Colex2Lang: Language Embeddings from Semantic Typology

Yiyi Chen[1], Russa Biswas[2], Johannes Bjerva[1]

[1]Aalborg University, Copenhagen, Denmark

[2] FIZ Karlsruhe, Karlsruhe, Germany

Semantic Typology (Evans 2013)

Definition

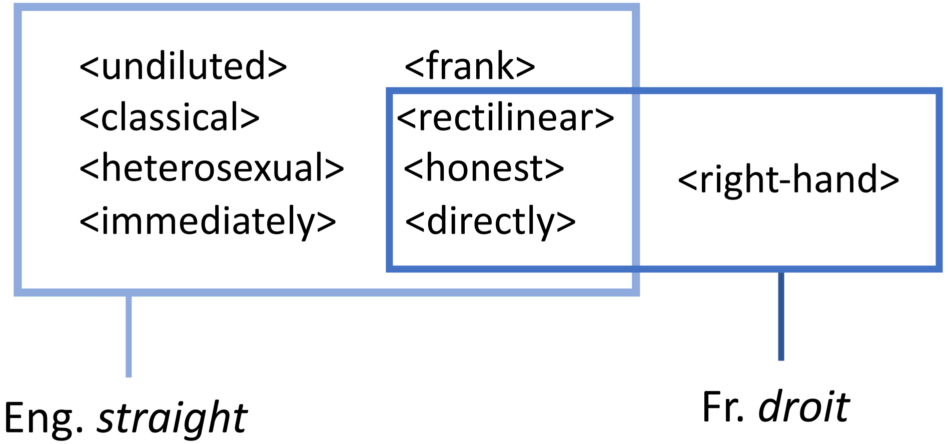
Semantic Typology *is the part of linguistic typology concerned with the expression of meaning in language and languages.*

It is thus the systematic cross-lingual study of how languages express meaning by way of signs.

Shaped by

Culture, human communication, the environment

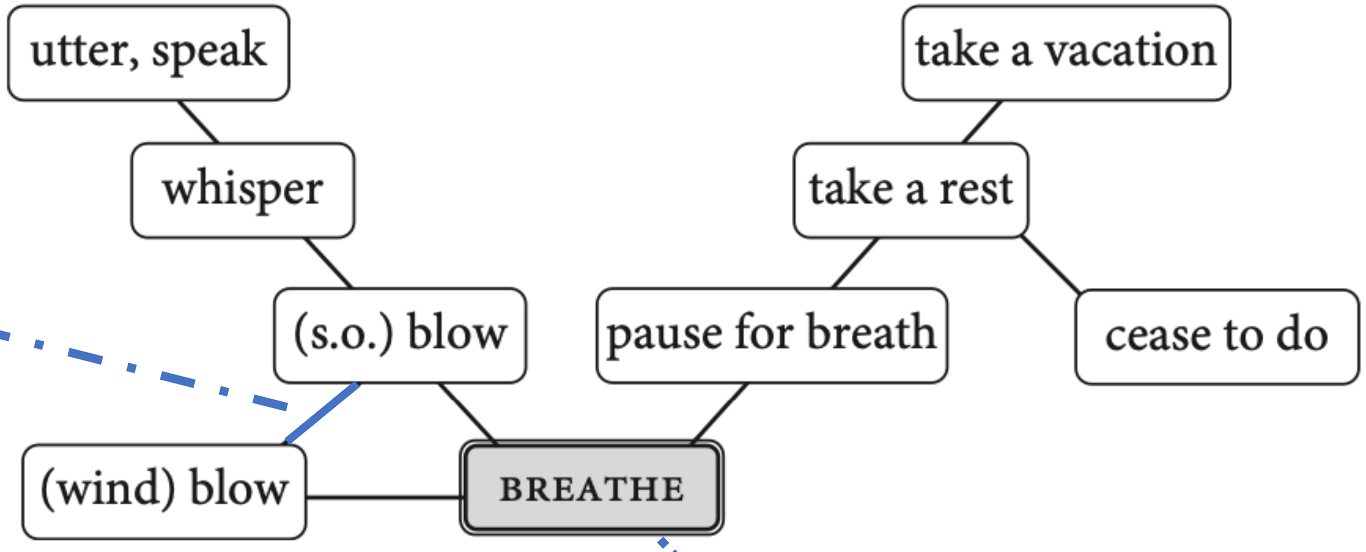
Colexification (François 2008)



A given language is said to **COLEXIFY** two functionally distinct senses *if, and only if*, it can associate them with the same lexical form.

Semantic Maps (François 2008)

