Lexical polysemy and intensity in contextualised representations

Marianna Apidianaki

University of Pennsylvania

NYU, NLP and Text as Data speaker series

Nov, 4

cnrs

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

Work done in collaboration
with Aina Garí Soler
who did her PhD in the
MULTISEM ANR project
(CNRS, University Paris-Saclay)

UNIVERSITY OF HELSINKI
FACULTY OF ARTS

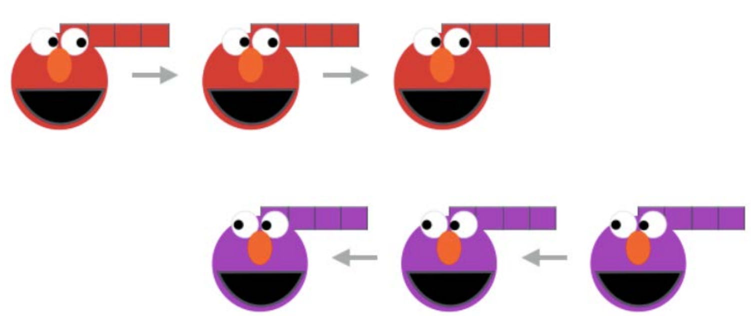
erc European Research Council

FOTRAN
Found in Translation

and while working on
the ERC project FoTraN
at the University of Helsinki

Pre-trained language models

- trained on massive amounts of unannotated data
- available in many languages
- deliver impressive performance in NLP and NLU tasks



ELMo (Peters et al., 2018)

BERT (Devlin et al., 2018)

ROBERTa (Liu et al., 2019)

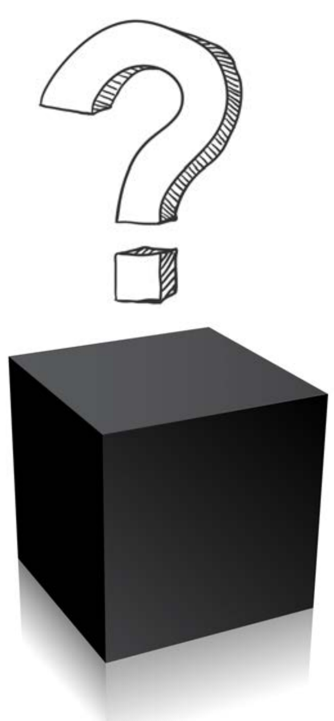
DistilBERT (Sanh et al., 2019)

ALBERT (Lan et al., 2020)

SpanBERT (Joshi et al., 2020)

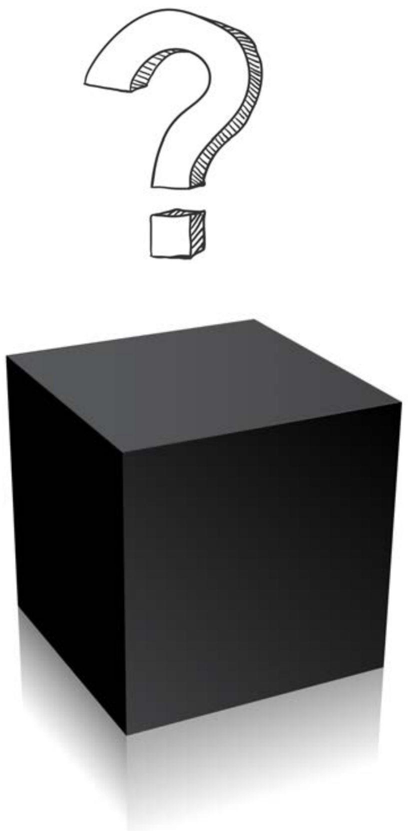
But what do these models really know about language?

- Does high performance reflect good knowledge of language and the world?
- Is this information encoded in the representations?



Bertology/interpretation studies are trying to answer this question

Looking inside the black box



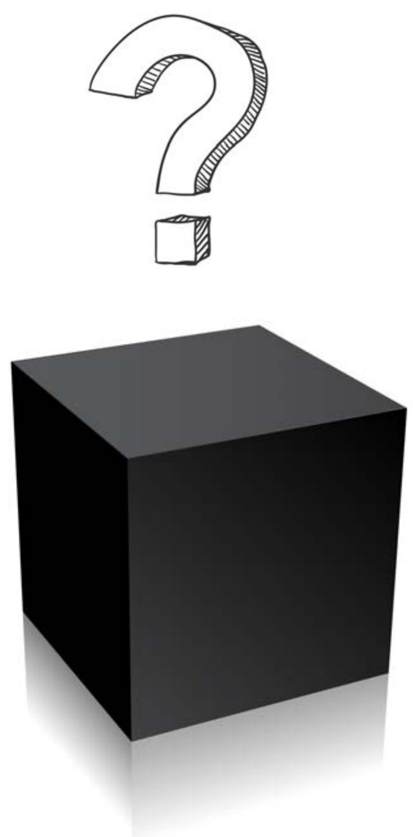
word order
number agreement

(Linzen, 2018; Goldberg 2019)

syntactic dependencies

(Shi et al., 2016; Linzen et al., 2016; Gulordava et al., 2018; Raganato and Tiedemann, 2018; Hewitt and Manning, 2019; Lakretz et al. 2019)

Looking inside the black box



word order
number agreement

(Linzen, 2018; Goldberg 2019)

syntactic dependencies

(Shi et al., 2016; Linzen et al., 2016; Gulordava et al., 2018; Raganato and Tiedemann, 2018; Hewitt and Manning, 2019; Lakretz et al. 2019)

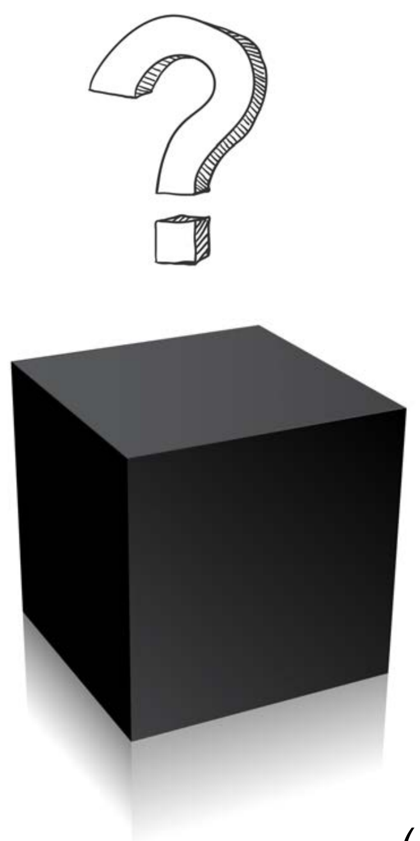
SRL and coreference

(Tenney et al., 2019; Kovaleva et al., 2019; Ettinger 2020)

negation

(Ettinger, 2020)

Looking inside the black box



word order
number agreement

(Linzen, 2018; Goldberg 2019)

syntactic dependencies

(Shi et al., 2016; Linzen et al., 2016; Gulordava et al., 2018; Raganato and Tiedemann, 2018; Hewitt and Manning, 2019; Lakretz et al. 2019)

SRL and coreference

(Tenney et al., 2019; Kovaleva et al., 2019; Ettinger 2020)

negation

(Ettinger, 2020)

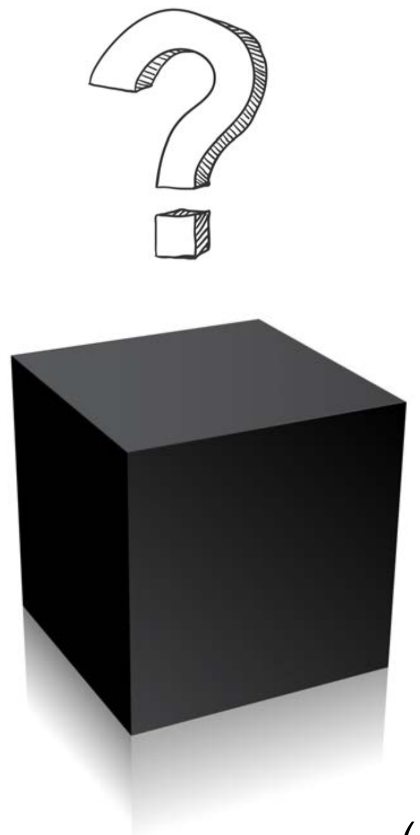
Hypernymy detection

(Ettinger, 2020; Ravichander et al., 2020)

factual and common-sense knowledge

(Petroni et al., 2019; Bouraoui et al., 2020; Ettinger, 2020)

Looking inside the black box



word order
number agreement

(Linzen, 2018; Goldberg 2019)

syntactic dependencies

(Shi et al., 2016; Linzen et al., 2016; Gulordava et al., 2018; Raganato and Tiedemann, 2018; Hewitt and Manning, 2019; Lakretz et al. 2019)

SRL and coreference

(Tenney et al., 2019; Kovaleva et al., 2019; Ettinger 2020)

negation

(Ettinger, 2020)

Hypernymy detection

(Ettinger, 2020; Ravichander et al., 2020)

factual and common-sense knowledge

(Petroni et al., 2019; Bouraoui et al., 2020; Ettinger, 2020)

WSD using sense annotations

(Reif et al., 2019; Wiedemann et al., 2019)

Contextual informativeness vs. ambiguity

(Pimentel et al., 2020)

In-context instance similarity

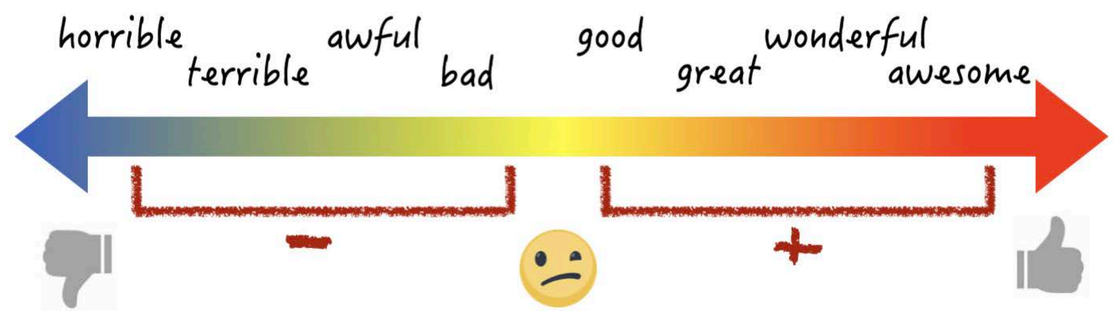
(Ethayarajh, 2019)

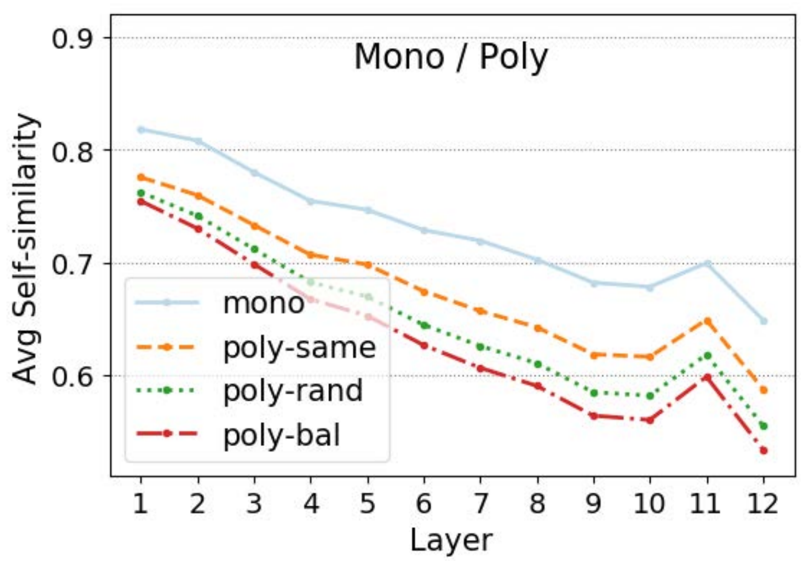
Out-of-context word similarity

(Vulić et al., 2020)

What BERT knows about..

Semantic relationships and intensity in particular?





Lexical polysemy and sense partitionability?

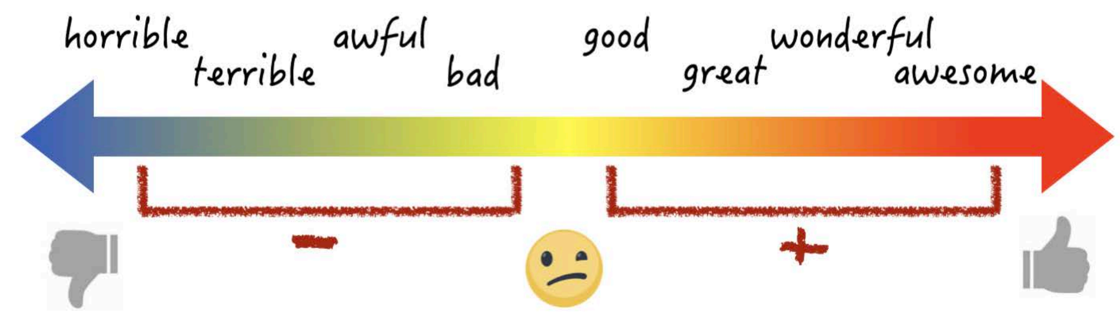
Noun properties and their prototypicality?

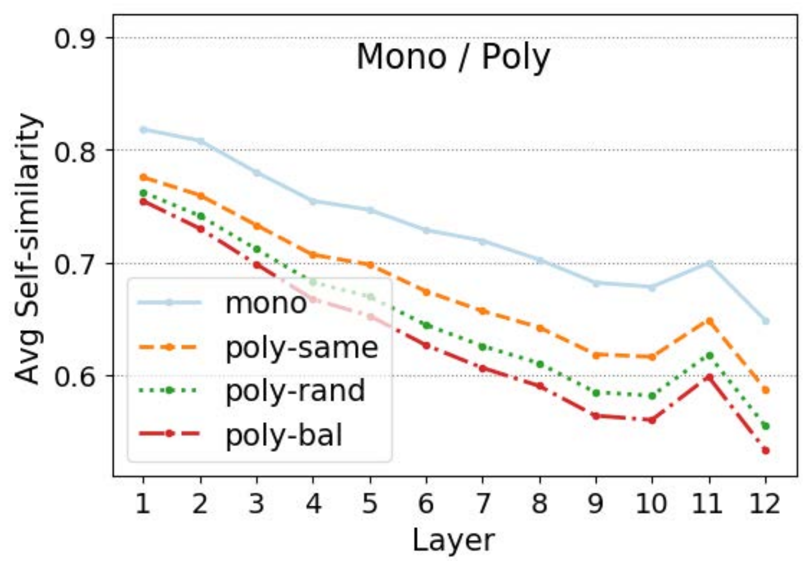
all strawberries are [MASK]

[MASK] balloons are colourful.

What BERT knows about..

Semantic relationships and intensity in particular?





Lexical polysemy and sense partitionability?

Noun properties and their prototypicality?

all strawberries are [MASK]

[MASK] balloons are colourful.

