UNIVERSITY OF GOTHENBURG

Language, action, and perception

Simon Dobnik
CLASP & FLOV, University of Gothenburg

AREA Workshop at LREC
Miyazaki, Japan, 7 May 2018

CLASP centre for linguistic theory and studies in probability

Situated conversational agents

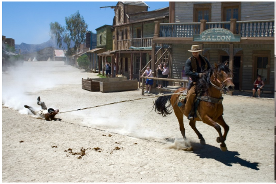
- Connect language, perception, and action
- Novel situations and uncertain environments
- Meanings are dynamic, adapted in interaction
  - Linguistic interaction between conversational partners (Clark, 1996; Fernández et al., 2011)
  - Interaction with the environment through perception (Skočaj et al., 2011; Matuszek et al., 2012)
- Several modalities are involved

Spatial language

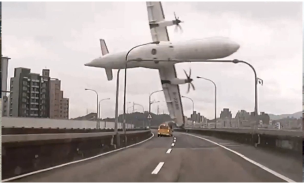
- The chair is to the left of the table.
- Go forward slowly until the next cross-road and then turn left.
- A: I see two blue cups on the left, one with a funny top. . . .
  B: OK, I also see the one with a funny top.

Pattern recognition is not enough

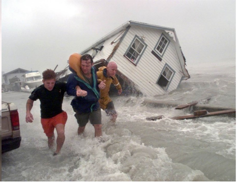
Generated by (Karpathy and Fei-Fei, 2015)



a woman riding a horse on a dirt road



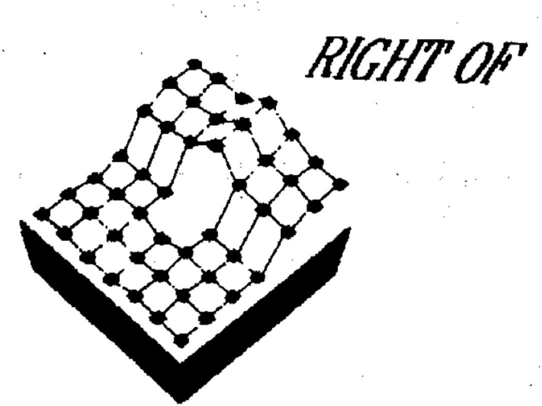
an airplane is parked on the tarmac at an airport



a group of people standing on top of a beach

"...without intuitive physics, intuitive psychology, compositionality, and causality." (Lake et al., 2016)

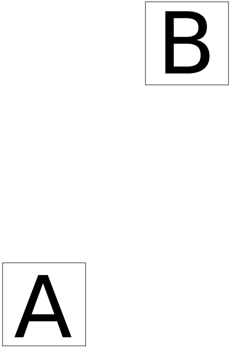
#1: Scene geometry



(Logan and Sadler, 1996)

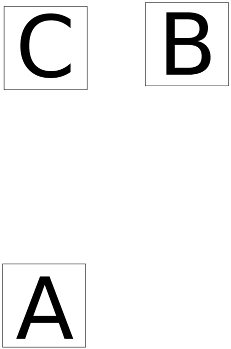
Context matters

Is B above A?



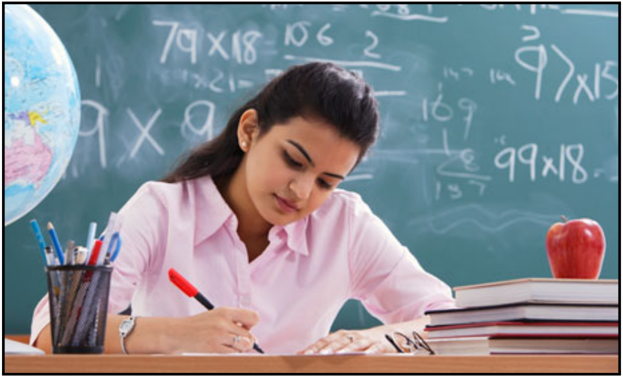
Context matters

Is B above A?



