

# **Polysemy in Spoken Conversations and Written Texts**

Aina Garí Soler, Matthieu Labeau, Chloé Clavel

June 21-23  
LREC 2022  
Marseille & online

# Textual Polysemy

We use polysemous words (i.e., words with multiple senses) all the time

} **POTENTIAL**  
polysemy

Sometimes even in different senses within one discourse

} **OBSERVED**  
polysemy

- This does not necessarily hinder communication
- For example, speakers may exploit polysemy for metaphors, or to showcase their sense of humor with irony and jokes... (Nerlich and Clarke, 2001)
- ... or also to illustrate a point in this presentation
- ... and to do it twice, to the point where the audience has had enough of it.

# Textual Polysemy: Background

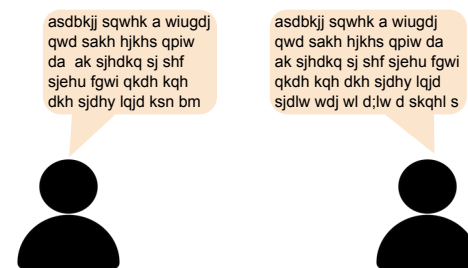
- **“One Sense per Discourse”** (OSD) hypothesis (Gale et al., 1992)
  - Used as a heuristic for Word Sense Disambiguation (WSD), Named Entity Recognition and Machine Translation (Cucerzan, 2007; Ture et al., 2012; Pilehvar and Navigli, 2015; Chaplot and Salakhutdinov, 2018...)
  - But also put into question (Krovetz 1998, Leacock et al., 1998)
- Potential polysemy of a corpus can be used as an indication of its **WSD-difficulty**
- Work on quantifying textual polysemy has focused on **written, monolog-like discourse** (McCarthy et al., 2007; Pasini and Camacho-Collados, 2020; Barba et al., 2021)



**What about spoken language and dialog?**

# Dialog vs Monolog

- Differences between speakers (background, world knowledge, idiolect, language level, opinion...)
- In dialog: misunderstandings, disagreements



- The monolog speaker may be unaware of potential ambiguities in their discourse
- In dialog, speakers often align in the way they talk (Pickering and Garrod, 2004)

# What we do

- We compare the observed polysemy level of texts of different nature
  - We perform **automatic WSD** on multiple different datasets,
  - we calculate different **polysemy measures** (and propose our own),
  - we **compare** the results for different kinds of texts.



**Data**

# (Spoken) Data

- 2020 US presidential **debate** (Joe Biden vs Donald Trump)



**15 Monologs**  
(longer topic-initial turns)



**6 Dialogs**  
(subsequent interaction about a topic)

- **lemocap** (Busso et al., 2008): hypothetical emotional conversations between actors



**71 Scripted**



**80 Spontaneous**

# (Spoken, Spontaneous, Dialog) Data

- **JUSThink** (Norman et al., 2021) 10 task-oriented conversations between children
- **Switchboard** (Stolcke et al., 2000) 1,126 conversations on a provided topic
- **HCRC MapTask Corpus** (Thompson et al., 1993) 128 task-oriented conversations
- **BT Oasis Corpus** (Leech and Weisser, 2003) 378 calls to British Telecom and Trainline operator services



# (Written) Data

Three WSD evaluation campaigns (Raganato et al., 2017)

- **Senseval-2** (Edmonds and Cotton, 2001) (3 texts)
- **Senseval-3** task 1 (Snyder and Palmer, 2004) (3 texts)
- **SemEval-15** task 13 (Moro and Navigli, 2015) (4 texts)

# Methodology

# Automatic WSD Annotation

1. Text preprocessing (tokenization, pos-tagging, lemmatization)
2. Automatic WSD annotation with ESCHER (Barba et al., 2021) of nouns, verbs and adjectives

Automatic Annotation Quality

**Accuracy**

**Written:** automatic evaluation

**73.4% - 79.9%**

**Spoken:** manual verification of 5 texts

**75.0% - 88.5%**

# Potential Polysemy Measures

(independent of sense annotations)

## Potential Ambiguity (PA)

average<sub>6</sub> number<sub>12</sub> of Wordnet senses<sub>5</sub> of instances<sub>2</sub> in a text<sub>4</sub>

$$\frac{6+12+5+2+4}{5} = 5.8$$

## Percentage of polysemous words (pct-poly)

proportion<sub>5</sub> of polysemous<sub>1</sub> word<sub>10</sub> instances<sub>2</sub> out of all word<sub>10</sub> instances<sub>2</sub> in a text<sub>4</sub>

$$\frac{6}{7} = 0.86$$

# Observed Polysemy Measures

## More than One Sense per Discourse (MOSD)

Percentage of repeated polysemous words used in more than one sense in a text  
(Krovetz, 1998)

## AVGSENSES

Average number of senses in which words in a discourse are used. It is similar to PA, but the number of senses of a word is the observed one and not the potential one taken from WordNet.