If you are interested in

noun properties and prototypicality

all strawberries are [MASK]
[MASK] peacocks are colourful.
[MASK] mittens are knitted
blueberries are [MASK]

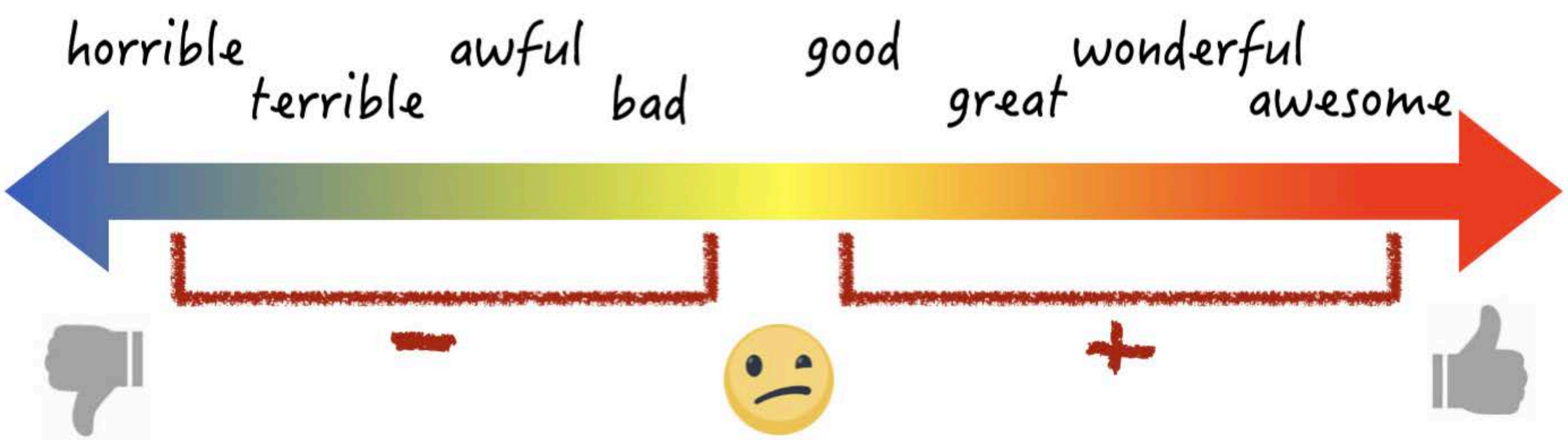
Check out our BlackBoxNLP paper *

ALL Dolphins Are Intelligent and SOME Are Friendly:
Probing BERT for Nouns' Semantic Properties and their Prototypicality

* (for the moment on arXiV, soon on ACL anthology)

What BERT knows about Semantic Relationships?

Scalar adjective ranking



BERT Knows Punta Cana is not just beautiful, it's gorgeous:
Ranking Scalar Adjectives with Contextualised Representations

EMNLP 2020

Previous work

Scalar Adjective Ranking

Pattern-based

(Sheinman and Tokunaga, '09; DeMelo and Bansal, '13)

"The show was **funny**, but not **hilarious**."

→ funny < hilarious

"It's not **freezing**, but still **cold**."

→ cold < freezing

Previous work

Scalar Adjective Ranking

Pattern-based

(Sheinman and Tokunaga, ’09; DeMelo and Bansal, ’13)

“The show was **funny**, but not **hilarious**.”

→ funny < hilarious

“It’s not **freezing**, but still **cold**.”

→ cold < freezing

Lexicon-based

Semantic Orientation CALculator (SOCAL)
Taboada et al. (2011)

Adjective	Score
exquisite	5
beautiful	4
appealing	3
above-average	2
okay	1
ho-hum	-1
pedestrian	-2
gross	-3
grisly	-4
abhorrent	-5

+

–

most intense

least intense

most intense

Previous work

Scalar Adjective Ranking

Pattern-based

(Sheinman and Tokunaga, '09; DeMelo and Bansal, '13)

"The show was **funny**, but not **hilarious**."

→ funny < hilarious

"It's not **freezing**, but still **cold**."

→ cold < freezing

Lexicon-based

Semantic Orientation CALculator (SOCAL)
Taboada et al. (2011)

Adjective	Score
exquisite	5
beautiful	4
appealing	3
above-average	2
okay	1
ho-hum	-1
pedestrian	-2
gross	-3
grisly	-4
abhorrent	-5

+

–

most intense

least intense

most intense

Paraphrase pair…			…is evidence that
particularly pleased	↔	*ecstatic*	*pleased < ecstatic*
quite limited	↔	*restricted*	*limited < restricted*
rather odd	↔	*crazy*	*odd < crazy*
so silly	↔	*dumb*	*silly < dumb*
completely mad	↔	*crazy*	*mad < crazy*
RB JJ_1	↔	*JJ_2*	*$JJ_1 < JJ_2$*

intensifying adverb

Paraphrase-based

(Cocos et al., 2018)

What BERT can do on this task?

Datasets

- DeMelo (87 half-scales)
  (de Melo and Bansal, 2013)

  [*soft* → *quiet* → *inaudible* → *silent*]
  [*thick* → *dense* → *impenetrable*]

- Crowd (79 half-scales)
  (Cocos et al., 2018)

  [*fine* → *remarkable* → *spectacular*]
  [*scary* || *frightening* → *terrifying*]

- Wilkinson (21 half-scales)
  (Wilkinson and Oates, 2016)

  [*damp* → *moist* → *wet*]
  [*dumb* → *stupid* → *idiotic*]

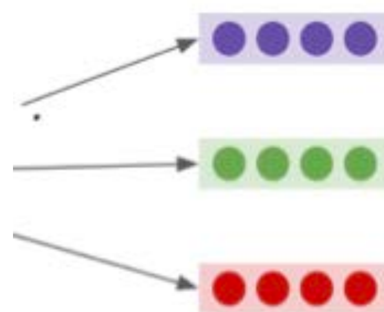
- Is intensity information encoded in BERT representations?
- Can we reproduce the ranking found in external resources using this information?

BERT representations

scale: [pretty => beautiful => gorgeous]

Punta Cana is **gorgeous** .

{ pretty
beautiful }

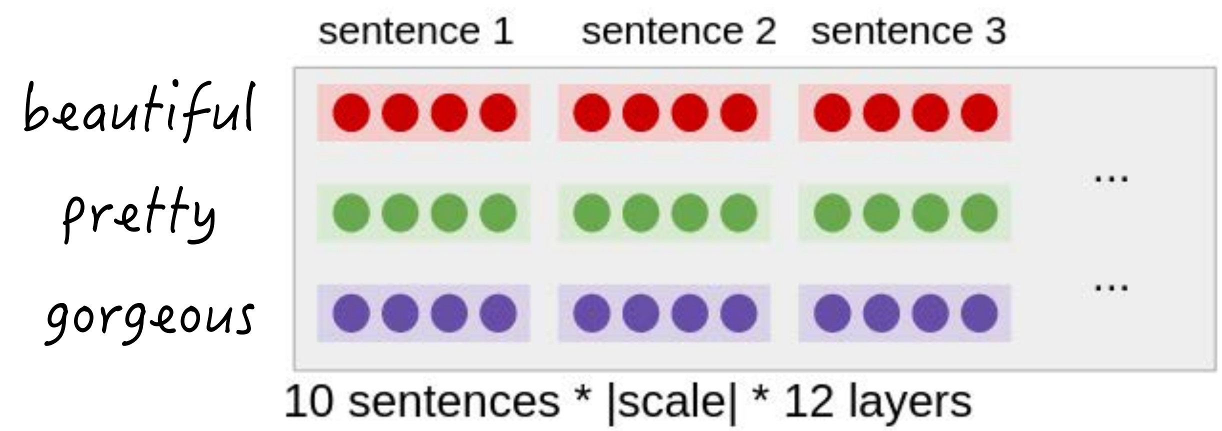


What a **beautiful** sunset!

{ pretty
gorgeous }

You look **pretty** today.

{ beautiful
gorgeous }



Similarity to the extreme adjective

$\{$ *big, large, huge, enormous,* ***gigantic*** $\}$

cos

$\{$ *good, great, wonderful,* ***awesome*** $\}$

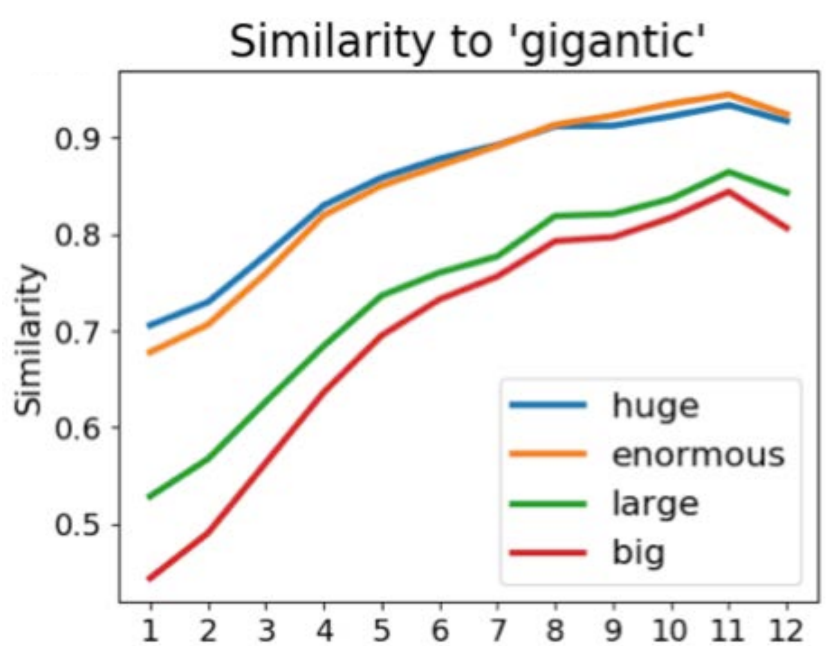
cos

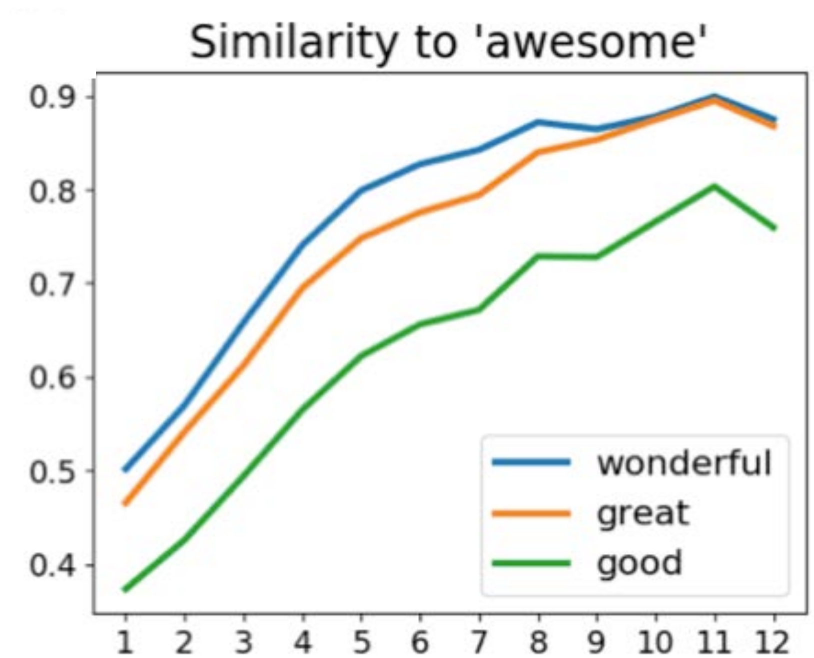
avg

$$cos(\overrightarrow{big_1}, \overrightarrow{gigantic_1})$$
$$cos(\overrightarrow{big_2}, \overrightarrow{gigantic_2})$$
$$\ldots$$
$$cos(\overrightarrow{big_{10}}, \overrightarrow{gigantic_{10}})$$

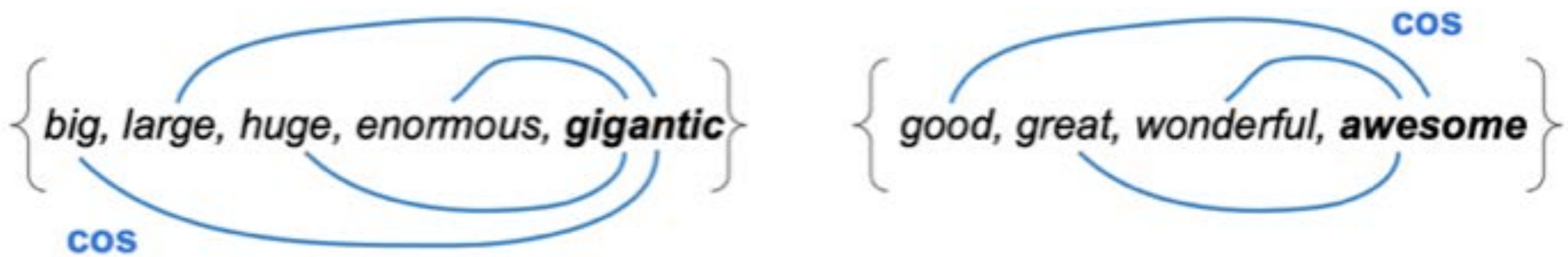
avg

$$cos(\overrightarrow{good_1}, \overrightarrow{awesome_1})$$
$$cos(\overrightarrow{good_2}, \overrightarrow{awesome_2})$$
$$\ldots$$
$$cos(\overrightarrow{good_{10}}, \overrightarrow{awesome_{10}})$$





Similarity to the extreme adjective



=> similarity to the "extreme" adjective seems to be a good feature

=> BUT we don't usually know which most intense word is

Dvec: a vector that represents intensity

Inspired by gender bias work (Bolukbasi et al., 2016)

