Word Meaning Representation and Negotiation

Aina Garí Soler

Postdoctoral Researcher at ALMAnaCH, INRIA
aina.gari-soler@inria.fr

Introduction: Personal background

Bachelor in Linguistics

LIBERTAS PERFVNDET · OMNIA LVCE

UNIVERSITAT DE BARCELONA

\+ Research assistant

upf. Universitat Pompeu Fabra Barcelona

Created with mapchart.net

Introduction: Personal background

Bachelor in Linguistics

UNIVERSITAT DE BARCELONA

St Petersburg University

Created with mapchart.net

Introduction: Personal background

Erasmus Mundus
Master in NLP

UNIVERSITÄT
DES
SAARLANDES

Created with mapchart.net

Introduction: Personal background

Erasmus Mundus
Master in NLP

UNIVERSITÄT
DES
SAARLANDES

L-Università
ta' Malta

Created with mapchart.net

Introduction: Personal background

PhD

LISN
LABORATOIRE INTERDISCIPLINAIRE
DES SCIENCES DU NUMÉRIQUE

université
PARIS-SACLAY

Postdoc 1
Postdoc 2

TELECOM
Paris

IP PARIS

Inria

Created with mapchart.net

Introduction: Research interests

Computational Lexical Semantics

How can **word meaning** be represented computationally?

out-of-context

compositional meaning

in-context

connotative meaning

semantic relationships

figurative meaning

Introduction: Research interests

Computational Lexical Semantics

🤔 But what even is (word) meaning??

Distributional semantics: A tangible, empirical solution!

(Imperfect, but it has taken us very far, in NLP and Computational Linguistics)

Introduction: Research directions

1. **Word Meaning Representation (in Neural Language Models)**

   How can word meaning be represented computationally?

2. **Word Meaning in Interaction**

   (How) do we manage to understand each other?

Word Meaning Representation

in Neural Language Models

~~The~~ An NLP Revolution

November 2017

BERT on arXiV (Devlin et al, 2018)

October 2018

June 2018, NAACL

Peters et al., (2018)

ELMo

https://www.wannapik.com

~~The~~ An NLP Revolution

November 2017
🥲

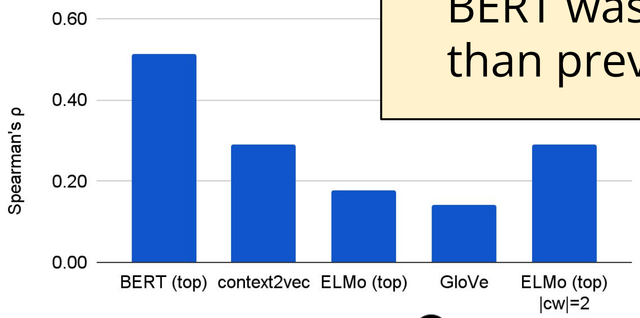
BERT on arXiV (Devlin et al, 2018)
October 2018

June 2018, NAACL
Peters et al., (2018)
ELMo

https://www.wannapik.com

How well do these models represent word meaning in context?

- Lexical substitution
- Word usage similarity

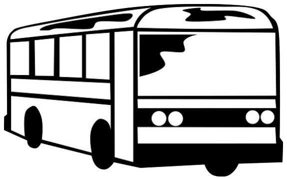


BERT was **way better** than previous models

This is also at the very essence or heart of being a **coach**.

1.3 / 10

We hopped back onto the **coach.**



bus
carriage
~~trainer~~

Garí Soler et al. (2019), IWCS
Garí Soler et al. (2019), *SEM

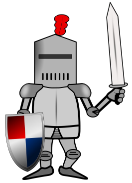
Does the semantic space built by contextual models reflect words' degree of polysemy?

Monosemous: one sense

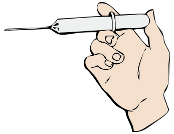
Polysemous: multiple senses

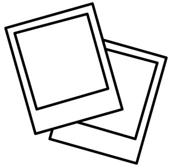
sofa

knight



shot





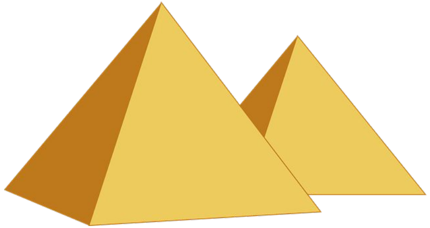
...

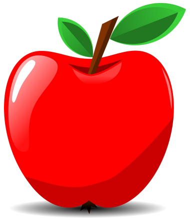
A word’s degree of polysemy is reflected in BERT representations, regardless of the context it is found in

Garí Soler and Apidianaki (2021), TACL

Semantic Relationships

old *vs* **ancient**





~~healthy~~ apple

healthy dessert

Garí Soler and Apidianaki (2020), EMNLP
Apidianaki and Garí Soler (2021), BlackBoxNLP

The Impact of Word Splitting on the Semantic Content of Contextualized Word Representations

Garí Soler, Labeau and Clavel (2024), TACL

Subword tokenization

Let's tokenize this sentence into subwords

let ' s token ##ize this sentence into sub ##words

- Rare / out-of-domain words

  `conjunctivitis` `[con ##jun ##ct ##iv ##itis]`

- Morphologically complex words

  `multiprocessor` `[multi ##pro ##ces ##sor]`

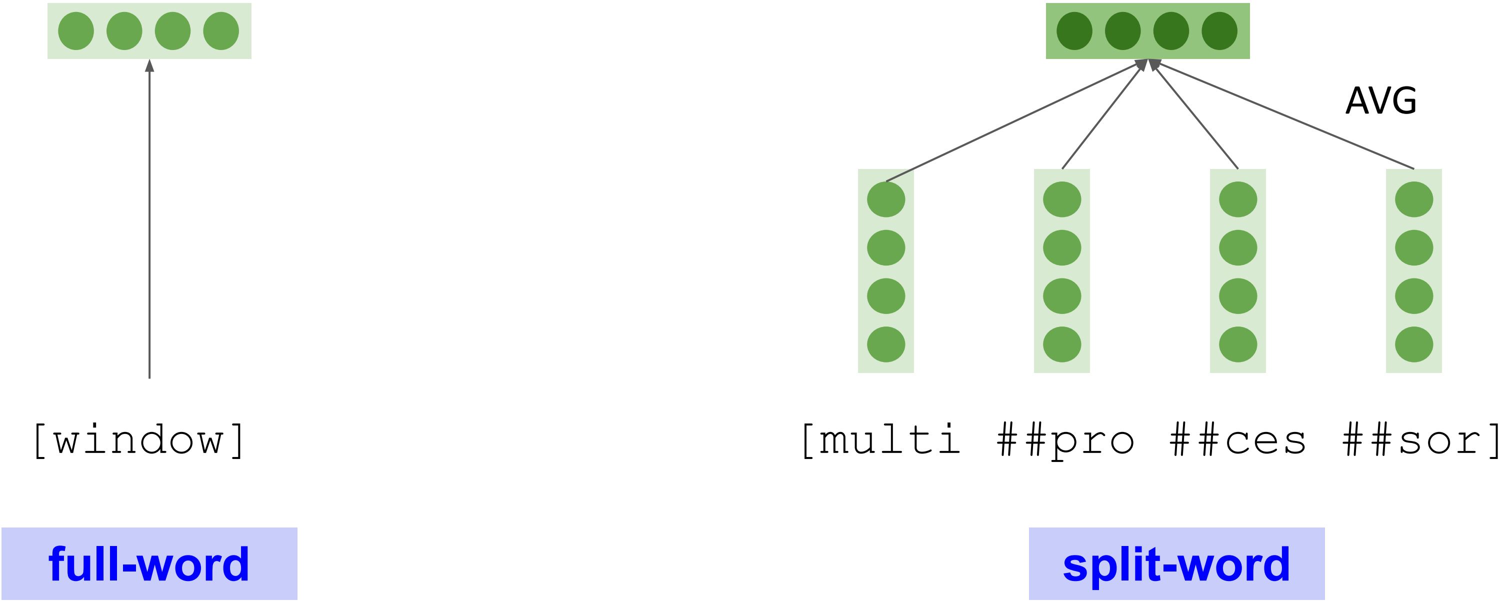
- Misspelled words

  `tabel, aaaaaand` `[tab ##el], ['aaa', '##aa', '##and']`

(Examples are obtained with bert-base-uncased)

Contextualized Word Representations

... But we often work at the word level!



Our questions

1. What is the best strategy to create a representation for split-words?

2. (Given a good strategy) how does the quality of the semantic content in split-word representations compare to that in full-word representations?

Our questions

The task

word similarity estimation

1. What is the best strategy to create a representation for split-words?

2. (Given a good strategy) how does the quality of the semantic content in split-word representations compare to that in full-word representations?

Our questions

1. What is the best strategy to create a representation for split-words?

2. (Given a good strategy) how does the quality of the semantic content in split-word representations compare to that in full-word representations?

Expectation: it's worse

Similarity and split-words

sim(w_1, w_2)

full-word vs **full**-word {accordion} vs {guitar} **0-SPLIT**

full-word vs **split**-word {ash, ##tray} vs {weather} **1-SPLIT**

split-word vs **split**-word {tom, ##fo, ##ole, ##ry} vs {loaf, ##ing} **2-SPLIT**

split-types

Inter-word similarity

➔ **Inter-word**

*... as an adult **adoptee**, this...*

*... she was a **descendant** of...*

In-context

Out-of-context

Datasets: CoSimLex (Armendariz et al., 2020), SWCS (Huang et al, 2012)

Inter-word similarity

➔ **Inter-word**

adoptee

descendant

In-context

Out-of-context

Datasets: SimLex-999 (Hill et al., 2015), WS535 (Agirre et al., 2009), CARD-660 (Pilehvar et al., 2018)...