# Observed Polysemy Measures

## More than One Sense per Discourse (MOSD)

Percentage of repeated polysemous words used in more than one sense in a text  
(Krovetz, 1998)

## AVGSENSES

Average number of senses in which words in a discourse are used. It is similar to PA, but the number of senses of a word is the observed one and not the potential one taken from WordNet.

# Sense Granularity

WordNet is very fine-grained: we can use WordNet supersenses

NOUN

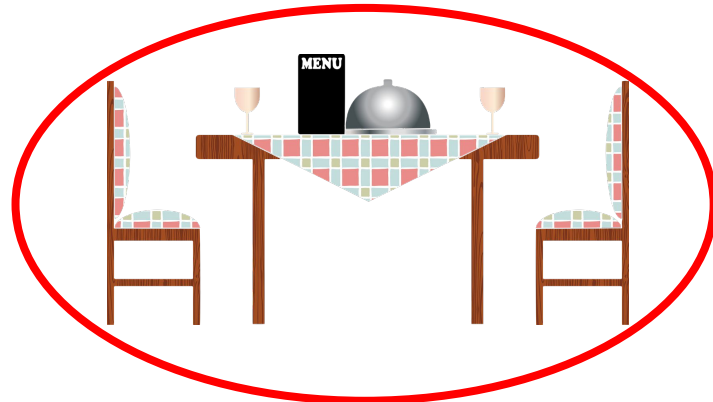
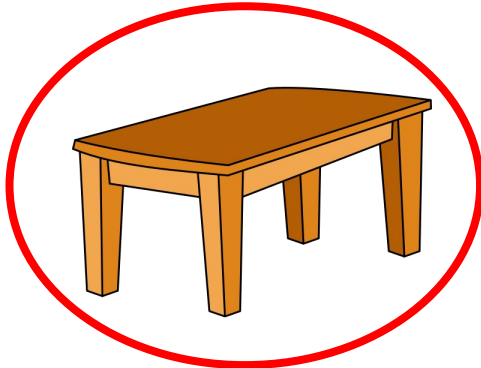
Tops, act, animal, artifact, attribute, body, cognition, communication, event, feeling, food, group, location, motive, object, person, phenomenon, plant, possession, process, quantity, relation, shape, state, substance, time

VERB

body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social, stative, weather

ADJ

all, pert, ppl



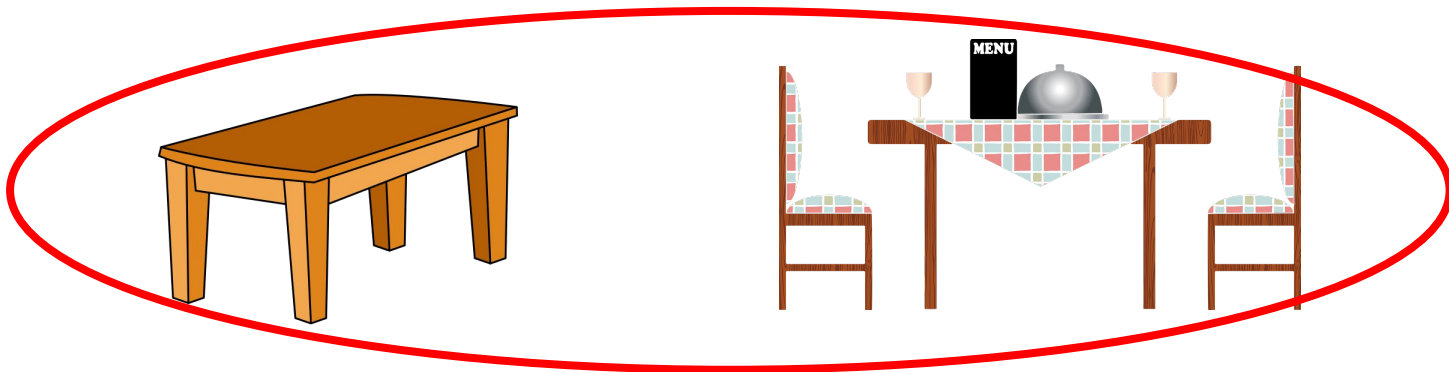
# Sense Granularity

WordNet is very fine-grained: we can use WordNet supersenses

NOUN    Tops, act, animal, **artifact**, attribute, body, cognition, communication, event, feeling, food, group, location, motive, object, person, phenomenon, plant, possession, process, quantity, relation, shape, state, substance, time

