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

Annika Tjuka
Max Planck Institute for the Science of Human History
Jena, Germany
tjuka@shh.mpg.de

Abstract

Cross-linguistic studies of concepts provide valuable insights for the investigation of the mental lexicon. Recent developments of cross-linguistic databases facilitate an exploration of a diverse set of languages on the basis of comparative concepts. These databases make use of a well-established reference catalog, the Concepticon, which is built from concept lists published in linguistics. A recently released feature of the Concepticon includes data on norms, ratings, and relations for words and concepts. The present study used data on word frequencies to test two hypotheses. First, I examined the assumption that related languages (i.e., English and German) share concepts with more similar frequencies than non-related languages (i.e., English and Chinese). Second, the variation of frequencies across both language pairs was explored to answer the question of whether the related languages share fewer concepts with a large difference between the frequency than the non-related languages. The findings indicate that related languages experience less variation in their frequencies. If there is variation, it seems to be due to cultural and structural differences. The implications of this study are far-reaching in that it exemplifies the use of cross-linguistic data for the study of the mental lexicon.

1 Introduction

The structure and functioning of the mental lexicon have been studied for many decades (Aitchison, 2012). The inner workings of the links and connections of the mental lexicon have been investigated in large scale studies and with non-invasive techniques such as EEG and fMRI. However, many of those studies focus solely on one language. Especially in experimental settings, creating a stimulus set across multiple languages that is controlled for the same variables such as frequency is difficult. But what if we could compare the properties of the same words in different languages to explore the similarities and differences that arise? We would need the word frequencies of translation equivalents for every word, for example, the first-person pronoun in English (*I*), German (*ich*), and Chinese (*wǒ* 我).

Although we have resources available that offer word frequencies for each of the three languages (Brysbaert and New, 2009; Brysbaert et al., 2011; Cai and Brysbaert, 2010), they lack a link between each other to make a comparison of the same word across languages possible. One solution would be to translate the words in the data set to a meta-language (e.g., English) and compare the translation equivalents. However, this comes with a risk of ignoring important information. An alternative is to link the words in the data sets to concepts. The Concepticon project (List et al., 2016) provides a list with 3,755 comparative concepts with links to elicitation glosses for various languages, including English, German, and Chinese. The decision of whether a word is mapped to a specific concept, for instance, the link between the word *tree* and the concept TREE, is based on elicitation glosses that are used in linguistic studies. Those studies often draw upon *Swadesh lists* which assess the genealogical relatedness between languages (Swadesh, 1955). The words in the list represent 'comparative concepts' (Haspelmath, 2010) that relate to basic meanings. The concept lists are linked to the concepts in Concepticon and provide the basis for the connection between a word and a concept. Tjuka et al. (2020b) used the Concepticon concept sets as a basis for the creation of the Database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (NoRaRe). This resource offers links to the Concepticon concept sets and various psycholinguistic values so that they can be easily compared across languages. The database also provides the opportunity to create and reproduce experiments on the basis of well-founded data curated by linguists.

The advantage of a cross-linguistic perspective on the mental lexicon is that we can discover general patterns and language-specific variation. The NoRaRe database promotes comparison of a shared part of the mental lexicon in different languages. The present study uses a set of three languages: English, German, and Chinese. On the one hand, English and German are related languages (both belong to the Germanic sub-branch of the Indo-European language family) and no large differences in the frequencies of the words in both languages are expected. English and Chinese, on the other hand, are genealogically different languages (Chinese belongs to the Sino-Tibetan language family). The assumption is that larger differences in the distribution of word frequencies can be found in the data of non-related languages. The aim of this study is to illustrate a database approach to language comparison on a large number of lexical items between multiple languages. The study sheds light on a cross-linguistic investigation of the mental lexicon.