Inter-word similarity

➔ **Inter-word**

adoptee

descendant

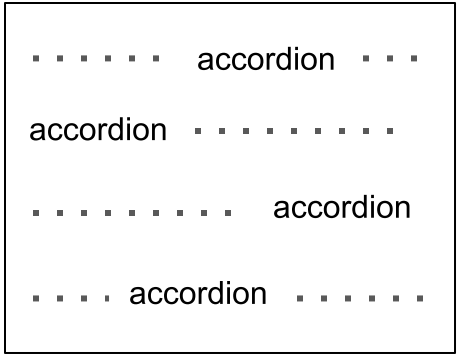
In-context

Out-of-context

Existing datasets have a weak representation of 1- and 2-SPLIT pairs

Inter-word similarity data: what we want

word1	word2	split-type	similarity
{accordion}	{guitar}	0-SPLIT	0.80
{tom, ##fo, ##ole, ##ry}	{loaf, ##ing}	2-SPLIT	0.63
{ethanol}	{fuel}	0-SPLIT	0.46
{ash, ##tray}	{weather}	1-SPLIT	0.24



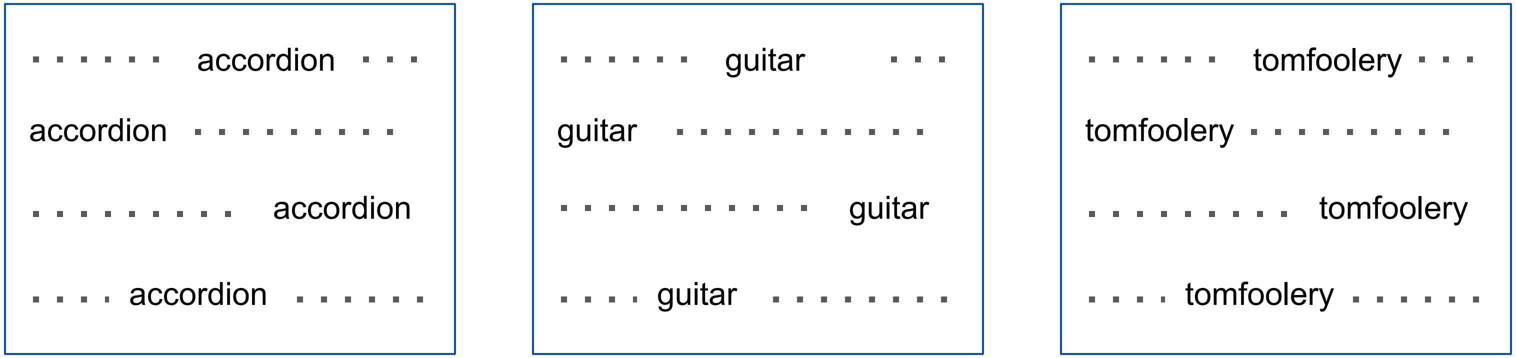
- Similarities vary with polysemy level and PoS: we separately analyze **monosemous/polysemous words and nouns/verbs**

Inter-word similarity data: words and sentences

1. Select all noun and verb lemmas in WordNet (Fellbaum, 1998)

    accordion, guitar, tomfoolery…

2. Extract at least 10 sentences per lemma in the c4 corpus (Raffel et al., 2020) (that contain the same lemma form & correct POS)



Inter-word similarity data: word pairs and similarities

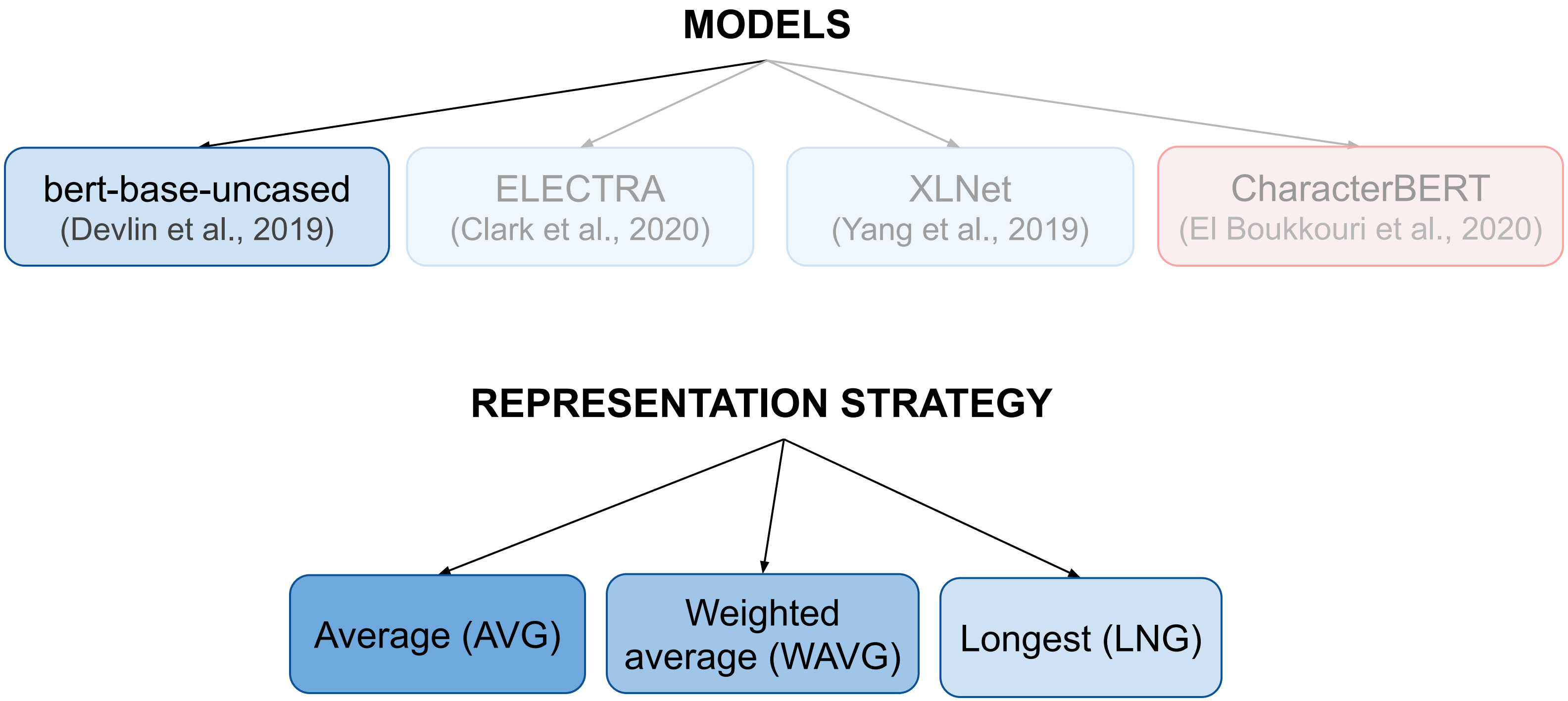
3. Exhaustively pair all lemmas and calculate their **WUP similarity** (Wu and Palmer, 1994)

(*accordion*, *guitar*)
(*accordion*, *tomfoolery*)
(*guitar*, *tomfoolery*)
...

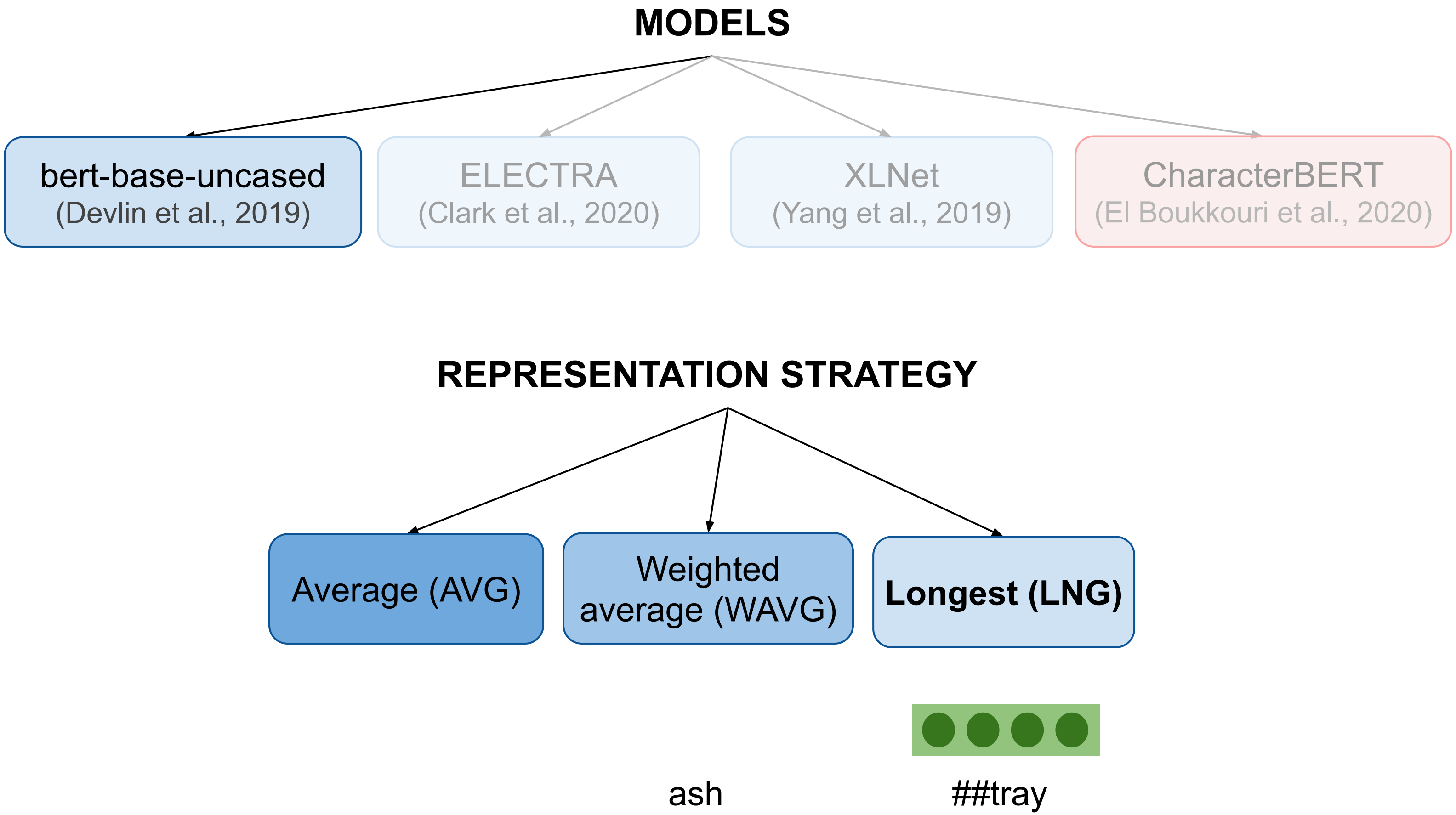
4. Select a subset ensuring a balanced representation of split-types and similarity ranges

SPLIT-SIM subset	**# pairs**
monosemous nouns (M-N)	67,500
monosemous verbs (M-V)	2,550
polysemous nouns (P-N)	15,000
polysemous verbs (P-V)	15,000

Experimental Setup



Experimental Setup



Experimental Setup

MODELS

bert-base-uncased (Devlin et al., 2019)