Word	Language
breath	English
dux	Russian
souffler	French

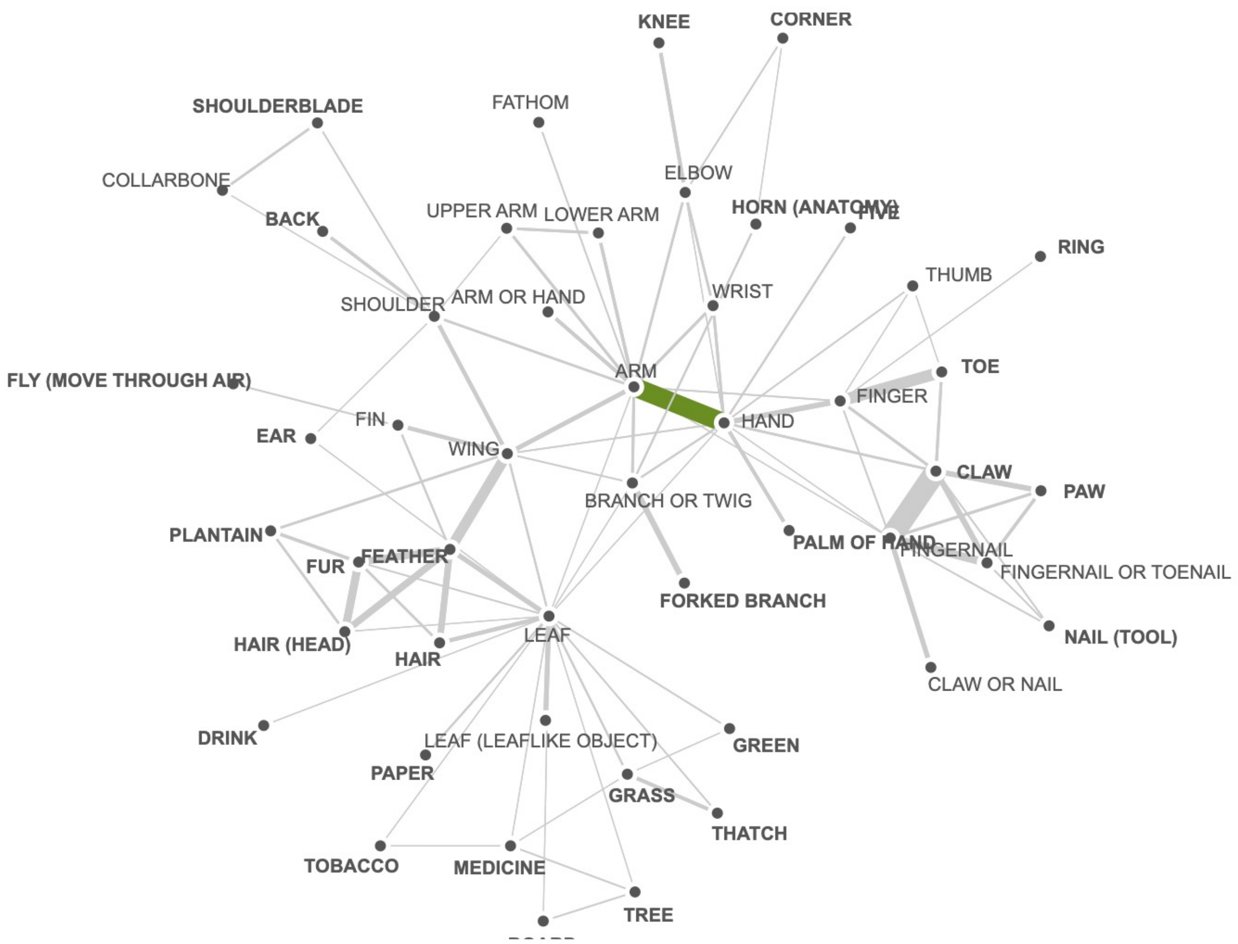


- 16 lexical words in 13 genetically diverse languages

The physiological activityof breathing characteristics of humans and animals

Database of Cross-Linguistic Colexifications (CLICS^3) (Rzymski et al., 2020)

Subgraph ARM



300 colexifications for "HAND" and "ARM":

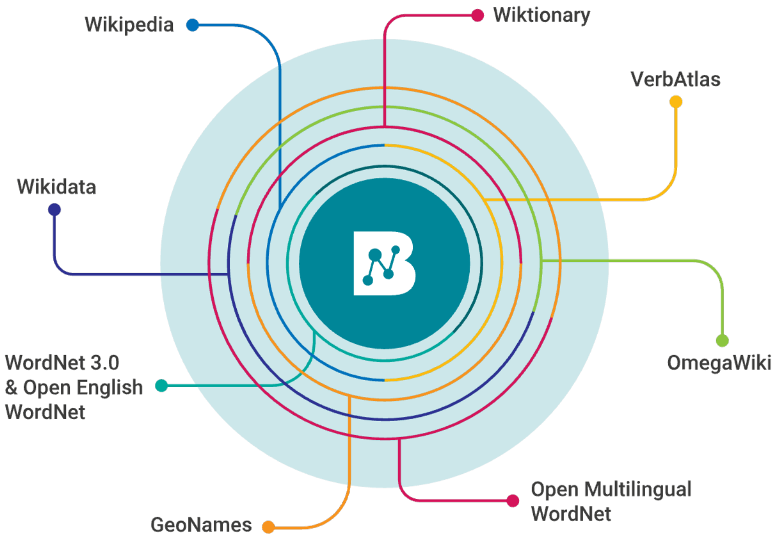
Language	Family	Form
Gawwada	Afro-Asiatic	hargo
Hausa	Afro-Asiatic	hannuu
Hausa	Afro-Asiatic	hannu
Iraqw	Afro-Asiatic	dawa1
Polci	Afro-Asiatic	aam
Tarifiyt Berber	Afro-Asiatic	fus
Hokkaido Ainu	Ainu	tek
Kimochi.unn	Atlantic-Congo	owoko
Kiseri.unn	Atlantic-Congo	kuoko
Lema.unn	Atlantic-Congo	kuwoko
Machame.unn	Atlantic-Congo	wako

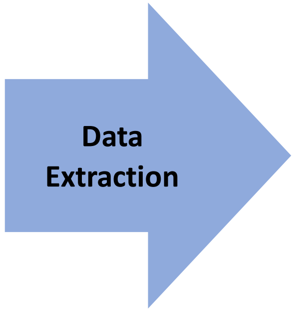
Hypotheses

Language representations learned using colexifications encapsulate a *distinctive language signal.*

