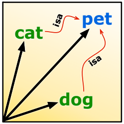
Distributional Semantic Models

Part 4: DS beyond NLP: Linguistic Issues

Stephanie Evert[1] & Gabriella Lapesa[4]
with Alessandro Lenci[2] and Marco Baroni[3]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]University of Pisa, Italy
[3]University of Trento, Italy
[4]University of Stuttgart, Germany

http://wordspace.collocations.de/doku.php/course:start



DSM similarity & Linguistic Theory

1. **Polysemy**
   - A textbook challenge, we will discuss the most intuitive solution
   - ☞ ... available in `wordspace`!
   - ☞ Code from the lecture and extensions in `hands_on_day4.R`
2. **Compositionality**
   - Above and below word level
   - ☞ Bonus evaluation dataset: derivational morphology in (Lazaridou *et al.* 2013)
   - ☞ Last part of `hands_on_day4.R`: perform your own standard tasks on `Lazaridou2013`
3. **Not all meaning is distributional**
   - Function words, proper names (literature pointers)

Great overview paper:

Distributional Semantics and Linguistic Theory (Boleda 2020)

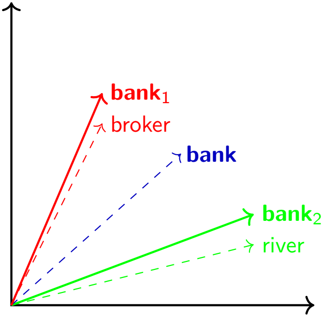
Outline

Polysemy in DSMs

- Problem: DSM vectors conflate contexts from different senses of a word
  - contexts of "bank": money, river, account, swim, ...
  - vectors are displaced suboptimally (far from everything)

Polysemy in DSMs

- Problem: DSM vectors conflate contexts from different senses of a word
  - contexts of “bank”: money, river, account, swim, . . .
  - vectors are displaced suboptimally (far from everything)



Polysemy in DSMs

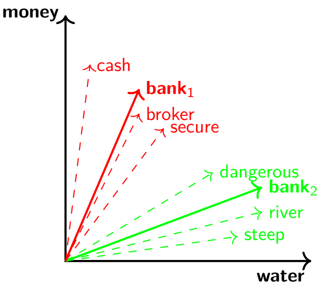
Observation: DSM vectors conflate contexts from word senses

- Solution: build a representation for each instance of the word we want to disambiguate (Schütze 1998)

sentence vectors

Target: bank

bank$_1$: *The broker went to the bank to secure his cash*
bank$_2$: *The river bank was steep and dangerous*



Application: word sense disambiguation

... can you think about another situation in which we may need it?

Context vectors: can we do it in `wordspace`?

Yes :D

```
library(wordspace)
# S1: ''Cats and dogs need their time''
s1 <- "cat and dog need their time"
# S2: ''Time is the cause not the effect''
s2 <- "time is the cause not the effect"
# Ingredients: vectors for individual words
>TT <- DSM_TermTermMatrix
>TT
       breed tail feed kill important explain likely
cat       84   17    8   38         0       2      0
dog      579   14   32   63         1       2      2
animal    45   11   86  136        13       5      4
time      19    8   29  134        94      44    100
reason     1    0    1   18        71     140     39
cause      0    1    0    3        55      35     51
effect     0    1    1    6        62      37     14
```

Context vectors: can we do it in `wordspace`?

Yes :D

"cats and dogs need their time"

```
> context.vectors(TT, s1)
     breed tail feed     kill important explain likely
1 227.3333   13   23 78.33333  31.66667      16     34
# context.vectors() is taking the average of the values in each cell
> (TT['cat','breed']+TT['dog','breed']+TT['time','breed'])/3
227.3333
```

"time is the cause not the effect"

```
round(context.vectors(TT, s2),3)
  breed  tail feed   kill important explain likely
1 6.333 3.333   10 47.667    70.333  38.667     55
```

Context vectors: can we do it in wordspace?