$\overrightarrow{she} - \overrightarrow{he}$

$\overrightarrow{her} - \overrightarrow{his}$

$\overrightarrow{woman} - \overrightarrow{man}$

$\overrightarrow{Mary} - \overrightarrow{John}$

$\overrightarrow{herself} - \overrightarrow{himself}$

$\overrightarrow{daughter} - \overrightarrow{son}$

$\overrightarrow{mother} - \overrightarrow{father}$

$\overrightarrow{gal} - \overrightarrow{guy}$

$\overrightarrow{girl} - \overrightarrow{boy}$

$\overrightarrow{female} - \overrightarrow{male}$

PCA

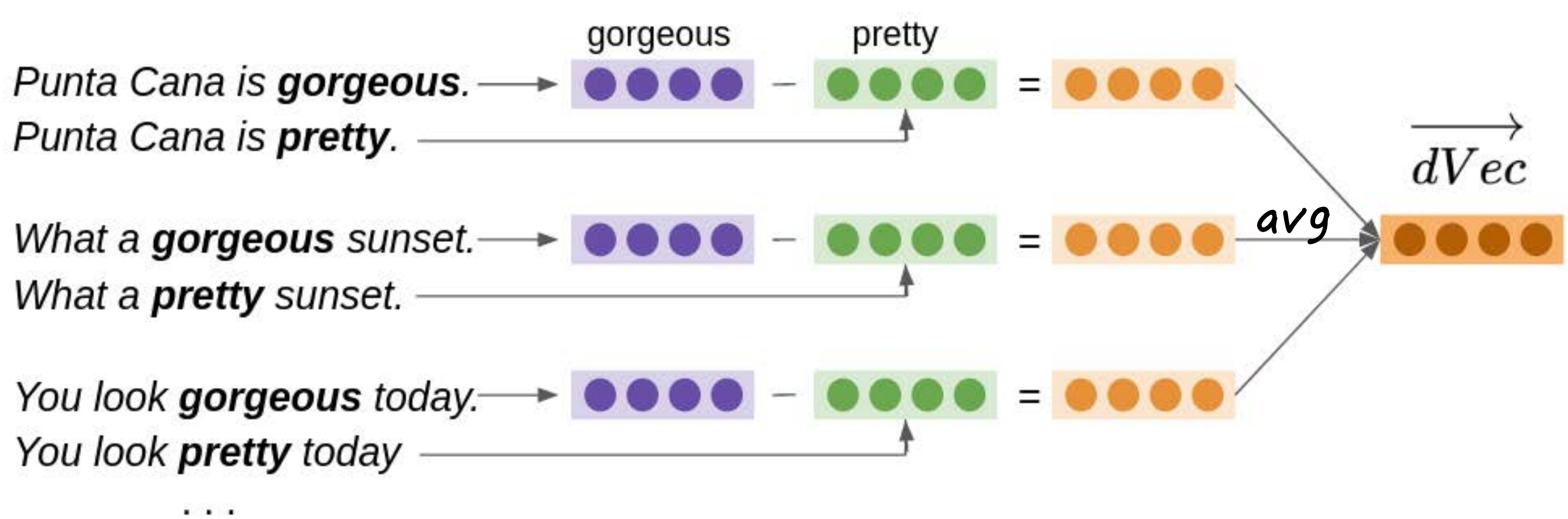
there is a single direction that explains the majority of variance in these vectors

$$\boxed{\overrightarrow{adj_{extreme}} - \overrightarrow{adj_{mild}}} \Rightarrow \overrightarrow{dVec}$$

representation of intensity

Dvec: the intensity vector

$\overrightarrow{dVec}$ for an adjective pair:

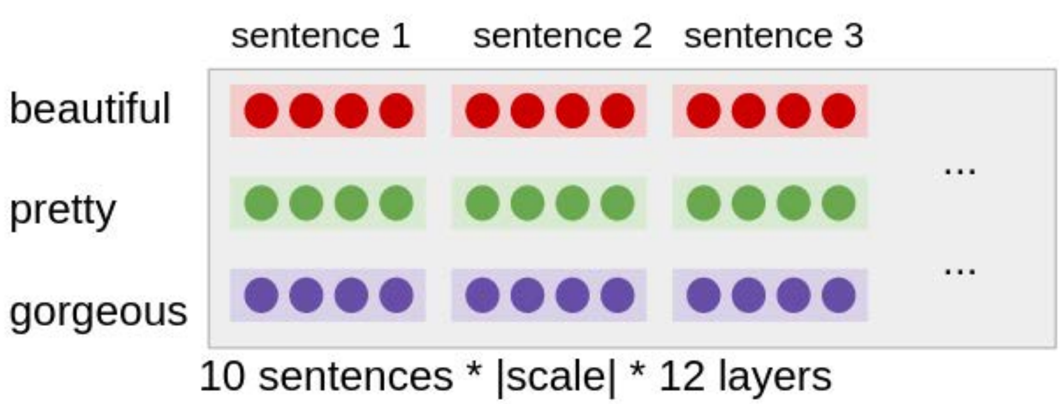


$\overrightarrow{dVec}$ for a dataset:

$$\text{avg} \begin{array}{l} \overrightarrow{dvec}(horrible - bad) \\ \overrightarrow{dvec}(awesome - good) \\ \overrightarrow{dvec}(gorgeous - pretty) \end{array} \Rightarrow \overrightarrow{dVec}$$

Adjective ranking using dvec

Average the representations obtained for an adjective.



$$avg(\overrightarrow{beautiful_1}, \overrightarrow{beautiful_2}, \ldots, \overrightarrow{beautiful_{10}}) \rightarrow \boxed{\overrightarrow{beautiful}}$$
$$avg(\overrightarrow{pretty_1}, \overrightarrow{pretty_2}, \ldots, \overrightarrow{pretty_{10}}) \rightarrow \boxed{\overrightarrow{pretty}}$$
$$avg(\overrightarrow{gorgeous_1}, \overrightarrow{gorgeous_2}, \ldots, \overrightarrow{gorgeous_{10}}) \rightarrow \boxed{\overrightarrow{gorgeous}}$$

Rank the adjectives in a scale using their cosine similarity score with dVec.

$$\boxed{\begin{array}{l} cos(\overrightarrow{gorgeous}, \overrightarrow{dVec}) \\ cos(\overrightarrow{beautiful}, \overrightarrow{dVec}) \\ cos(\overrightarrow{pretty}, \overrightarrow{dVec}) \end{array}}$$

the closer an ADJ is to dVec, the more intense it is!

Baselines

FREQ: frequency from Google Ngrams

- mild ADJs more frequent than extreme ADJs
- extreme ADJs denote more exceptional properties of nouns and restrict their denotation to a smaller class of referents (e.g., *a good view* vs. *a fantastic view*)

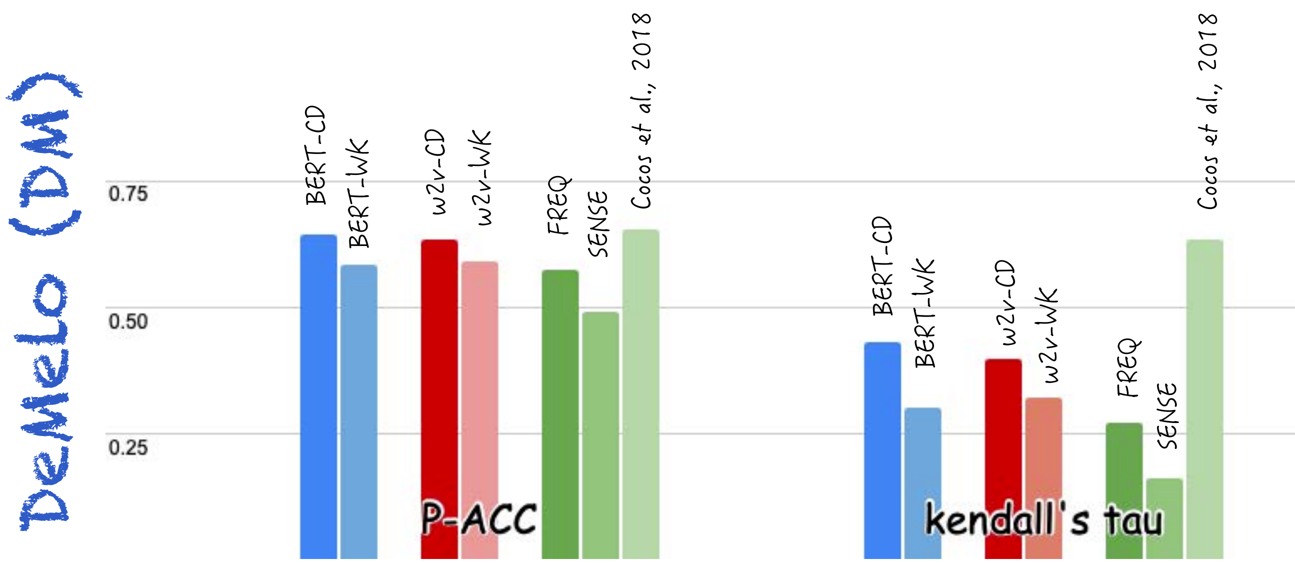
SENSE: # of senses from WordNet

- higher frequency -> higher number of senses (Zipf, 1945)

$\overrightarrow{dVec}$ from static embeddings

- difference between the word2vec embeddings of adj_{mild} and $adj_{extreme}$

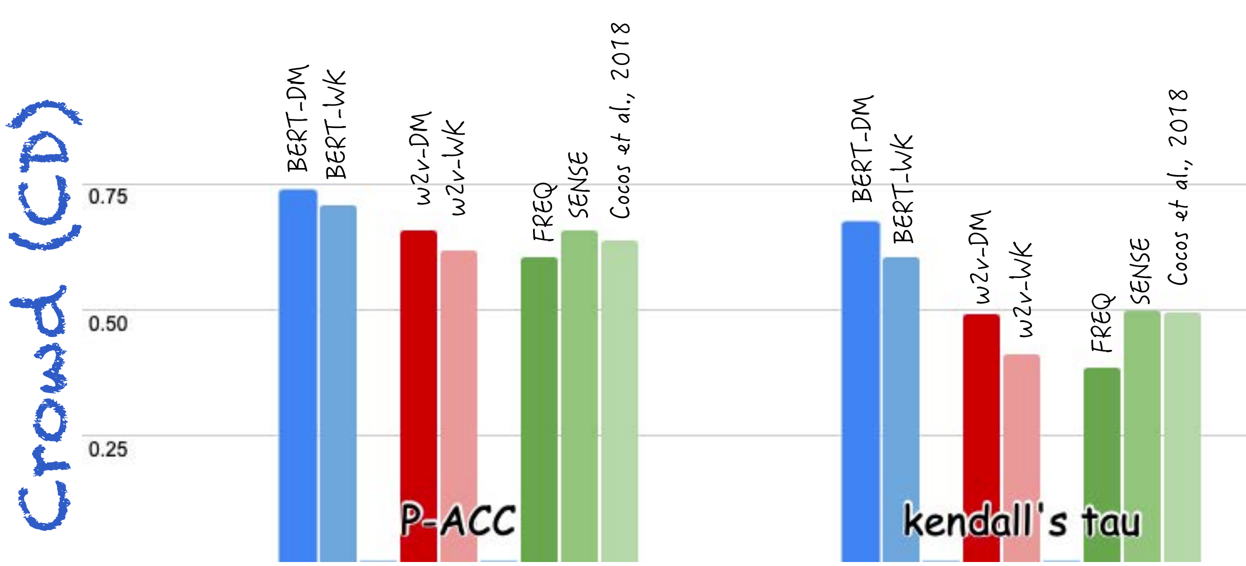
Ranking results

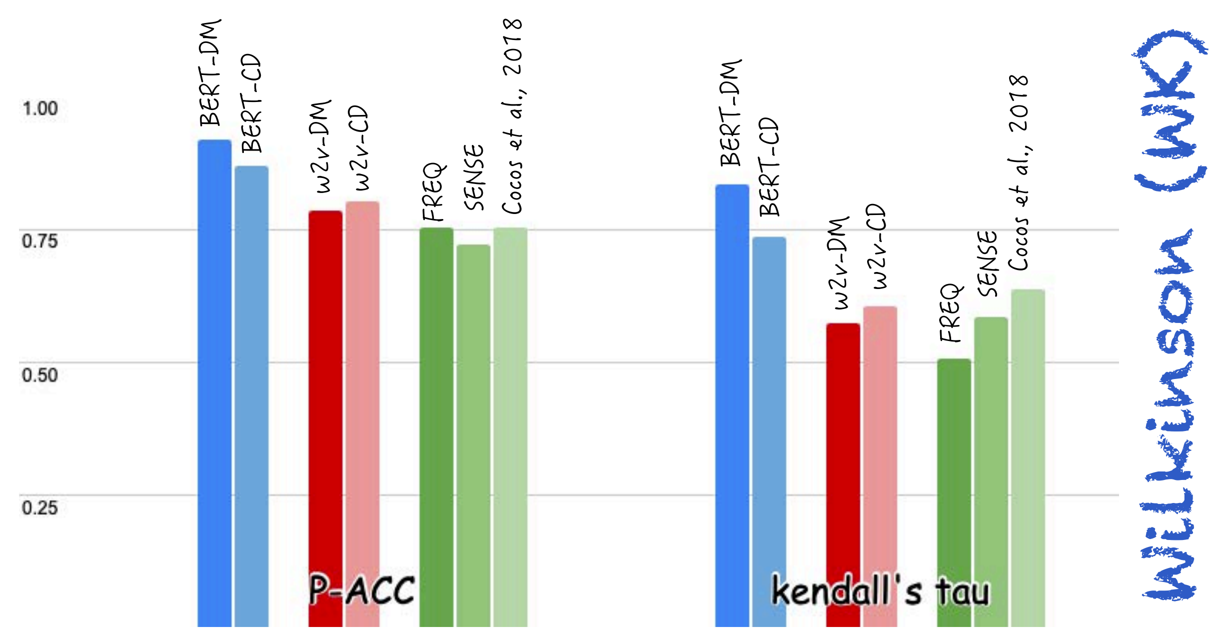


BERT
word2vec
baselines

Pair-wise accuracy: whether the relative intensity for each adjective pair was correctly predicted

Kendall's τ correlation of the produced ranking with the gold standard ranking for a scale





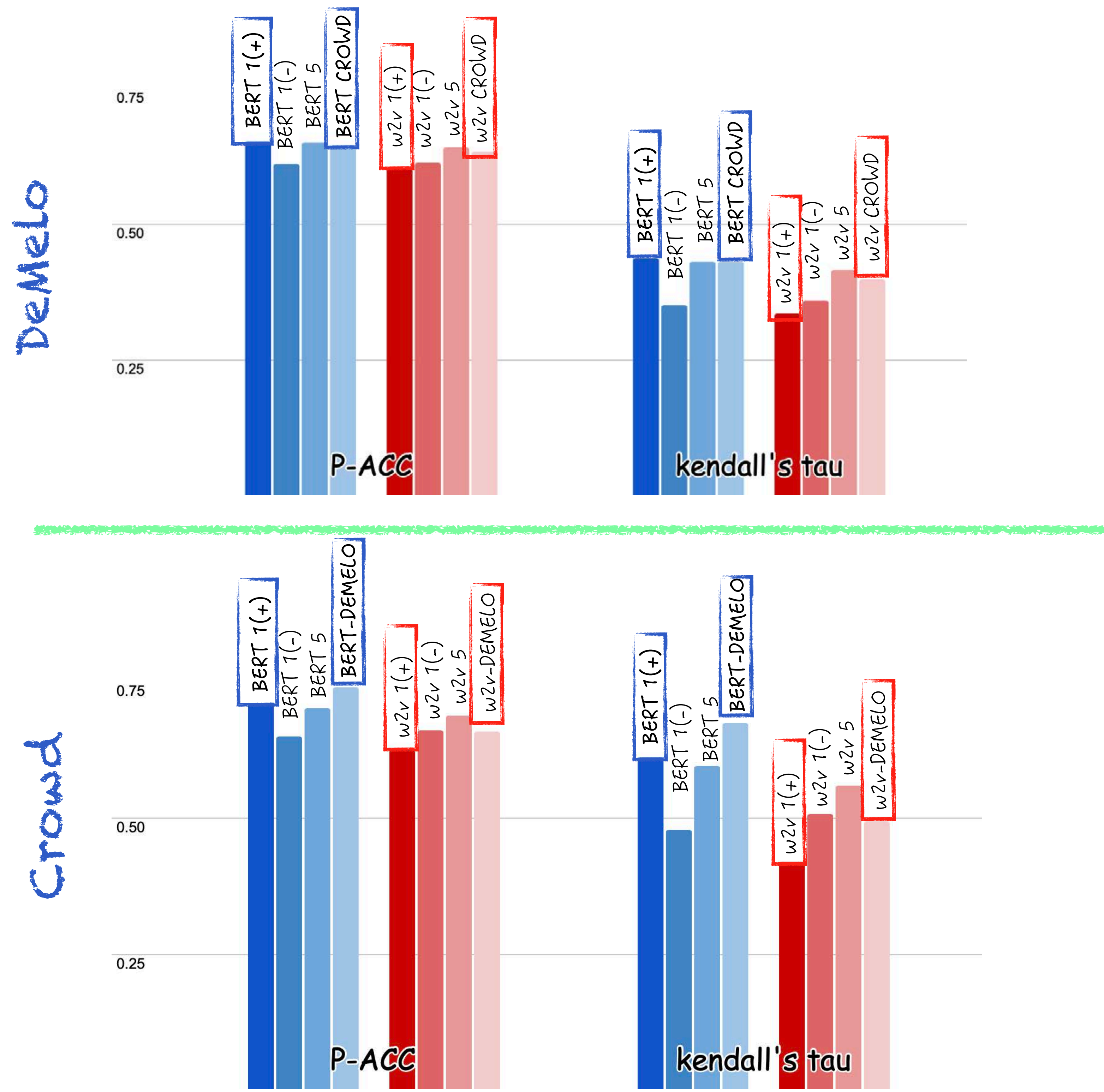
Ranking results

How many pairs to use?