#2: Interaction between objects

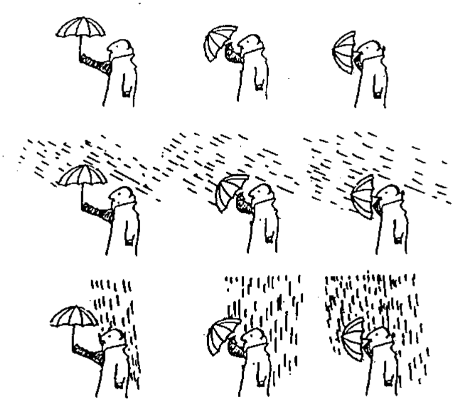
“Alex is at her desk.”



(Coventry and Garrod, 2004)

Dynamic kinematic routines

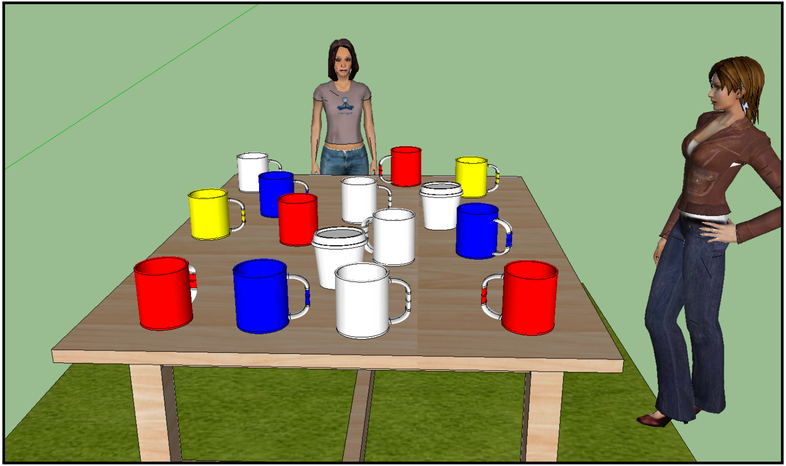
over/*under* and *above*/*below*



(Coventry et al., 2001, 2005)

#3 Interaction between speakers

Where is the yellow mug?



#1: Grounding as interaction

- Connecting perceptual and conceptual representations
- Classifiers with a fixed set of features: Harnad (1990); Roy (2005); Dobnik (2009); Schlangen et al. (2016)
- Feature salience and selection: generating referring expressions (GRE): (Dale and Reiter, 1995; Deemter, 2016)

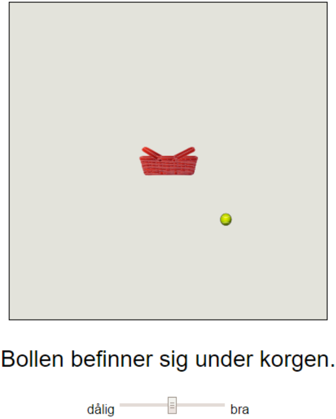
#1: Grounding as interaction

- Connecting perceptual and conceptual representations
- Classifiers with a fixed set of features: Harnad (1990); Roy (2005); Dobnik (2009); Schlangen et al. (2016)
- Feature salience and selection: generating referring expressions (GRE): (Dale and Reiter, 1995; Deemter, 2016)
- Dobnik and Åstbom (2017): feature selection is dynamic, dependent on
  - the feature richness of the perceptual scene
  - the task that an agent is engaged with

#1: Grounding as interaction

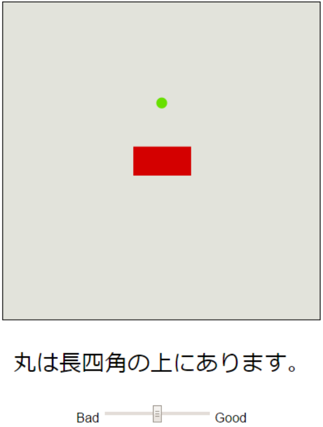
- Connecting perceptual and conceptual representations
- Classifiers with a fixed set of features: Harnad (1990); Roy (2005); Dobnik (2009); Schlangen et al. (2016)
- Feature salience and selection: generating referring expressions (GRE): (Dale and Reiter, 1995; Deemter, 2016)
- Dobnik and Åstbom (2017): feature selection is dynamic, dependent on
  - the feature richness of the perceptual scene
  - the task that an agent is engaged with
- Generative lexicon and lexical semantics: (Pustejovsky, 1995)

The effect of context on grounding?



