Variability Across Languages in Zero-Shot Multilingual Learning

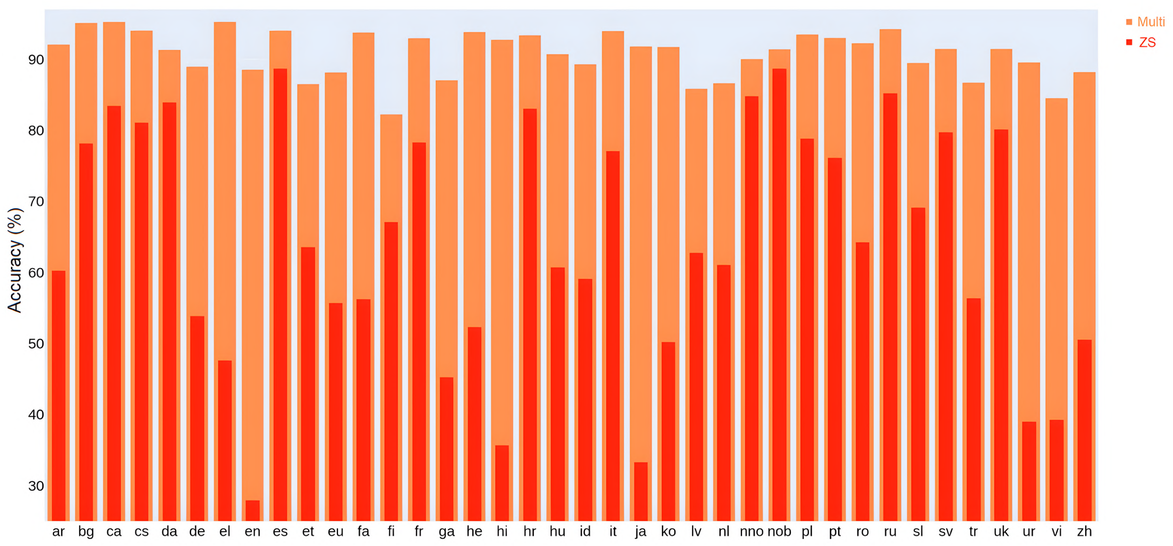
Manon Scholivet – first.last@lis-lab.fr

Aix Marseille Univ, CNRS, LIS, Marseille, France

Contributions

- The **variability** in zero-shot predictions is mainly explained by the **presence of a close language** in the training corpus, which has a major impact on results for zero-shot-type experiments.
- The more **isolated** a language is, according to the *World Atlas of Language Structures* (WALS) but especially in the empirical sense, the lower the zero-shot results will be.

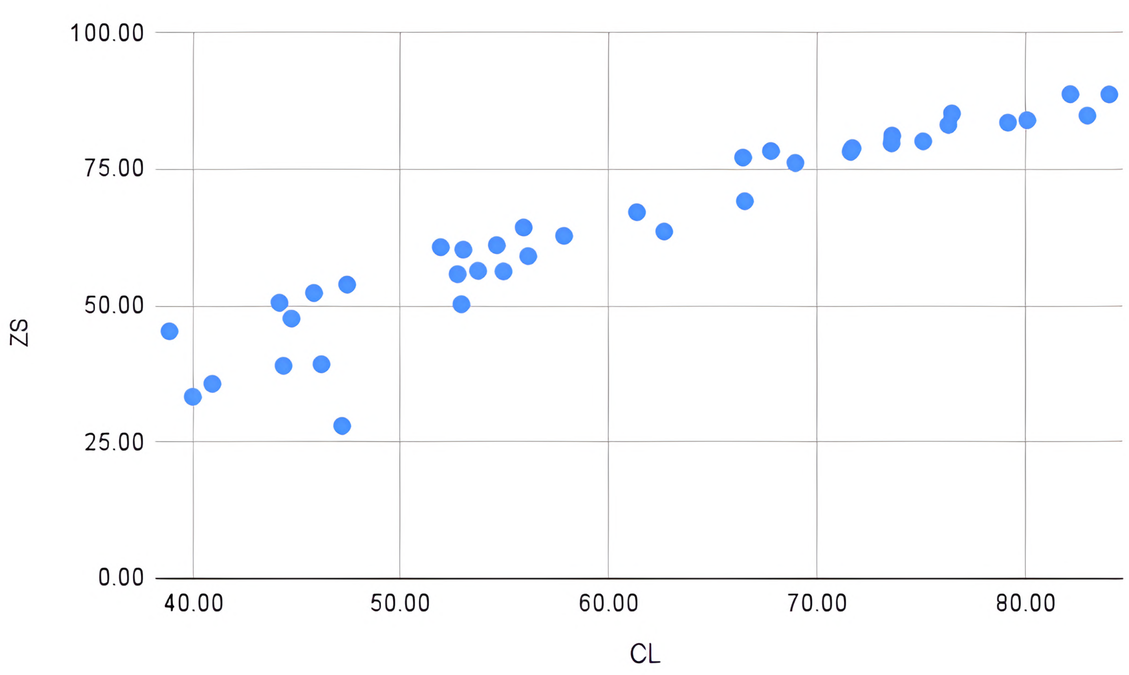
The zero-shot variability : a problem ?