Almost there...

```
# context.vectors() can also take a list as an input
contexts <- round(context.vectors(TT, c(s1, s2)),2)
# The output is a matrix, let's give it better rownames first
rownames(contexts) <- c("s1", "s2")
# ...and then append it to our original matrix
TT <- rbind(TT, contexts)
TT
        breed  tail feed   kill important explain likely
cat     84.00 17.00    8  38.00      0.00    2.00      0
dog    579.00 14.00   32  63.00      1.00    2.00      2
animal  45.00 11.00   86 136.00     13.00    5.00      4
time    19.00  8.00   29 134.00     94.00   44.00    100
reason   1.00  0.00    1  18.00     71.00  140.00     39
cause    0.00  1.00    0   3.00     55.00   35.00     51
effect   0.00  1.00    1   6.00     62.00   37.00     14
s1     227.33 13.00   23  78.33     31.67   16.00     34
s2       6.33  3.33   10  47.67     70.33   38.67     55
```


Context vectors: can we do it in `wordspace`?

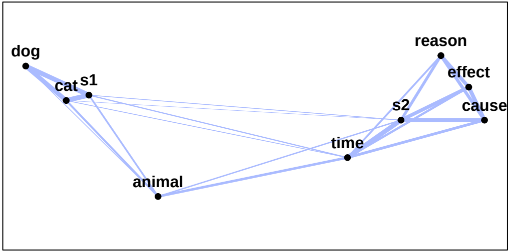
And what now?

```
# We can do all the cool things we are used to do with DSM matrices
# Nearest neighbors...
nearest.neighbours(TT, c("s1", "s2"), n=6)
$s1
     cat      dog   animal     time       s2    cause
14.31016 17.16200 55.27587 62.66470 67.81707 77.90557

$s2
    time    cause   effect   reason   animal       s1
18.85097 25.19348 31.51682 40.83768 60.61621 67.81707
```

Context vectors: can we do it in `wordspace`?

```
# And a semantic map!
plot(dist.matrix(TT))
```



`hands_on_day_4.R` also contains an example for the *bank* polysemy, with word2vec vectors. If you fell in love with centroids the bonus exercise in `schuetze1998.R` (word sense disambiguation, advanced) is perfect for you!

Polysemy in DSMs: contextualized word embeddings

A little detour in embeddingland: BERT

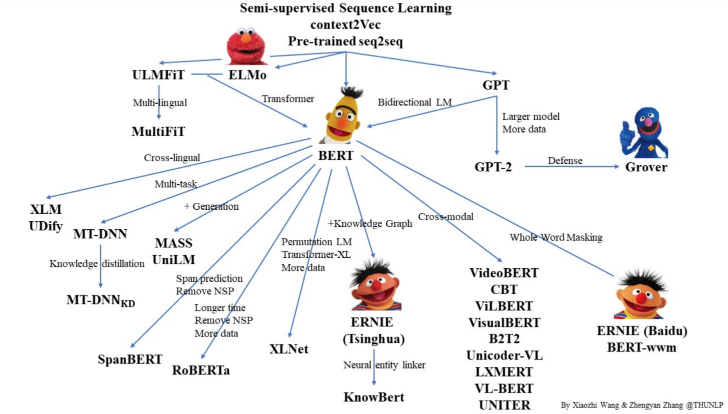
Next step: one contextualized representation per token

The_1, broker_1, went_1, to_2, the_1, bank_1, I_2, swam_2, to_2, the_2, bank_2, The_3, river_3, bank_3, is_3, steep_3

- Bidirectional Encoder Representations from Transformers
- **Most popular embeddings right now. Why?**
  - Multilingual and easily fine-tuned for specific tasks (e.g., question answering, sentiment analysis)
  - Google open-source NLP framework (2018) (`https://github.com/google-research/bert`)
    - Pre-trained on Wikipedia (2.5B tokens) + Google Books (800M tokens)

Polysemy in DSMs: contextualized word embeddings

BERT & other Animals



Problem: some tasks (e.g., those from) require lemma-level representations, which need to be reconstructed "backwards"

Outline

Compositionality

Can we capture it in DS?