Bollen befinner sig under korgen.

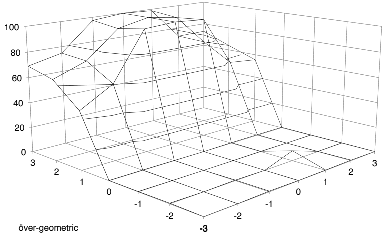
dålig bra

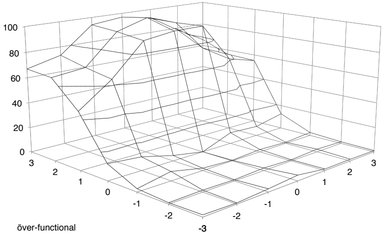


丸は長四角の上にあります。

Bad Good

G vs F: över

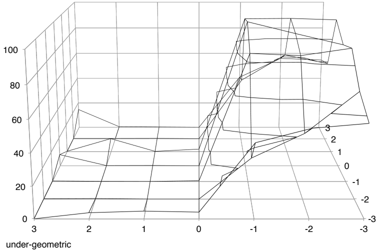


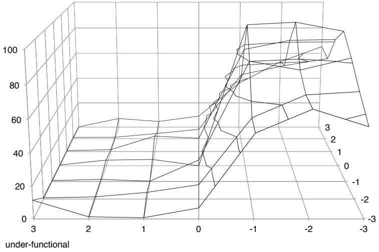


A Wilcoxon signed-rank test: $V = 481, p = 0.383$

Correlation: $r(46) = 0.995, p < 0.001$

G vs F: under

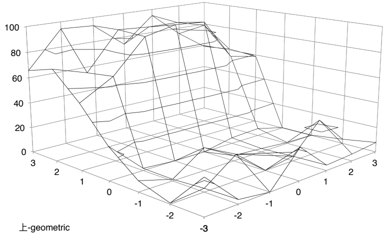


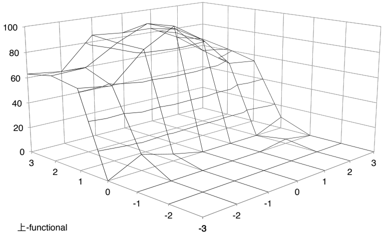


rotated by 90° anticlockwise

A Wilcoxon signed-rank test: $V = 445, p = 0.145$
Correlation: $r(46) = 0.969, p < 0.001$

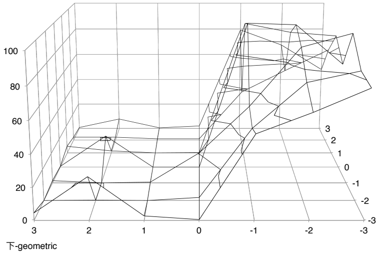
G vs F: 上 ue

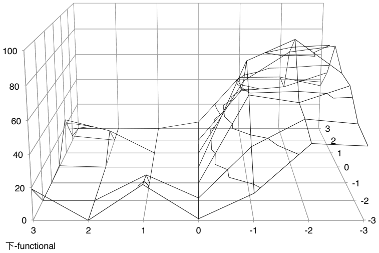




Wilcoxon signed-rank test: $V = 867, p < 0.001$
Correlation: $r(46) = 0.961, p < 0.001$

G vs F: 下 shita





Wilcoxon signed-rank test: $V = 785, p < 0.001$

Correlation: $r(46) = 0.923, p < 0.001$

The effect of perceptual context: summary

- Effect in Japanese but not Swedish
- Contrary to Hörberg (2008) and Coventry et al. (2001, 2005): *över* is sensitive to function

