Summary of findings

Main analyses of Wikipedia data.

See the main text for a summary of the main analyses using the Wikipedia data.

Analyses of the estimates from the Common Crawl data

Mixed effects model: The correlation between semantic alignment and cultural similarity was significant (β= 0.182, $\chi^2(1)$= 4.94, p=0.026). See the figure 1 below:

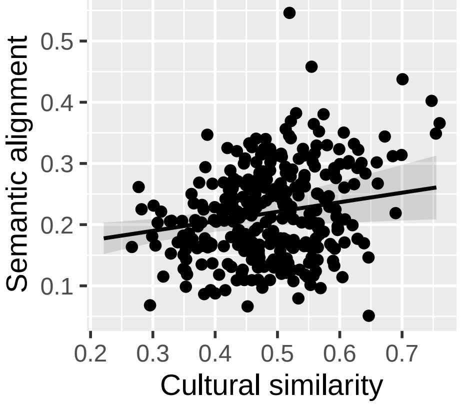

Figure 1: Semantic alignment and cultural similarity for data using the Common Crawl alignments

MRM results:

	Estimate	p-value
Intercept	0.306	0.031
Cultural distance	0.272	0.0373
Language family	-0.273	0.0981
Geographic distance	0.148	0.0976
Comparison count	0.110	0.138

Table 1: MRM analysis predicting semantic alignment (Common Crawl), with family control. R^2=0.0999

	Estimate	p-value
Intercept	0.190	0.0001
Cultural distance	0.285	0.0405
ASJP	-0.464	0.0001
Geographic distance	0.229	0.0103
Comparison count	0.087	0.244

Table 2: MRM analysis predicting semantic alignment (Common Crawl), with ASJP control. R^2=0.185

Mantel tests:

	Var1	Var2	r	llim	ulim	p
2	Cultural	Linguistic	0.18	0.0131	0.286	0.12
3	Cultural	Historical	-0.315	-0.447	-0.204	0.0211
4	Cultural	Geographic	-0.461	-0.586	-0.312	0.00295
5	Linguistic	Historical	-0.404	-0.523	-0.194	0.00104
6	Linguistic	Geographic	-0.0837	-0.202	-0.00552	0.262
7	Historical	Geographic	0.405	0.31	0.519	0.00101
71	Linguistic	Cultural **	0.106	0.000414	0.191	0.226

Table 3: Mantel tests (Common Crawl). ** = partial Mantel test, controlling for historical and geographical distance.

Analyses of the estimates from the Subtitles data

Mixed effects model: The correlation between semantic alignment and cultural similarity was not significant (β= 0.0606, $\chi^2(1)$= 0.74, p=0.39). See the figure 2 below:

MRM results:

	Estimate	p-value
Intercept	0.073	0.517
Cultural distance	0.125	0.413
Language family	-0.041	0.838
Geographic distance	-0.013	0.887
Comparison count	0.806	0.0002

Table 4: MRM analysis predicting semantic alignment (Subtitles), with family control. R^2=0.744

	Estimate	p-value
Intercept	0.083	0.297
Cultural distance	-0.023	0.884
ASJP	-0.282	0.0083
Geographic distance	0.019	0.854
Comparison count	0.831	0.0003

Table 5: MRM analysis predicting semantic alignment (Subtitles), with ASJP control. R^2=0.803

Mantel tests:

	Var1	Var2	r	llim	ulim	p
2	Cultural	Linguistic	0.351	0.241	0.538	0.0884
3	Cultural	Historical	-0.17	-0.286	-0.0153	0.155
4	Cultural	Geographic	-0.34	-0.57	-0.172	0.0276
5	Linguistic	Historical	-0.352	-0.503	-0.0822	0.0301
6	Linguistic	Geographic	-0.273	-0.462	-0.0818	0.0806
7	Historical	Geographic	0.346	0.181	0.522	0.01
71	Linguistic	Cultural **	0.281	0.129	0.466	0.135

Table 6: Mantel tests (Subtitles). ** = partial Mantel test, controlling for historical and geographical distance.

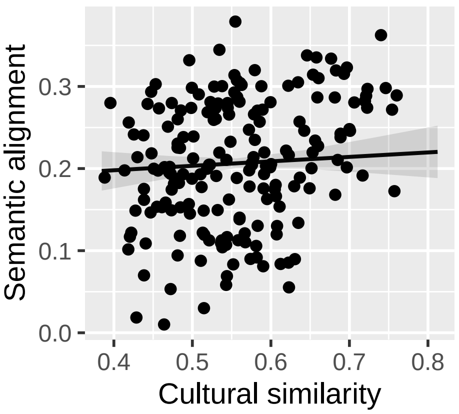

Figure 2: Semantic alignment and cultural similarity for data using the Subtitles alignments

Analysis of numerals

The analyses of numerals found:

- 1 and 2 have lower alignment due to often being grammaticalised as indefinite or dual marker (Givon, 1981).
- Numbers 3-12 generally have high alignment (mean local alignment = 0.87), and higher numbers decline in alignment up to 1000.
- There are also language-specific differences due to how numerals are constructed (e.g. base, combination rules, see Calude & Verkerk, 2016), or for irregular forms (e.g. 50, 60, 70, 80 and 90 in Danish).
- Some number words have alternative associations due to homophones (e.g. the Hungarian 7 is used directly to mean 'week', and 'neuf' in French means '9' or 'new').
- The historical distance between languages did not predict much of the variation.

See the main text for a discussion.