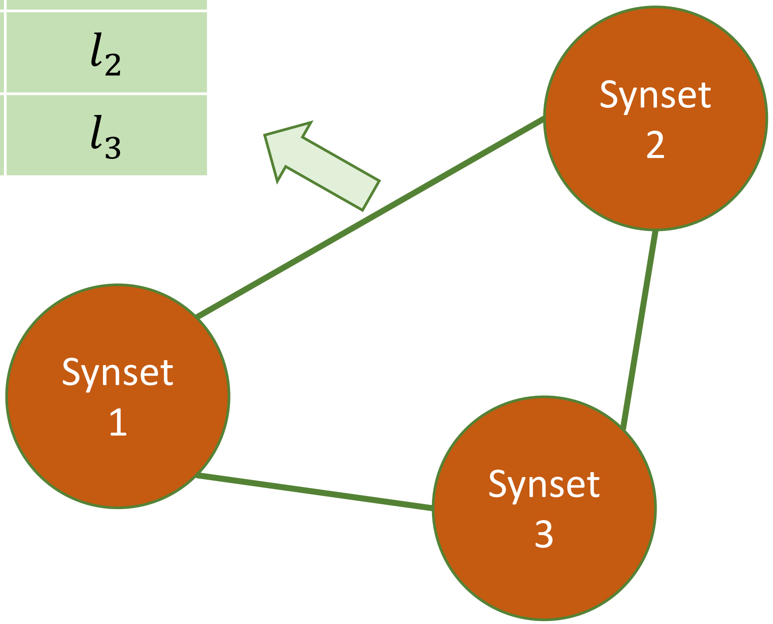
The *data size* of colexification networks has a *positive* impact on the learned language representations and the modelled language similarities.