ELECTRA (Clark et al., 2020)

XLNet (Yang et al., 2019)

CharacterBERT (El Boukkouri et al., 2020)

REPRESENTATION STRATEGY

Average (AVG)

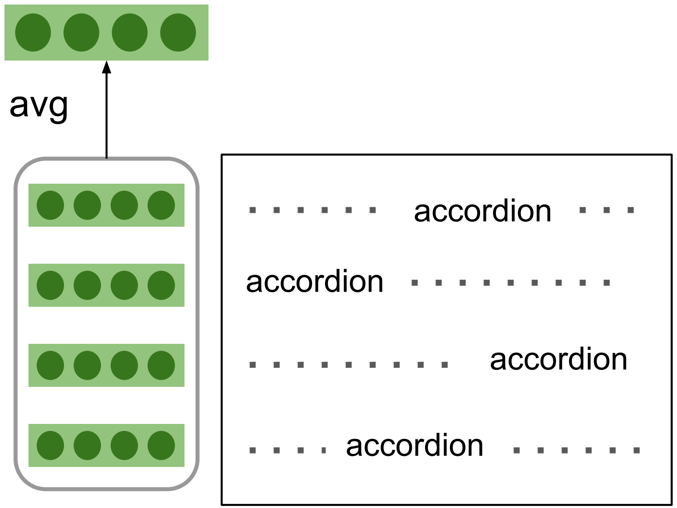
Weighted average (WAVG)

Longest (LNG)

3/7 * ash + 4/7 * ##tray

Experimental Setup

CONTEXTS: 10 sentences

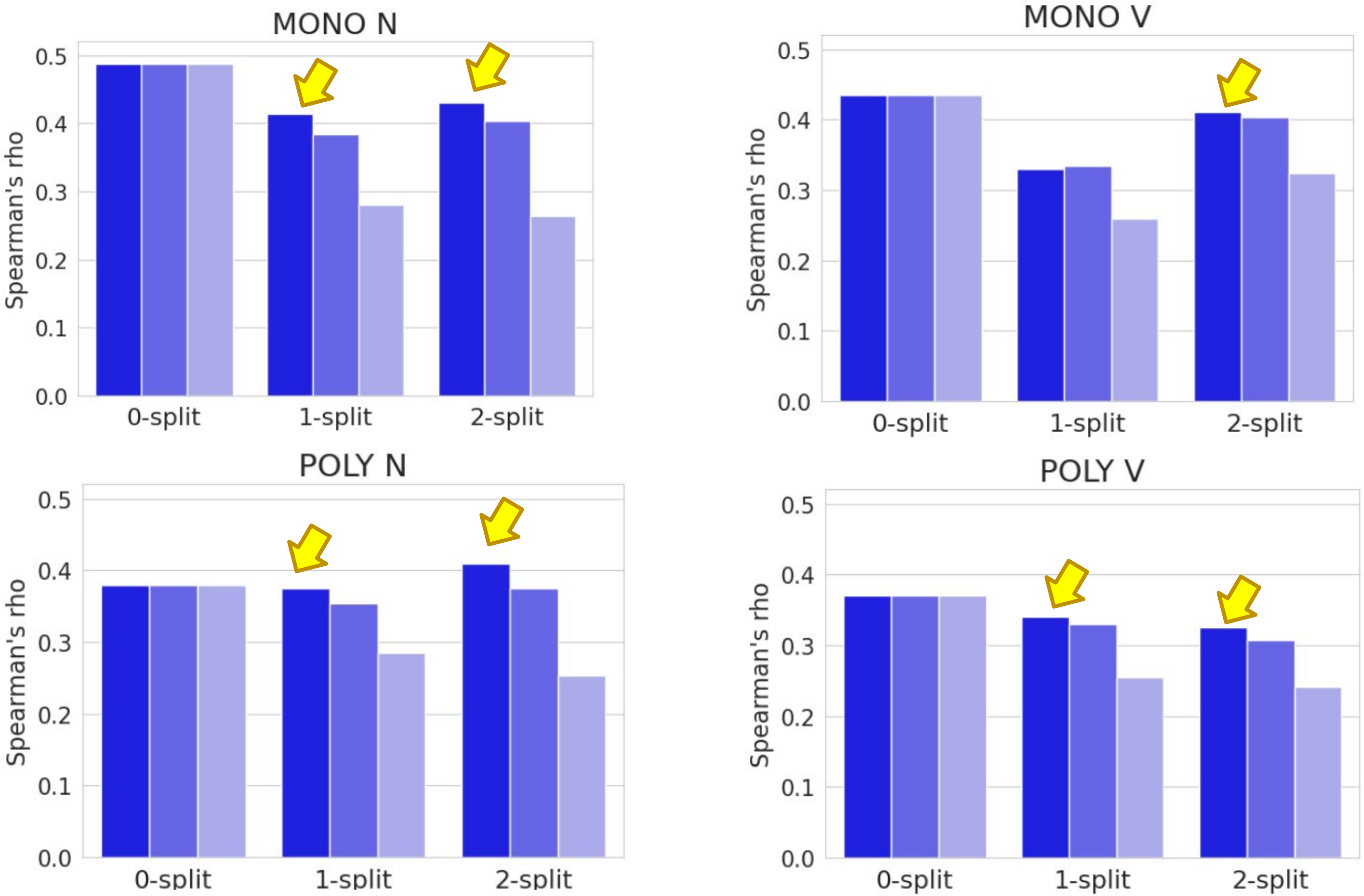


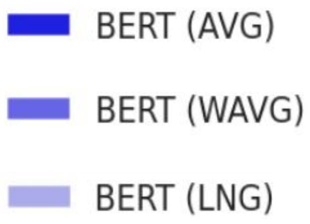
EVALUATION

Spearman's ρ between WUP similarity and cosine similarity

Results

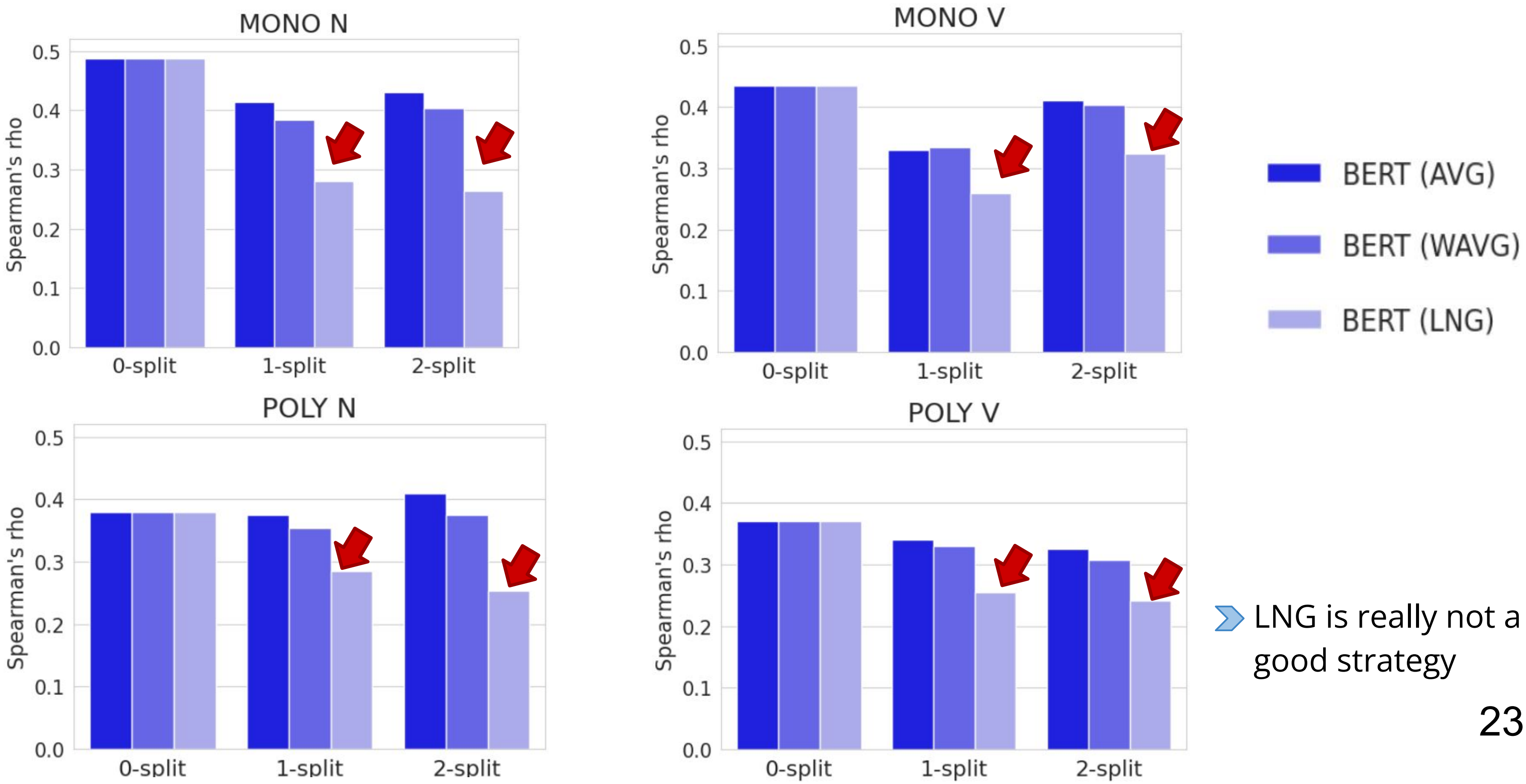
What is the best representation strategy?





