Haber, J., Poesio, M. (2021). Patterns of Polysemy and Homonymy in Contextualised Language Models. In *Findings of the Association for Computational Linguistics : EMNLP 2021* (pp. 2663-2676)

Lucie Barque (Université Paris 13 & LLF)

Morphology Reading Group - 11 February 2022

Introduction

General aims of the paper : Investigating how well contextualised language models capture

- Graded word sense similarity as observed in human annotations
- The distinction between homonymy and polysemy

Ongoing studies on ambiguous word forms

- Effects of polysemy regularity and lexical figure on neological intuition (with A. Lombard, R. Huyghe, D. Gras)
- Effect of sense compatibility on metonymy processing (with J. Salvadori, R. Huyghe)

Ongoing studies on ambiguous suffixes

- Effect of form similarity on sense similarity (with M. Wauquier, O. Bonami, D. Tribout)

Homonymy vs Polysemy

- Homonyms : entirely unrelated distinct meanings

(1) a. *The **match** burned my fingers.*
b. *The **match** ended without a winner.*

- Polysemous words : distinct but related senses

(2) a. *They agreed to meet at the **school**.* [*building*]
b. *The **school** has prohibited drones.* [*institution*]
c. *The **school** called Toms parents.* [*administration*]

Method

Target words

- Dataset : annotated sample contexts for different sense interpretations of 28 English polysemic nouns

- Types of logical metonymy

  animal/meat : lamb, chicken, pheasant, seagull ;
  food/event : lunch, dinner ;
  container-for-content : glass, bottle, cup ;
  content-for-container : beer, wine, milk, juice ;
  opening/physical : window, door ;
  process/result : building, construction, settlement ;
  physical/information : book, record ;
  physical/ information/organisation : newspaper, magazine ;
  physical/information/medium : CD, DVD ;
  building/pupils/directorate/institution : school, university

  - (28 vs 26) + homonyms ?
    - 14 homonyms in (Haber and Poesio 2020) : bat, match, club, bank, mole, etc.

Sample sentences

- Custom samples were created such that
  - (i) the ambiguous target expression is the subject of the sentence,
  - (ii) the context is kept as short as possible,
  - (iii) the context invokes a certain sense as clearly as possible without mentioning that sense explicitly.

- Each of a polysemous word senses is invoked in two different contexts

(3) a. *The **newspaper** fired its editor in chief.* [*organisation*]
b. *The **newspaper** was sued for defamation.*

(4) a. *The **newspaper** lies on the kitchen table.* [*physical object*]
b. *The **newspaper** got wet from the rain.*

(5) a. *The **newspaper** wasnt very interesting.* [*information content*]
b. *The **newspaper** is rather satirical today.*

Method

Sample sentences

- For co-predication, two contexts are combined into a single sentence by conjunction reduction (Zwicky and Sadock, 1975)

  (6) *The newspaper fired its editor in chief and was sued for defamation*

- Besides polysemic alternations, some of the targets also allow for homonymic alternations (e.g *magazine*)

Sample sentences

Two conditions

- Same

(7) a. *The **newspaper** wasnt very interesting.* [*information content*]
b. *The **newspaper** is rather satirical today.* [*information content*]

(8) *The newspaper fired its editor in chief and was sued for defamation*

- Cross

(9) a. *The **newspaper** fired its editor in chief.* [*organisation*]
b. *The **newspaper** is rather satirical today.* [*information content*]

(10) *The newspaper fired its editor in chief and is rather satirical today.*

Method

Human annotation

- Amazon Mechanical Turk (AMT) to collect human annotations online
- Participants were asked to rate
  (1) The similarity in meaning of a target word shown in two different contexts
  (2) The acceptability of a co-predication structure combining two contexts with the same target
- Graded word sense similarity judgement
  (1) Slider from "completely different meaning" to "completely the same meaning"
  (2) Slider from "absolutely unacceptable" to "absolutely acceptable"

Method

Contextualized Language Models

- Assessing word sense similarity encoded in contextualised embeddings
- Extraction of target word embeddings from the different disambiguating contexts and calculated their cosine similarity (1-cosine)
- Models
  - ELMo
  - Bert base (12 layers, hidden state size of 768)
  - Bert large (24 layers, hidden state size of 1024)
  - Baseline : by averaging over the static Word2Vec (Mikolov et al., 2013) encodings of all words in a sample context to create a naive contextualised embedding