1(+) : awesome - good

1(-) : horrible - bad

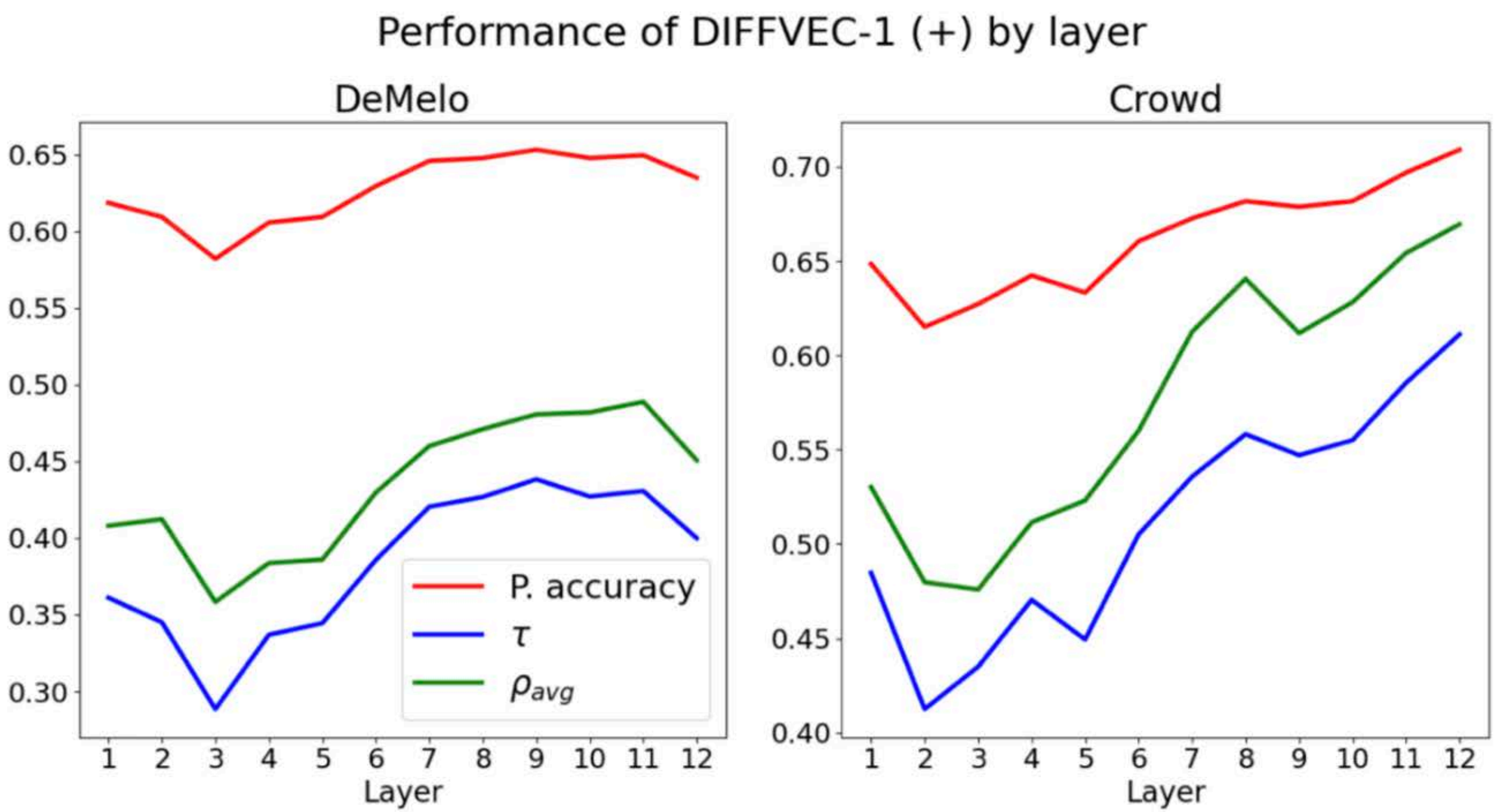
awesome - good (1(+))
horrible - bad (1(-))
ancient - old
gorgeous - pretty
hideous - ugly



BERT

word2vec

Performance by layer



Multilingual Ranking

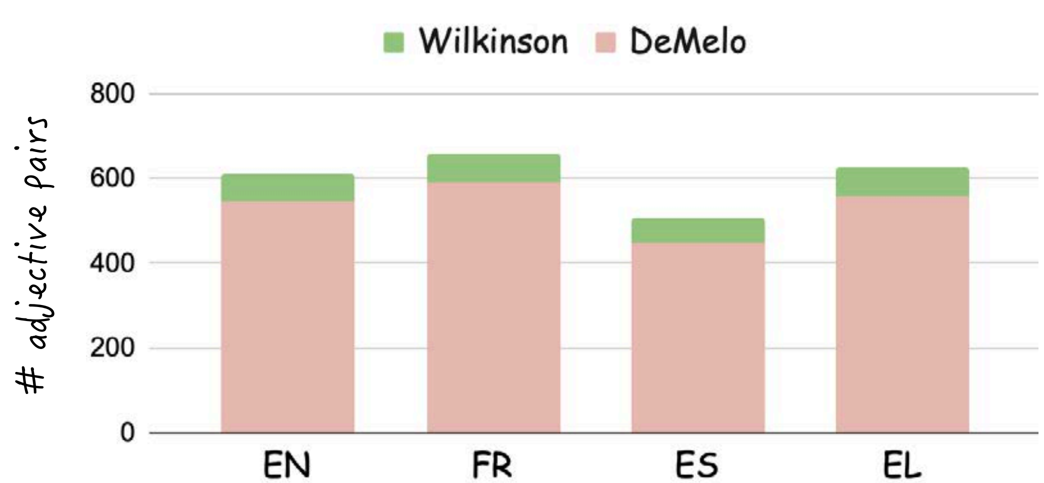
The MULTI-SCALE dataset

(paper @NAACL-HLT 2021)

- Translations into French, Spanish and Greek.
- Sentences from OSCAR (UkWaC for English).

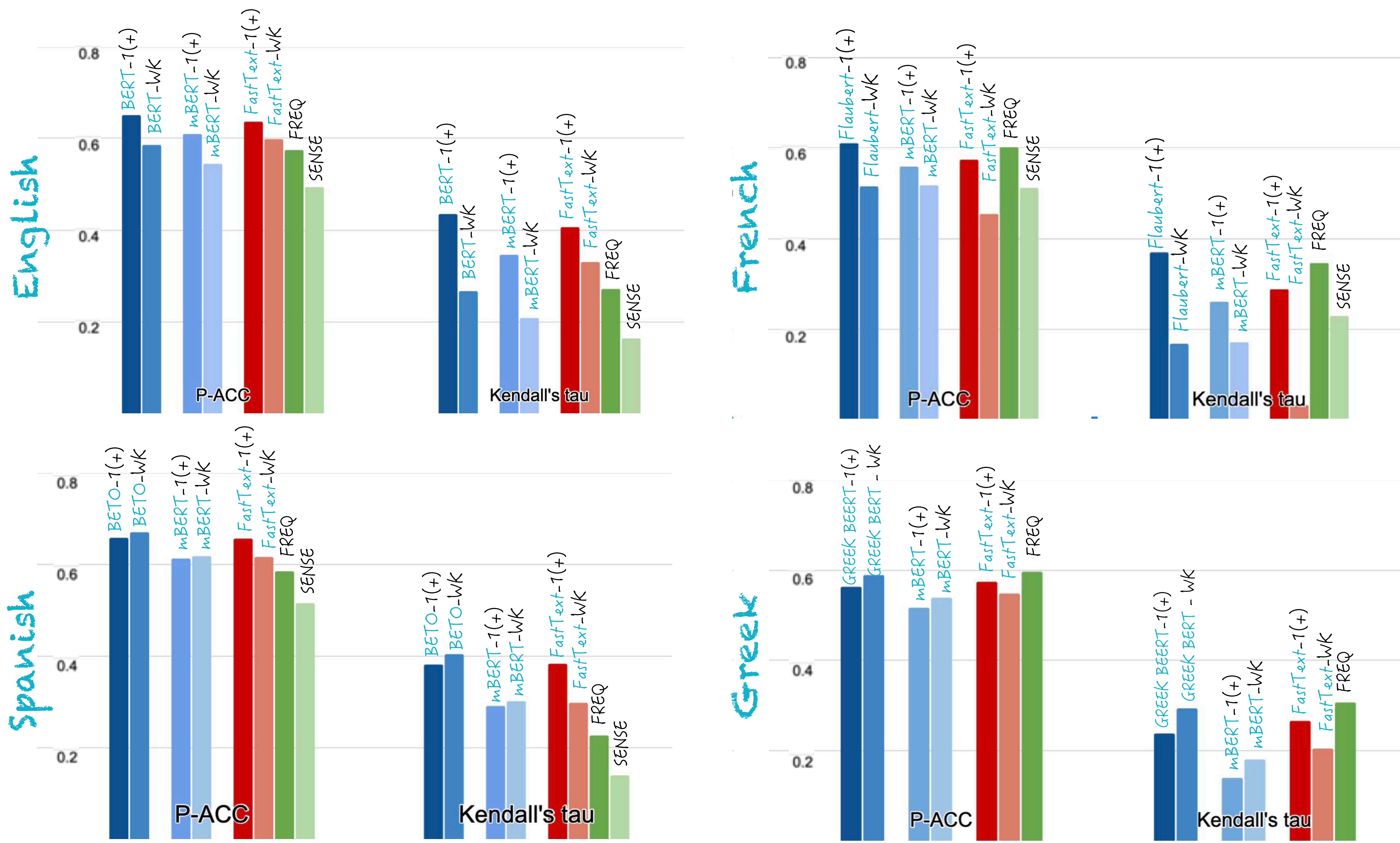
Models

- **EN**: BERT base (Devlin et al., 2019), **FR**: Flaubert (Le et al., 2020), **SP**: BETO (Cañete et al., 2020), **GR**: Greek BERT (Koutsikakis et al., 2020)
- Multilingual BERT



	DeMelo
EN	dim → gloomy → dark → black
FR	terne → sombre → foncé → noir
ES	sombrío → tenebroso → oscuro → negro
EL	αμυδρός || αχνός → μουντός → σκοτεινός→ μαύρος
	Wilkinson
EN	bad → awful → terrible → horrible
FR	mauvais → affreux → terrible → horrible
ES	malo → terrible → horrible → horroroso
EL	κακός → απαίσιος → τρομερός → φρικτός

Results on DeMelo



Indirect Question Answering

Q: Was he a successful ruler?

A: Oh, a tremendous ruler.

(YES!)

Q: Does it have a large impact?

A: It has a medium-sized impact.

(NO!)

Indirect Question Answering

adjq

Q: Was he a successful ruler?

A: Oh, a tremendous ruler.

adja

(YES!)

adjq

Q: Does it have a large impact?

A: It has a medium-sized impact.

adja

(NO!)

Indirect Question-Answer Pairs (IDQA) Dataset (deMarneffe et al., 2010)

- 123 Q-A pairs
- decision procedure for using pairwise intensity scores to predict the polarity of the answer

.

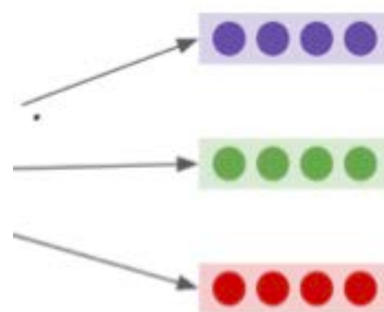
- compute BERT embeddings for adj_q and adj_a
- if $int(adj_a) >= int(adj_q)$, predict YES
- else predict NO
- in the presence of negation, switch YES to NO

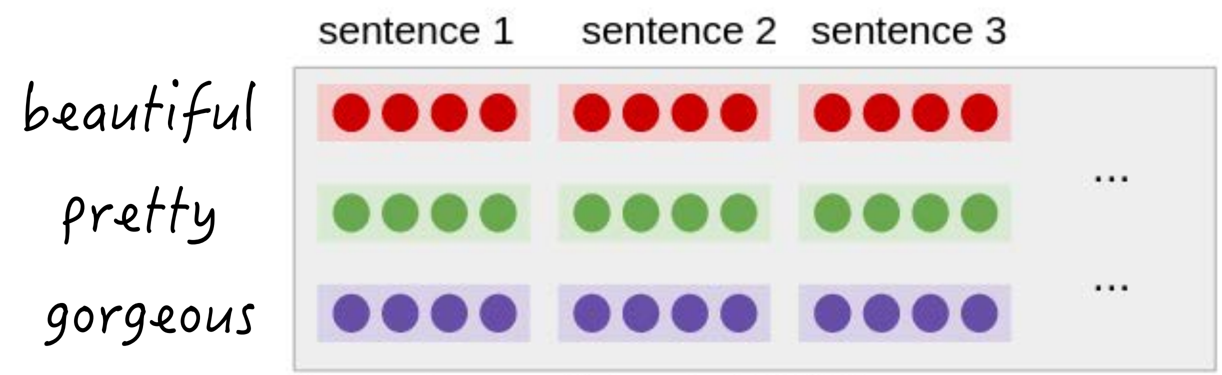
BERT representations

scale: [pretty => beautiful => gorgeous]

Punta Cana is **gorgeous** .

{ pretty, beautiful }





10 sentences * |scale| * 12 layers

What a **beautiful** sunset!

{ pretty, gorgeous }

You look **pretty** today.