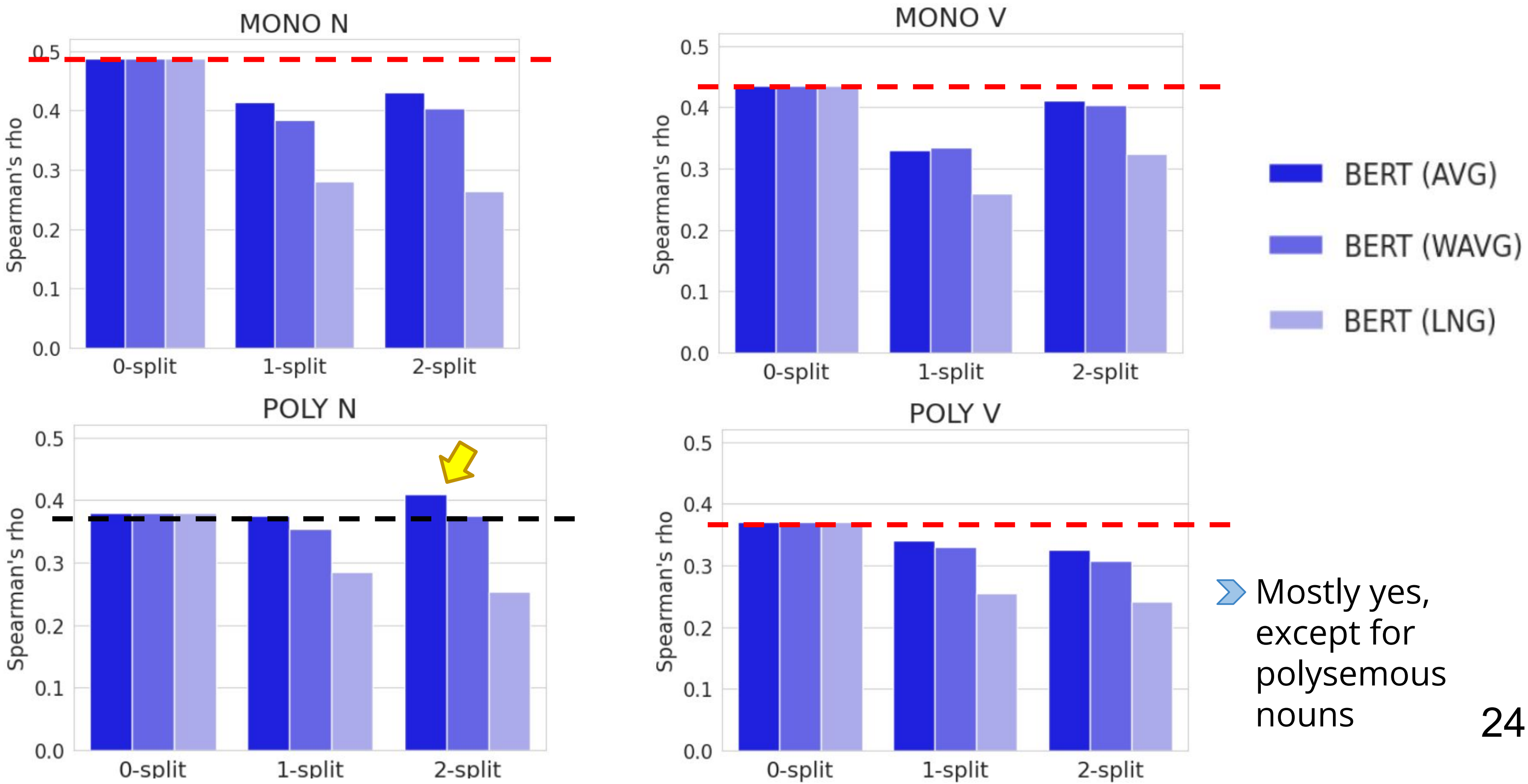
- The simple average

What is the best representation strategy?



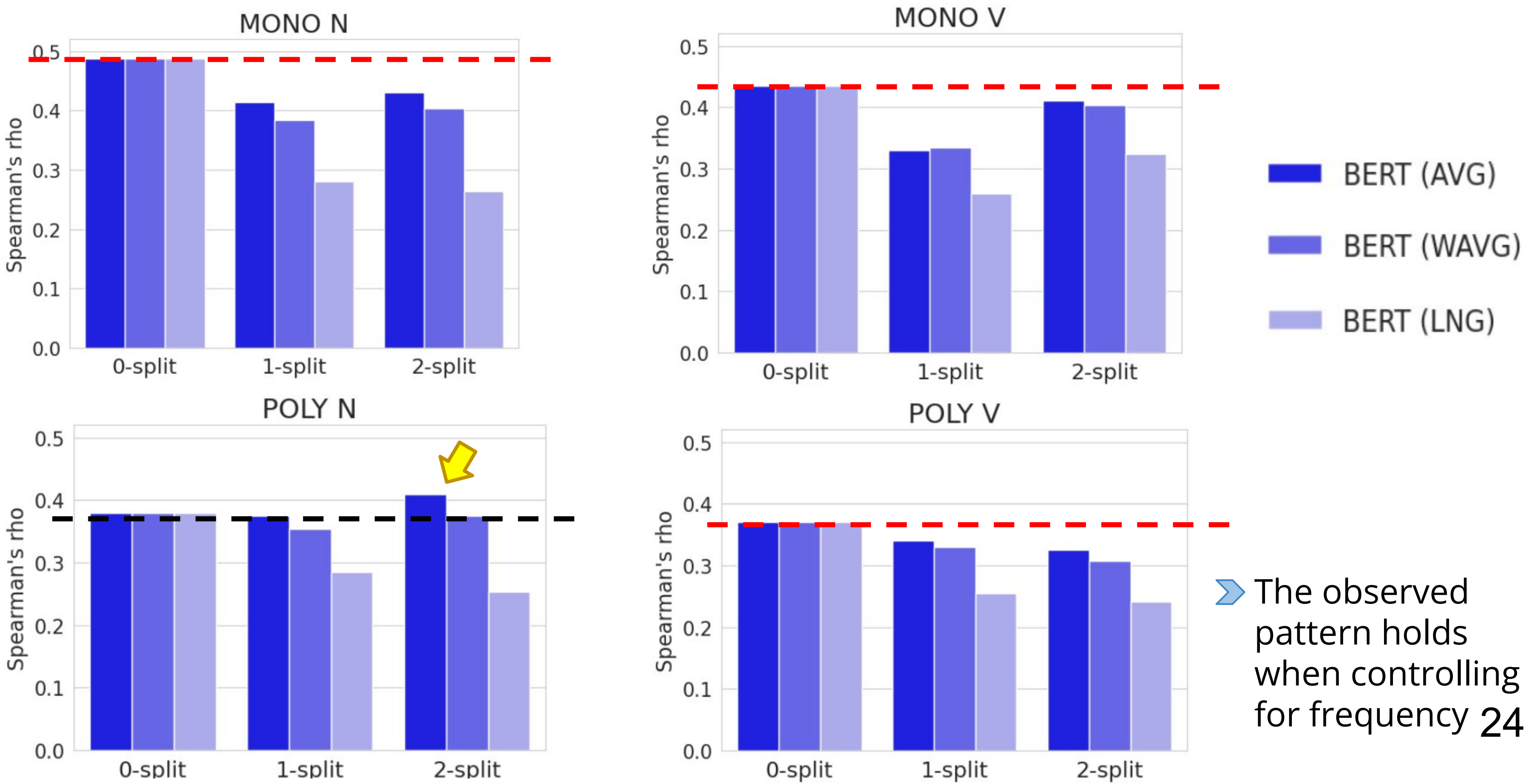
- LNG is really not a good strategy

Is performance on pairs involving split-words worse than on 0-SPLIT pairs?



- Mostly yes, except for polysemous nouns

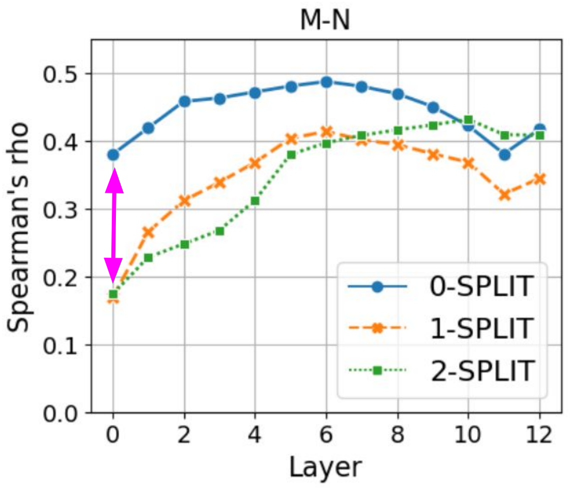
Is performance on pairs involving split-words worse than on 0-SPLIT pairs?

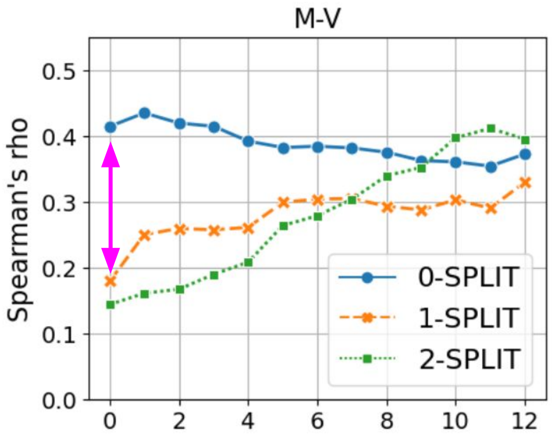


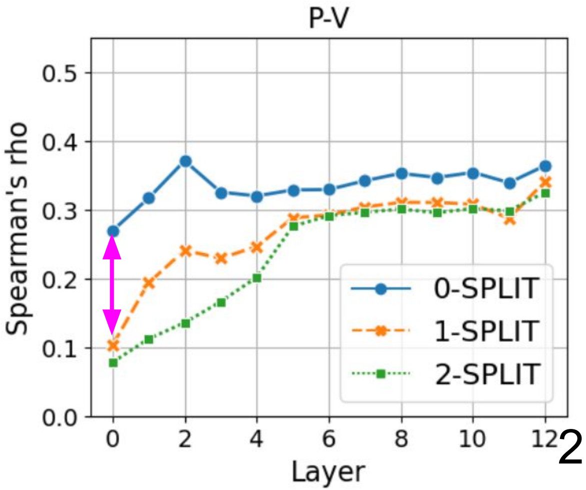
- The observed pattern holds when controlling for frequency

How do results change across layers for every split-type?

- At earlier layers the quality of 1- and 2-split similarity estimations is much lower than that of 0- SPLIT pairs.