Results

1. Analysis of similarity and acceptability ratings based on the collected annotations
2. Analysis of how the different contextualised language models target embeddings correlate with either of the human annotation
3. Analysis of the contextualised embbedings themselves, for a preliminary assessment of how well these off-the-shelf word sense encodings fare in clustering samples based on their sense interpretation

Results

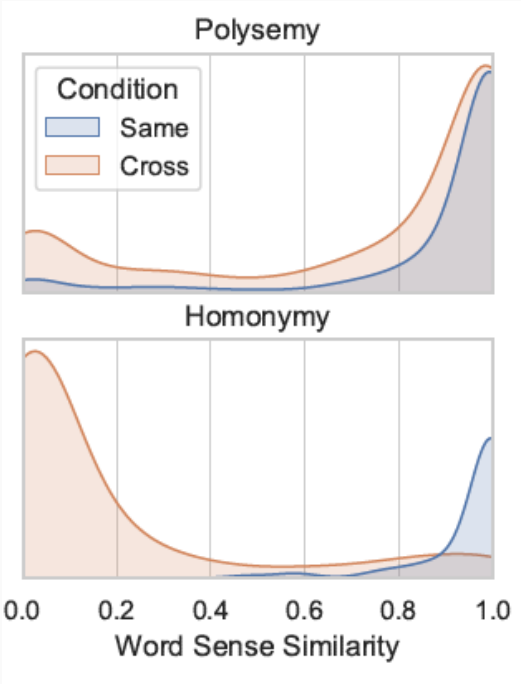
1. Word Sense Similarity Ratings

- Data
  - 5862 judgements (after filtering)
  - 16.5 annotations per item on average (minimum 7)
  - IAA rate of 0.62 (Krippendorffs alpha, Artstein and Poesio, 2008)

- Neglibible effects of predicate ordering

Results

1. Word Sense Similarity Ratings



Measure	Same-Sense Pol.	Hom.	p	Cross-Sense Pol.	Hom.	p
Similarity	0.89	0.96	0.03	0.73	0.17	<0.05
Acceptability	0.83	0.86	0.10	0.64	0.41	<0.05
Word2Vec	0.60	0.65	0.12	0.55	0.58	0.06
ELMo	0.90	0.87	0.14	0.87	0.82	<0.05
BERT Base	0.91	0.93	0.22	0.88	0.78	<0.05
BERT Base (L4)	0.93	0.95	0.27	0.91	0.82	<0.05
BERT Large	0.79	0.85	0.15	0.72	0.44	<0.05
BERT Large (L4)	0.88	0.91	0.18	0.84	0.64	<0.05

Table 1: Word sense similarity distribution means for the different measures investigated in this study. p-values calculated through Mann-Whitney *U*.

- These results support the traditional view that polysemy occupies a distinctive middle ground between identity of meaning and homonymy

Results

1. Word Sense Similarity Ratings

- Same-sense samples are quite consistently rated
  - 6.90% of the 58 pairwise comparisons passed the Boneferroni correction
- Cross-sense samples are less consistently rated than same-sense samples, and polysemic alternations are rated less consistently than homonymic ones
  - Homonymic cross-senses : 14.71% of the 34 pairwise comparisons
  - Polysemic cross-senses : 23.44% of the 337 pairwise comparisons
- The results also provide a novel type of empirical evidence against a uniform treatment of polysemic senses