- Formally: compositionality implies some operator $\bigoplus$ such that

$$\text{meaning}(w_1 w_2) = \text{meaning}(w_1) \bigoplus \text{meaning}(w_2)$$

- CDSM recipe
  - Distributional vectors for meaning(w_1) and meaning(w_2)
  - Operators: mathematical stategies to combine w_1 and w_2 to *predict* a vector representation for $w_1 w_2$
    - ⋆ vector addition
    - ⋆ vector multiplication
    - ⋆ nonlinear operations learned by neural networks
- Problem: some words (e.g., not) are themselves more like operators than points in space

Great overview paper: Frege in space: a program for compositional distributional semantics (Baroni *et al.* 2014)

Compositionality with distributional vectors

Additive and Multiplicative Models (Mitchell and Lapata, 2010)

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0
problem	2	15	7	9	1

Compositionality with distributional vectors

Additive and Multiplicative Models (Mitchell and Lapata, 2010)

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0
problem	2	15	7	9	1

$$p = u + v$$

predicted(practical difficulty) = **practical** + **difficulty** = [1 14 6 14 4]

Compositionality with distributional vectors

Additive and Multiplicative Models (Mitchell and Lapata, 2010)

	music	solution	economy	craft	create
practical	0	6	2	10	4
difficulty	1	8	4	4	0
problem	2	15	7	9	1

$$p = u + v$$

predicted(practical difficulty) = **practical** + **difficulty** = [1 14 6 14 4]

$$p = u \odot v$$

predicted(practical difficulty) = **practical** $\odot$ **difficulty** = [0 48 8 40 0]

What is your intuition about the effect of multiplication? Have you already seen it as an ingredient of something else?

How do I know my composed representations are "good"?

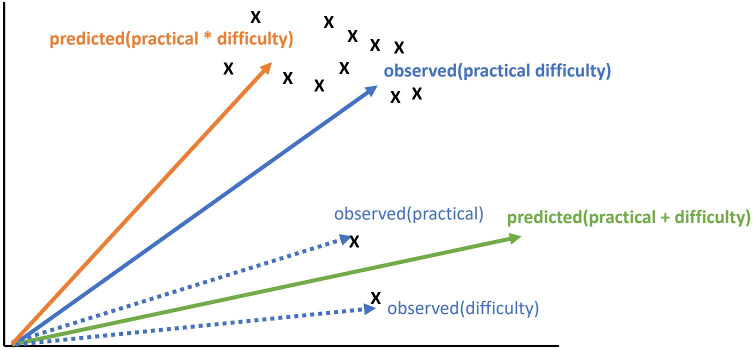
Evaluation, again :)

1. **Qualitative inspection of nearest neighbors**
   - Which neighbors "make more sense" ?
     - practical $+$ difficulty or practical $\odot$ difficulty ?

2. **Quantitative evaluation**
   - Collect a vector for "practical difficulty" in (obviously the same) corpus: **observed(practical difficulty)**
   - observed(practical difficulty) $\approx$ predicted(practical difficulty)
     - Which of the two produces a better approximation?
     - practical $+$ difficulty or practical $\odot$ difficulty
   - Evaluation metric
     - distance(predicted,observed) (Lazaridou *et al.* 2013)
     - rank(predicted,observed) (Baroni & Zamparelli 2010; Padó *et al.* 2016)

How do I know my composed representations are “good”?

Observed vs. Predicted vector



rank(predicted(practical + difficulty)) = 5 < rank(predicted(practical * difficulty)) = 10

distance(predicted(practical * difficulty)) < distance(predicted(practical + difficulty))

Adjective-noun composition (Baroni & Zamparelli 2010)

Starting point: observed AN vectors

- **Input**: triples of {observed(AN), A, N}
  - {bad luck, bad, luck}, {red cover, red, cover}, etc.
  - 36 adjectives (size, color, temporal, etc.)

bad luck	*electronic communities*	*historical map*
bad	electronic storage	topographical
bad weekend	electronic transmission	atlas
good spirit	purpose	historical material
important route	*nice girl*	*little war*
important transport	good girl	great war
important road	big girl	major war
major road	guy	small war
red cover	*special collection*	*young husband*
black cover	general collection	small son
hardback	small collection	small daughter
red label	archives	mistress

- **Methods**: increasing computational complexity
  - No learning (additive, multiplicative)
  - ☞ heavy learning: learns matrix A by comparing AN and N

Adjective-noun composition in Baroni & Zamparelli (2010)