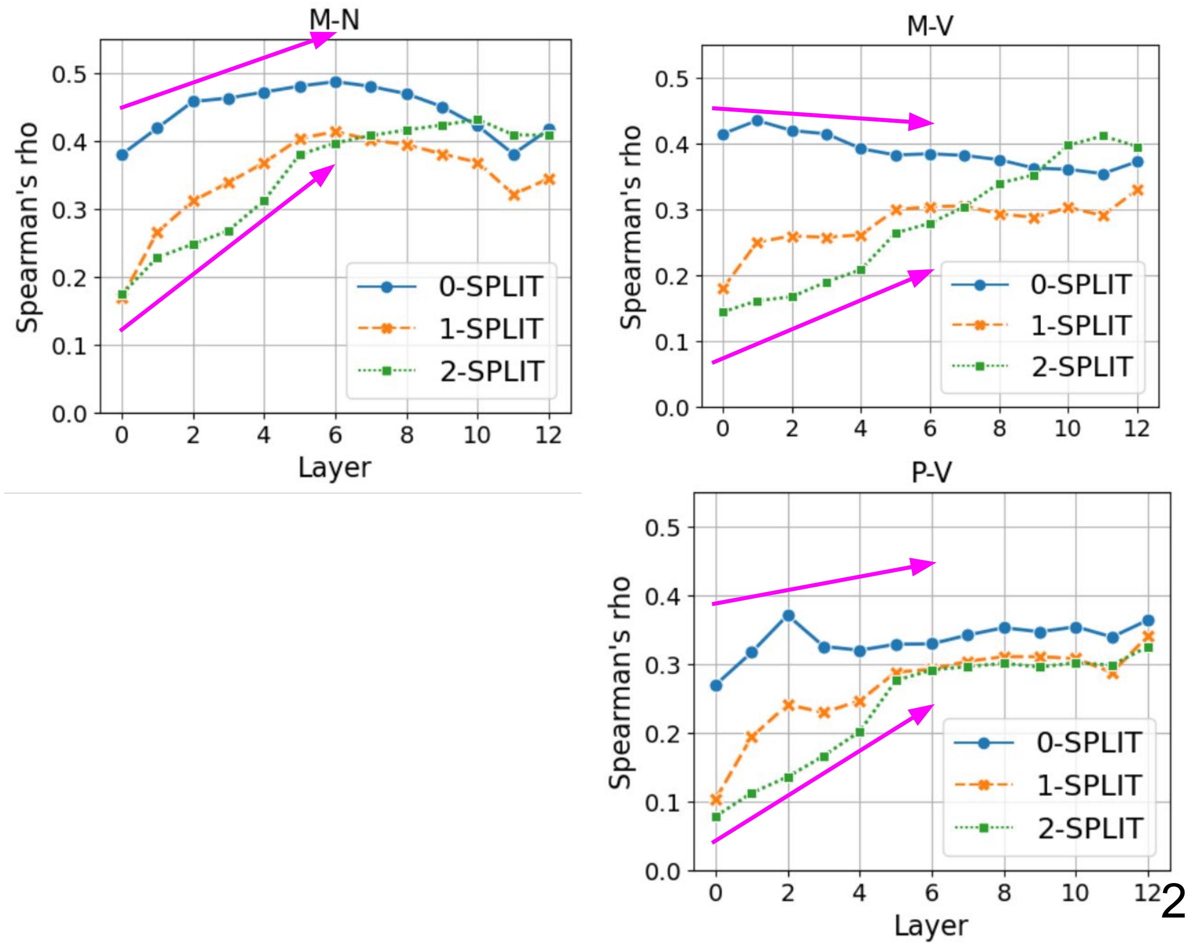






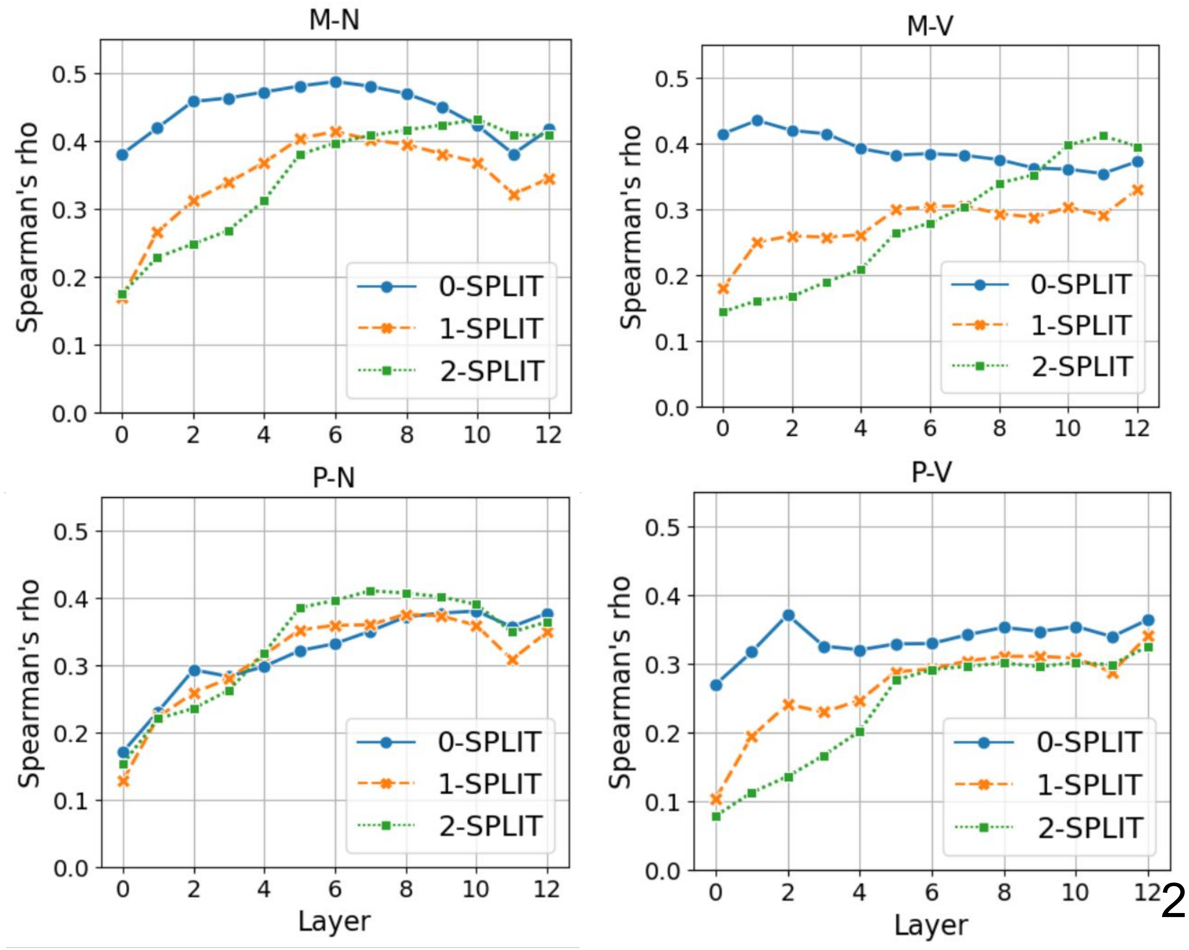
How do results change across layers for every split-type?

- At earlier layers the quality of 1- and 2-split similarity estimations is much lower than that of 0- SPLIT pairs.
- However, their quality improves at a higher rate than that of 0- SPLIT, which remains more stable.



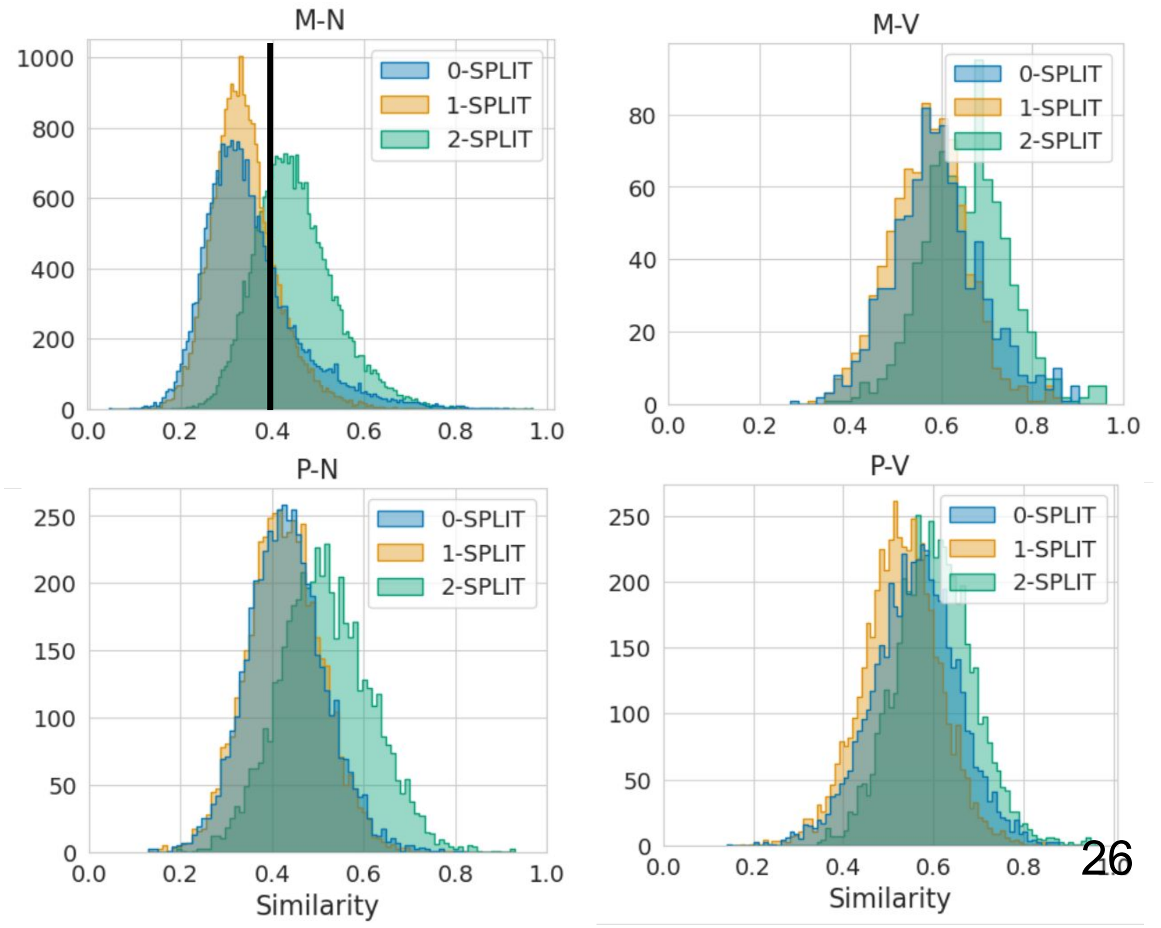
How do results change across layers for every split-type?

- For polysemous nouns, instead, 0- SPLIT pairs behave in a similar way as 1- and 2- SPLIT pairs from the very first layers.



Do similarity predictions vary across split-types?

- Similarities in 2-SPLIT pairs are in a different range: comparison of similarity values across split-types can be misleading



Does the number of subwords have an impact on the representations' semantic content?