Results

1. Co-Predication Acceptability Ratings

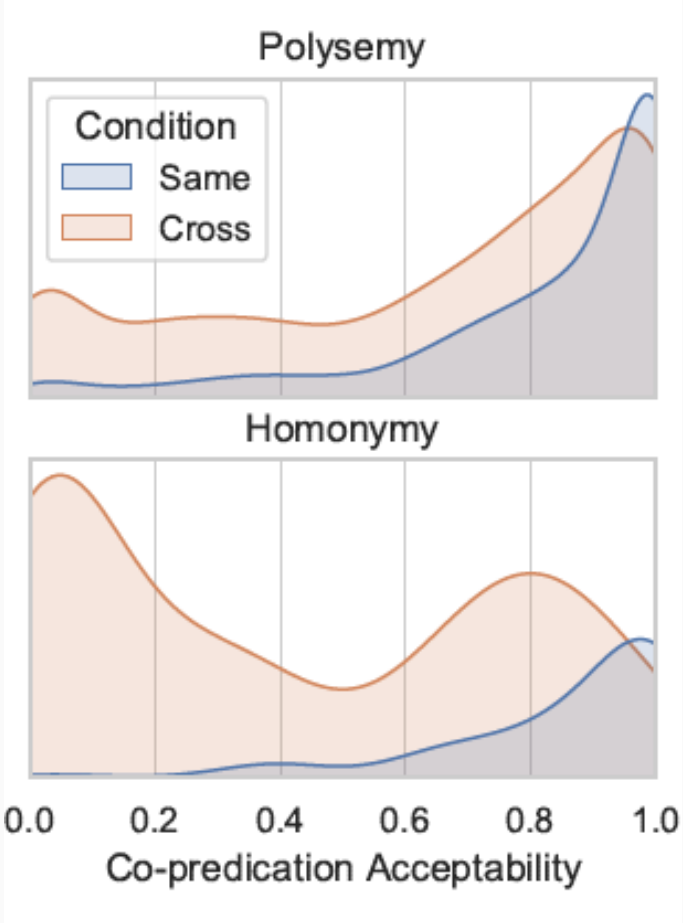
- Data
  - 7379 judgements
  - 16.75 annotations per target word on average (minimum 12)
  - IAA here only reached a Krippendorffs alpha rating of 0.34, indicating stronger individual differences

- Co-predication acceptability is meant to provide a more ecological signal of word sense similarity than the explicit similarity ratings

- Order effect : samples are free from any secondary acceptability factors based on predication order (Murphy 2021)

Results

1. Co-Predication Acceptability Ratings



Measure	Same-Sense Pol.	Same-Sense Hom.	Same-Sense p	Cross-Sense Pol.	Cross-Sense Hom.	Cross-Sense p
Similarity	0.89	0.96	0.03	0.73	0.17	<0.05
Acceptability	0.83	0.86	0.10	0.64	0.41	<0.05
Word2Vec	0.60	0.65	0.12	0.55	0.58	0.06
ELMo	0.90	0.87	0.14	0.87	0.82	<0.05
BERT Base	0.91	0.93	0.22	0.88	0.78	<0.05
BERT Base (L4)	0.93	0.95	0.27	0.91	0.82	<0.05
BERT Large	0.79	0.85	0.15	0.72	0.44	<0.05
BERT Large (L4)	0.88	0.91	0.18	0.84	0.64	<0.05

Table 1: Word sense similarity distribution means for the different measures investigated in this study. p-values calculated through Mann-Whitney *U*.

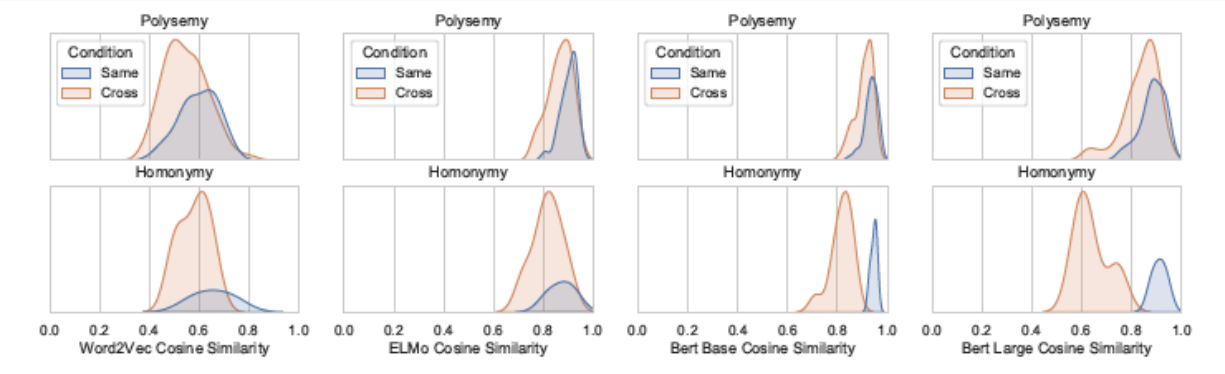
Results

1. Co-Predication Acceptability Ratings

- Support previous observations of co-predication acceptability being a non-binary signal but rather forming a continuum (Lau et al., 2014)
- Provide an additional challenge to co-predication as a linguistic test to distinguish polysemy from homonymy
  - This test is used to distinguish vagueness from ambiguity (Cruse 1986)
- Polyseme samples again show some degree of inconsistency [. . .] which provide additional evidence for the non-uniformity in interpreting polysemic samples