Colex2Lang

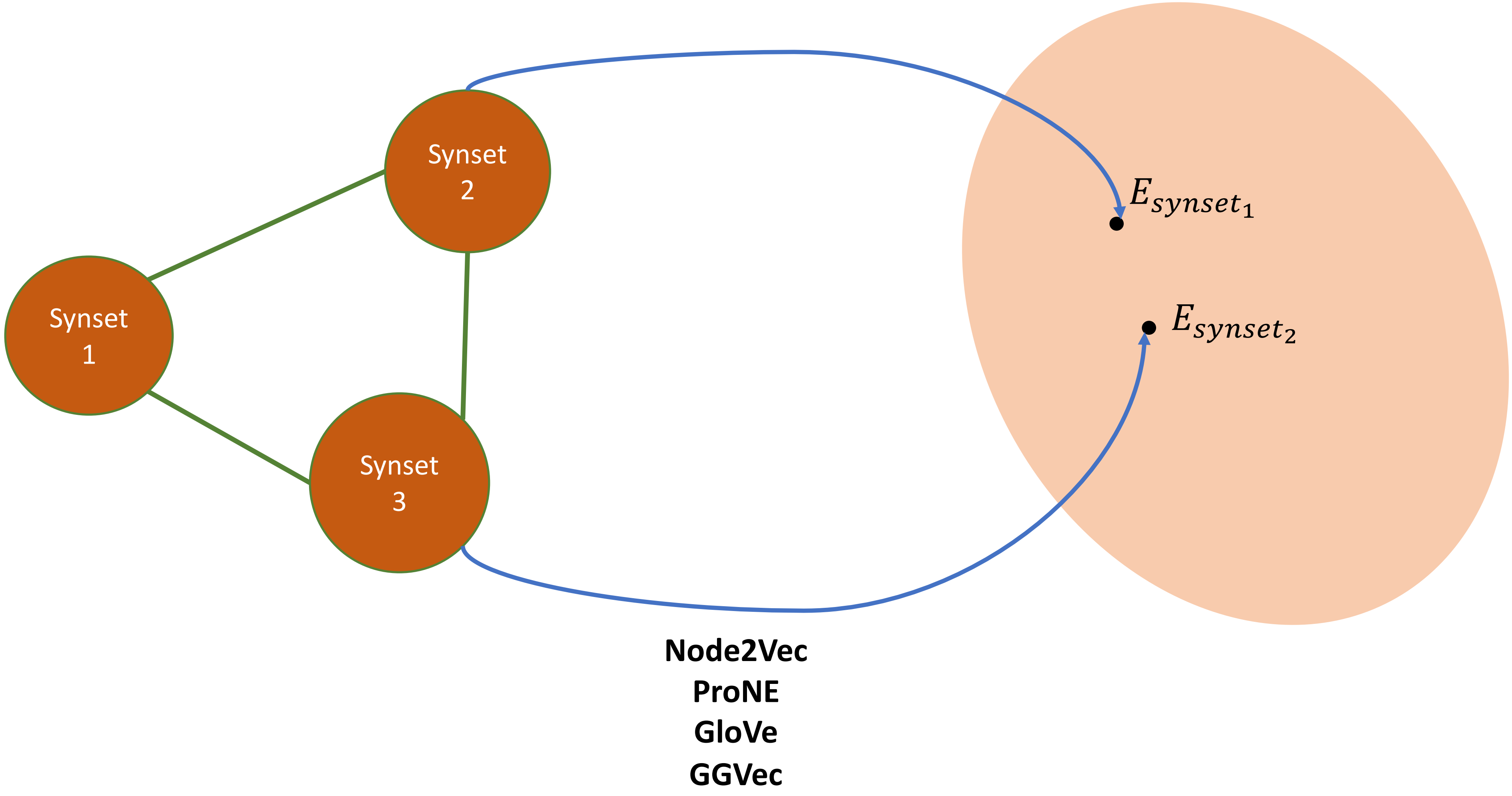




Word	Language
x_1	l_1
x_2	l_2
x_3	l_3

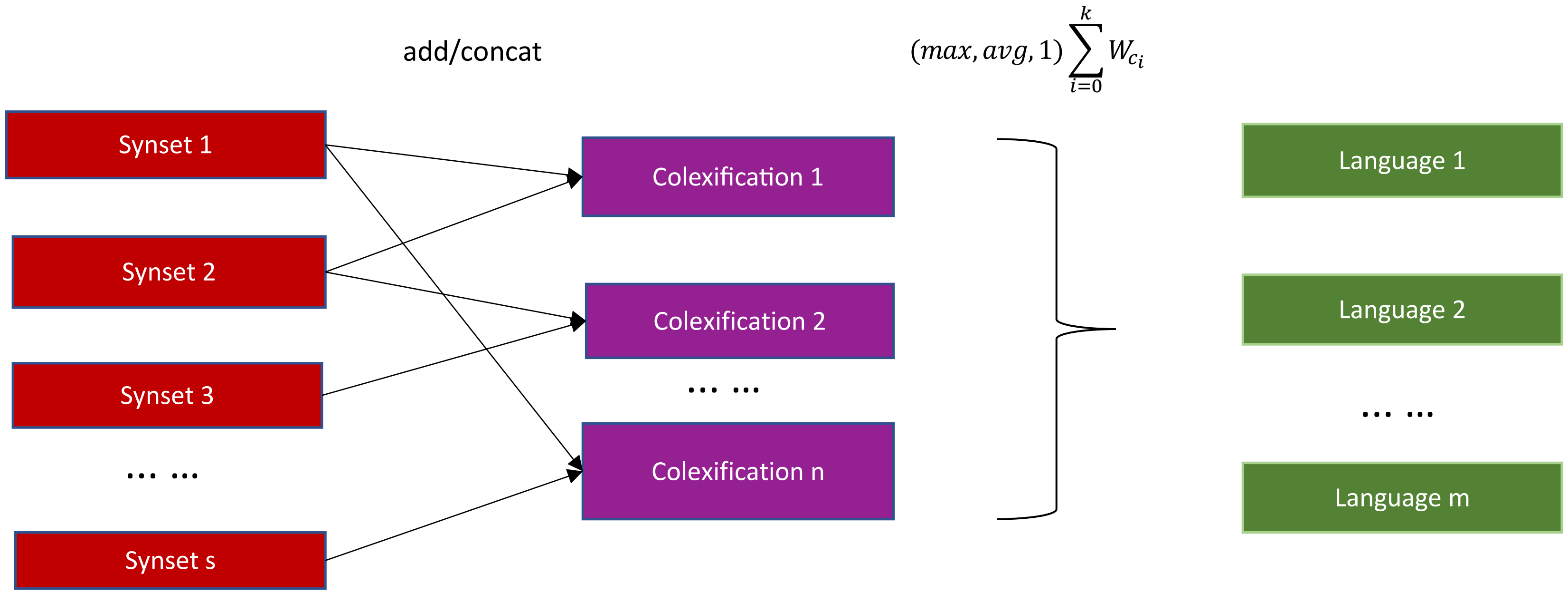


Colex2Lang



Synset 2
Synset 1
Synset 3
E_{synset_1}
E_{synset_2}
Node2Vec
ProNE
GloVe
GGVec

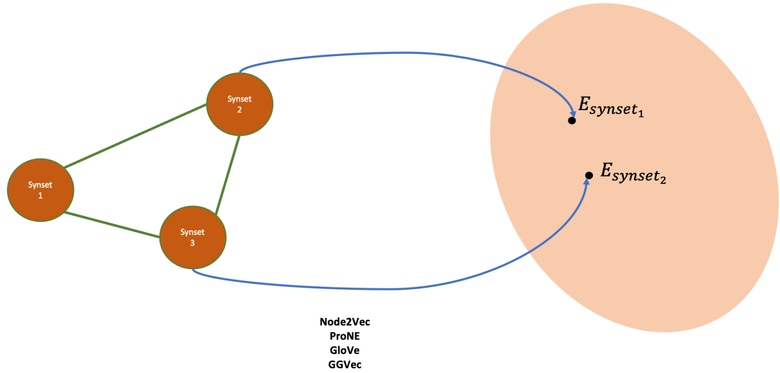
Colex2Lang



Node2Vec
ProNE
GloVe
GGVec

Statistics of Colexification Datasets/Networks

Dataset	#(C,X,L)	Colexifications(C)	Lexicalizations (X)	Synsets/Concepts	#Language (L) (Pair)
WordNet	6,199,897	2,525,591	974,346	105,827	519 (134421)
WordNet Concept	6,075,413	2,486,485	920,031	99,817	519 (134421)
CLICS	68,560	4,228	53,259	1,647	1609 (332783)



Typological Features – Word Order

Feature 81A: Order of Subject, Object and Verb (WALS)

Values

●	SOV	564
●	SVO	488
○	VSO	95
◇	VOS	25
◆	OVS	11
◆	OSV	4
○	No dominant order	189

- SOV (Japanese)

  Watashitachi wa Nihongo o hanasu.
  we TOP Japanse OBJ speak.
  “We speak Japanese.”