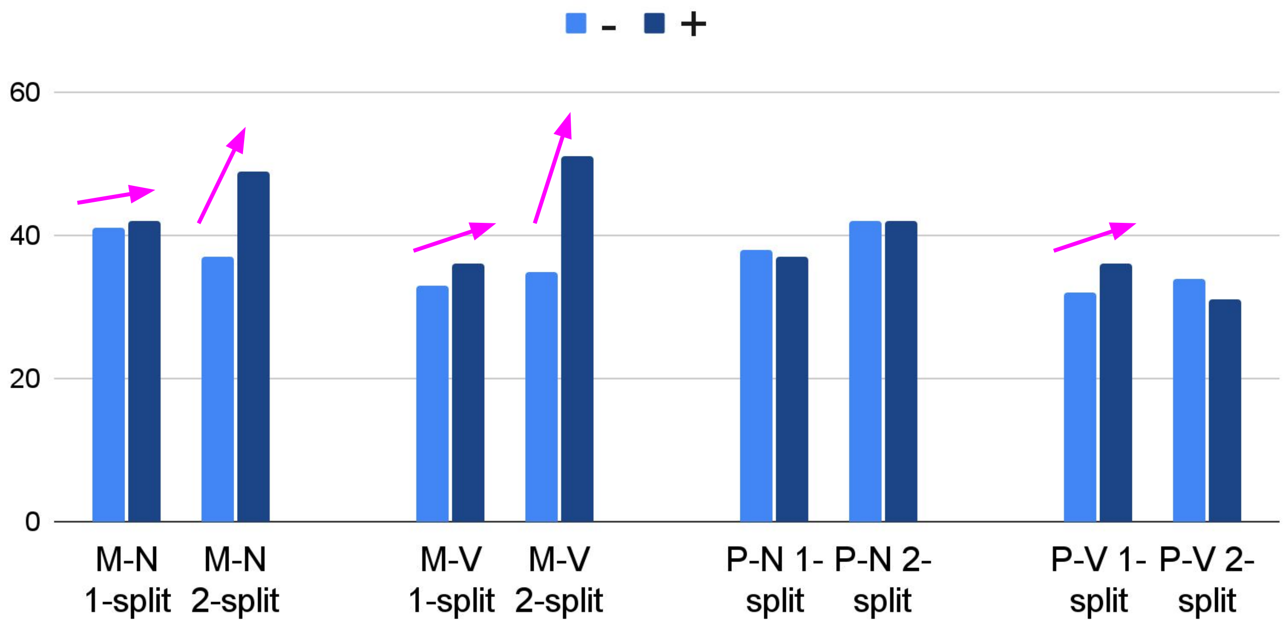
{ash, ##tray} {weather} ⇨ 3 (2+1)

{tom, ##fo, ##ole, ##ry} {loaf, ##ing} ⇨ 6 (4+2)

	-	+
1-split	3	>3
2-split	≥5	>5

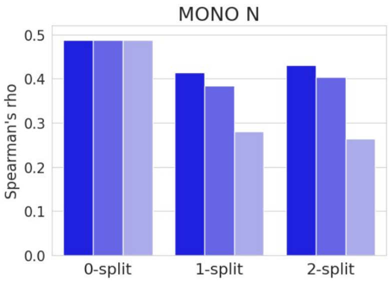
Expectation: more subwords ⇨ worse quality

Does the number of subwords have an impact on the representations' semantic content?

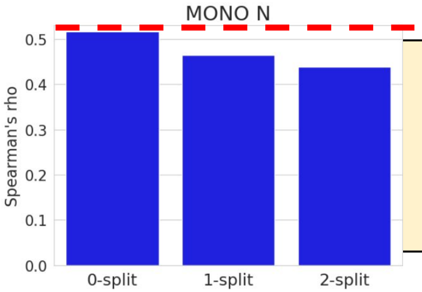


- Most of the time, more tokens was better!

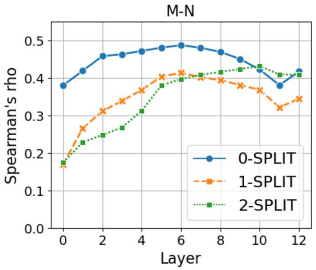
Conclusion



Averaging representations of all tokens is the best strategy to represent split-words

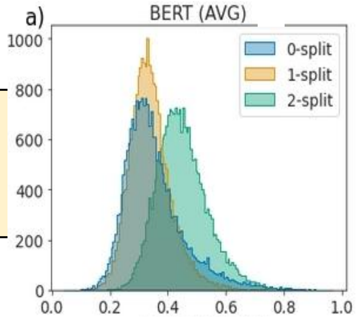


The **quality of split-word representations is worse than that of full-words**, but it depends on the kind of words considered

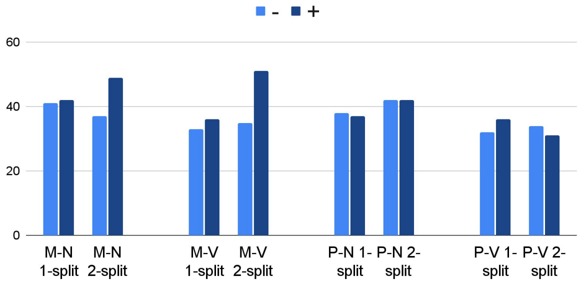


The best layers to use differ across split-types

Conclusion



Similarity values obtained **between two split-words are** generally **higher** than similarities involving full-words



A **higher number of tokens does not decrease representation quality**.

Word Meaning in Interaction

Conversational alignment

Alignment (or entrainment): phenomenon by which people mimic each other in conversations.

It can happen at different levels: lexical, syntactic, prosodic, postural…

Conceptual alignment

Conceptual alignment: The extent to which two dialog participants "mean the same things when using the same words" (Schober et al., 2005)

Knowing the meaning of a word does not guarantee conceptual alignment:

- Different mental representations of words (connotation, associations, detail)
- Ambiguity
- Novel usages

Do language models <u>understand</u>?

Is she <u>tall</u>?

Lexico-semantic alignment

We can't access people's mental representations of words…

We propose a more restricted notion of conceptual alignment:

Lexico-semantic alignment

"The convergence of word meaning inferrable from textual information alone"

Are dialogs more "polysemous" than monologs?

(Are words used in more different senses in dialogs than in monologs?)

- Inter-personal differences in dialogs (backgrounds, world knowledge, idiolect, language level, opinion...) can lead to misunderstandings and disagreements

- **One Sense per Discourse** hypothesis: only tested on monolog-like data

It's even less true of dialog

Garí Soler et al., (2022), LREC

Do contextualized word representations reflect stance?

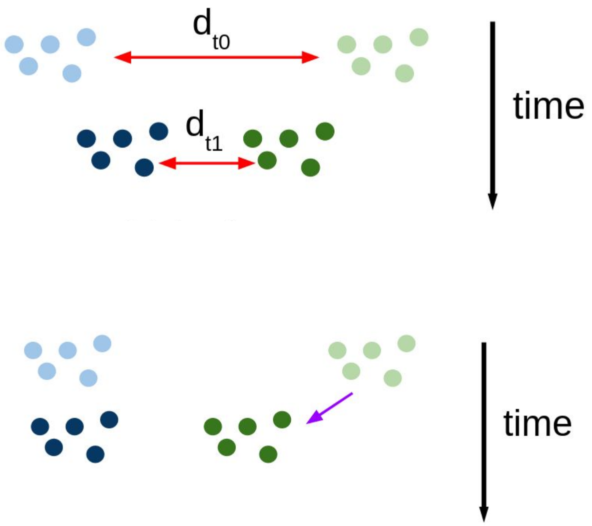
Differences in opinion are a likely source of misalignment

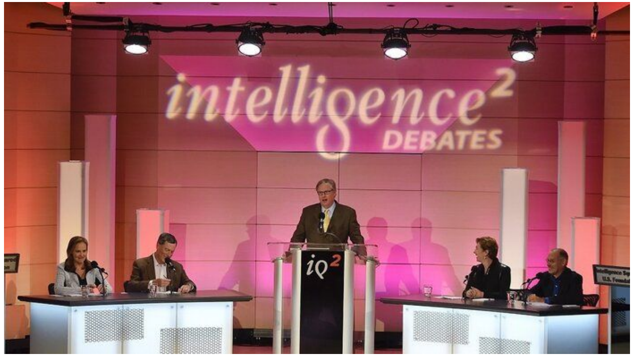
Yes.

Garí Soler et al., (2022), COLING

Can we measure lexico-semantic alignment?

We propose measures capturing different aspects of lexico-semantic alignment and relying on contextualized word representations





Garí Soler et al., (2023), SICon

Lexico-semantic alignment

Our measures reflected multiple semantic phenomena that characterize the way each side of a debate uses specific words…

…But we can’t evaluate them!

Let’s find examples of cases where speakers signal misalignment explicitly

Word Meaning Negotiation: The NeWMe Corpus

Garí Soler, Myrendal, Clavel and Larsson (under review at LRE)

Word Meaning Negotiation (WMN)

3 components:

J3VPSUNK
Okay good . Erm a few hard facts just to confirm first of all

Trigger

PS3PB
Hard , what 's a hard fact ?

Indicator

J3VPSUNK
Your name and your address .

Negotiation

PS3PB
Oh right . So why are they called hard facts ?

Word Meaning Negotiation (WMN)

2 main WMN types:

- NONs (originating from non-understanding)
- **DIN**s (originating from disagreement)

Mavericgamer_##_t1_cnfd02s_##_rt-t1_cnfcl4z -

In reply to the edit: Yes, I could do the same. I could also do a "football/soccer flop" and fake pain when I'm not in it. I could even make the robot do that for important things like when it was detecting structural damage, because if you can invoke empathy, it is an important survival trait. But at that point, philosophically, isn't it just that we have created a being that can feel pain?

FarkCookies_##_t1_cnfd9en_##_rt-t1_cnfd02s -

Not by any meaningful definition of pain. You equate pain and response to negative stimulus.

Mavericgamer_##_t1_cnfdeva_##_rt-t1_cnfd9en -

What else is pain than a response to a negative stimulus? I believe that is the *only* meaningful definition of pain.

FarkCookies_##_t1_cnfdkqj_##_rt-t1_cnfdeva -

Ah it is cool then. Your rules, your game.

The NeWMe (**Ne**gotiating **W**ord **Me**aning) corpus

Annotation of WMN in existing conversational corpora:

- **Switchboard Dialog Act Corpus** (Stolcke et al., 2000)

  Oral - dyadic phone conversations

- **British National Corpus** (BNC Consortium, 2007)

  Oral - lectures, meetings, interviews...