The effect of perceptual context: summary

- Effect in Japanese but not Swedish
- Contrary to Hörberg (2008) and Coventry et al. (2001, 2005): *över* is sensitive to function
- Functional interaction was not a salient feature in our context
- (Logan and Sadler, 1996) no functional features – here a choice
- Participants select features in each context as part of their interaction

The effect of perceptual context: summary

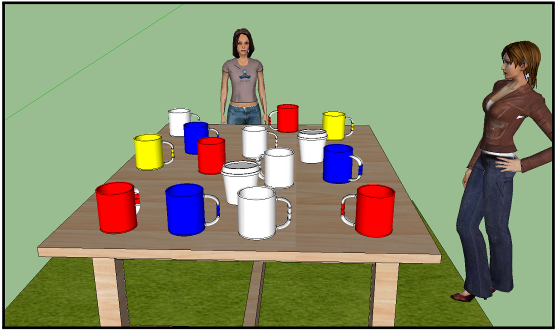
- Effect in Japanese but not Swedish
- Contrary to Hörberg (2008) and Coventry et al. (2001, 2005): *över* is sensitive to function
- Functional interaction was not a salient feature in our context
- (Logan and Sadler, 1996) no functional features – here a choice
- Participants select features in each context as part of their interaction
- Perceptional grounding is dynamically adaptable to contexts

The effect of perceptual context: summary

- Effect in Japanese but not Swedish
- Contrary to Hörberg (2008) and Coventry et al. (2001, 2005): *över* is sensitive to function
- Functional interaction was not a salient feature in our context
- (Logan and Sadler, 1996) no functional features – here a choice
- Participants select features in each context as part of their interaction
- Perceptional grounding is dynamically adaptable to contexts
- A functional scene as a game of Pong

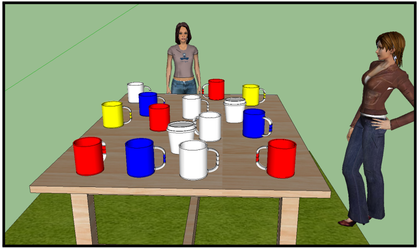
#3: Interaction between speakers

Where is the yellow mug?

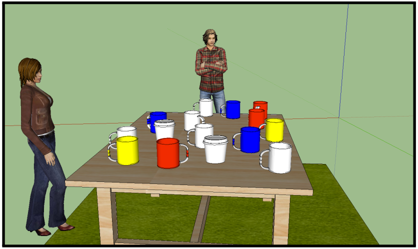


(Dobnik, Howes, and Kelleher, 2015; Dobnik, Howes, Demaret, and Kelleher, 2016)

The views



View for participant 1



View for participant 2

Example

20 *P1:* from her right I see yell, white, blue red
spatial, relative-katie, explicit

21 and the white has a funny thing around the top

22 *P2:* then you probably miss the white i see

23 *P1:* and is between yel and bl but furhter away from katie
spatial, relative-katie, explicit

24 *P2:* because i see a normal mug too, right next to the yellow one, on the left **spatial, relative-katie**

25 *P1:* ok, is your white one closer to katie than the yellow and blue?
spatial, relative-katie

26 *P2:* yes

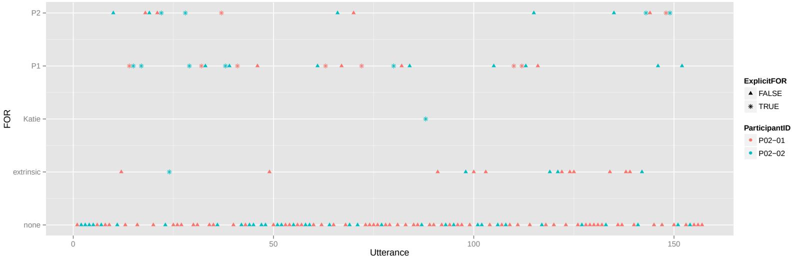
27 closest to me, from right to left:
spatial, relative-p2

