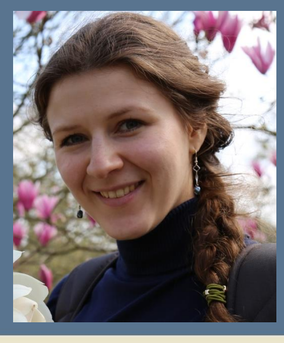
Contact: Anna Samoilenko
anna.samoilenko@gesis.org
@AnnySml



gesis
Leibniz Institute for the Social Sciences

Linguistic neighbourhoods:

Explaining cultural borders on Wikipedia through multilingual co-editing activity

Research question:

- What socio-linguistic features explain common editing interests between language communities on Wikipedia?

Data:

- 4,5M edits in 110 languages to 2,591,644 Wikipedia articles created between 2005-2013
- 200,748 unique concepts, random stratified 1%-sample, bots excluded

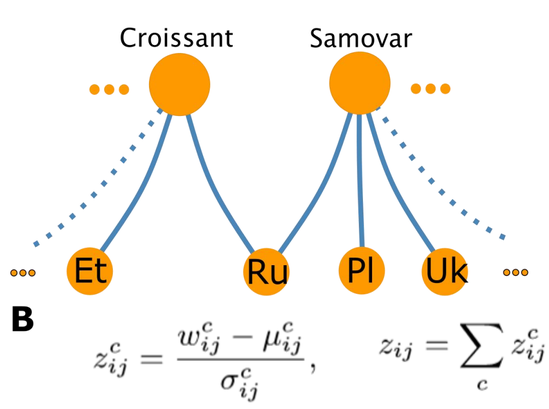
Main findings:

- Co-editing similarity of language communities on Wikipedia is best explained by bilingualism, shared religion, and demographic attraction of communities.
- Geographical distance is a weak, although significant factor.
- No global dominance of English; instead local interconnections come to the forefront.

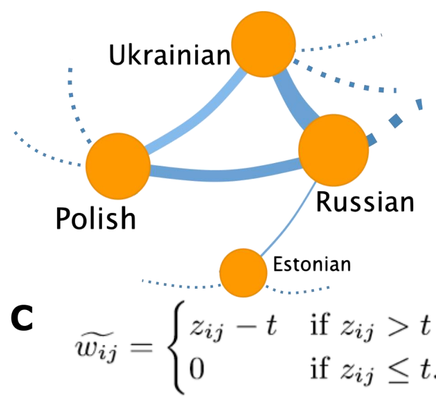
Contributions:

- Explored and quantified a socio-linguistic definition of cultural similarity
- Presented a large-scale network of similarities between 110 language-speaking communities
- Presented a generalisable approach to quantifying and explaining similarity, which
  - scales well in terms of number of hypotheses and communities that could be analysed;
  - does not require understanding of a language;
  - is applicable for any example of collaborative production of common good where individual activity is recorded.