Results

2. Computational Ratings



Measure	Same-Sense Pol.	Same-Sense Hom.	Same-Sense p	Cross-Sense Pol.	Cross-Sense Hom.	Cross-Sense p
Similarity	0.89	0.96	0.03	0.73	0.17	<0.05
Acceptability	0.83	0.86	0.10	0.64	0.41	<0.05
Word2Vec	0.60	0.65	0.12	0.55	0.58	0.06
ELMo	0.90	0.87	0.14	0.87	0.82	<0.05
BERT Base	0.91	0.93	0.22	0.88	0.78	<0.05
BERT Base (L4)	0.93	0.95	0.27	0.91	0.82	<0.05
BERT Large	0.79	0.85	0.15	0.72	0.44	<0.05
BERT Large (L4)	0.88	0.91	0.18	0.84	0.64	<0.05

- All computational models assign a much narrower range of similarity scores to the ambiguous samples (Ethayarajh 2019)
- All BERT models produce clearly distinct distributions for polysemic, homonymic and same-sense samples (all p-values <0.05)

Results

2. Computational ratings (vs judgments)

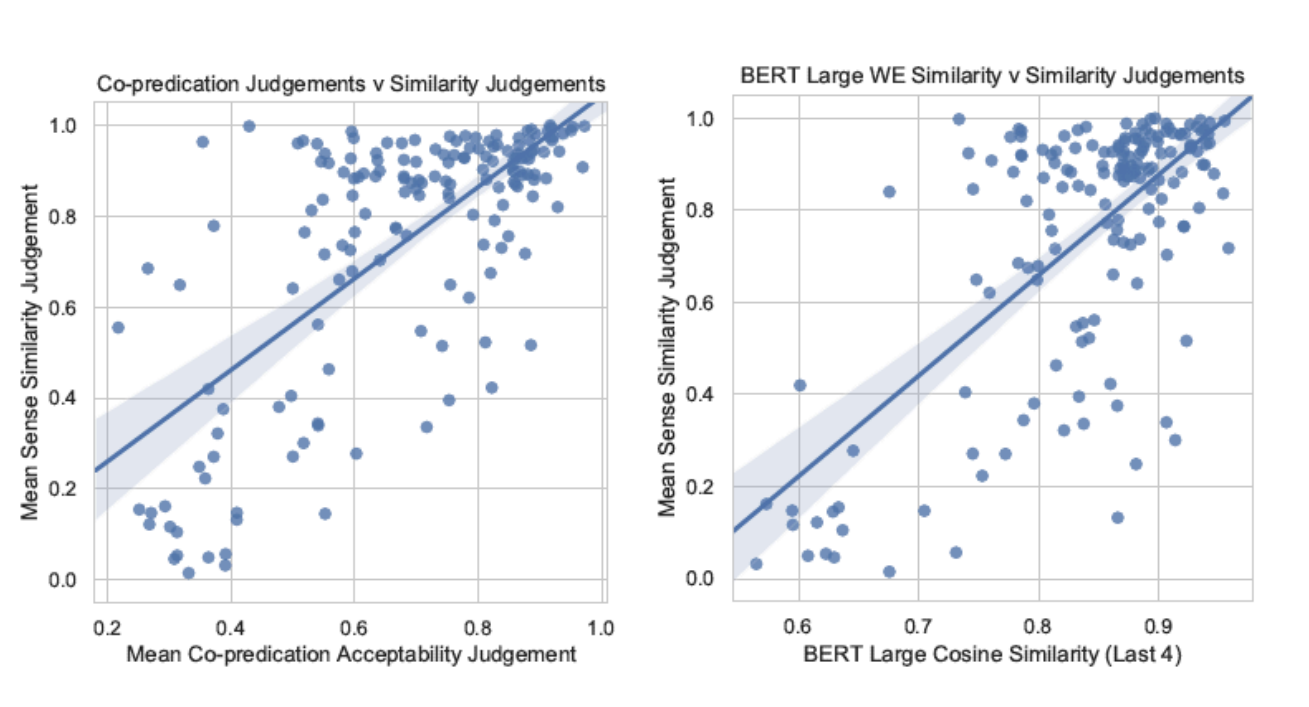
Combination		Correlation		Ordinary Least Squares (OLS) Regression Analysis					
First Measure	Second Measure	r	p	Coef.	R^2	F-stat.	Prob.	Omnib.	Prob.
Similarity	Acceptability	0.698	1.09E-25	0.484	0.487	156.571	1.09E-25	9.733	0.008
Acceptability	Similarity	0.698	1.09E-25	1.005	0.487	156.571	1.09E-25	0.967	0.617
Word2Vec	Similarity	0.206	0.008	0.675	0.042	7.309	0.008	31.562	0
Word2Vec	Acceptability	0.311	4.39E-05	0.707	0.097	17.625	4.39E-05	9.668	0.008
ELMo	Similarity	0.515	1.11E-12	2.863	0.265	59.475	1.11E-12	10.43	0.005
ELMo	Acceptability	0.523	4.39E-13	2.018	0.273	61.973	4.39E-13	6.552	0.038
BERT Base	Similarity	0.641	1.02E-20	4.070	0.411	115.185	1.02E-20	3.496	0.174
BERT Base	Acceptability	0.560	3.43E-15	2.469	0.314	75.521	3.43E-15	2.07	0.355
BERT Large	Similarity	0.687	1.22E-24	2.181	0.472	147.361	1.22E-24	15.96	0
BERT Large	Acceptability	0.550	1.40E-14	1.212	0.302	71.520	1.40E-14	5.324	0.07