28 *P1:* ok, got it

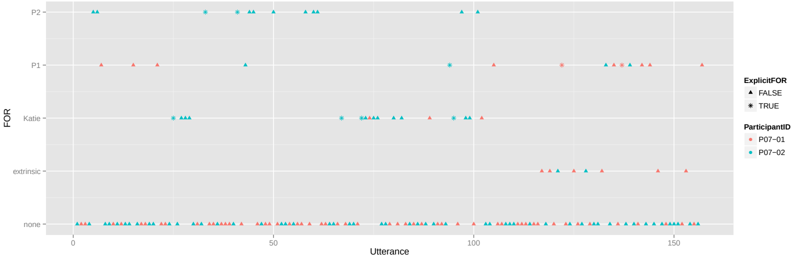
Overview of results

Category	English		Swedish	
	Turns	%	Turns	%
Contains a spatial desc.	245	40.97	273	34.38
FoR=P1	88	35.92	122	44.69
FoR=P2	66	26.94	83	30.40
FoR=speaker	81	33.06	107	39.19
FoR=addressee	72	29.39	98	35.90
FoR=Katie	15	6.12	52	19.05
FoR=extrinsic	61	24.90	38	13.92
Topological description	44	17.96	52	19.05
Total turns	598		794	

Results: Alignment?



English P2



Swedish P7

ory
n probability

Results: Local alignment

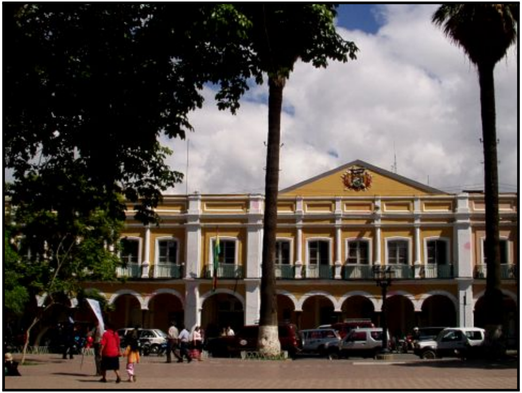
- Participants tend to align to FoR over several turns
- Partial auto-correlations on each binary FoR variable: P1, P2, Katie and Extrinsic
  - Each correlates positively with itself ($p < 0.05$) at 1-3 (English) and 1-2 (Swedish) turns lag
  - use of a particular FoR makes reuse of that FoR more likely
- Adopting an effective communicative strategy within a dialogue game

#2: Extraction knowledge about object interaction

- Encoded in the language model, cf. the success of distributional semantics
- Predict the bias of a spatial relation to functional or geometric knowledge:
  - A functional spatial relation is more selective of their target and landmark objects
  - A geometric relation will occur with any kind of objects.

(Dobnik and Kelleher, 2013, 2014)

Corpora of image descriptions



a yellow building with white columns in the background; two palm trees in front of the house; cars parked in front of the house; a woman and a child are walking over the square;

Choosing a relation

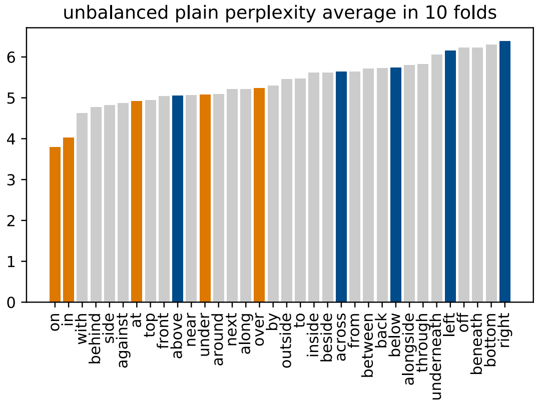
FG	Prep	$-2 log \lambda$	H_2 vs. H_1
people*square	on	655.66*	2.37×10^{142}
people*square	in	133.63*	1.04×10^{29}
people*square	at	1.81	2.47
people*umbrella	with	16.06*	3076.878
boy*umbrella	under	12.16*	436.788
table*umbrella	under	9.39*	109.447
child*umbrella	under	8.35*	65.006
sculpture*umbrella	with	6.88*	31.25
woman*umbrella	with	6.83*	30.428
woman*umbrella	under	6.78*	29.592
girl*umbrella	with	4.59*	9.921
man*umbrella	with	2.29	3.15
child*umbrella	with	1.53	2.153