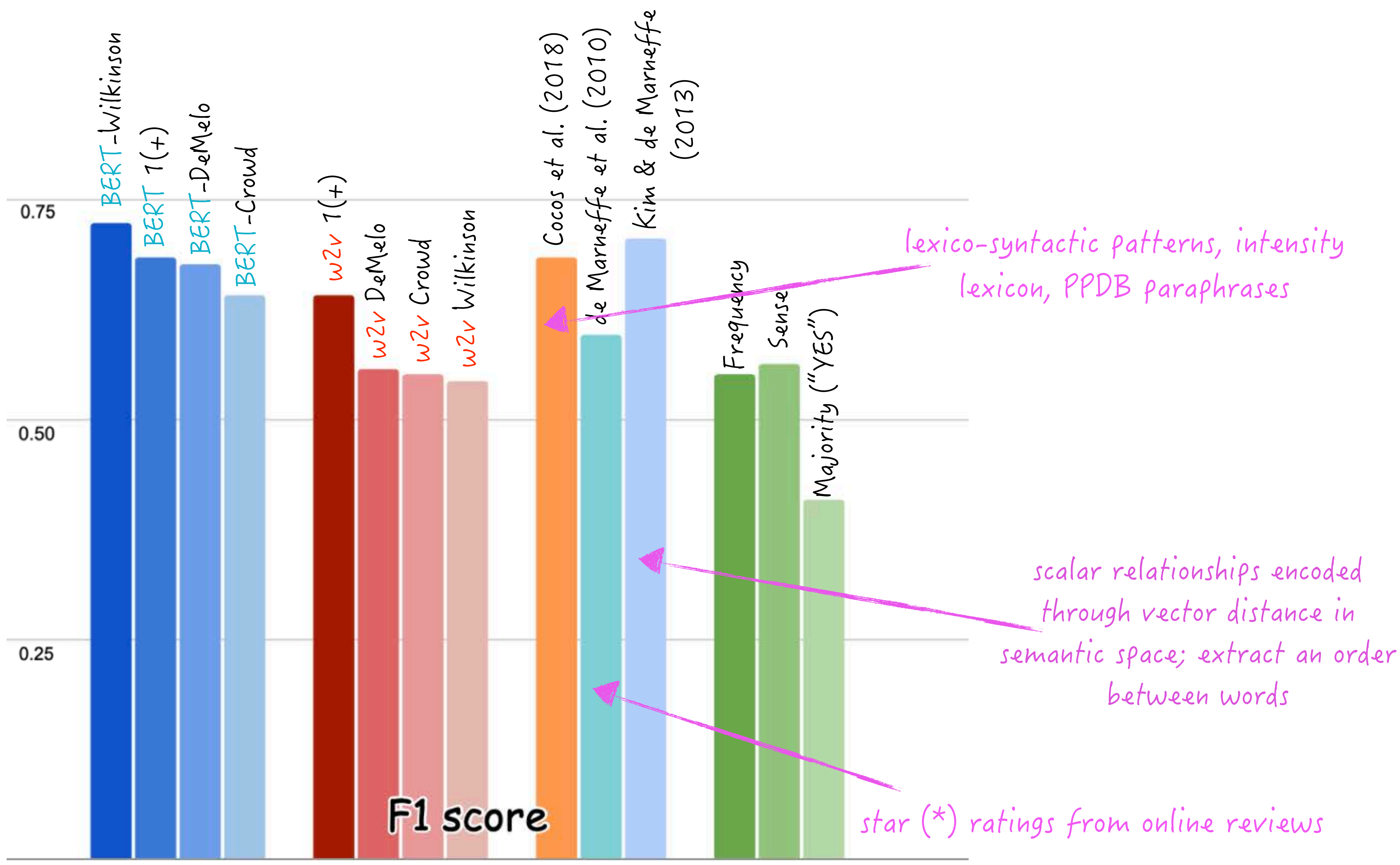
{ beautiful, gorgeous }

$$avg(\overrightarrow{beautiful_1}, \overrightarrow{beautiful_2}, \ldots, \overrightarrow{beautiful_{10}}) \rightarrow \boxed{\overrightarrow{beautiful}}$$

$$avg(\overrightarrow{pretty_1}, \overrightarrow{pretty_2}, \ldots, \overrightarrow{pretty_{10}}) \rightarrow \boxed{\overrightarrow{pretty}}$$

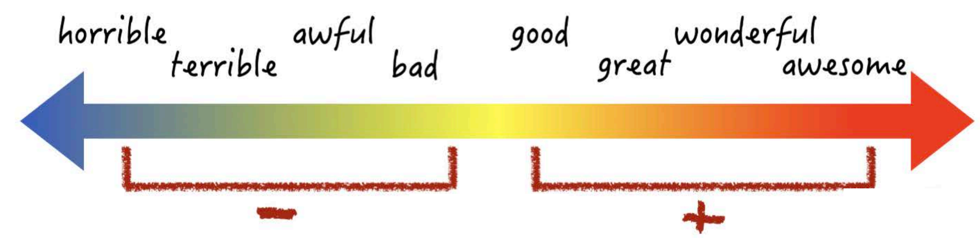
$$avg(\overrightarrow{gorgeous_1}, \overrightarrow{gorgeous_2}, \ldots, \overrightarrow{gorgeous_{10}}) \rightarrow \boxed{\overrightarrow{gorgeous}}$$

Indirect QA results



Take away message

* Contextualised representations encode abstract semantic notions, such as intensity.
* A <u>single adjective pair</u> is sufficient for obtaining good results in different languages!



Q: Was he a *successful* ruler?

A: Oh, a *tremendous* ruler.

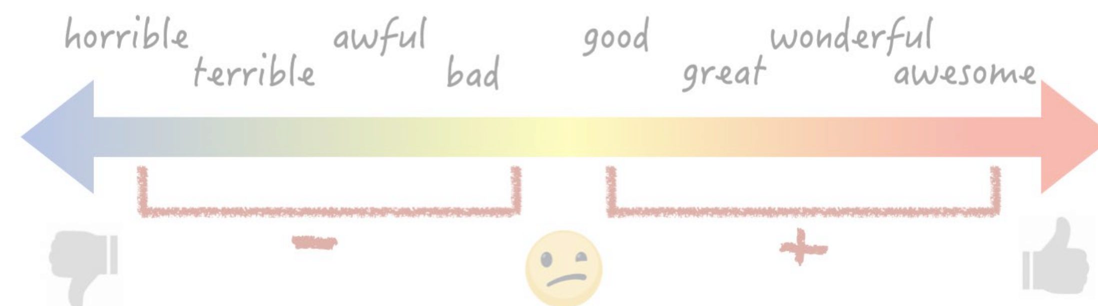
(YES!)

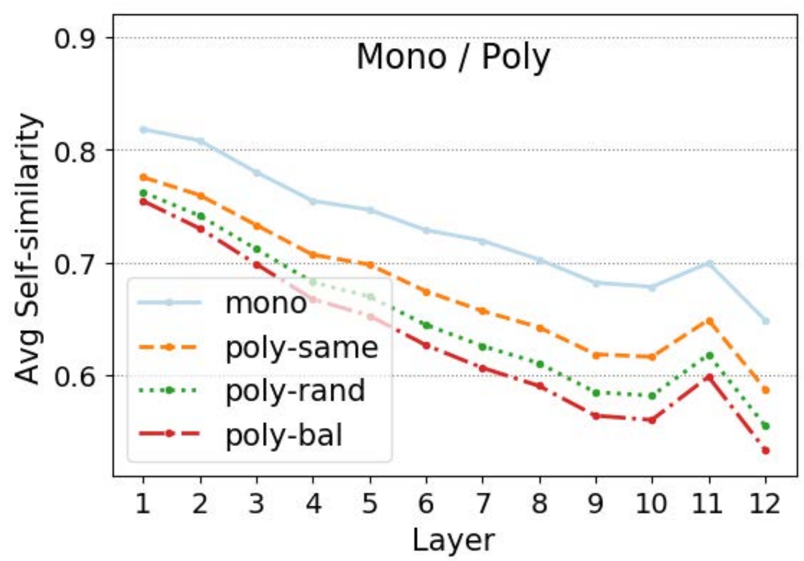
* Intensity is useful for product review analysis and recommendation systems, emotional chatbots and QA. But also for fake news, hate speech or subjectivity detection.

* Are other semantic notions encoded in the space? For example emotions, polarity, formality, or complexity?

What BERT knows about..

Semantic relationships and intensity in particular?





Lexical polysemy and sense partitionability?

Noun properties and their prototypicality?

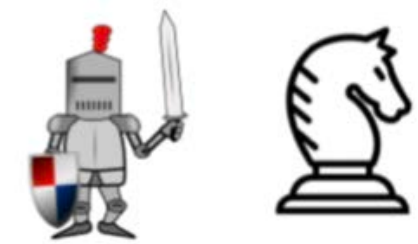
all strawberries are [MASK]

[MASK] balloons are colourful.

Let's play mono-poly!

sofa

- Can BERT models distinguish **mono**semous from **poly**semous words?
- When is knowledge about polysemy acquired? (pre-training? new contexts?)
- What is the influence of word frequency and grammatical category?



knight

...

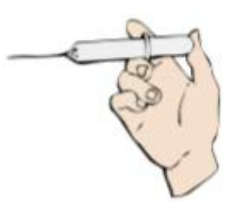
shot

Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses (TACL 2021)

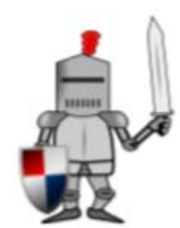
Data

Sentences from sense annotated corpora illustrating word usages

- English: SemCor (Miller et al., 1993)
- French, Spanish, Greek: EuroSense (Delli Bovi et al., 2017)



flu **shot**



knights in the middle ages

firing a **shot**

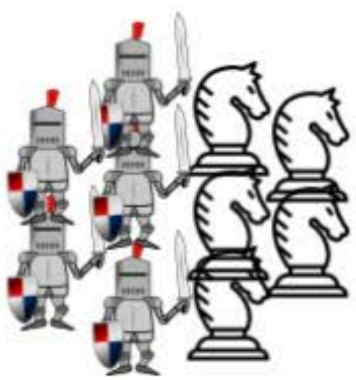
Important note: Annotations only serve to control for the composition of the sentence pools used in the experiments (not used for training!)

Sentence pools

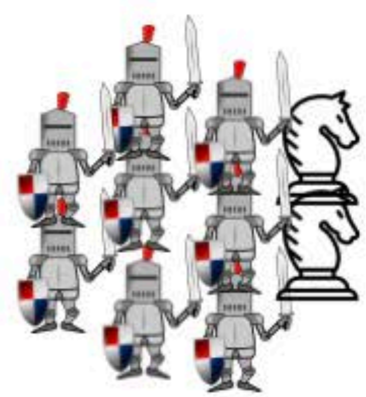
Sentences are grouped controlling for **sense distribution**

- **418 monosemous** words: 10 **random** instances
- **418 polysemous** words: 10 instances each, 3 sense distributions

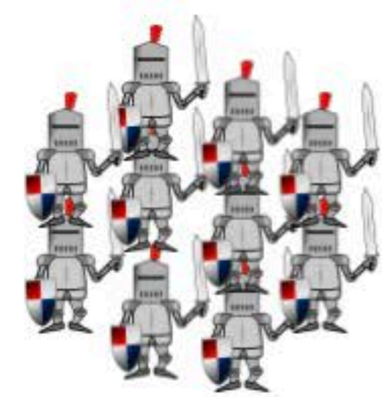
mono



poly-bal
(balanced)



poly-rand
(random)



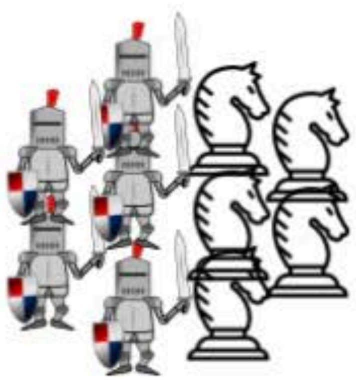
poly-same
(one sense)

Sentence pools

Sentences are grouped controlling for **sense distribution**

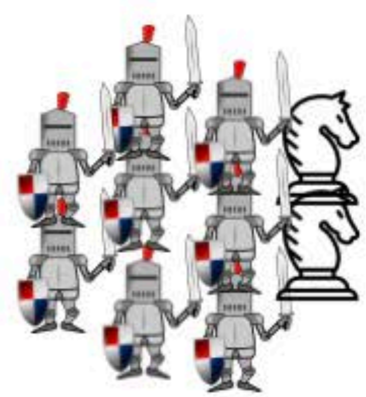
- **418 monosemous** words: 10 **random** instances
- **418 polysemous** words: 10 instances each, 3 sense distributions

mono



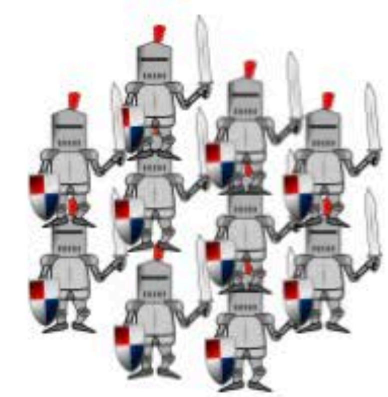
poly-bal

(balanced)



poly-rand

(random)



poly-same

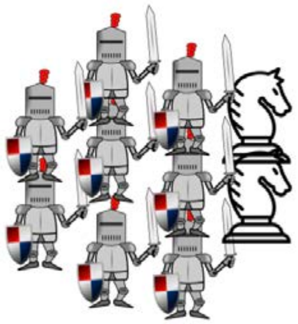
(one sense)

room.n

CHAMBER	(. . .) he left the room, walked down the hall (. . .)
SPACE	It gives them room to play and plenty of fresh air.
OPPORTUNITY	Even here there is room for some variation, for metal surfaces vary (. . .)

CHAMBER	The room vibrated as if a giant hand had rocked it.
CHAMBER	(. . .) Tell her to come to Adam's room (. . .)

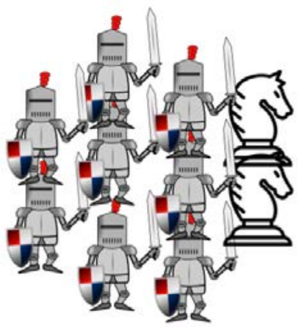
Sentence pools



poly-rand
(random)

- Strongly biased towards the MFS due to the skewed frequency distribution of word senses (Kilgarriff, 2004)
- Closer to the expected natural occurrence of senses in a corpus
- Serves to estimate the behaviour of the models in a real-world setting

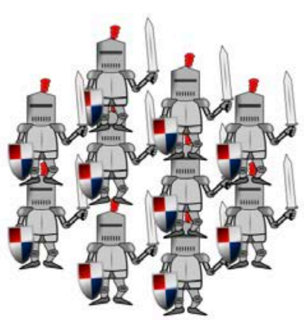
Sentence pools



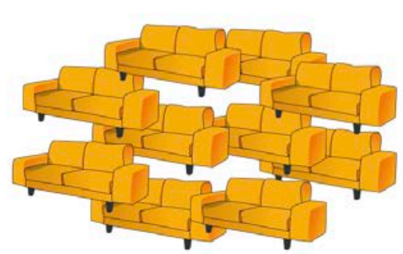
poly-rand
(random)

- Strongly biased towards the MFS due to the skewed frequency distribution of word senses (Kilgarriff, 2004)
- Closer to the expected natural occurrence of senses in a corpus
- Serves to estimate the behaviour of the models in a real-world setting

a key comparison!



vs.



poly-same
(one sense)

- Pools with similar composition: just one sense
- No meaning variation inside the pool: serves to explore whether BERT can distinguish mono from poly words using information from pre-training.