Table 2: Correlations between measures of contextualised word sense similarity. The first set of columns displays pairwise correlation based on Pearson's *r*, the second set shows the key statistics obtained from an OLS regression analysis. BERT results for summing over the last four hidden states.

- ELMo and BERT show similar performance in predicting co-pred acceptability
- BERT Large is the best performing model when predicting similarity scores

Results

2. Computational ratings (vs judgments)



- BERT Large seems to be able to capture nuanced word sense distinctions in a similar way as human annotators

3. Sense similarity patterns

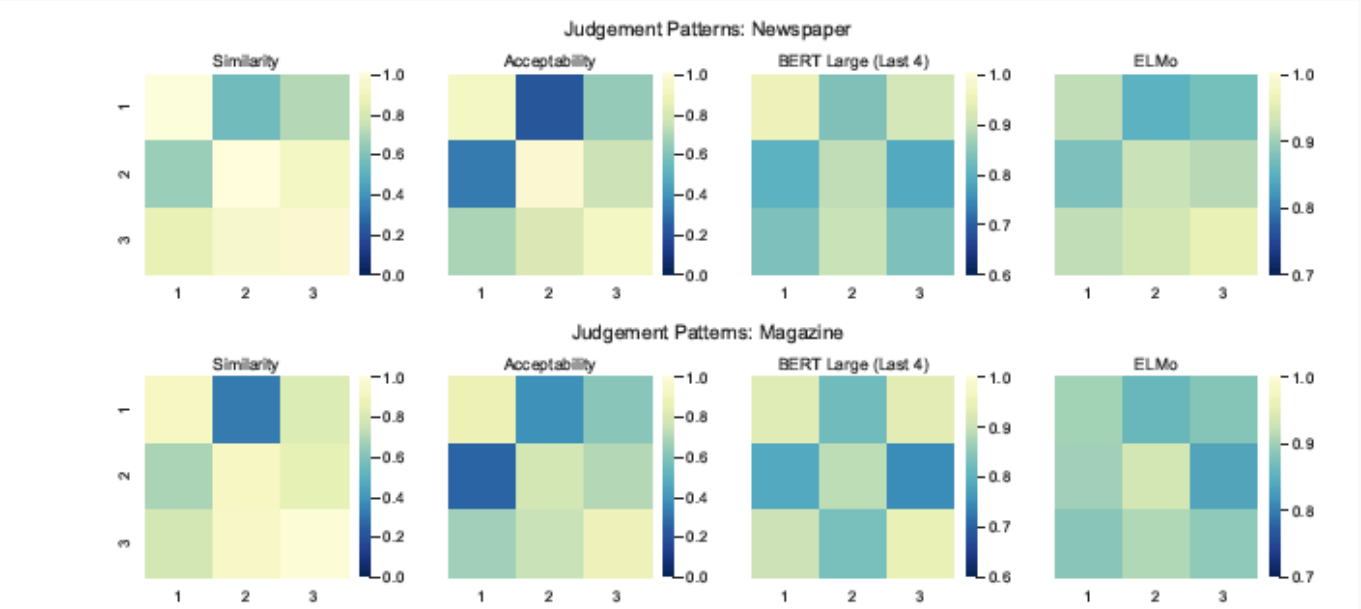


Figure 4: Similarity patterns in the sense similarity ratings for polysemes *newspaper* and *magazine*. Senses: 1-physical, 2-information, 3-organisation. Colour scales adjusted for computational measures.

