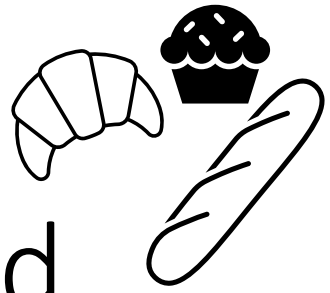
PASTRIE: A Corpus of Prepositions Annotated with Supersense Tags in Reddit International English



Michael Kranzlein
Emma Manning
Siyao Peng
Shira Wein
Aryaman Arora
Nathan Schneider

Georgetown University

Linguistic Annotation Workshop
@ COLING ‘20

nert

Overview

We've created a new corpus with preposition supersense annotations following the SNACS schema.

Overview

The corpus contains English text from Reddit.

Overview

The text was produced by **presumed** native speakers of 4 languages: English, French, German, and Spanish.

Outline

Motivation

Annotation schema

Data and the annotation process

Analysis

Motivation

- Prepositions are difficult for learners (Littlemore and Low, ‘06; Mueller, ‘12)
- Prepositions are highly polysemous (Hwang et al., *SEM ‘17)
    a) It is **at**/LOCATION 123 Main St.
    b) We met him **at**/TIME 7pm.
    c) Everyone pointed **at**/GOAL him.
    d) She laughed **at**/STIMULUS my acting.
    e) He held her **at**/INSTRUMENT gunpoint.

Motivation

- Prepositions make up a substantial portion of English usage—about 10%
- We need ways to study how preposition use and meaning varies for:
  - Understanding acquisition
  - Improving teaching methodologies
  - Creating NLP systems that do better with prepositions

Motivation

- Building a corpus is one way to study this
- Preposition supersense corpora exist in several languages
  - English (Schneider et al., ACL ’18)
  - Chinese (Peng et al., LREC ’20)
  - Korean (Hwang et al., DMR ’20)
- Existing corpora contain relatively homogenous data from sources like:
  - English Web Treebank
  - *The Little Prince*

Motivation

We want to look at data from speakers with varying L1 backgrounds.

Motivation

1. Build a resource that facilitates study of preposition use by high-proficiency L2 English speakers.
2. Conduct initial analysis of preposition and supersense distributions by L1.

SNACS: Semantic Network of Adpositions and Case Supersenses v2.5 (Schneider et al., '20)

SNACS is a hierarchical inventory of supersenses to capture the semantic contribution of preposition-like tokens.

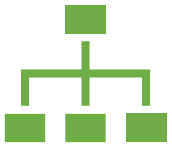
SNACS offers us different groupings of meanings for prepositions.

Guidelines: https://arxiv.org/abs/1704.02134

- Circumstance
  - Temporal
    - Time
      - StartTime
      - EndTime
    - Frequency
    - Duration
    - Interval
  - Locus
    - Source
    - Goal
  - Path
    - Direction
    - Extent
  - Means
  - Manner
  - Explanation
    - Purpose
- Participant
  - Causer
    - Agent
  - Theme
    - Topic
  - Ancillary
  - Stimulus
  - Experiencer
  - Originator
  - Recipient
  - Cost
  - Beneficiary
  - Instrument
- Configuration
  - Identity
  - Species
  - Gestalt
    - Possessor
    - Whole
    - Org
    - QuantityItem
  - Characteristic
    - Possession
    - PartPortion
      - Stuff
    - OrgMember
    - QuantityValue
      - Approximator
  - Ensemble
  - ComparisonRef
  - RateUnit
  - SocialRel

SNACS: Semantic Network of Adpositions and Case Supersenses

SNACS covers more than strictly prepositions.



Prepositions	**for, over, in, above**
English possessive markings	**'s, my, their**
Prepositional multiword expressions (MWEs)	**in_front_of, at**_large
Intransitive particles	He flew **away**

SNACS: Semantic Network of Adpositions and Case Supersenses

Annotation targets are marked with 1 supersense.

The menu had plenty of options even **for**/BENEFICIARY picky eaters .

