General patterns and language variation: Word frequencies across English, German, and Chinese

Annika Tjuka, tjuka@shh.mpg.de

Max Planck Institute for the Science of Human History

Cognitive Aspects of the Lexicon (CogALex-VI)
December 12th, 2020

Aim

A cross-linguistic comparison of word frequencies.

For example, what are the frequencies of the first-person pronoun across

- English: *I*
- German: *ich*
- Chinese: *wǒ* 我

SUBTLEX

SUBTLEX data for

- English: Brysbaert and New (2009)
- German: Brysbaert et al. (2011)
- Chinese: Cai and Brysbaert (2010)

`http://crr.ugent.be/programs-data/subtitle-frequencies`

Concepticon

The Concepticon project links concept sets consisting of a standardized identifier, a concept label, and a description, to elicitation glosses used in concept lists for research in linguistics (List et al., 2016; List et al., 2020).

In its current version (2.4.0.), the Concepticon offers 3,743 concept sets based on 353 concept lists.

The concept lists exist for a variety of glossing languages and the Concepticon currently supports mappings for common languages such as English, Spanish, Russian, German, French, Portuguese, and Chinese.

`https://concepticon.clld.org/`

NoRaRe

Cross-Linguistic **No**rms, **Ra**tings, and **Re**lations for Words and Concepts (Tjuka, Forkel and List 2020b)

- **Norms**: e.g., word frequency, reaction time
- **Ratings**: e.g., age-of-acquisition, discrete emotions, sensory modality
- **Relation**: e.g., semantic field, polysemy

`https://digling.org/norare/`

Hypotheses

1. Related languages (i.e., belonging to the same language family) have more similar frequencies across a set of shared concepts than non-related languages (i.e., belonging to different language families).
2. In related languages, there are fewer concepts that have a large difference between frequencies than in non-related languages.