- **Winning Arguments** (ChangeMyView) **Corpus** (Tan et al., 2016)

  Written (Reddit) - debate-like

Data collection

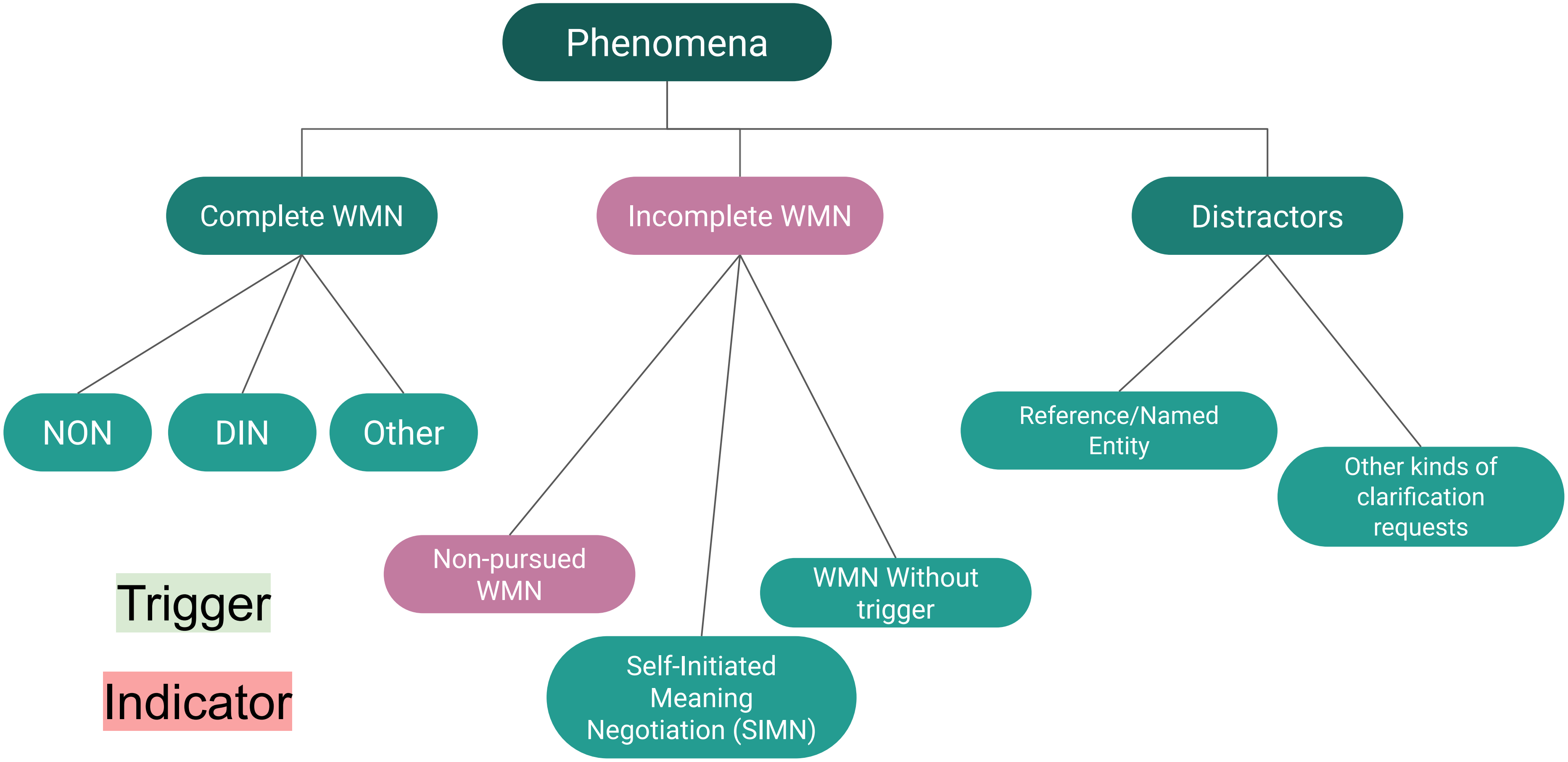
- Focus on indicators: the part of a WMN with least variability

- Regular expression matching

  - ❑ (what|wtf) do you (actually) mean by
  - ❑ this is not X
  - ❑ meaning of
  - ❑ definition of
  - ❑ what is the difference between
  - ❑ S1: …hard facts… S2: hard facts?
  - ❑ ...

Result: 8313 potential indicators

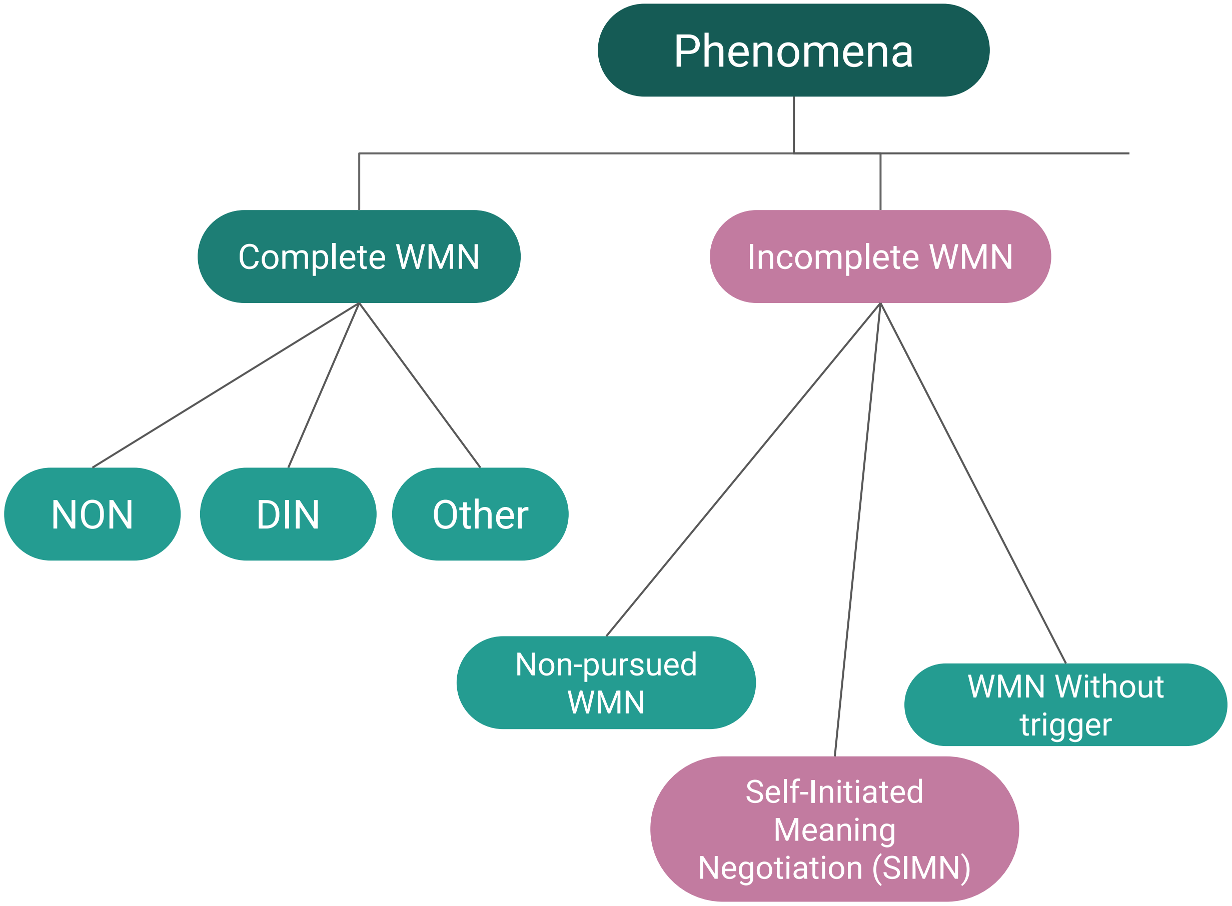
Annotation schema



Trigger

Indicator

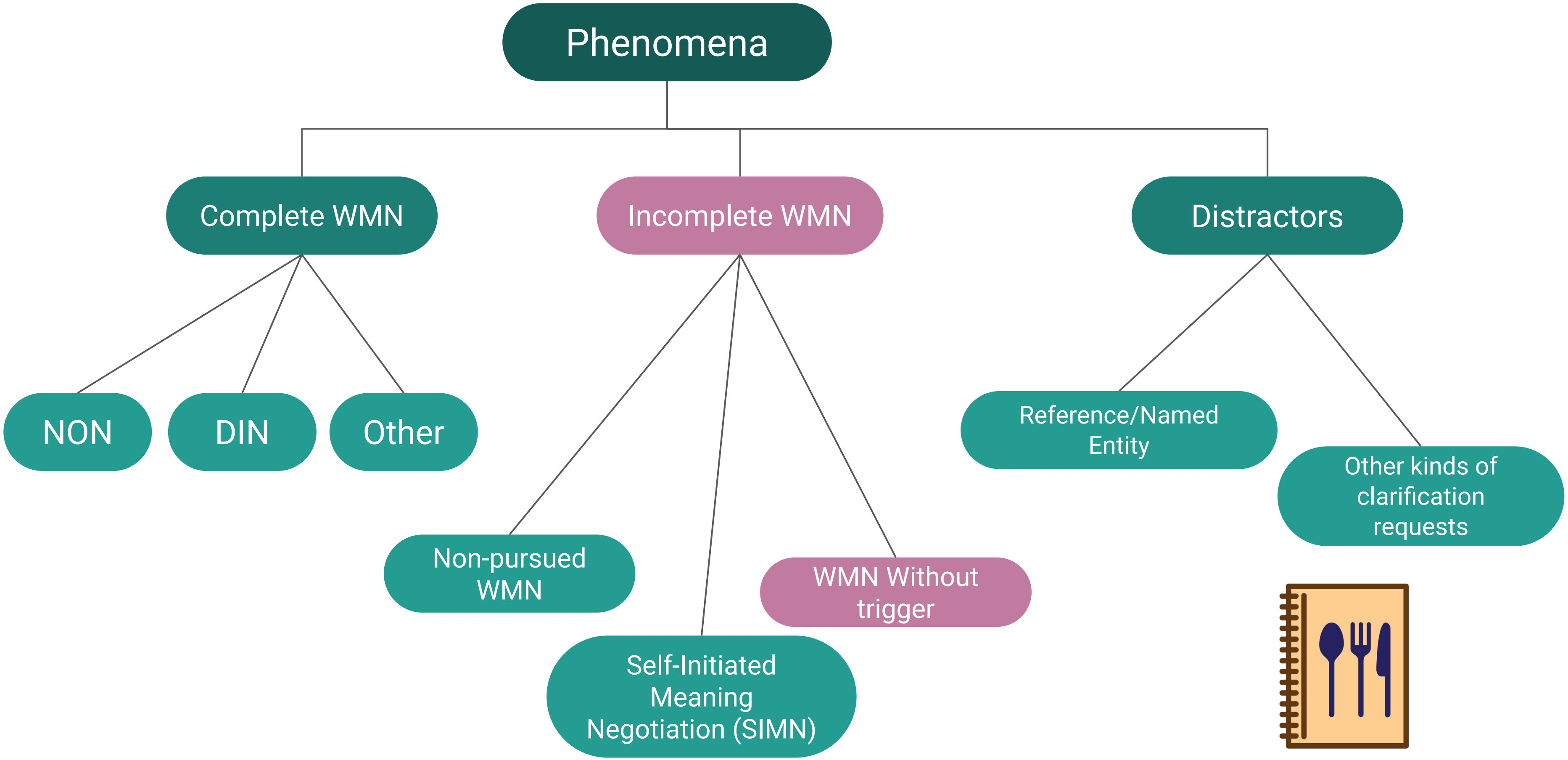
...