- 1-physical, 2-information, 3-organisation
  - ! (Haber & Poesio 2020) 1-organisation, 2-physical object, 3-information
- Correlation for human ratings - *sim*:0.89 (p=0.001), *accept*:0.95 (p=6.88e-05)
- Correlation for computational ratings - *BERT*:0.65 (p=0.06), *ELMo*:0.34 (p=0.37)

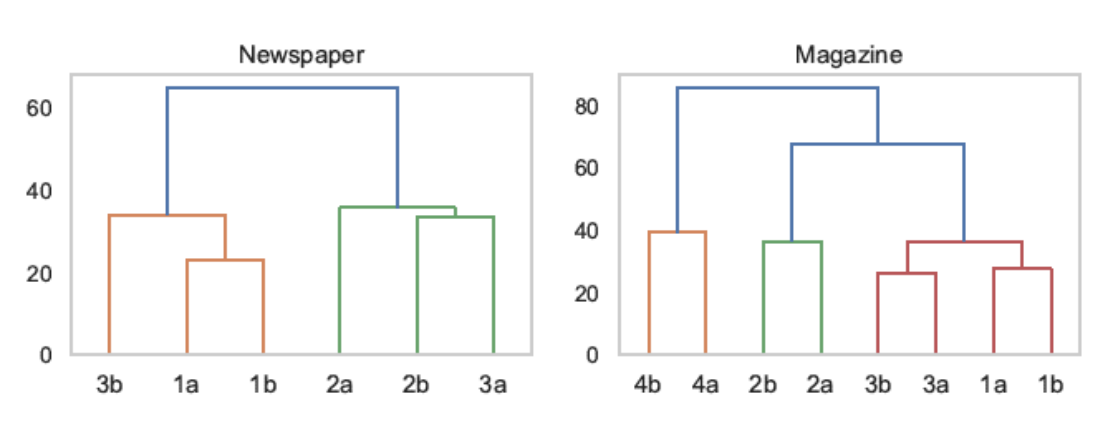
3. Sense similarity patterns

Measure	Pairwise		Overall	
	r	p <0.05	*r*	p
Similarity	0.44	3/24 (12.5%)	0.53	8.260e-10
Acceptability	0.44	4/24 (16.7%)	0.62	5.306e-14
ELMo	0.14	0/24 (0%)	0.21	0.025
BERT Large	0.28	1/24 (4.2%)	0.27	0.003

Table 3: Mean Pearson correlation of polysemic word sense similarity patterns across different target words allowing the same alternation of senses, number of significant comparisons, and overall pattern correlation.

- Correlations accross target words of the same polysemy type
- Sense similarity patterns are best to be investigated within a given type of alternation
    - Consistent similarity patterns (eg. *animal/meat*)
    - Inconsistent similarity patterns (eg. *content/container*)

3. Sense clustering



- Investigating how well BERTs contextualised embeddings can be used to cluster our polysemous targets according to their interpretation
- Hierarchical clustering of BERT Large's contextualised target encodings
    - 1-organisation, 2-physical object, 3-information (Haber & Poesio 2020) ?
- The clustering of alternations like *food/event*, *animal/meat* and *process/result* appears work consistently well, while others like the *content-for-container* alternation lead to consistently wrong sense groupings

Related work

- Datasets that capture graded similarity judgements
  - Word pairs in isolation (Taieb et al., 2019)
  - Small number of items (Erk et al., 2013)
  - Binary classification (Pilehvar and Camacho- Collados 2019)
  - Distinct target forms (Armendariz et al., 2020)

- (Nair et al 2020) 32 polysemic and homonymic word types extracted from the Semcor corpus
  - Polysemic senses are rated significantly more similar to one another in both the human annotations and BERT Base embeddings
  - Strong correlation between the cosine distance of BERT sense centroids and aggregated relatedness judgements.

- (Trott and Bergen 2021) 112 polysemes and homonyms
  - One noticeable difference can be found in the distribution of cross-sense polyseme ratings (almost even distribution of similarity scores)
    - Regular metonymic polysemes *vs* metaphoric polysemy in Trott and Bergen's data
    - Use of compound noun to disambiguate target words