Models

- BERT (Devlin et al., 2019; bert-base-uncased/cased)
- ELMo (Peters et al., 2018)
- context2vec (Melamud et al., 2016)

- Flaubert (Le et al., 2020)
- BETO (Cañete et al., 2020)
- Greek BERT (Koutsikakis et al., 2020)

- Multilingual BERT (mBERT) (Devlin et al., 2019)

Models

- BERT (Devlin et al., 2019; bert-base-uncased/cased)
- ELMo (Peters et al., 2018)
- context2vec (Melamud et al., 2016)

- Flaubert (Le et al., 2020)
- BETO (Cañete et al., 2020)
- Greek BERT (Koutsikakis et al., 2020)

- Multilingual BERT (mBERT) (Devlin et al., 2019)

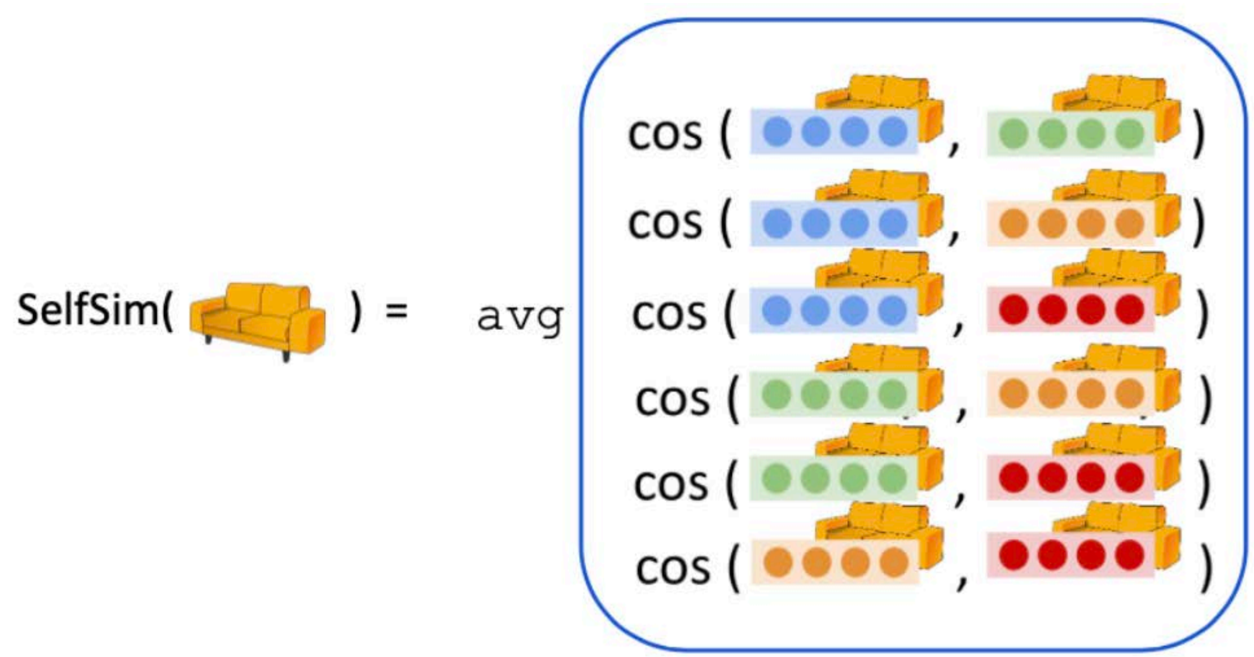
Mono-poly approach

- Similarity of contextualised instances/representations (Erk et al., 2009; 2013)
- For each instance i of a word w, a representation is extracted from the 12 BERT layers.
- **Self-similarity** (*SelfSim*) of w in a sentence pool p and a layer ℓ
  - the average of the pairwise cosine similarities of its representations in ℓ (Ethayarajh, 2019)

$$SelfSim_l(w) = \frac{1}{|I|^2 - |I|} \sum_{i \in I} \sum_{\substack{j \in I \\ j \neq i}} cos(x_{wli}, x_{wlj})$$

Mono-poly approach

- Similarity of contextualised instances/representations (Erk et al., 2009; 2013)
- For each instance i of a word w, a representation is extracted from the 12 BERT layers.
- **Self-similarity** ($SelfSim$) of w in a sentence pool p and a layer ℓ
    - the average of the pairwise cosine similarities of its representations in ℓ (Ethayarajh, 2019)

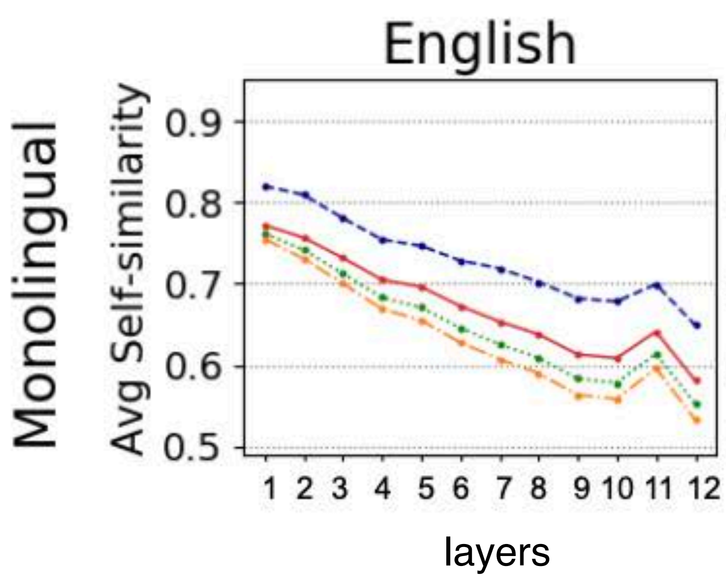


SelfSim

- Average *SelfSim* for all words in a pool p (`mono`, `poly-same/bal/rand`)

- We expect *SelfSim* to be
  - higher for `mono` words, lower for words with many senses
  - higher in the `poly-same` pool than in the other `poly` pools which contain instances of different senses
  - to be lower in layers where the impact of context variation is stronger

Mono-poly distinctions



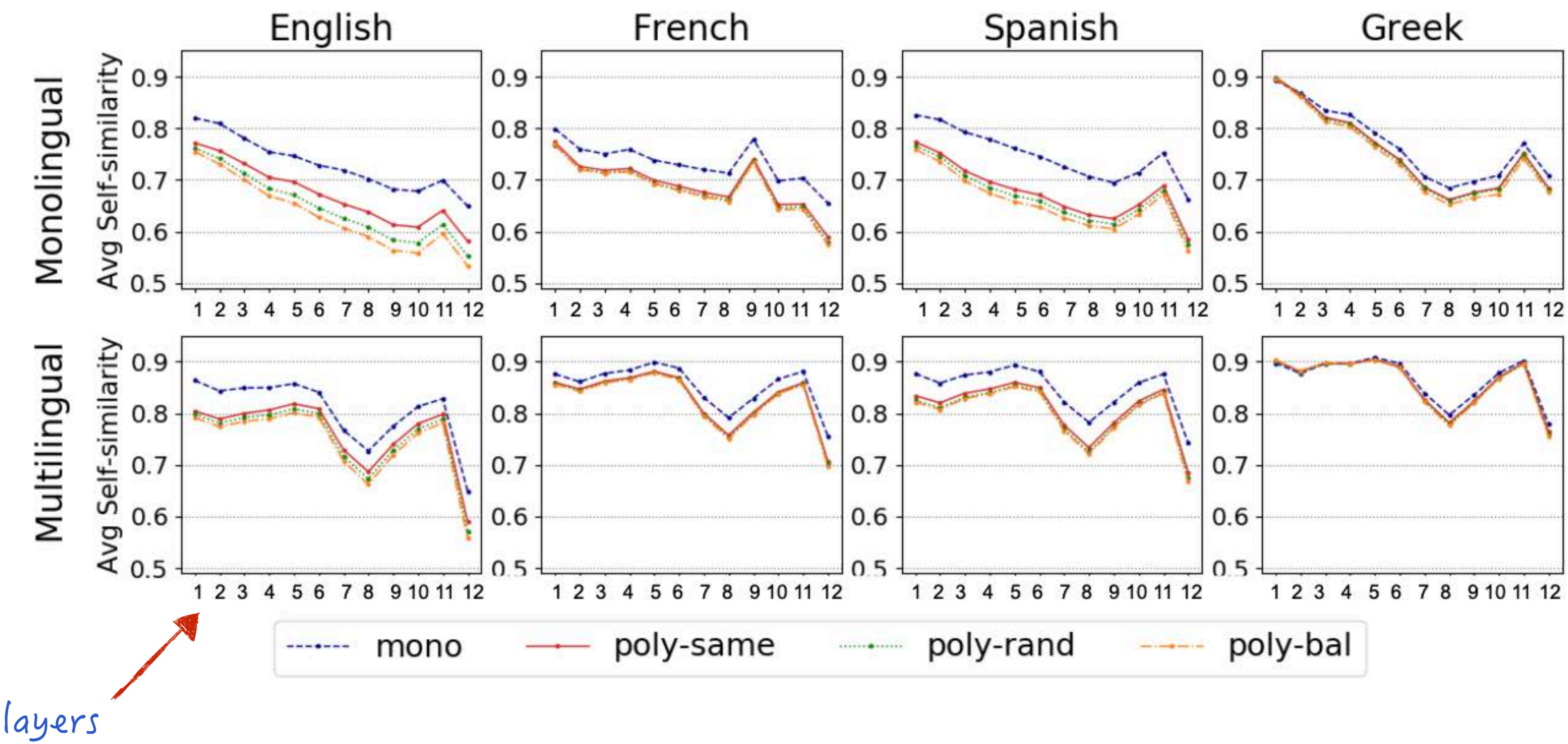
Differences are significant across all layers

BERT encodes two types of lexical knowledge!

- Information <u>acquired through pre-training</u>, as reflected in the `mono/poly-same` distinction

- Information from the <u>particular instances</u> used to extract the representations, as shown by `poly` distinctions (SelfSim in `poly-bal` < SelfSim in `poly-rand` < SelfSim in `poly-same`)

Mono-poly distinctions



Differences between mono and poly-rand are significant across all layers of all models, except for mBERT for Greek (significant in 10 layers).

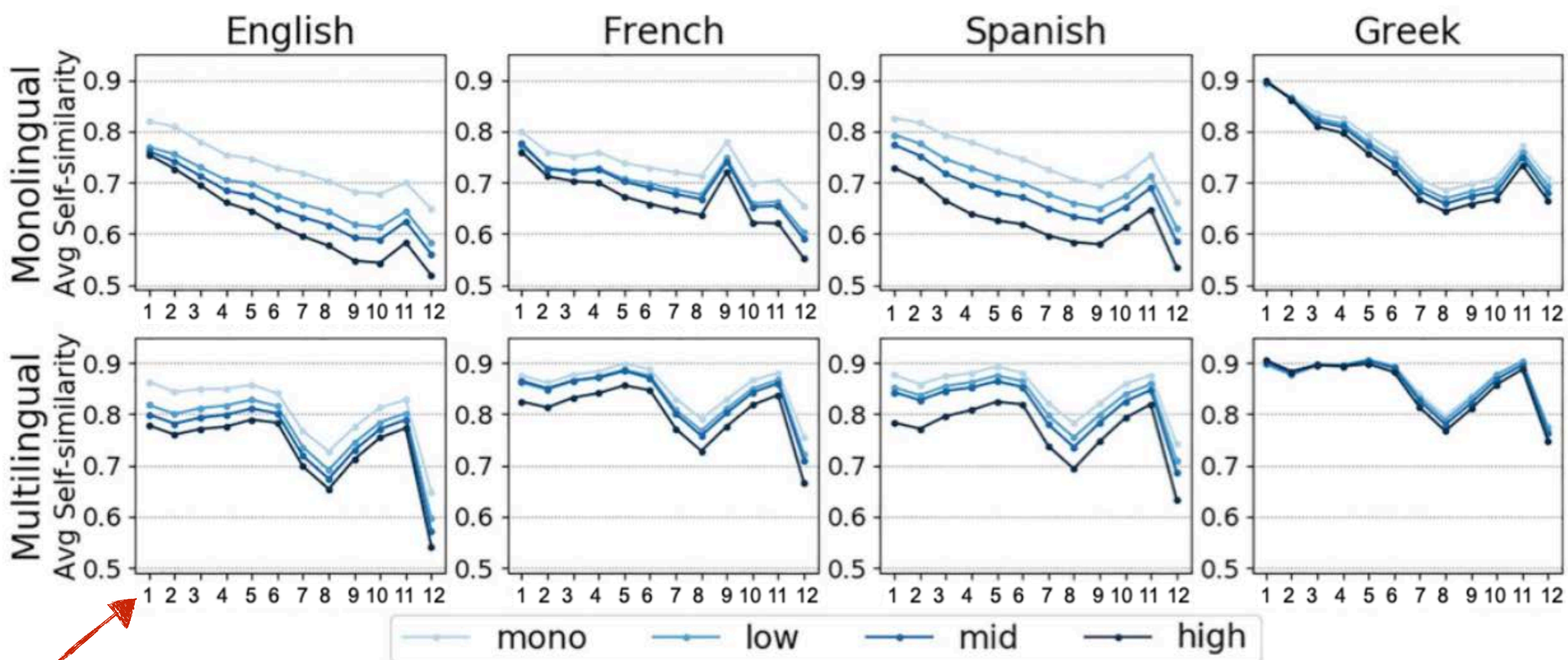
Polysemy bands

We group words into 3 polysemy bands according to their number of senses in WordNet (Fellbaum, 1998) and in BabelNet (Navigli and Ponzetto, 2012)

- **low**: $2 \leq k \leq 3$ senses
- **mid**: $4 \leq k \leq 6$ senses
- **high**: $k > 6$ senses