Observed(AN) vs. predicted(AN): neighbors

SIMILAR			DISSIMILAR		
adj N	*obs. neighbor*	*pred. neighbor*	*adj N*	*obs. neighbor*	*pred. neighbor*
common understanding	common approach	common vision	American affair	Am. development	Am. policy
different authority	diff. objective	diff. description	current dimension	left (a)	current element
different partner	diff. organisation	diff. department	good complaint	current complaint	good beginning
general question	general issue	*same*	great field	excellent field	gr. distribution
historical introduction	hist. background	*same*	historical thing	different today	hist. reality
necessary qualification	nec. experience	*same*	important summer	summer	big holiday
new actor	new cast	*same*	large pass	historical region	large dimension
recent request	recent enquiry	*same*	special something	little animal	special thing
small drop	droplet	drop	white profile	chrome (n)	white show
young engineer	young designer	y. engineering	young photo	important song	young image

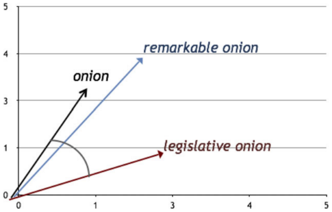
Table 4: Left: nearest neighbors of observed and *alm*-predicted ANs (excluding each other) for a random set of ANs where rank of observed w.r.t. predicted is 1. Right: nearest neighbors of predicted and observed ANs for random set where rank of observed w.r.t. predicted is $\geq 1K$.

How about unattested AN combinations?

Capturing Semantically Deviant AN Combinations (Vecchi *et al.* 2017)

Can we use compositional DSMs to tell, among equally unattested AN, which one is semantically less plausible?

The *composed vectors* for semantically deviant (human rated) combinations will be farther away from the head noun than the acceptable ones



... they test other measures (e.g., neighbors density, vector length) as well as different composition methods: have a look at the paper!

How about unattested AN combinations?

Capturing Semantically Deviant AN Combinations (Vecchi *et al.* 2017)

Can we use compositional DSMs to tell, among equally unattested AN, which one is semantically less plausible?

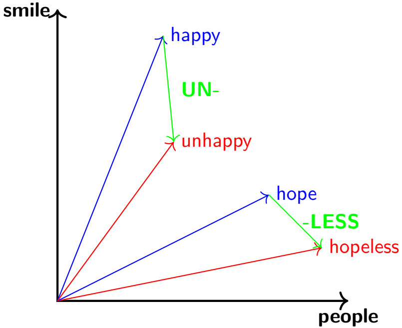
Qualitative inspection: the *composed vectors* of semantically acceptable pairs have plausible nearest neighbors

a.	*angry lamp	{ *shocked*, *fearful*, *angry*, *defiant* }
b.	*nuclear fox	{ *nuclear*, *nuclear arm*, *nuclear development*, *nuclear expert* }
c.	warm garlic	{ *green salad*, *wild mushroom*, *sauce*, *green sauce* }
d.	spectacular striker	{ *goal*, *crucial goal*, *famous goal*, *amazing goal* }

`hands_on_day_4.R` (part 2) contains an implementation of vector addition and multiplication in `wordspace`. Have fun chasing the strangest AN combinations! And other combinations, as well

Compositionality below word level

Can we use compositional DSMs to investigate the meaning of derivational patterns?



- Starting point: vectors for base and derived words.
- Two strategies:
  - ☞ learn the **semantic shifts** with compositional methods
  - investigate **properties** of the patterns → semantic relations
    - ★ zero-nominalizations as hyponyms of the base verb (Varvara *et al.* 2021)
    - ★ un- as antonyms of the base nouns

The DS of Derivational Morphology (Lazaridou *et al.* 2013)

1. **Input**: derived/stem vector pairs for each affix
   - un-: unfaithful/faithful, unbiased/biased, unwell/well
   - -ly: true/truly, mad/madly, deep/deeply

The DS of Derivational Morphology (Lazaridou *et al.* 2013)

1. **Input**: derived/stem vector pairs for each affix
   - un-: unfaithful/faithful, unbiased/biased, unwell/well
   - -ly: true/truly, mad/madly, deep/deeply
2. **Goal: build one representation per affix**
   - No (well, little) learning (additive and multiplicative)
     - un- = centroid(unfaithful, unbiased, unwell, etc.)
   - Increasingly complex learning
     - Parameters set during training to optimize composition, affixes as matrices (cf. adjectives)

The DS of Derivational Morphology (Lazaridou *et al.* 2013)

1. **Input**: derived/stem vector pairs for each affix
   - un-: unfaithful/faithful, unbiased/biased, unwell/well
   - -ly: true/truly, mad/madly, deep/deeply
2. **Goal: build one representation per affix**
   - No (well, little) learning (additive and multiplicative)
     - un- = centroid(unfaithful, unbiased, unwell, etc.)
   - Increasingly complex learning
     - Parameters set during training to optimize composition, affixes as matrices (cf. adjectives)
3. **Prediction & Evaluation**
   - Apply affix to unseen base: predicted(derived) vs. observed(derived). Who did it best?
     - Simplest (additive) & most complex (lexical functional, theoretically motivated): comparable
     - Cf. Padó *et al.* (2016) for German: simplest composition methods work better!