- SVO (English)

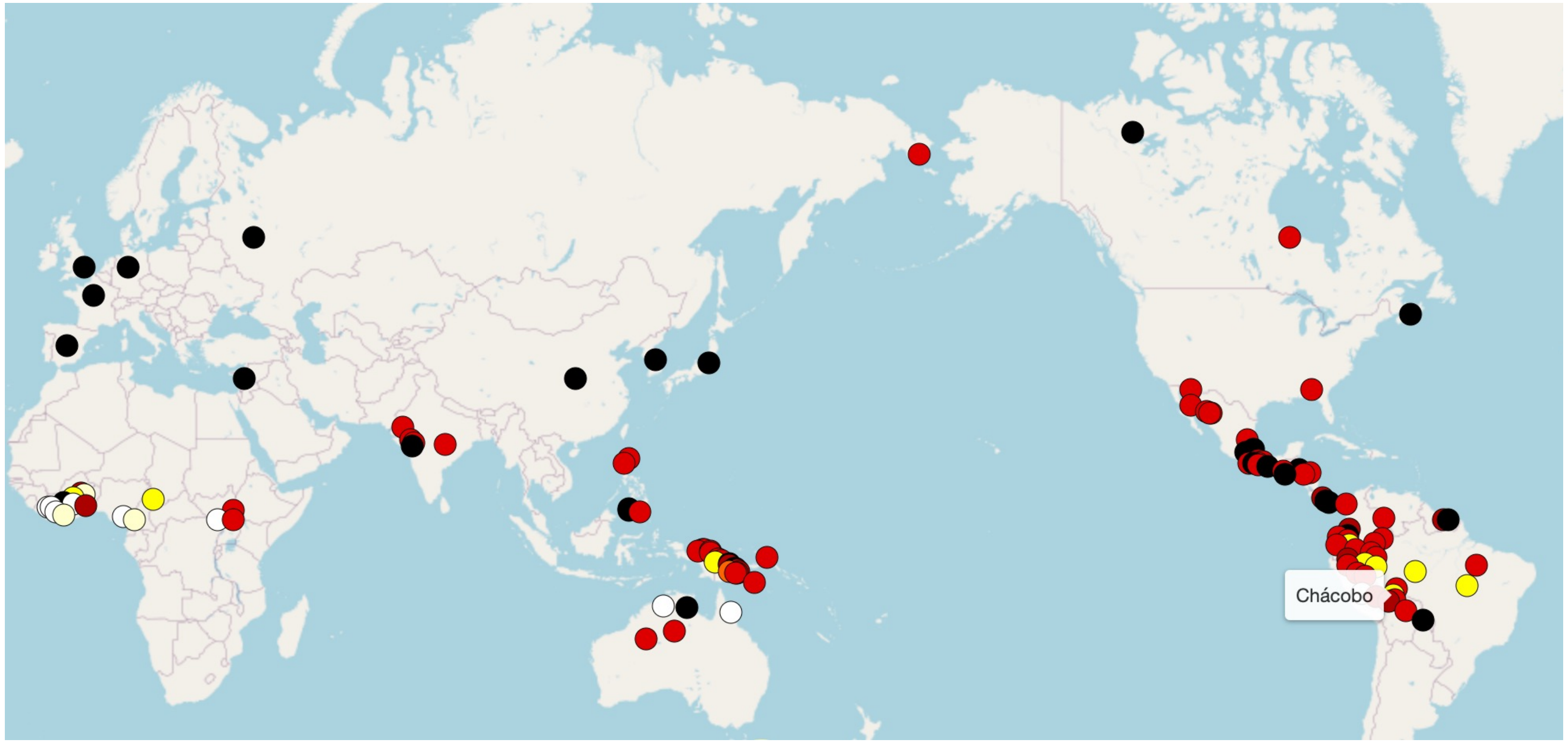
  He ate the pudding.

- VSO (Arabic)

  Qatala l- malik-u l- malikat-a
  kill DEF king NOM+DEF DEF queen ACC
  “The king killed the queen.”

Typological Feature - Lexicon

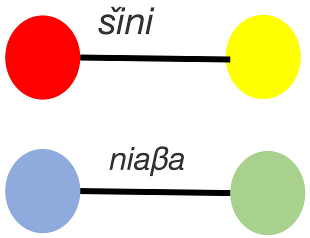
Feature 132A: Number of Non-Derived Basic Colour Categories



Value	Representation
3 categories	10
Between 3 and 4 categories	3
4 categories	9
Between 4 and 5 categories	1
5 categories	56
Between 5 and 6 categories	11
6 categories	29
Total:	119

- **Six colour primaries (Berlin&Kay):**
  Black, White, Red, Yellow, Green, Blue

- **Chácobo**
  4 Categories



Typology Feature Prediction Datasets

∩ WALS	#Lang	Lexicon			Simple Clauses			10 Feature Areas		
		#F	#V	#D	#F	#V	#D	#F	#V	#D
CLICS	737	13	4	93	26	4	142	188	9	288
WordNet (Concept)	330	13	2	58	26	4	89	185	8	166
Malaviya et al. (2017)	624	13	4	92	26	4	117	190	9	238
Östling & Tiedemann (2017)	597	13	4	85	26	4	109	190	9	219

Under each feature area and in all ten feature areas, #F represents the total number of features, #V represents the average number of feature values, #D represents the average number of data samples.

