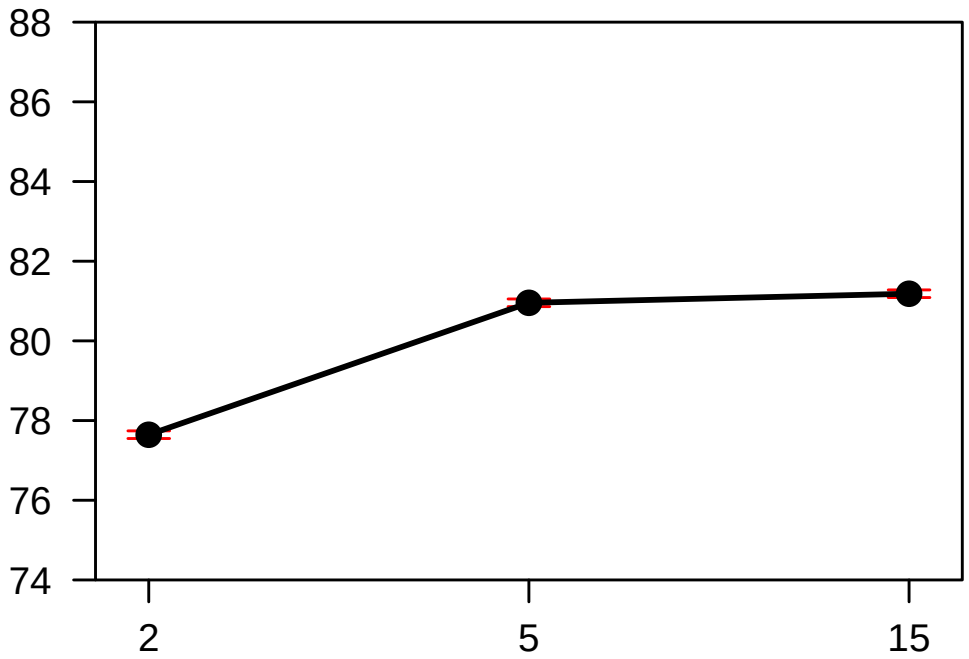
Step 2: Window

Noun-Verb, Patient



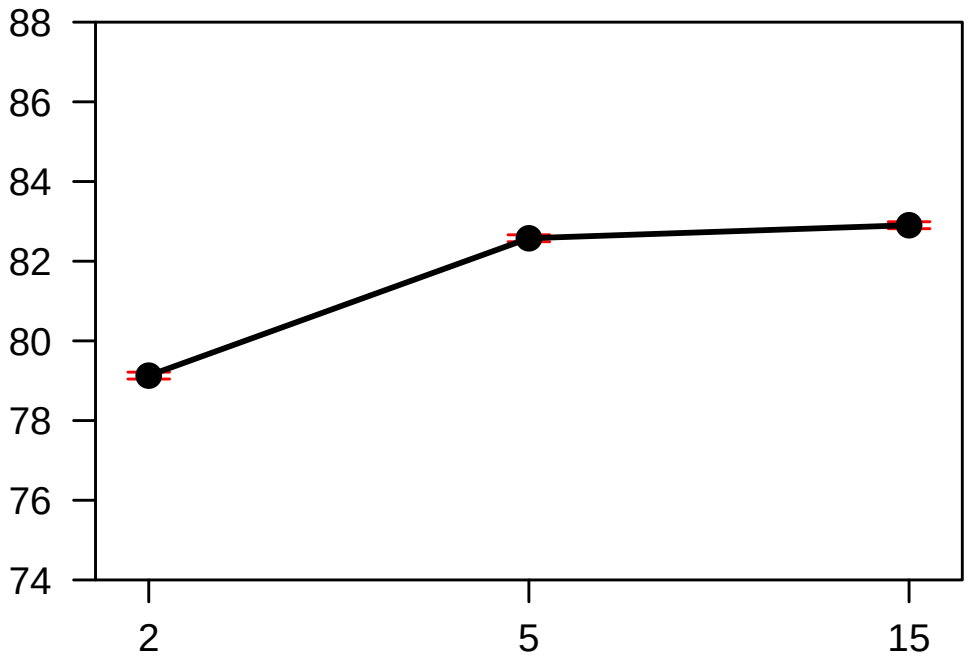
Step 2: Window

Noun-Verb, Instrument



Step 2: Window

Noun-Verb, Location



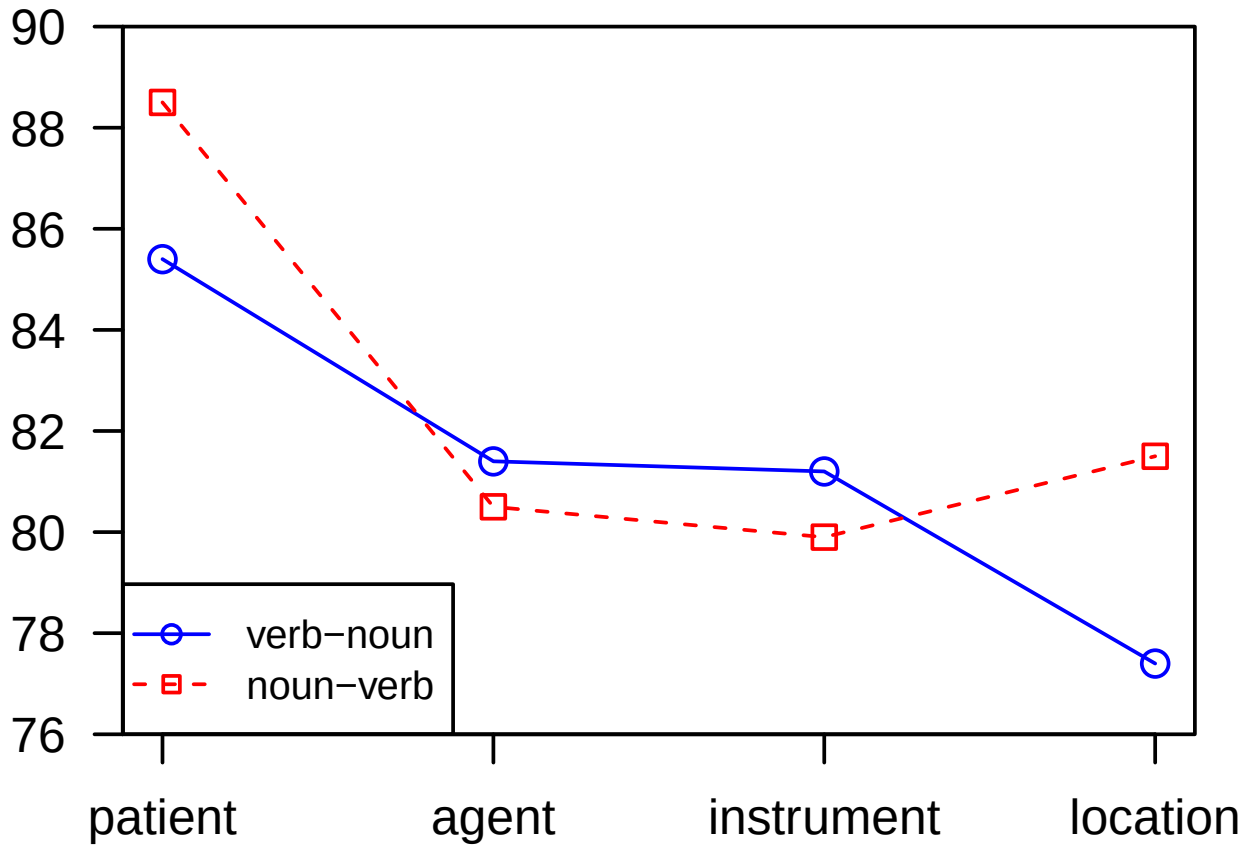
Step 3: Thematic Roles and DSM Performance

Model Parameters and Interactions (R^2:0.72)

Parameter	df	R^2	signif
corpus	4	3.42	***
window	2	2.02	***
pos	2	0.96	***
score	5	4.28	***
transformation	3	2.43	***
distance	2	14.73	***
dimensionality reduction	2	6.25	***
relatedness index	3	7.81	***
relation	7	**8.48**	***
score:transformation	15	3.10	***
window:dim.reduction	4	1.48	***
distance:dim.reduction	4	1.45	***
distance:relatedness index	12	1.42	***

Step 3: Thematic Roles and DSM Performance

Thematic Relation: Partial Effect



Step 3: Thematic Roles and DSM Performance

Relation and its Interactions (R^2: 0.16)

Parameter	df	R^2	signif
relation	7	**8.48**	***
relation:corpus	28	2.35	***
relation:window	14	0.47	***
relation:pos	14	0.86	***
relation:score	35	0.84	***
relation:transformation	21	0.62	***
relation:distance	14	1.25	***
relation:dimensionality reduction	14	0.24	***
relation:relatedness index	21	1.35	***

Summary

- DSMs that make no use of syntax show good performances in a task related to selectional preference
  - The representation responsible for the effects is stable across relations
- The distribution of DSMs' performance across thematic relations shows patterns which are compatible with some general assumptions in theoretical linguistics
  - Some relations are more salient than others in the semantic space, and more subject to typicality effects (prototypical fillers are closer than non prototypical ones)

Work in progress

- We are currently evaluating syntax-based models (dependency filtered/structured, prototype-based)