BENEFICIARY: Animate or personified undergoer that is (potentially) advantaged or disadvantaged by the event or state.

SNACS: Semantic Network of Adpositions and Case Supersenses

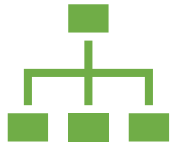
Annotation targets are marked with **1?** supersense.

The menu had plenty of options even **for**/BENEFICIARY picky eaters .

This was a flavorful , enjoyable meal **for**/? both of us .

EXPERIENCER: Animate who is aware of a bodily sensation, perception, emotion, or mental state.

SNACS: Semantic Network of Adpositions and Case Supersenses



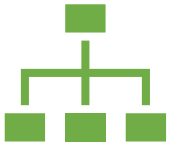
Annotation targets are marked with ~~1~~ 2 supersenses.

The menu had plenty of options even **for**/BENEFICIARY picky eaters .

This was a flavorful , enjoyable meal **for**/EXPERIENCER↝BENEFICIARY both of us .

Scene role	Function

SNACS: Semantic Network of Adpositions and Case Supersenses



Annotation targets are marked with a **construal**—a pair of supersenses—in the format SCENEROLE↝FUNCTION. (Hwang et al., *SEM ’17)

The **scene role** marks the contribution of the preposition in the specific context.

The **function** describes the typical lexical meaning of the preposition.

If the scene role and function are the same, they’re only shown once.

This is why **my**/ORGMEMBER↝GESTALT employer has just finished updating 50 k users **from**/SOURCE XP **to**/GOAL Windows 7 .

Example: An annotated sentence from PASTRIE (annotation targets bolded)

Reddit-L2 Corpus
(Rabinovich et al., TACL '18)

- Source corpus for PASTRIE
- 3.8B tokens
- Heuristically identified user L1s via flairs
    1. Identified user with self-specified country flairs on subreddits like r/AskEurope
    2. Gathered rest of user's posts and comments
- Corpus mostly used for native language identification, but a good fit for our purposes

PASTRIE

- 1,100 sentences
- 22,000 tokens
- 2,200 annotation targets
- 4 L1s and 10 countries represented

L1	Countries Sampled
English (24%)	Australia, New Zealand, UK, US
French (24%)	France
German (28%)	Austria, Germany
Spanish (24%)	Argentina, Mexico, Spain

Annotation Process

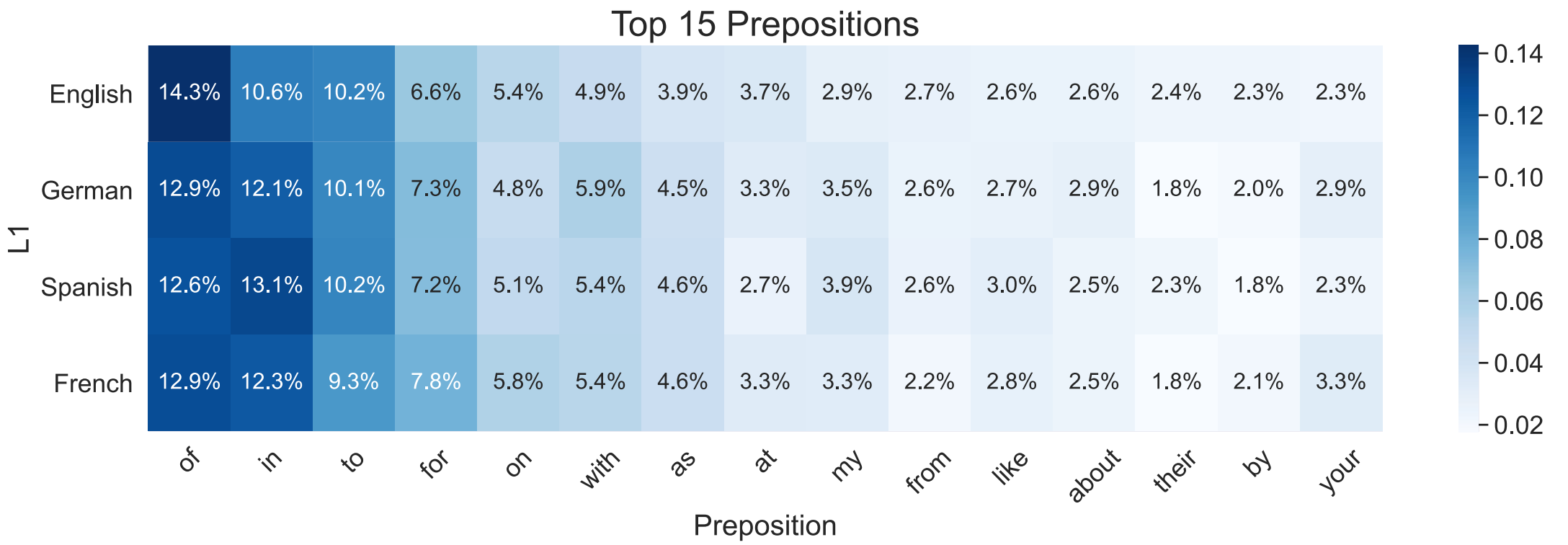
- We broke our sample of the Reddit-L2 corpus into smaller chunks for annotation
- Sample contained 15 Reddit posts or replies (“documents”)
- A document from all 4 L1s appeared at least once in every sample
- Each document was double-annotated and then adjudicated with an additional person