*: $p < 0.05$

(Normalised) entropy and object variation

#	Preposition	FG-Types	Tokens	Norm FG ent
1	on_left_side_of	5	31	0.35448
2	underneath	31	74	0.65535
3	in	7584	34846	0.6714
4	onto	49	86	0.79109
5	down	83	142	0.81099
6	over	440	736	0.83106
7	at	1393	2726	0.83148
8	on_top_of	61	87	0.83409
9	against	50	68	0.85171
10	on	4897	10085	0.852
11	on_side_of	46	63	0.87644
...	...	...	...	...
15	on_back_of	9	11	0.89489
16	through	179	245	0.89738
17	in_front_of	1278	1938	0.90998
...	...	...	...	...
22	under	167	220	0.92096
23	above	145	190	0.9228
...	...	...	...	...
26	below	13	14	0.96248
...	...	...	...	...

Neural language models and perplexity

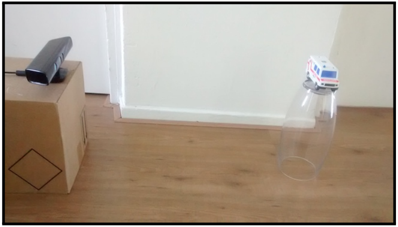


(Dobnik, Ghanimifard, and Kelleher, 2018)

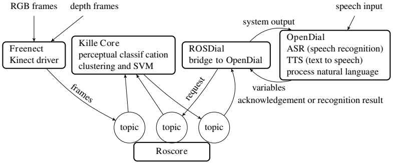
KILLE: Kinect Is Learning LanguagE

Proof of concept incremental learning

Hardware



Software



RGB frames

depth frames

Freenect Kinect driver

Kille Core perceptual classification clustering and SVM

ROSDial bridge to OpenDial

OpenDial ASR (speech recognition) TTS (text to speech) process natural language

system output

speech input

variables

acknowledgement or recognition result

frames

request

topic

topic

topic

Roscore

(Dobnik and de Graaf, 2017)

Conclusions

- Language, action, and perception
  - Action as dynamic world
  - Action as how meaning is assigned to words
- Computational modelling:
  - Situated dialogue systems such as KILLE
  - Deep neural networks which allow integration of different knowledge (Ghanimifard and Dobnik, 2017)
  - . . . modularisation (Dobnik and Kelleher, 2017)
  - . . . integration of existing knowledge (Adouane, Dobnik, Bernardy, and Semmar, 2018)
  - . . . incremental training.

References I

Adouane, W., S. Dobnik, J.-P. Bernardy, and N. Semmar (2018, June 1–6). A comparison of character neural language model and bootstrapping for language identification in multilingual noisy texts. In *Proceedings of the Second Workshop on Subword and Character Level Models in NLP (SCLeM) at 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, New Orleans, Louisiana, USA, pp. 1–10. Association for Computational Linguistics.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

References II

Coventry, K. R., A. Cangelosi, R. Rajapakse, A. Bacon, S. Newstead, D. Joyce, and L. V. Richards (2005). Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, and T. Barkowsky (Eds.), *Spatial Cognition IV. Reasoning, Action, Interaction*, Volume 3343 of *Lecture Notes in Computer Science*, pp. 98–110. Springer Berlin Heidelberg.

Coventry, K. R. and S. C. Garrod (2004). *Saying, seeing, and acting: the psychological semantics of spatial prepositions*. Hove, East Sussex: Psychology Press.

Coventry, K. R., M. Prat-Sala, and L. Richards (2001). The interplay between geometry and function in the apprehension of Over, Under, Above and Below. *Journal of Memory and Language 44*(3), 376–398.

References III

Dale, R. and E. Reiter (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science 19*(2), 233–263.

Deemter, K. v. (2016). *Computational models of referring: a study in cognitive science*. Cambridge, Massachusetts and London, England: The MIT Press.