a. Set of languages in which a concept exists

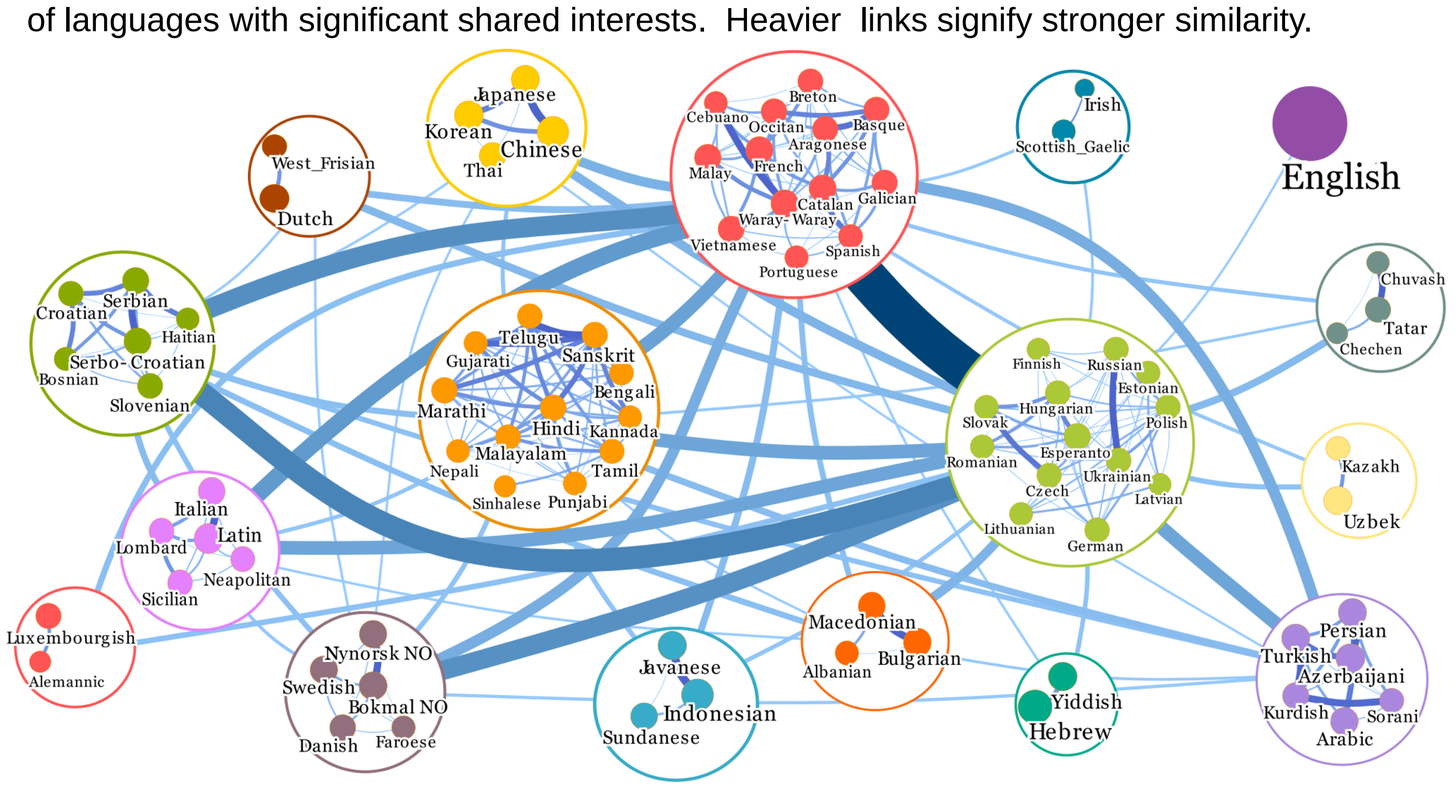


b. Bipartite network of concept co-editing

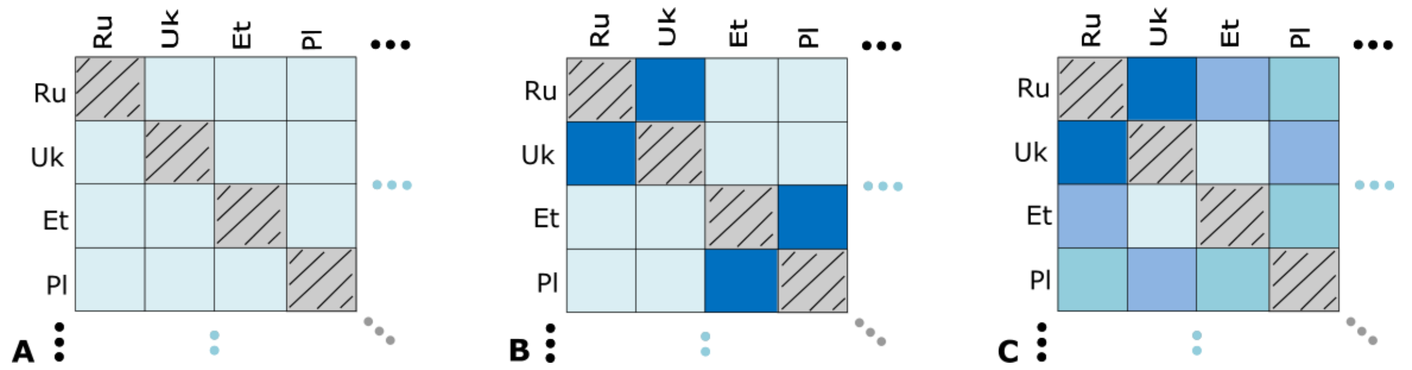


c. Network of significant links

Illustration of the *z*-score-based filtering method. The method requires 3 steps: (a) to retrieve all edits to each concept in all linked language editions; (b) to compare the empirical and expected probabilities of each language pair to co-edit a concept; (c) to create a filtered network of languages with significant shared interests. Heavier links signify stronger similarity.

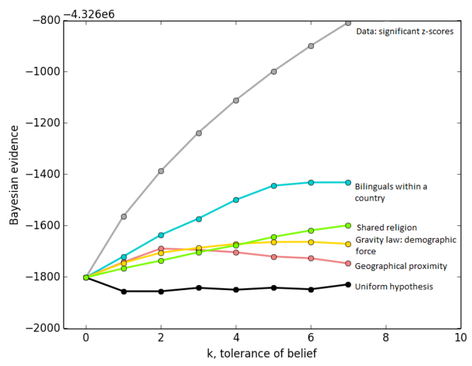


Network of significant co-editing ties between language pairs on Wikipedia. Nodes are coloured according to the clusters found by the Infomap algorithm, link weights within clusters represents significant z-scores; links are significant at the 99% level. The inter-cluster links show the aggregated z-scores between all nodes of a pair of clusters. For visualisation purposes we display only 16 clusters and the 113 strongest inter-cluster links in the network.



Uniform hypothesis | **Shared religion hypothesis** | **Geographical proximity hypothesis**

Expressing hypotheses through a transition probability matrix. The diagonal is empty since the data do not allow self-loops. Cells with more likely transitions are coloured in darker shades of blue. (a) all transitions are equally possible; (b) based on the dominant religion of the language-speaking community; (c) the shorter the distance between languages, the stronger belief in the transition