Analysis

L1	Prepositions / Token
English	10.70%
French	10.18%
German	10.69%
Spanish	11.00%

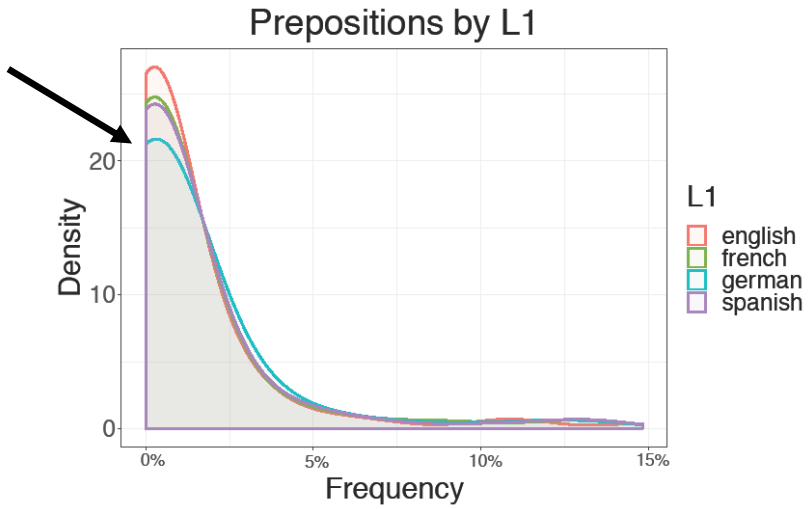
- Similar rates of raw preposition usage (10-11%)
- Similar ordering of most frequent prepositions
- Some differences in frequency