Correlations of Frequencies

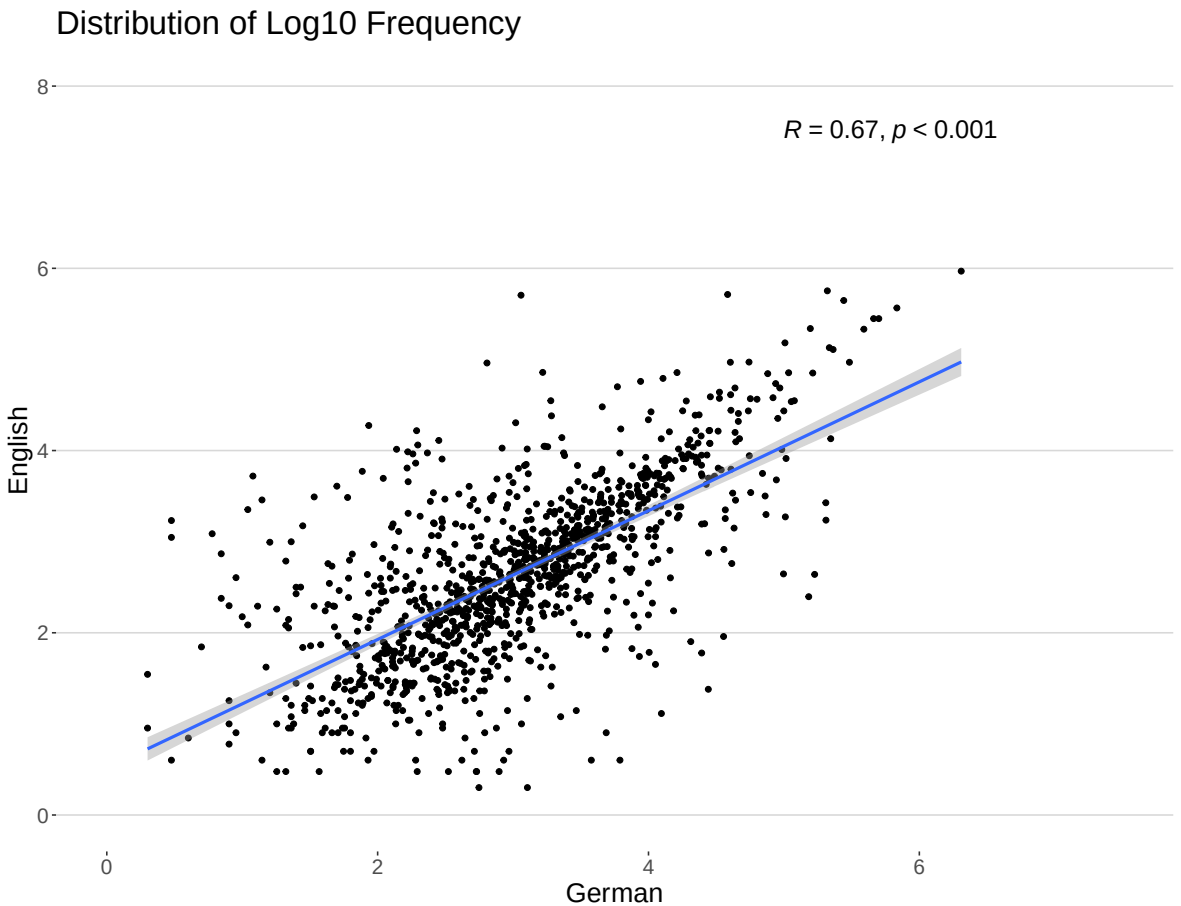

Figure 1: Distribution of the $\log_{10}$ word frequencies across the language pair English–German.