Experimental Setup

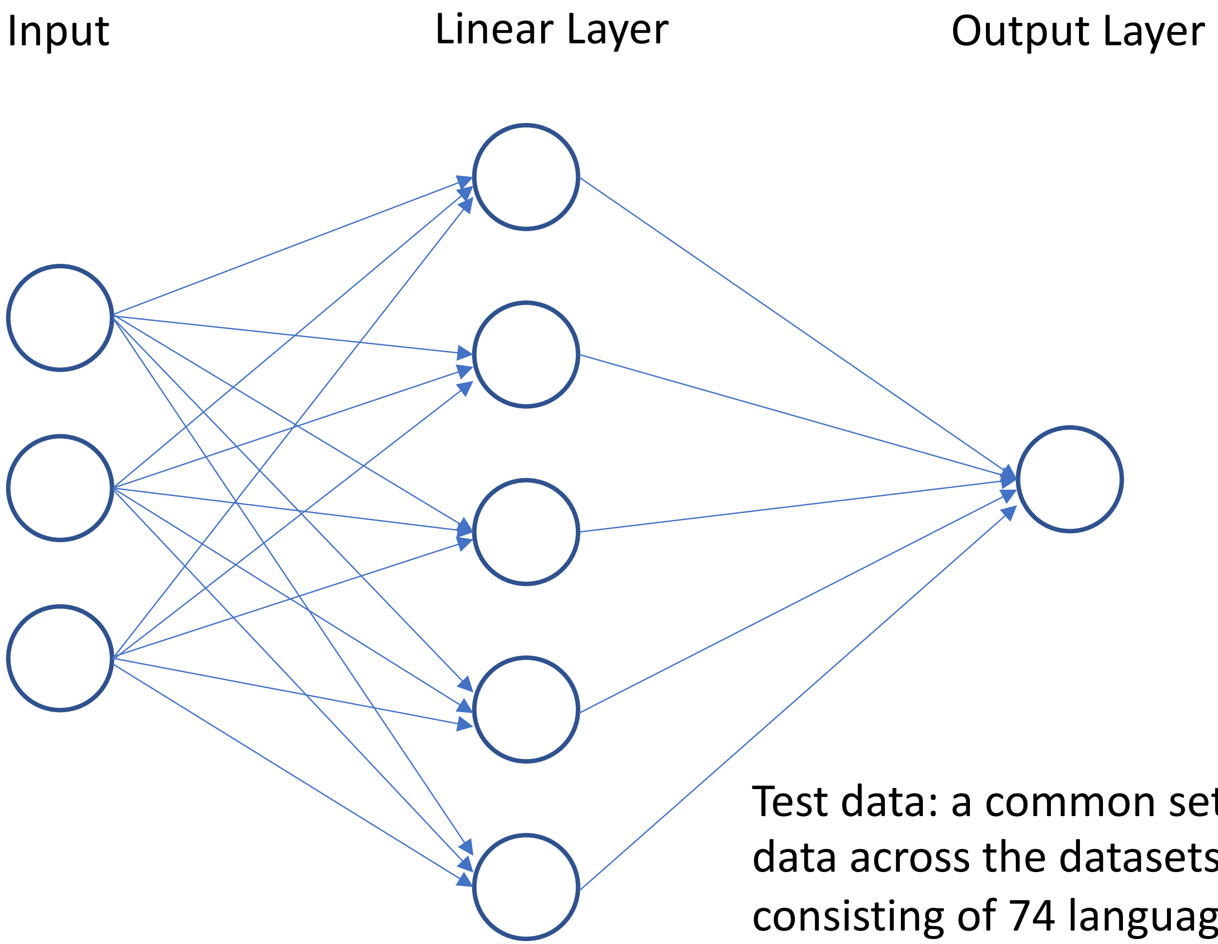
Randomly initialized

CLICS Language Embeddings

WN (Concept) Language Embeddings

Malaviya et al. (2017)

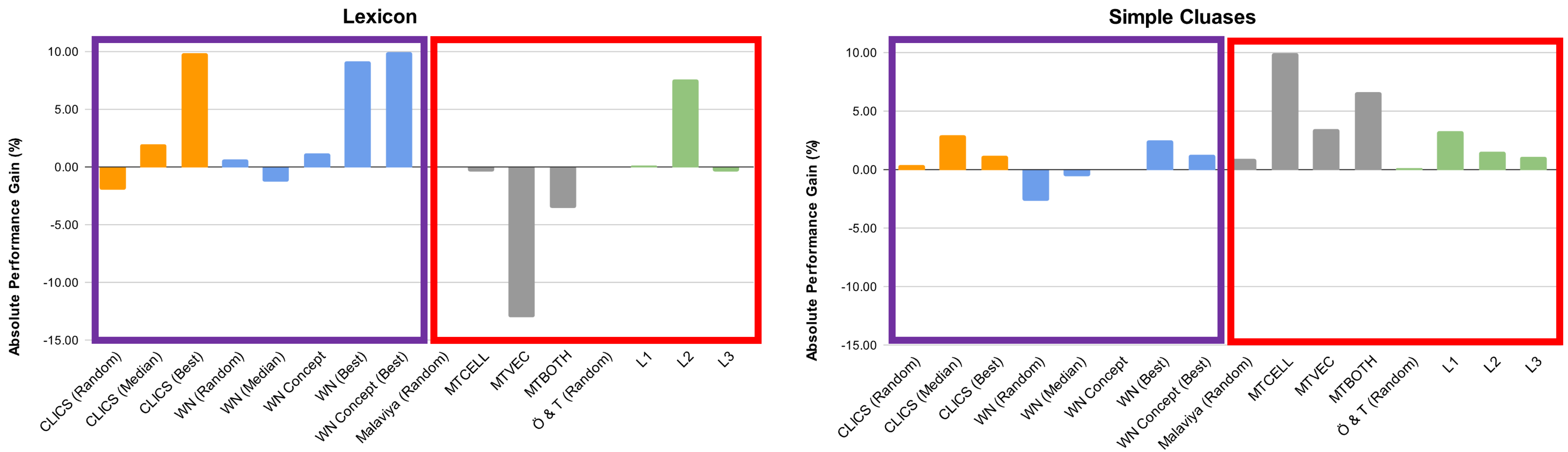
Östling & Tiedemann (2017)