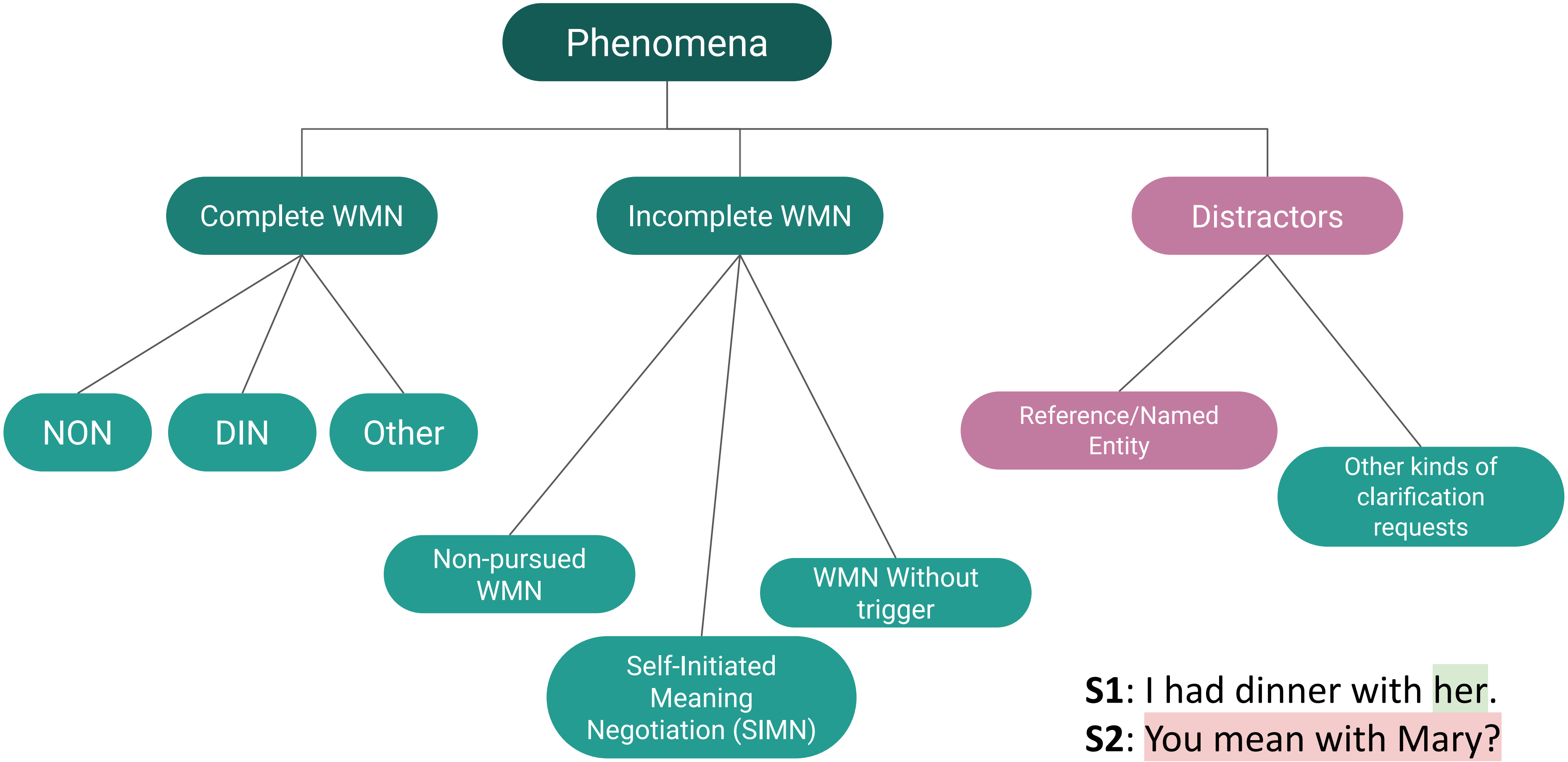
S1: We need to calibrate the machine before we start using it. Do you understand what calibrate means?

S2: It means to adjust it properly.

...



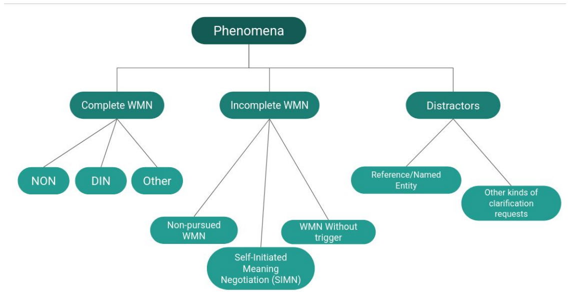
What’s tortellini?



S1: I had dinner with her.
S2: You mean with Mary?

Annotation schema

- Phenomenon label



- Spans

trigger, indicator, negotiation

Annotation procedure

2 expert annotators

1st round

Regular meetings to discuss difficult cases and refine the annotation schema

Annotation guidelines write-up

2nd round

Double-checking all phenomena for consistency

3rd round

Inter-annotator agreement

Statistics

	BNC	Reddit	Switchboard	Total
NONs	116	66	33	215
DINs	11	158	0	169
WMN: Other	14	3	3	20
Non-pursued	4	197	2	203
SIMN	37	2	3	42
Without trigger	10	0	2	12
Reference/NE	7	3	18	28
Other kinds of clarification requests	15	109	49	173
Nothing	3746	2353	1188	7287
Total	3984	2892	1298	8174
Total phenomena	214	538	110	721

Statistics

	BNC	Reddit	Switchboard	Total
NONs	116	66	33	215
DINs	11	158	0	169
WMN: Other	14	3	3	20
Non-pursued	4	197	2	203
SIMN	37	2	3	42
Without trigger	10	0	2	12
Reference/NE	7	3	18	28
Other kinds of clarification requests	15	109	49	173
Nothing	3746	2353	1188	7287
Total	3984	2892	1298	8174
Total phenomena	214	538	110	721

404
WMNs

Statistics

	BNC	Reddit	Switchboard	Total
NONs	116	66	33	215
DINs	11	158	0	169
WMN: Other	14	3	3	20
Non-pursued	4	197	2	203
SIMN	37	2	3	42
Without trigger	10	0	2	12
Reference/NE	7	3	18	28
Other kinds of clarification requests	15	109	49	173
Nothing	3746	2353	1188	7287
Total	3984	2892	1298	8174
Total phenomena	214	538	110	721

Statistics

	BNC	Reddit	Switchboard	Total
NONs	116	66	33	215
DINs	11	158	0	169
WMN: Other	14	3	3	20
Non-pursued	4	197	2	203
SIMN	37	2	3	42
Without trigger	10	0	2	12
Reference/NE	7	3	18	28
Other kinds of clarification requests	15	109	49	173
Nothing	3746	2353	1188	7287
Total	3984	2892	1298	8174
Total phenomena	214	538	110	721

Inter-annotator agreement

Expert annotation

256 instances

86-90% total agreement ➡ 94-96% after discussion

Can we obtain reliable results by training annotators with our annotation guidelines?

Inter-annotator agreement

3 Master’s students in Computational Linguistics with an advanced level of English

LEARNING

- Annotation guidelines
- 2 training videos
- Common meeting for Q&A

TRAINING

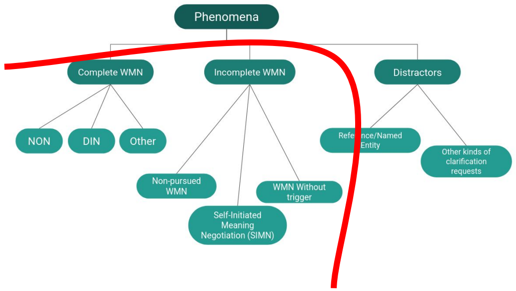
- Two 15-instance annotation samples
- Individual meetings with feedback

Guidelines revision

ANNOTATION

Annotation of the same sample without feedback (704 instances)

Inter-annotator agreement



Lessons learned:

- Moderate* agreement on a higher-level distinction is reachable (by some annotators, on some corpora)
    - Reddit data was harder to annotate
    - Subjectivity, recurrent mistakes…

- We need more training, examples and feedback, with an emphasis on Reddit

*Krippendorff's alpha ≥ 0.67

NeWMe: Next steps

- First corpus of its kind
- But it could be bigger: working on its semi-automatic extension

It will enable

- Characterizing and detecting problematic word usages
- Studying signaling behavior and negotiation strategies
- Determining the success of a negotiation

Later...

- Writing assistants
- Human-machine interaction

Concluding thoughts

Word Meaning Representation

- Complex

- (Still) relevant
  - NLP: less mainstream tasks, domains and languages
  - (Computational) Linguistics, Lexicography
  - Social sciences
  - Often more light-weight and computationally cheaper

Concluding thoughts

Word Meaning Negotiation

- Every speaker has their own “semantic network” - word-related misunderstandings are a window into inter-personal differences and language variation
- We only collected cases of **detected and signaled** conceptual misalignment
- To learn about how communication works, we need to study how and why it fails
- A model that succeeds at communicating needs to be able to avoid, detect and/or navigate word-related misunderstandings and disagreements

Thank you!