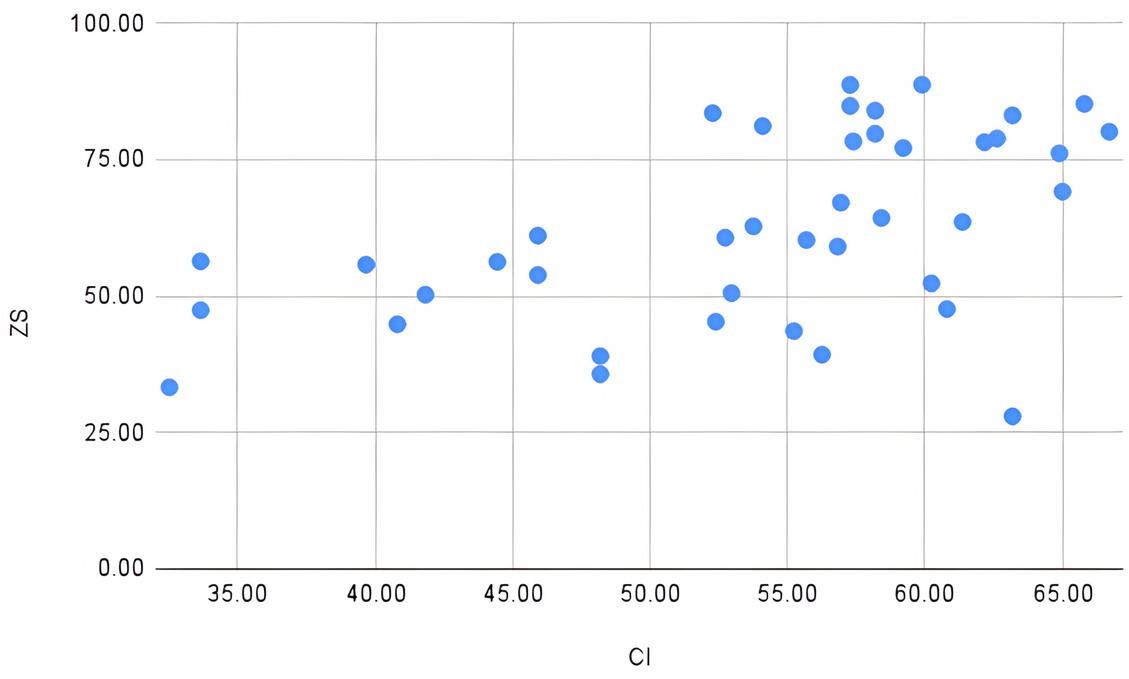
The results **vary significantly** from one language to another in a zero-shot setting, with **17.06** of std. In the monolingual *Mono* experiments : **2.72** of std, and **3.23** in the multilingual *Multi* experiments.

Correlation with the Zero-shot results

Closest Language (CL)



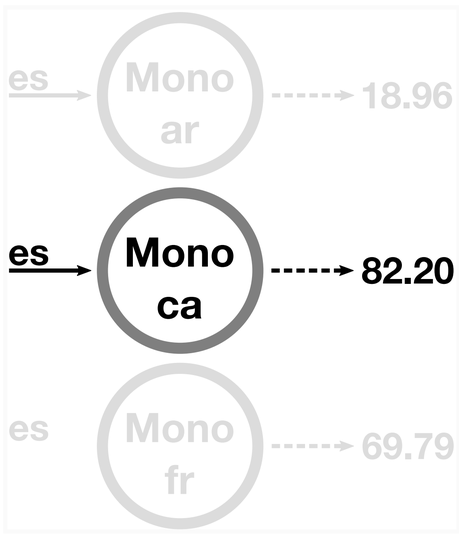
Connectedness Index (CI)



Measuring Language Isolation

The higher the values, the less isolated is the language!

The **Closest Language (CL)** : the empirical measure



Lang.	CL	CL Score	Lang.	CL	CL Score
ar	fa	53.00	it	es	66.46
bg	ru	71.63	ja	pl	39.99
bxr	cs	45.98	kmr	tr	38.12
ca	es	79.20	ko	cs	52.91
cs	sl	73.63	lv	sl	57.85
da	nob	80.12	nl	nob	54.61
de	nl	47.42	nno	nob	83.02
el	cs	44.74	nob	nno	84.07
en	fr	47.18	pl	cs	71.72
es	ca	82.20	pt	es	68.98
et	fi	62.67	ro	fr	55.91
eu	fi	52.73	ru	bg	76.50
fa	sl	54.94	sl	hr	66.54
fi	et	61.36	sme	fi	39.15
fr	ca	67.80	sv	da	73.60
ga	pl	38.87	tr	eu	53.71
he	cs	45.82	uk	ru	75.12
hi	sl	40.94	ur	sl	44.36
hr	sl	76.33	vi	ko	46.18
hu	pt	51.92	zh	ja	44.16
id	fi	56.12			

The **Connectedness Index (CI)** : the measure based on the WALS vectors. 22 features describing each language

$$CI(L) = \frac{100}{k} \sum_{f=1}^{k} \frac{1}{N-1} \sum_{L' \neq L} \delta(W(L', f), W(L, f))$$

Lang.	CI	Lang.	CI	Lang.	CI
ar	57.00	fr	58.85	nob	59.09
bg	63.64	ga	53.69	pl	64.25
ca	53.93	he	61.67	pt	66.34
cs	54.55	hi	48.03	ro	59.95
da	59.83	hr	64.37	ru	67.32
de	47.17	hu	53.19	sl	66.34
el	62.16	id	68.30	sv	59.83
en	65.11	it	48.03	tr	31.57
es	61.79	ja	30.71	uk	68.30
et	62.04	ko	41.65	ur	48.03
eu	39.07	lv	54.05	vi	57.49
fa	43.24	nl	47.17	zh	53.69
fi	57.13	nno	59.09		

Future Work

- Additional experiments using **WALS as input** could help in knowledge sharing between languages.
- Using languages from the **same language family**.
- Check the variability in the ratio of **unknown words** for each language.

Funded by the European Union