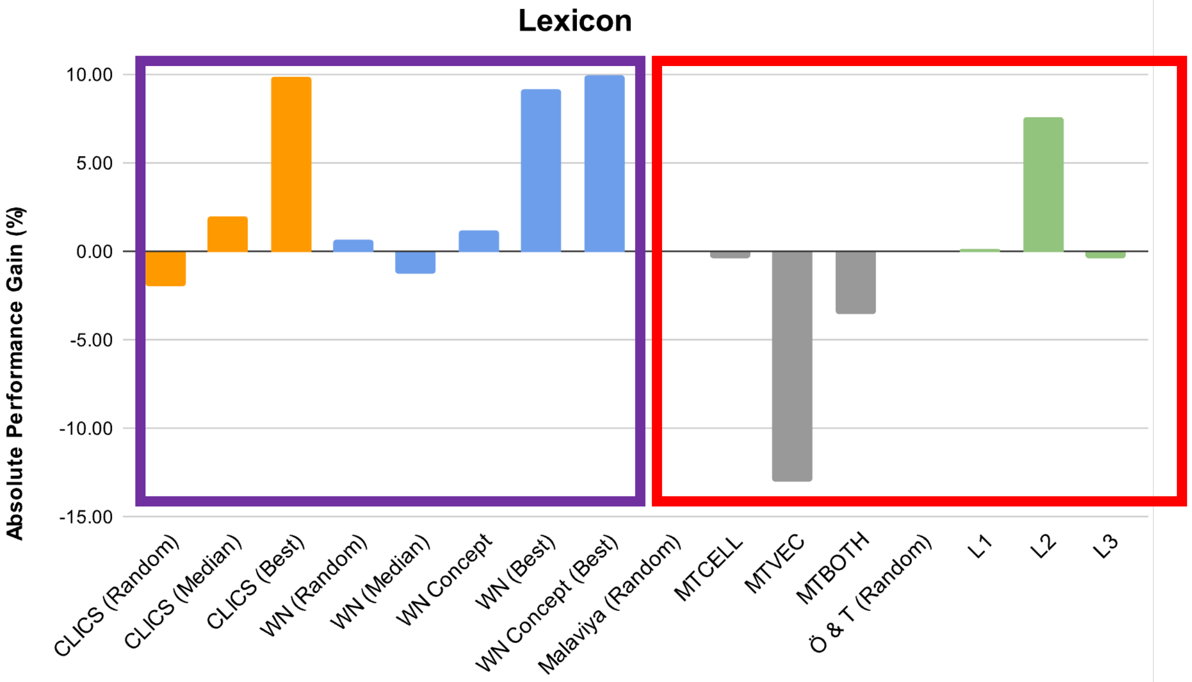
Test data: a common set of data across the datasets consisting of 74 languages.