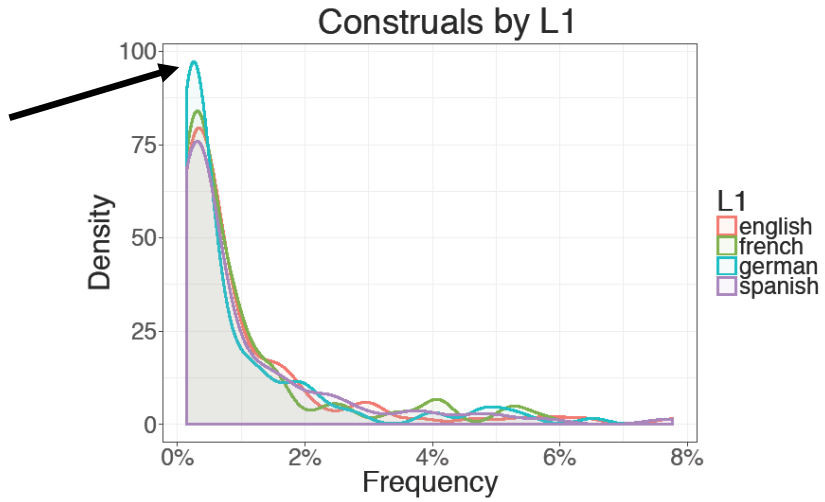
Analysis


Top 15 Preposition
English 14.3% 10.6% 10.2% 6.6% 5.4% 4.9% 3.9% 3.7% 2.9% 2.7% 2.6% 2.6% 2.4% 2.3% 2.3%
German 12.9% 12.1% 10.1% 7.3% 4.8% 5.9% 4.5% 3.3% 3.5% 2.6% 2.7% 2.9% 1.8% 2.0% 2.9%
Spanish 12.6% 13.1% 10.2% 7.2% 5.1% 5.4% 4.6% 2.7% 3.9% 2.6% 3.0% 2.5% 2.3% 1.8% 2.3%
French 12.9% 12.3% 9.3% 7.8% 5.8% 5.4% 4.6% 3.3% 3.3% 2.2% 2.8% 2.5% 1.8% 2.1% 3.3%
L1
of in to for on with as at my from like about their by your
Preposition
0.14 0.12 0.10 0.08 0.06 0.04 0.02

Analysis

- We see German L1 speakers have fewer low-frequency prepositions, and more low-frequency construals
- English L1s have the most low-frequency prepositions
- All L1 speakers' use is dominated by a handful of prepositions





Analysis

- The most common supersenses are similar for each L1
- The most common construals have the same scene role and function

	All		English		French		German		Spanish	
Scene Role	Locus	168	`i	45	Topic	35	Topic	49	Locus	51
	Topic	155	Locus	41	Theme	35	Locus	41	Time	39
	Theme	139	Topic	39	Locus	35	CompRef.	38	Theme	38
	`i	137	Gestalt	38	Gestalt	30	Gestalt	37	Goal	36
	Gestalt	127	Theme	37	Circum.	28	Circum.	35	CompRef.	35
Function	Gestalt	325	Gestalt	88	Gestalt	74	Gestalt	87	Gestalt	76
	Locus	242	Locus	55	Locus	55	Locus	60	Locus	72
	Topic	154	Goal	49	Topic	33	Topic	51	Topic	36
	Goal	153	`i	45	CompRef.	30	Goal	44	Time	35
	`i	137	Topic	34	Goal	28	CompRef.	35	Goal	32
Construal	Locus	146	`i	45	Topic	31	Topic	44	Locus	46
	Topic	137	Gestalt	37	Locus	29	Locus	37	Time	35
	`i	137	Locus	34	Gestalt	28	`i	35	Topic	31
	Gestalt	121	Topic	31	`i	28	Gestalt	34	`i	29
	Time	106	Goal	27	`d	22	Circum.	32	Theme	27
	Total	2395	**Total**	579	**Total**	539	**Total**	675	**Total**	602

Top scene roles, functions, and construals by L1

Analysis

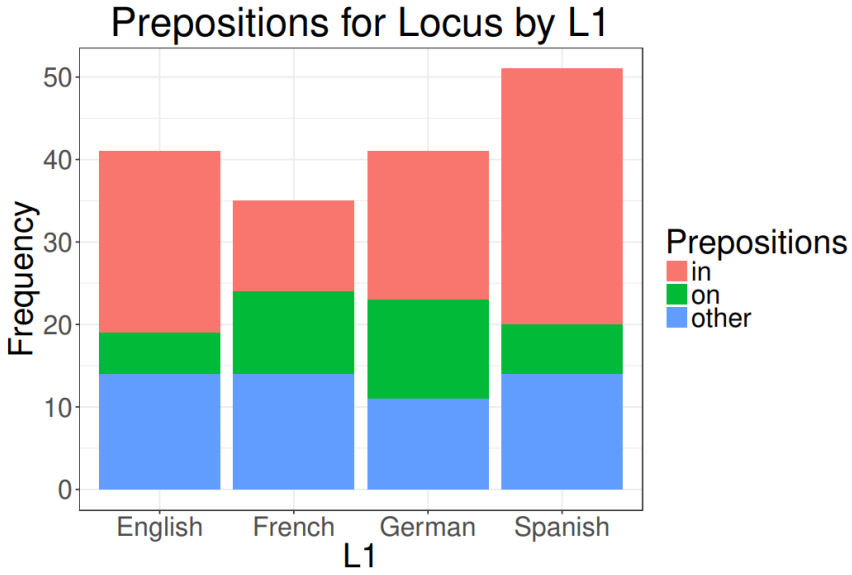
- Jaccard similarity scores suggest similar general set of meanings filled by prepositions