Polysemy bands

✓ poly-rand pool

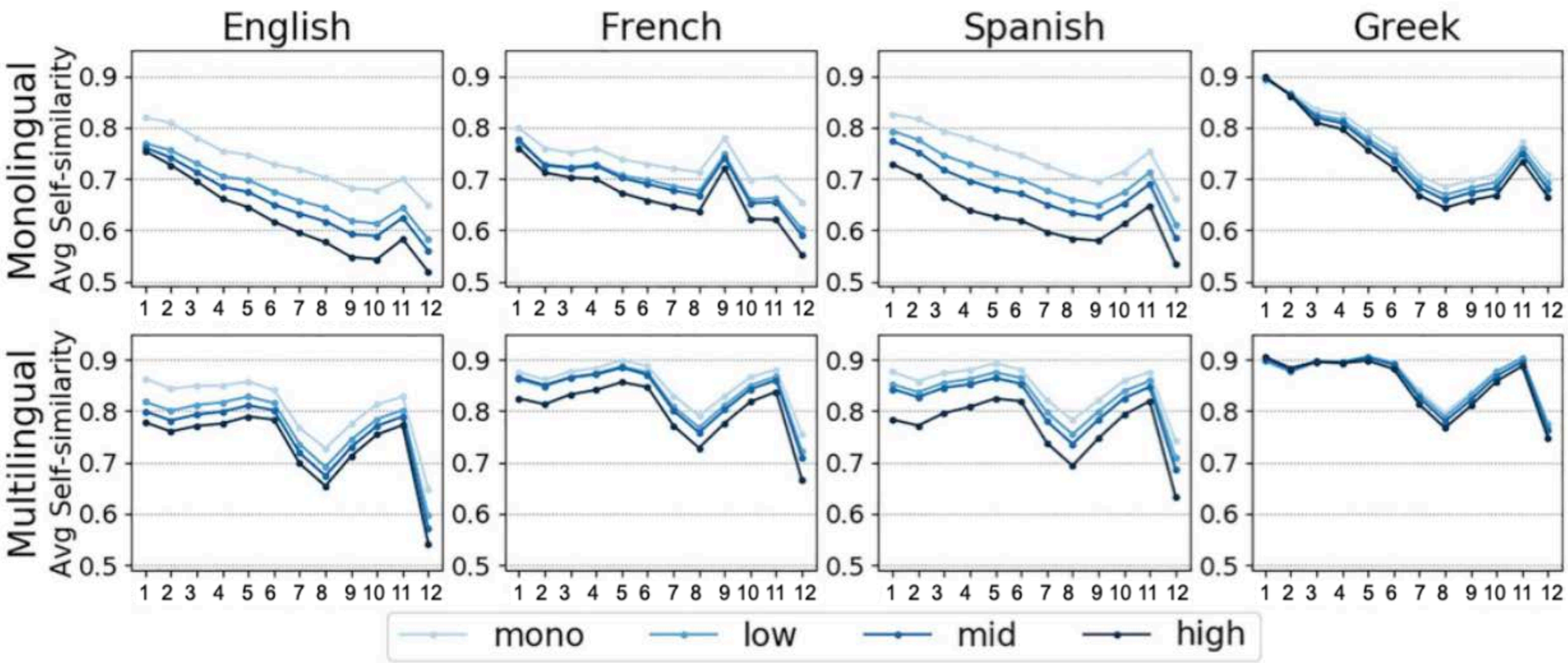


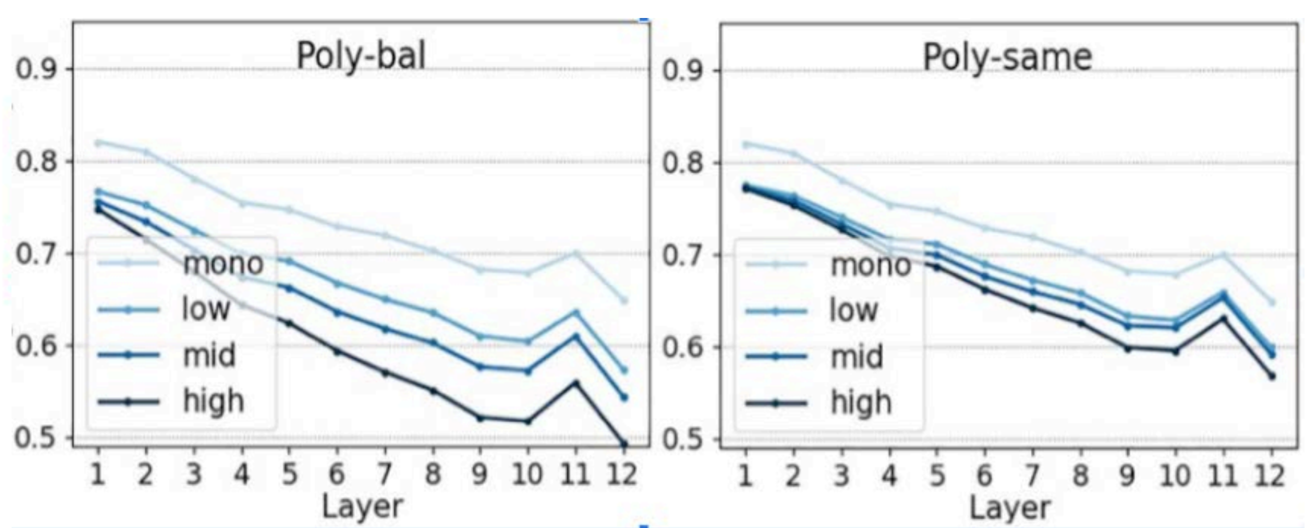
layers

- low: 2 ≤ k ≤ 3 senses
- mid: 4 ≤ k ≤ 6 senses
- high: k > 6 senses

Distinctions are less clear but inter-band differences are significant in all but a few layers of the models.

Polysemy bands





Observations

Why are English BERT and BETO better than other models?

- Might be due to the quality and quantity of the training data

Why is mBERT worse than the monolingual models?

- The “curse of multilinguality” (Conneau et al., 2020)
- Not enough training data?
- English-centric tokenization
- Higher anisotropy?

γιγάντιος → γ - ι - γ - άν - τιος

Anisotropy analysis

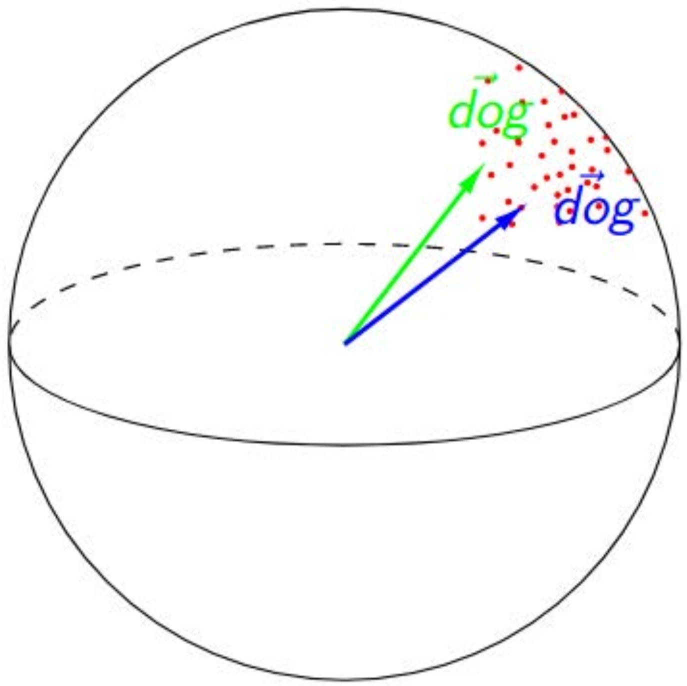


Figure from
Ethayarajh (2019)

High anisotropy

- representations occupy a narrow cone in the vector space
- lower quality similarity estimates

Anisotropy analysis

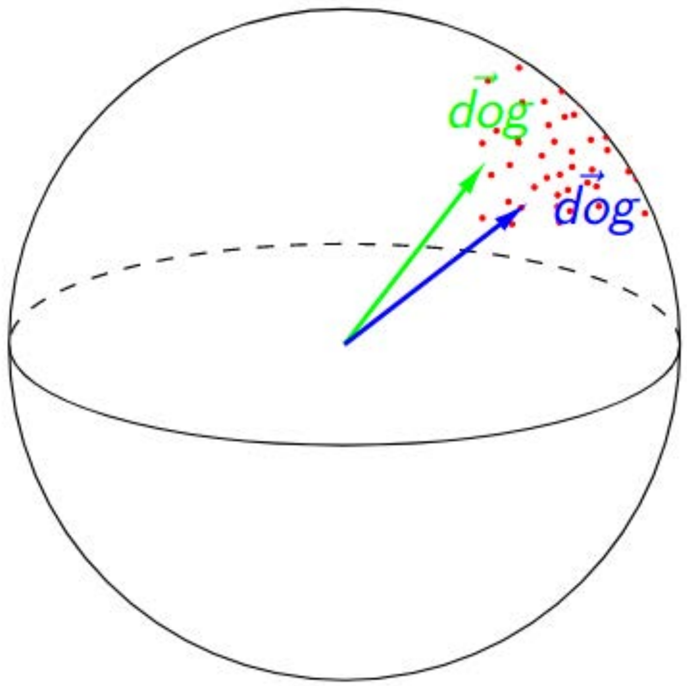


Figure from
Ethayarajh (2019)

High anisotropy

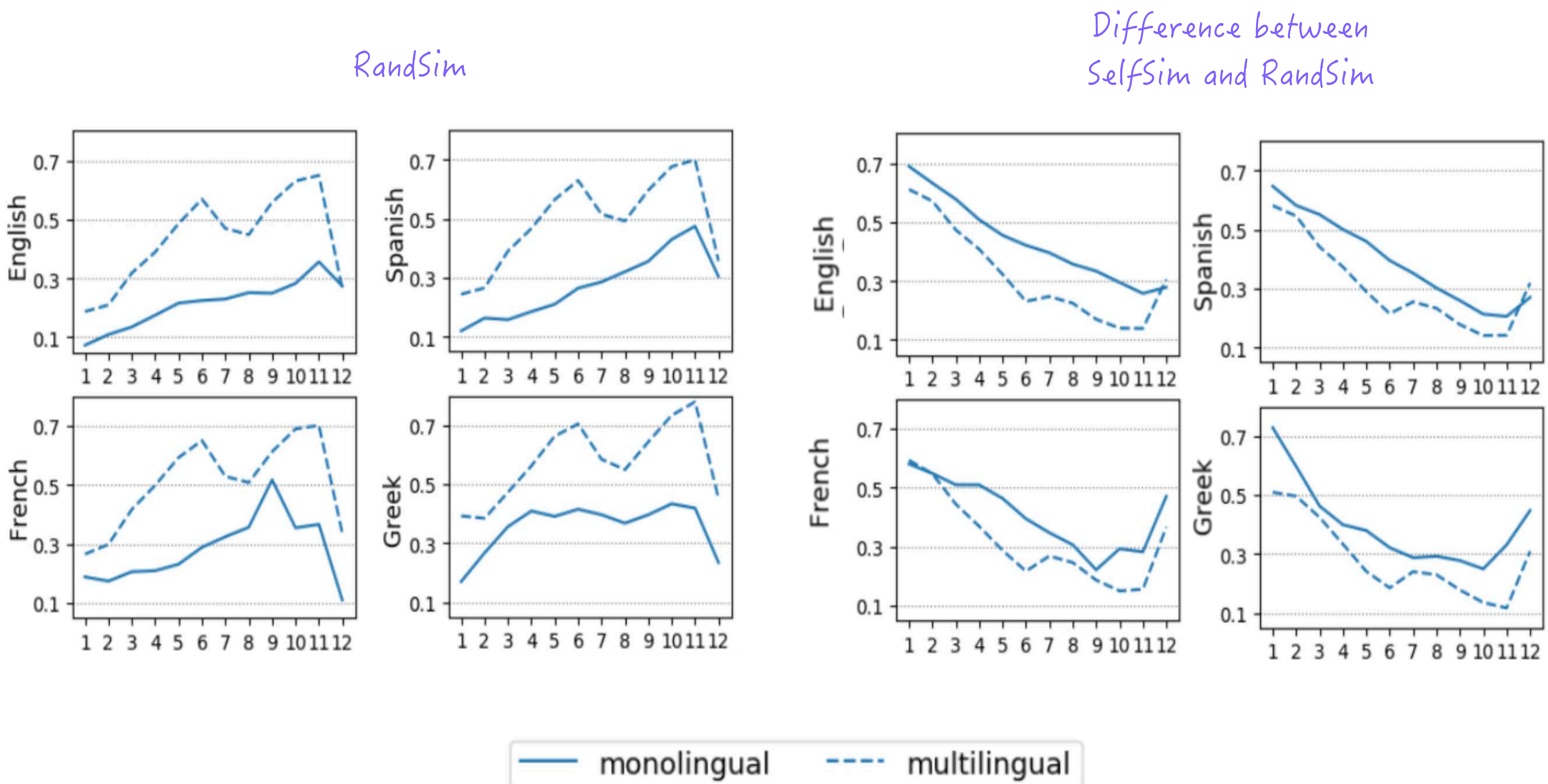
- representations occupy a narrow cone in the vector space
- lower quality similarity estimates

SelfSim: `cos(knight`$_1$`,knight`$_2$`)`

Similarity of random words (RandSim): `cos(knight`$_1$`,sofa`$_1$`)`

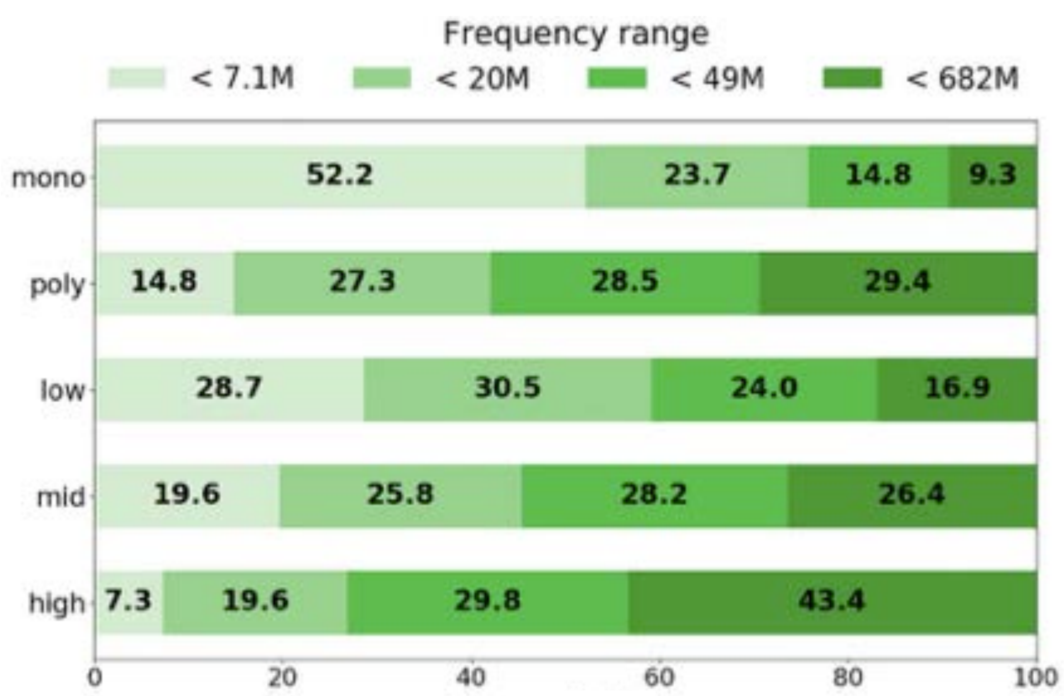
- 2,183 random EN word pairs, 1,318 in other languages
- calculate the similarity between two random instances of the words in each pair
- take the average over all pairs (RandSim)