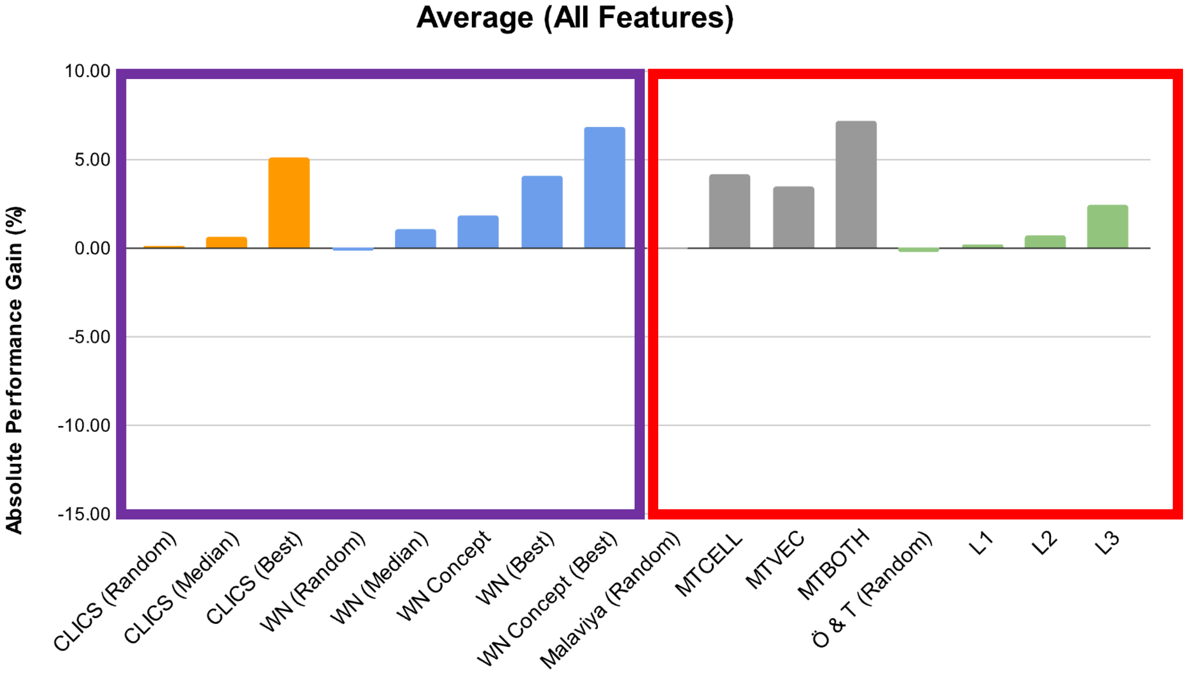
Baseline model: majority vote

Results – Typology Feature Prediction



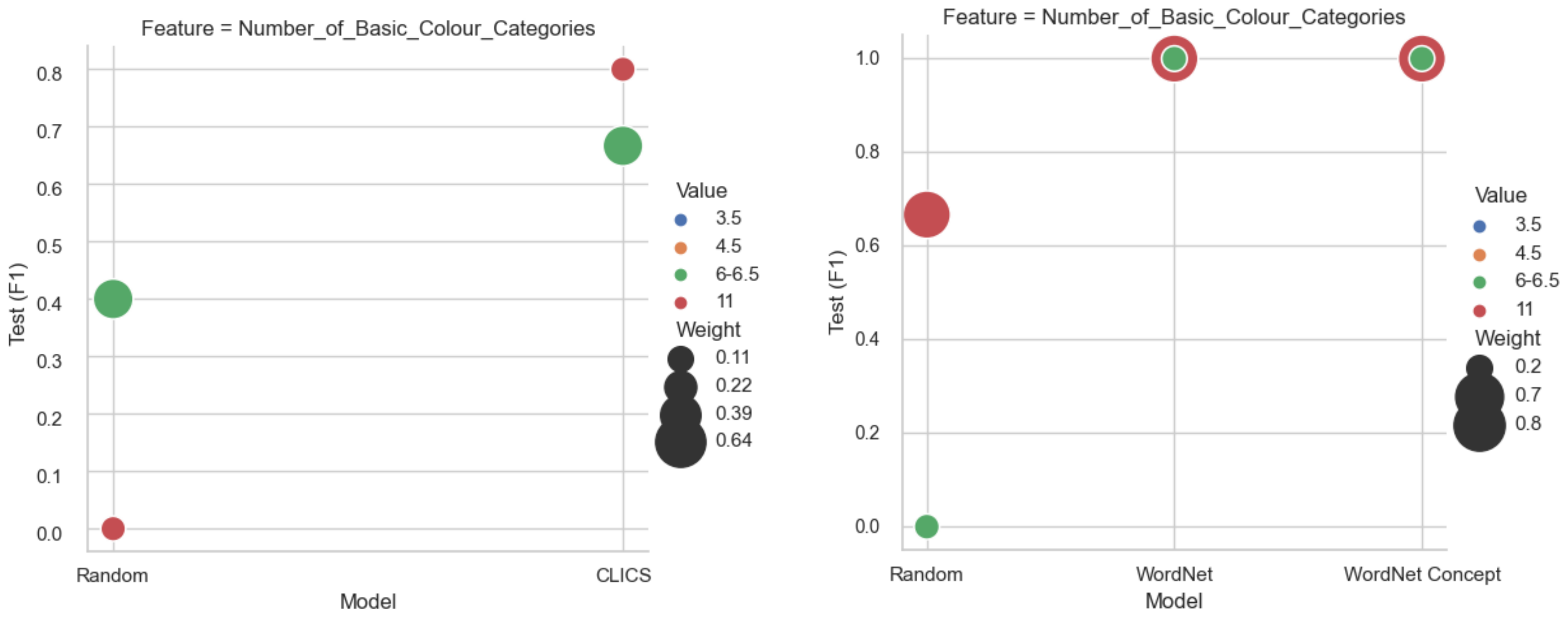
Results – Typology Feature Prediction





Colexification-informed language embeddings capture a distinct signal, especially in lexico-semantic typological features, compared to more general language embeddings.

Capturing Lexicon Typological Features



Performance of Predicing Lexicon Typological Features. The test results are in macro F1- scores, the colour of the circle represents the feature values, and the size of the circles indicates the size of the data samples for the regarding values in the train data.

Language Similarities

Language Embeddings	#Language (Pair)	Correlation Coefficient (P-Value)	#Language (Pair)*	Correlation Coefficient (P-Value)
CLICS	343 (58653)	- 0.049 (4.436e-33*)	8 (28)	- 0.0876 (0.6575)
WordNet	216 (23220)	**0.1469 (3.525e-112*)**	8 (28)	0.7679 (1.838e-06*)
WordNet Concept	216 (23220)	0.1274 (1.339e-84*)	8 (28)	**0.8515 (9.210e-09*)**

Correlation between Language Similarities represented by Lexicon Typological Features and Colexification-informed Language Embeddings. 8 Languages* : Danish, Estonian, Finnish, Greenlandic, Icelandic, Latvian, Lithuanian, and Swedish.

➡ Language embeddings learned on large-scale WordNet datasets present stronger semantic typological signals than the ones trained on CLICS.

Conclusion and Future Work

- Explored the potential of using semantic typology in NLP, specifically colexifications.
- Demonstrated:
  - Colexification-informed language embeddings capture a distinctive aspect of languages;
  - Data scopes of the curated synsets affects the performance.
- The framework provides a new benchmark for further research in this direction.

- Apply colexifications further in multilingual NLP, assisting low-resource languages, e.g., in cross-cultural transfer learning.

Thanks for your attention!