L1	vs. L1	Scene	Fxn.	Cons.
English	vs. German	0.71	0.73	0.61
English	vs. French	0.71	0.76	0.61
French	vs. Spanish	0.70	0.76	0.59
English	vs. Spanish	0.70	0.73	0.61
French	vs. German	0.69	0.71	0.60
German	vs. Spanish	0.67	0.72	0.61

Jaccard similarity scores for each L1 pair

A closer look at Locus

- In the British National Corpus, “in” is 2.8x more frequent than “on”
- In PASTRIE, this drops to 2.3x
- High use of Locus “on” by L1 French and German speakers with
  - 1.1x for French
  - 1.5x for German
- Transfer likely plays a role (Šeškauskienė and Juknevičienė, ’20)



Conclusion

New corpus: PASTRIE

A new supersense-annotated English corpus enabling future study of preposition use by high-proficiency English speakers from 4 L1 backgrounds

Initial Analysis

Explored similarities and differences in prepositions and their supersenses by L1

Conclusion

The corpus is publicly available, opening the door to future analysis.

Feel free to use it!

https://github.com/nert-nlp/pastrie

Acknowledgments

We thank our annotators Tripp Maloney, Ryan Mannion, and Sasha Slone.

We thank Shuly Wintner, Ella Rabinovich, and Liat Nativ for helping us sample data from the Reddit-L2 corpus.

References

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. 2017. Double trouble: the problem of construal in semantic annotation of adpositions. In *Proc. of *SEM*, pages 178–188, Vancouver, Canada.

Jena D. Hwang, Hanwool Choe, Na-Rae Han, and Nathan Schneider. 2020. K-SNACS: Annotating Korean Adposition Semantics. In *Proc. of 2nd International Workshop on Designing Meaning Representations*.

Jeannette Littlemore and Graham D Low. 2006. Figurative thinking and foreign language learning. *Springer*.

Charles M Mueller. 2012.Comparison of an Integrative Inductive Approach, Presentation-and-Practice Approach, and Two Hybrid Approaches to Instruction of English Prepositions. Ph.D. thesis, University of Maryland.

Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao, and Nathan Schneider. 2020. A Corpus of Adpositional Supersenses for Mandarin Chinese. In *Proc. of 12th Language Resources and Evaluation Conference*, pages 5986-5994, Marseille, France.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, viram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of English prepositions and possessives. In *Proc. of ACL*, pages 185–196, Melbourne, Australia.

Nathan Schneider, Jena D. Hwang, Archna Bhatia, Vivek Srikumar, Na-Rae Han, Tim O'Gorman, Sarah R.Moeller, Omri Abend, Adi Shalev, Austin Blodgett, and Jakob Prange. 2020. Adposition and CaseSupersenses v2.5: Guidelines for English. *arXiv:1704.02134v6 [cs]*.

Inesa Šeškauskienė and Rita Juknevičienė. 2020. Prepositions in L2 written English, or why on poses more difficulties than in. *Nordic Journal of English Studies*, 19(1)

If you're interested in the SNACS annotation schema, check out these other SNACS papers being presented at COLING 2020 workshops:

Korean corpus (DMR)
Proxy annotation tasks (LAW)
Accompaniment and purpose (LAW)