Dobnik, S. (2009, September 4). *Teaching mobile robots to use spatial words*. Ph. D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom. http://www.dobnik.net/simon/documents/thesis.pdf.

Dobnik, S. and A. Åstbom (2017, August 15–17). (Perceptual) grounding as interaction. In V. Petukhova and Y. Tian (Eds.), *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, Saarbrücken, Germany, pp. 17–26.

References IV

Dobnik, S. and E. de Graaf (2017, 22–24 May). KILLE: a framework for situated agents for learning language through interaction. In J. Tiedemann and N. Tahmasebi (Eds.), *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, Gothenburg, Sweden, pp. 162–171. Northern European Association for Language Technology (NEALT): Association for Computational Linguistics.

Dobnik, S., M. Ghanimifard, and J. D. Kelleher (2018, June 6). Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, New Orleans, Louisiana, USA, pp. 1–11. Association for Computational Linguistics.

References V

Dobnik, S., C. Howes, K. Demaret, and J. D. Kelleher (2016, 17–18 November). Towards a computational model of frame of reference alignment in Swedish dialogue. In J. Björklund and S. Stymne (Eds.), *Proceedings of the Sixth Swedish language technology conference (SLTC)*, Umeå, pp. 1–3. Umeå University.

Dobnik, S., C. Howes, and J. D. Kelleher (2015, 24–26th August). Changing perspective: Local alignment of reference frames in dialogue. In C. Howes and S. Larsson (Eds.), *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, Gothenburg, Sweden, pp. 24–32.

Dobnik, S. and J. D. Kelleher (2013, 31 July). Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, Berlin, Germany, pp. 1–6.

References VI

Dobnik, S. and J. D. Kelleher (2014, August). Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language*, Dublin, Ireland, pp. 33–37. Dublin City University and the Association for Computational Linguistics.

Dobnik, S. and J. D. Kelleher (2017, June 12–14). Modular networks: An approach to the top-down versus bottom-up dilemma in natural language processing. *Forthcoming in Post-proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML) 1*(1), 1–8.

Fernández, R., S. Larsson, R. Cooper, J. Ginzburg, and D. Schlangen (2011). Reciprocal learning via dialogue interaction: Challenges and prospects. In *Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT)*, Barcelona, Catalonia, Spain.

References VII

Ghanimifard, M. and S. Dobnik (2017, September 19–22). Learning to compose spatial relations with grounded neural language models. In C. Gardent and C. Retoré (Eds.), *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics*, Montpellier, France, pp. 1–12. Association for Computational Linguistics.

Harnad, S. (1990, June). The symbol grounding problem. *Physica D 42*(1–3), 335–346.

Hörberg, T. (2008). Influences of form and function on the acceptability of projective prepositions in swedish. *Spatial Cognition & Computation 8*(3), 193–218.

Karpathy, A. and L. Fei-Fei (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137.

References VIII

Lake, B. M., T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman (2016). Building machines that learn and think like people. *arXiv 1604.00289v3 [cs.AI]*, 1–58.

Logan, G. D. and D. D. Sadler (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. A. Peterson, L. Nadel, and M. F. Garrett (Eds.), *Language and Space*, pp. 493–530. Cambridge, MA: MIT Press.

Matuszek, C., N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox (2012, June 27th - July 3rd). A joint model of language and perception for grounded attribute learning. In J. Langford and J. Pineau (Eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, Edinburgh, Scotland.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, Mass.: MIT Press.

References IX

Roy, D. (2005, September). Semiotic schemas: a framework for grounding language in action and perception. *Artificial Intelligence 167*(1-2), 170–205.

Schlangen, D., S. Zarrieß, and C. Kennington (2016, August 7–12, 2016). Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany, pp. 1213–1223. Association for Computational Linguistics.

Skočaj, D., M. Kristan, A. Vrečko, M. Mahnič, M. Janíček, G.-J. M. Kruijff, M. Hanheide, N. Hawes, T. Keller, M. Zillich, and K. Zhou (2011, 25-30 September). A system for interactive learning in dialogue with a tutor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2011*, San Francisco, CA, USA.