Correlations of Frequencies

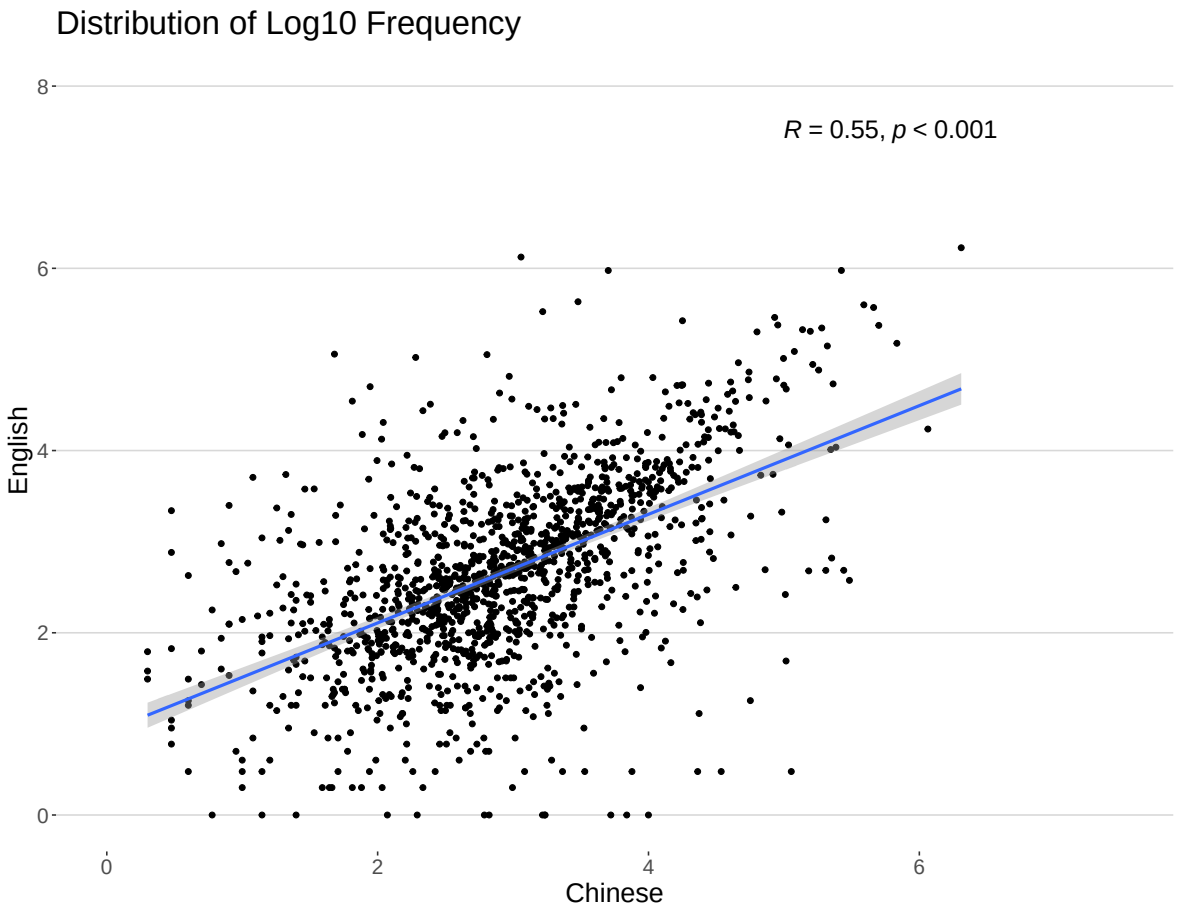

Figure 2: Distribution of the $\log_{10}$ word frequencies across the language pair English–Chinese.

Correlations of Frequencies

	Overlap concept sets	**Pearson coefficient**	**p-value**
English–German	1,149	0.67	$p < .001$
English–Chinese	1,313	0.55	$p < .001$

Language Variation

Table 1: Comparison of the differences in the $\log_{10}$ frequencies across English and German. The list includes the concept sets which vary greatly in their frequencies (difference greater than 3).

ID	**Label**	**English $\log_{10}$**	**German $\log_{10}$**	**Difference**
1301	FOOT	3.79	0.60	3.19
492	THREE	4.44	1.38	3.06

Language Variation

Table 2: Comparison of the differences in the $\log_{10}$ frequencies across English and Chinese. The list includes the concept sets which vary greatly in their frequencies (difference greater than 3). Subset of the original list with 16 concept sets.

ID	**Label**	**English $\log_{10}$**	**Chinese $\log_{10}$**	**Difference**
1235	WHO	5.05	0.48	4.58
2483	COLD (OF WEATHER)	4.00	0.00	4.00
930	VILLAGE	3.23	0.00	3.23

Conclusion

- The correlation of the frequencies between the language pair English–German was slightly higher than between English–Chinese. (supports hypothesis 1)
- The findings of the study indicate that frequencies of the same concepts can differ greatly across languages. (supports hypothesis 2)
- The detailed examination of the individual concepts showed that several processes may lead to the differences in frequencies:
  1. capitalization of first letter (*drei* vs. *Drei*)
  2. space between compounds (*metrical foot* vs. *Versfuß*)
  3. cultural diversity (COLD (OF WEATHER))
  4. use of two word-forms (*cūnzhài* 'village' vs. *cūnzi* 'village')

Contact info

tjuka@shh.mpg.de

www.annikatjuka.com

References

- Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4):977–990.
- Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. Experimental Psychology, 58:412–424.
- Qing Cai and Marc Brysbaert. 2010. SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. Plos ONE, 5(6):1–8.
- Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon. A resource for the linking of concept lists. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation, pages 2393–2400, Luxem- bourg. European Language Resources Association (ELRA).
- Johann-Mattis List, Christoph Rzymski, Simon J. Greenhill, Nathanael Schweikhard, Kristina Pianykh, Annika Tjuka, Mei-Shin Wu, and Robert Forkel. 2020. Concepticon. A resource for the linking of concept lists (Version 2.4.0-rc.1). Max Planck Institute for the Science of Human History, Jena.
- Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2020a. Database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (Version 0.1). Max Planck Institute for the Science of Human History, Jena.
- Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2020b. Linking norms, ratings, and relations of words and concepts across multiple language varieties. PsyArXiv:10.31234. version 1.