The DS of Derivational Morphology (Lazaridou *et al.* 2013)

Dataset

Affix	Stem/Der. POS	Training Items	HQ/Tot. Test Items	Avg. SDR
-able	verb/adj	177	30/50	5.96
-al	noun/adj	245	41/50	5.88
-er	verb/noun	824	33/50	5.51
-ful	noun/adj	53	42/50	6.11
-ic	noun/adj	280	43/50	5.99
-ion	verb/noun	637	38/50	6.22
-ist	noun/noun	244	38/50	6.16
-ity	adj/noun	372	33/50	6.19
-ize	noun/verb	105	40/50	5.96
-less	noun/adj	122	35/50	3.72
-ly	adj/adv	1847	20/50	6.33
-ment	verb/noun	165	38/50	6.06
-ness	adj/noun	602	33/50	6.29
-ous	noun/adj	157	35/50	5.94
-y	noun/adj	404	27/50	5.25
in-	adj/adj	101	34/50	3.39
re-	verb/verb	86	27/50	5.28
un-	adj/adj	128	36/50	3.23
tot	*/*	6549	623/900	5.52

7000 base/derived pairs from CELEX, 18 patterns, training vs. test (further annotated for base/derived relatedness and vector quality)

Outline

Not all Semantic Knowledge is Distributional

Not all Semantic Knowledge is Distributional

Proper names "answer the purpose of showing what thing it is that we are talking about but not of telling anything about it" (Mill, 1843)

- Intuition: instances of categories such as PER, ORG, etc.
- Herbelot (2015), standard DSMs: category → instance
  - "… upon encountering the name *Mr Darcy* for the first time in the novel, a reader will attribute it the representation of the concept `man` and subsequently specialise it as per the linguistic contexts in which the name appears"
- Westera *et al.* (2021), embeddings: instance → category

Not all Semantic Knowledge is Distributional

Proper names "answer the purpose of showing what thing it is that we are talking about but not of telling anything about it" (Mill, 1843)

- Intuition: instances of categories such as PER, ORG, etc.
- Herbelot (2015), standard DSMs: category → instance
  - "... upon encountering the name *Mr Darcy* for the first time in the novel, a reader will attribute it the representation of the concept `man` and subsequently specialise it as per the linguistic contexts in which the name appears"
- Westera *et al.* (2021), embeddings: instance → category

Function words: some pointers

- Baroni *et al.* (2012) on quantifiers/entailment, Bernardi *et al.* (2013) on determiners, Hole & Padó (2021) on the polysemy of the German reflexive *sich*

Wrapping up

Wrapping up

- Distributional semantics allows us to represent (and compare) a quite heterogeneous selection of "linguistic objects":
  - Subword units (e.g., derivational affixes)
  - Words (content words, proper names, function words)
  - Phrases (e.g., AN)
  - Entire sentences

Wrapping up

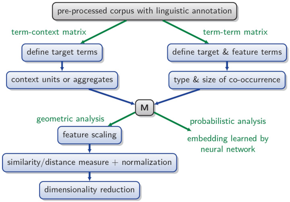
- Distributional semantics allows us to represent (and compare) a quite heterogeneous selection of "linguistic objects":
  - Subword units (e.g., derivational affixes)
  - Words (content words, proper names, function words)
  - Phrases (e.g., AN)
  - Entire sentences

- This is fascinating and promising, but also challenging

Wrapping up

- Distributional semantics allows us to represent (and compare) a quite heterogeneous selection of "linguistic objects":
  - Subword units (e.g., derivational affixes)
  - Words (content words, proper names, function words)
  - Phrases (e.g., AN)
  - Entire sentences
- This is fascinating and promising, but also challenging
  - On top of the DSM parameters, also other experimental choices (e.g., composition. methods)
- . . . and this is exactly the fun of distributional semantics (at least for us :))
  - ☞ Now it is finally your turn to have fun

It is practice session time!