VERB    body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social, stative, weather

ADJ    all, pert, ppl





# Sense Granularity

WordNet is very fine-grained: we can use WordNet supersenses

NOUN

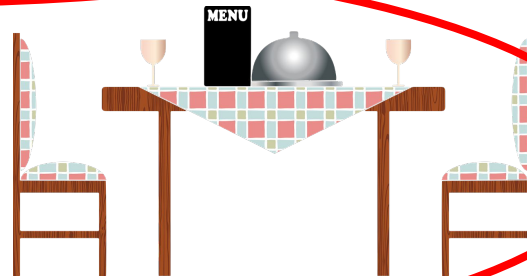
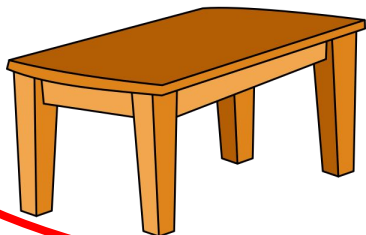
Tops, act, animal, **artifact**, attribute, body, cognition, communication, event, feeling, food, **group**, location, motive, object, person, phenomenon, plant, possession, process, quantity, relation, shape, state, substance, time

VERB

body, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social, stative, weather

ADJ

all, pert, ppl



# Discourse Length and Lexical Diversity

- **Discourse length** correlates **positively** with **observed polysemy**  
(longer text -> more occasions to reuse a word in a different sense)
- **Lexical diversity** correlates **negatively** with **observed polysemy**  
(lower diversity -> words are reused more often -> higher chances of encountering different senses for a word)

# Results

# Results

	POTENTIAL		OBSERVED		
Discourse type	PA	PCT-POLY	MOSD	AVGSENSES	ENTROPY
Spontaneous Spoken Dialog	10.6	91.9	37.1	<b>1.75</b>	0.24
Scripted Spoken Dialog	9.9	93.1	<b>44.1</b>	1.51	<b>0.32</b>
Spontaneous Spoken Monolog	<b>10.9</b>	<b>93.6</b>	33.1	1.28	0.23
Scripted <b>Written</b> Monolog	<b>6.9</b>	<b>85.1</b>	<b>28.2</b>	<b>1.19</b>	<b>0.17</b>

# Results

	POTENTIAL		OBSERVED		
Discourse type	PA	PCT-POLY	MOSD	AVGSENSES	ENTROPY
Spontaneous Spoken <b>Dialog</b>	10.6	91.9	37.1	<b>1.75</b>	0.24
Scripted Spoken <b>Dialog</b>	9.9	93.1	<b>44.1</b>	1.51	<b>0.32</b>
Spontaneous Spoken <b>Monolog</b>	<b>10.9</b>	<b>93.6</b>	33.1	1.28	0.23
Scripted <b>Written</b> <b>Monolog</b>	<b>6.9</b>	<b>85.1</b>	<b>28.2</b>	<b>1.19</b>	<b>0.17</b>

Written text has a lower polysemy level than spoken text

# Results

	POTENTIAL		OBSERVED		
Discourse type	PA	PCT-POLY	MOSD	AVGSENSES	ENTROPY
Spontaneous Spoken <b>Dialog</b>	10.6	91.9	37.1	<b>1.75</b>	0.24
Scripted Spoken <b>Dialog</b>	9.9	93.1	<b>44.1</b>	1.51	<b>0.32</b>
Spontaneous Spoken <b>Monolog</b>	<b>10.9</b>	<b>93.6</b>	33.1	1.28	0.23
Scripted <b>Written</b> <b>Monolog</b>	<b>6.9</b>	<b>85.1</b>	<b>28.2</b>	<b>1.19</b>	<b>0.17</b>

# Conclusion and Future Work

- ❖ **Spoken** texts tend to present a **higher level of observed polysemy** than **written** discourse
- ❖ **Useful for WSD** of different kinds of text: the OSD heuristic may work less well on this kind of data
- ❖ What kind of words tend to be used in multiple senses in a discourse?
- ❖ Modeling the introduction of new senses into a conversation and the conceptual alignment between speakers

**Thank you!**