Anisotropy analysis



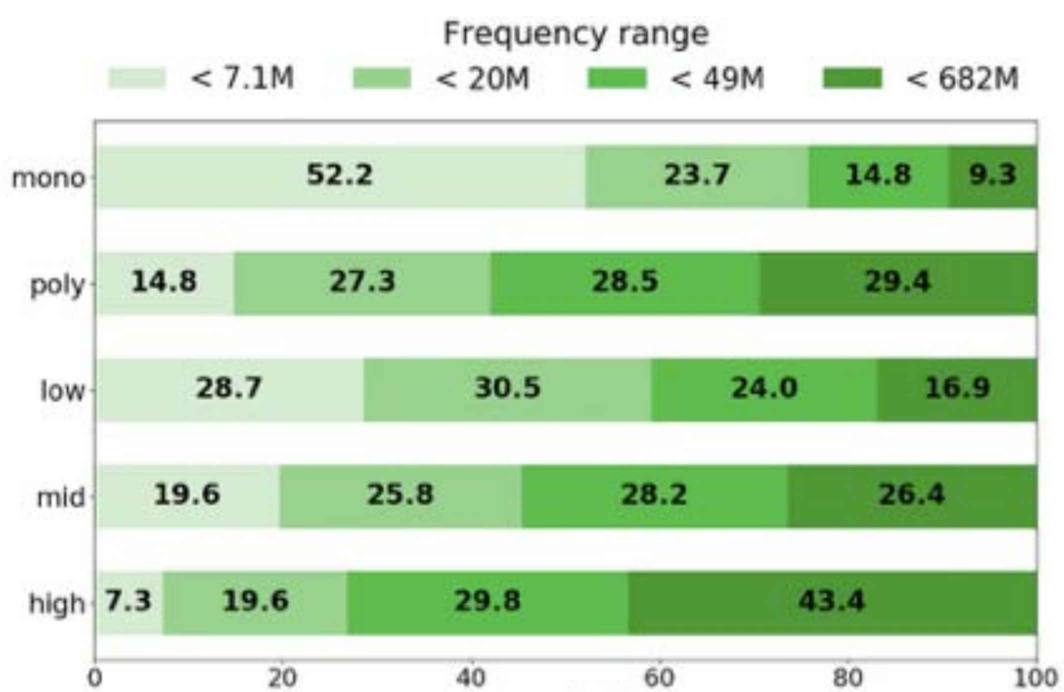
Frequency and polysemy

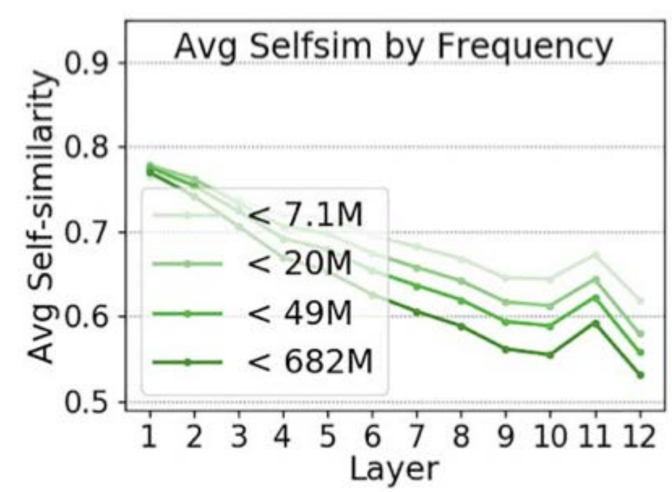
- Strong correlation between word frequency and number of senses (Zipf, 1945)
- Frequencies from Google Ngrams and the Oscar corpus (Suárez et al., 2019)



Frequency and polysemy

- Strong correlation between word frequency and number of senses (Zipf, 1945)
- Frequencies from Google Ngrams and the Oscar corpus (Suárez et al., 2019)

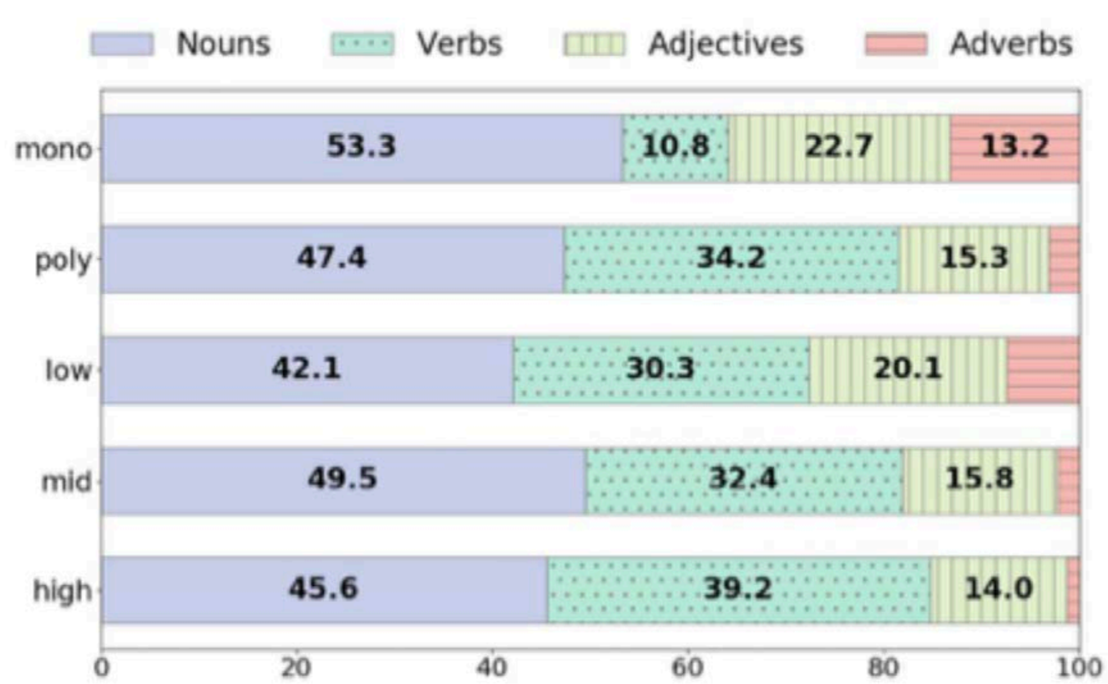




- Clear ordering by range
- BERT can distinguish words by frequency
- Same trend for monolingual models in the other languages

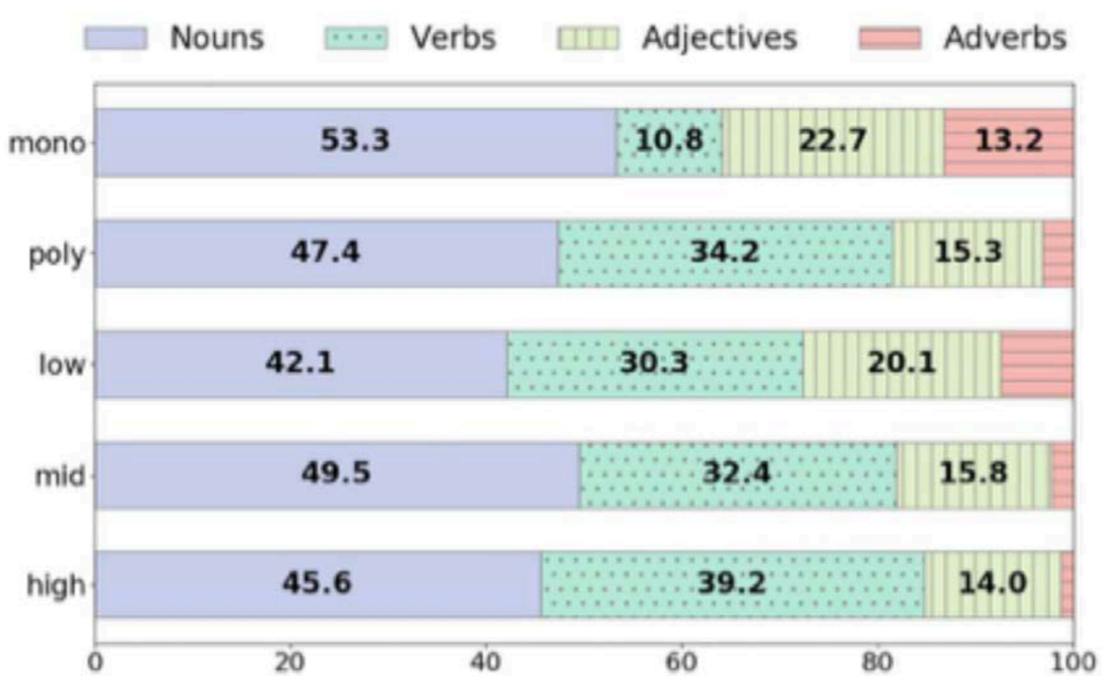
PoS distribution in each band

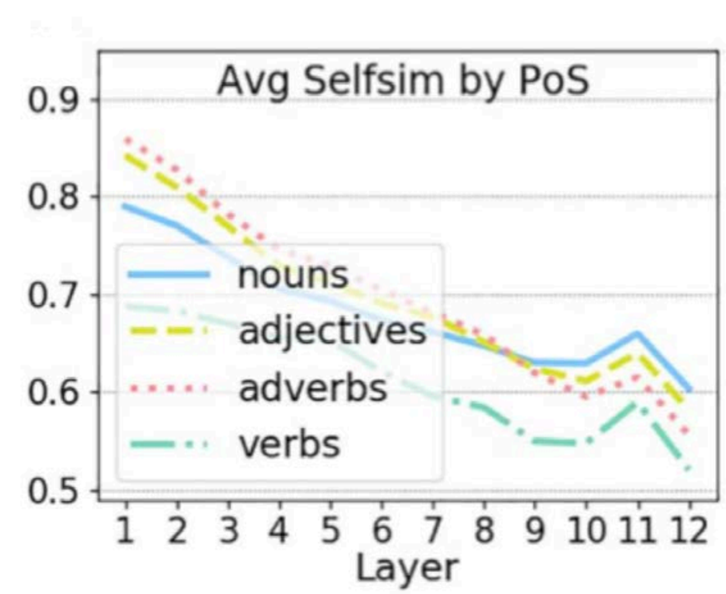
- Strong correlation between word frequency and number of senses (Zipf, 1945)
- Frequencies from Google Ngrams and the Oscar corpus (Suárez et al., 2019)



PoS distribution in each band

- Strong correlation between word frequency and number of senses (Zipf, 1945)
- Frequencies from Google Ngrams and the Oscar corpus (Suárez et al., 2019)

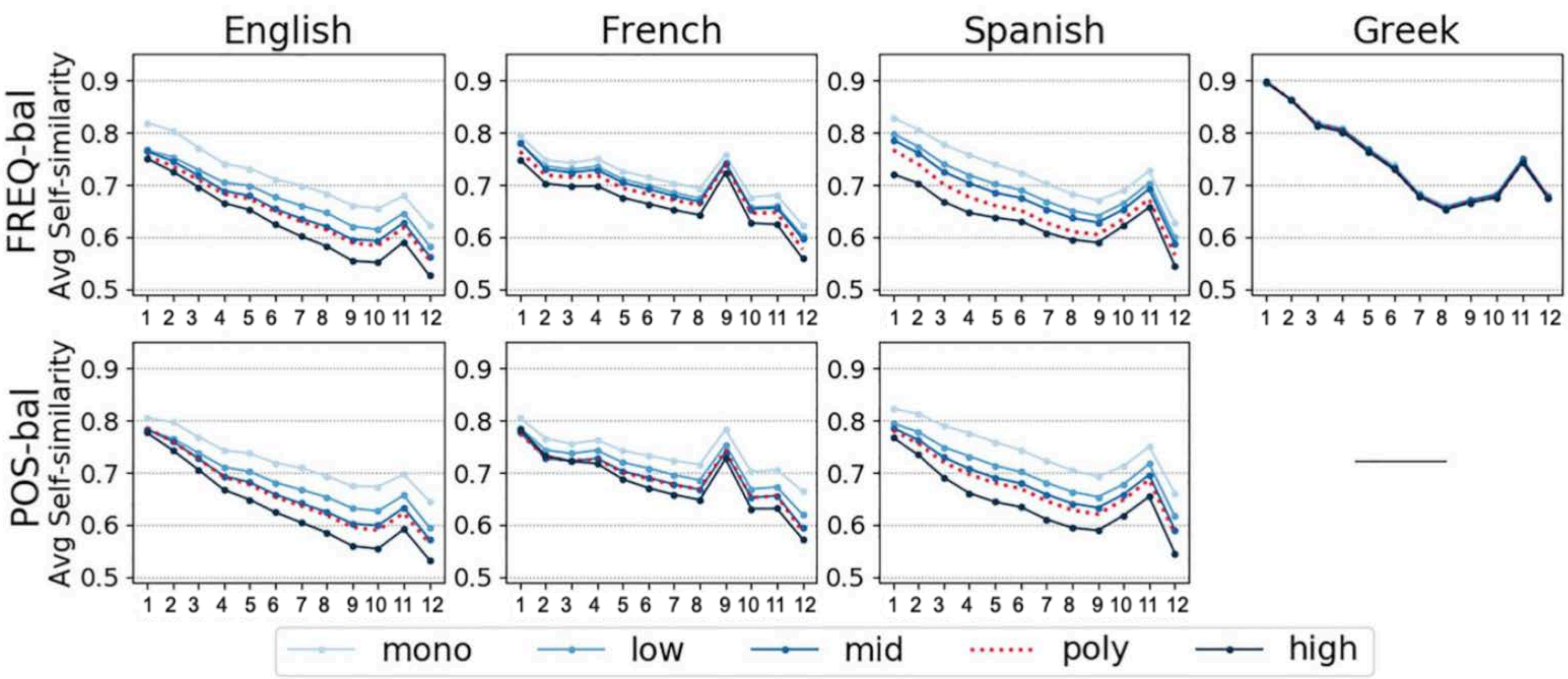




- Verbs have the lowest *SelfSim* due to polysemy
- Same trend for monolingual models in the other languages

Balancing for frequency and PoS

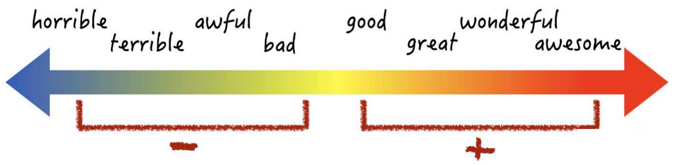
- **POS-bal** bands contain the same number of words of a specific PoS
- **FREQ-bal** bands contain the same number of words in a specific frequency range



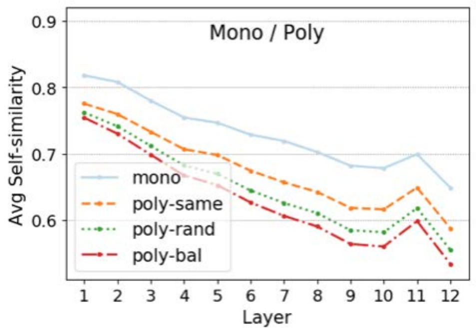
Do BERT models encode knowledge about abstract semantic notions and polysemy?

Yes!

- semantic notions such as intensity can be discovered through simple operations in vector space



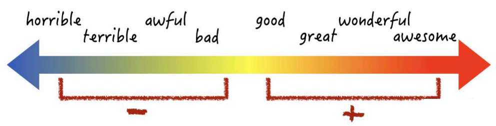
- knowledge about polysemy acquired during pre-training is being combined with information from new contexts of use



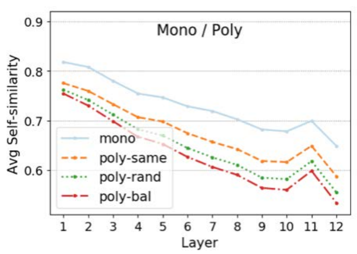
- the two types of information are encoded in BERT-type models in the four languages of study, but seem to be of higher quality in English BERT

Why is this information useful?

* Knowledge about intensity



  ‣ product review analysis and recommendation systems, emotional chatbots, QA systems. But also for fake news, hate speech or subjectivity detection.



* Knowledge about polysemy

  ‣ help lexicographers define words' number of senses
  ‣ study lexical semantic change
  ‣ plan the time and effort needed in semantic annotation tasks
  ‣ identify words with stable semantics that can be safe cues for WSD
  ‣ determine needs in terms of context size for WSD (e.g., in queries, chatbots)
  ‣ guide cross-lingual transfer using unambiguous words as anchors

appreciative < thankful < grateful