TOEFL dataset

Target: **consume** - Choices: eat, breed, catch, supply
Target: **constant** - Choices: continuing, instant, rapid, accidental
Target: **concise** - Choices: succinct, powerful, positive, free

Almuhareb Poesio

402 nouns, 21 classes
day ⟹ TIME
kiwi ⟹ FRUIT
kitten ⟹ ANIMAL
volleyball ⟹ GAME

BATTIG set

83 nouns, 10 classes
chicken ⟹ BIRD
bear ⟹ LAND_MAMMAL
pot ⟹ KITCHENWARE
oak ⟹ TREE

ESSLLI categorization task

44 nouns, 6 classes
potato ⟹ GREEN
hammer ⟹ TOOL
car ⟹ VEHICLE
peacock ⟹ BIRD

MITCHELL set

60 nouns, 12 classes
ant ⟹ INSECT
carrot ⟹ VEGETABLE
train ⟹ VEHICLE
cat ⟹ ANIMAL

Rubenstein and Goodenough

65 pairs, rated from 0 to 4
gem, jewel: 3.94
grin, smile: 3.46
fruit, furnace: 0.05

WordSim

353 pairs, rated from 1 to 10
announcement, news: 7.56
weapon, secret: 6.06
travel, activity: 5.00

Affix	Stem/Der. POS	Training Items	HQ/Tot. Test Items	Avg. SDR
-able	verb/adj	177	30/50	5.96
-al	noun/adj	245	41/50	5.88
-er	verb/noun	824	33/50	5.51
-ful	noun/adj	53	42/50	6.11
-ic	noun/adj	280	43/50	5.99
-ion	verb/noun	637	38/50	6.22
-ist	noun/noun	244	38/50	6.16
-ity	adj/noun	372	33/50	6.19
-ize	noun/verb	105	40/50	5.96

References I

Baroni, Marco and Zamparelli, Roberto (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Baroni, Marco; Bernardi, Raffaella; Do, Ngoc-Quynh; Shan, Chung-chieh (2012). Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32, Avignon, France. Association for Computational Linguistics.

Baroni, Marco; Bernardi, Raffaelle; Zamparelli, Roberto (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology (LiLT)*, **9**(6), 5–109.

Bernardi, Raffaella; Dinu, Georgiana; Marelli, Marco; Baroni, Marco (2013). A relatedness benchmark to test the role of determiners in compositional distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 53–57, Sofia, Bulgaria. Association for Computational Linguistics.

References II

Boleda, Gemma (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, **6**(1), 213–234.

Herbelot, Aurélie (2015). Mr darcy and mr toad, gentlemen: distributional names and their kinds. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 151–161, London, UK. Association for Computational Linguistics.

Hole, Daniel and Padó, Sebastian (2021). Distributional analysis of function words. To appear in Proceedings of the 13th International Tbilisi Symposium on Language, Logic and Computation.

Lazaridou, Angeliki; Marelli, Marco; Zamparelli, Roberto; Baroni, Marco (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1526, Sofia, Bulgaria. Association for Computational Linguistics.

Padó, Sebastian; Herbelot, Aurélie; Kisselew, Max; Šnajder, Jan (2016). Predictability of distributional semantics in derivational word formation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1285–1296, Osaka, Japan. The COLING 2016 Organizing Committee.

References III

Schütze, Hinrich (1998). Automatic word sense discrimination. *Computational Linguistics*, **24**(1), 97–123.

Varvara, Rossella; Lapesa, Gabriella; Padó, Sebastian (2021). Grounding semantic transparency in context: A distributional semantic study on German event nominalizations. *Morphology*.

Vecchi, Eva M.; Marelli, Marco; Zamparelli, Roberto; Baroni, Marco (2017). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, **41**(1), 102–136.

Westera, Matthijs; Gupta, Abhijeet; Boleda, Gemma; Padó, Sebastian (2021). Distributional models of category concepts based on names of category members. *Cognitive Science*. Accepted for publication. Preprint available at https://nlpado.de/ sebastian/pub/papers/WesteraEtal2021.pdf.