A. HypTrails

	Model	Bilinguals	Religion	Gravity	Distance[1]	R^2 adj.	F-stat.	df	Intercept
1	Estimate	0.0333	0.0687	0.0404	−0.0833	**0.1974**	389.8	6317	0.0086
	t-statistic	**21.1397**	**19.6471**	**17.9675**	**−8.6126**				
2	Estimate	–	0.0715	0.0478	−0.0749	0.1407	351.8	6427	0.0088
	t-statistic	–	19.9069	20.9998	−7.5951				
3	Estimate	0.0326	0.0656	0.0411	–	0.1881	489.3	6318	0.0079
	t-statistic	20.6007	18.7577	18.1914	–				
4	Estimate	0.0378	0.0796	–	−0.0897	0.1566	392.1	6318	0.0089
	t-statistic	23.7089	22.5576	–	−9.0517				
5	Estimate	0.0341	–	0.0474	−0.0592	0.1471	377.0	6536	0.0090
	t-statistic	21.4719	–	21.2093	−6.0945				
6	Estimate	0.0409	0.0771	–	–	0.1134	691.1	10788	0.0081
	t-statistic	30.4130	20.3477	–	–				
7	Estimate	0.0418	–	–	–	0.07986	955.3	11007	0.0088
	t-statistic	30.9071	–	–	–				
8	Estimate	–	0.0807	–	–	0.03559	434.3	11770	0.0084
	t-statistic	–	20.8410	–	–				

[1] primary language

B. MRQAP (Multiple Regression Quadratic Assignment procedure)

Frequentist and Bayesian approaches to hypotheses testing. Both methods suggest that bilingualism and shared religion are most plausible explanatory factors, followed by population attraction and geographical distance. (a) Ranking of hypotheses should be compared for the same k. (b) Model 1 explains 20% variation in the data, significant at 0.05 level.

References:

A. Samoilenko, F. Karimi, D. Edler, J. Kunegis, M. Strohmaier. Linguistic neighbourhoods: Explaining cultural borders on Wikipedia through multilingual co-editing activity (submitted)