- Test additional parameters and parameter values
- Include standard tasks in evaluation (TOEFL, . . .)
- Evaluate other types of DSMs (term-context)
- Item-based prediction of RTs, based on different types of corpus-based information (first order, DSMs)
- Context-dependent priming for agent-verb-patient triples (Bicknell *et al.* 2008) and verb-instrument-patient triples (McRae and Matsuki 2009)
- Any other ideas?

References I

Baroni, Marco and Lenci, Alessandro (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 1–49.

Bicknell, Klinton; Elman, Jeffrey L.; Hare, Mary; McRae, Ken; Kutas, Marta (2008). Online expectations for verbal arguments conditional on event knowledge. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society, Volume 1*, pages 2220–2225.

Erk, Katrin; Padó, Sebastian; Padó, Ulrike (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, **36**(4), 723–763.

Ferretti, Todd; McRae, Ken; Hatherell, Ann (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, **44**(4), 516–547.

Hare, Mary; Jones, Michael; Thomson, Caroline; Kelly, Sarah; McRae, Ken (2009). Activating event knowledge. *Cognition*, **111**(2), 151–167.

Harris, Zelig (1954). Distributional structure. *Word*, **10**(23), 146–162.

References II

McRae, Ken and Matsuki, Kazunaga (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, **3**(6), 1417–1429.

McRae, Ken; Hare, Mary; Elman, Jeffrey L.; Ferretti, Todd (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, **33**(7), 1174–1184.

Michelbacher, Lukas; Evert, Stefan; Schütze, Hinrich (2011). Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, **7**(2), 245–276.

Sahlgren, Magnus (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, University of Stockolm.