Another aim of the study is to examine the variation of individual word frequencies in more detail and explore two patterns that could lead to different distributions. The first option is cultural differences in the structure of the mental lexicon. Cross-linguistic studies showed that languages vary in how they structure certain semantic domains such as color (Gibson et al., 2017) or emotion (Jackson et al., 2019). The second pattern that seems to emerge is a correlation between word frequency and the number of meanings (Zipf, 1945). If this is a valid principle, differences in frequencies of the same word in two languages might be due to differences in the number of meanings of the word in the two languages. For example, the word *back* (as a noun) seems to have more meanings in English than *Rücken* "back" in German, based on a search in the Extended Open Multilingual WordNet (Bond and Foster, 2013). A few studies demonstrated that Zipf's meaning-frequency law seems to hold across multiple languages, for instance, English, Turkish, Spanish, Dutch, among others (Ilgen and Karaoglan, 2007; Hernández-Fernández et al., 2016; Casas et al., 2019; Bond et al., 2019). Each of those studies compared the frequencies of words in different corpora with the number of meanings for the words in an individual language, which was taken from the respective WordNet (Fellbaum, 2012). They did not show particular words in each of the languages that gave rise to the correlation. Therefore, the pattern emerged on the basis of a black box. If one would analyze the words in a given data set in more detail, the Zipf's meaning-frequency law might only be true for specific words (e.g., high-frequency words) and could vary across word types (e.g., verbs, adjectives, nouns).[1]

By using the NoRaRe data (Tjuka et al., 2020b), the basis for the word frequencies can be uncovered and a well-established set of concepts in the Concepticon curated by linguists provides a solid basis for cross-linguistic comparison. The NoRaRe database facilitates a quantitative analysis in that frequency values for many concepts across languages can be correlated. Furthermore, the link to the concept sets in Concepticon offers the possibility for a qualitative analysis of outliers that show language-specific variation. The pattern of more frequent words having more meanings can thus be investigated based on individual cases to test its validity as an explanation for language specificities.

In the next section (Sect. 2), the materials and methods used for preparing the data sets of the present study are discussed. Section 3 shows the results of the correlation between the three languages as well as cases of cross-linguistic variation. Finally, in Section 4 and Section 5, the study is summarized and its implications for the investigation of the structure and functioning of the mental lexicon with cross-linguistic data are illustrated.

[1] For a detailed discussion of the limitations of Zipf's meaning-frequency law see Piantadosi (2014).

2 Material and Method

The foundation of this study is the concepts listed in the Concepticon resource (List et al., 2016; List et al., 2020). The Concepticon project[2] 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 such as Swadesh lists (Swadesh, 1955). 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. For example, the glosses of the first-person pronoun in the languages English (*I*), German (*ich*), and Chinese (*wǒ* 我) are linked to the concept set with the ID 1209 and the label "I." The concepts in Concepticon represent comparative concepts (Haspelmath, 2010) that are commonly used to assess the relatedness of languages. The words linked to the concepts are based on elicitations from linguists either working in language documentation or historical linguistics to study basic meanings across languages. It is therefore assumed that a cross-linguistic comparison between the words that are linked to a specific Concepticon concept can be carried out. The mapping of elicitation glosses to concept sets is based on a manual workflow in which the Concepticon editors (a group of linguists) review and discuss each list that is integrated into the database. The Concepticon offers information on more than 3,500 concept sets linked to more than 300 concept lists.[3] It is also used as a reference catalog to add specialized data collections such as the NoRaRe data (Tjuka et al., 2020b) or data on colexifications (Rzymski et al., 2020).

The NoRaRe database[4] links additional information to the concept sets in Concepticon (Tjuka et al., 2020b). This information includes norms, ratings, and relations on words and concepts. The data come from studies in psychology and linguistics and currently include more than 70 data sets (Tjuka et al., 2020a). The 'norms' category consists of data on word frequencies or reaction times. The 'ratings' category provides participant judgments for psycholinguistic criteria such as age-of-acquisition, arousal, valence, among others. The category of 'relations' comprises, for instance, semantic field categorization and semantic networks. In the case of the NoRaRe data, the words *I*, *ich*, and *wǒ* 我, as well as the values for a property in a given data set, are linked to the Concepticon concept set 1209 I. The NoRaRe database also incorporates the word frequencies in subtitles for film and TV-series in English (Brysbaert and New, 2009), German (Brysbaert et al., 2011), and Chinese (Cai and Brysbaert, 2010) for several Concepticon concept sets.[5]

Another data collection that is based on the Concepticon is the Database of Cross-Linguistic Colexifications (CLICS) (Rzymski et al., 2020).[6] The term 'colexification' was established by François (2008) and refers to one lexeme having multiple meanings. It is an umbrella term for instances of vagueness, homonymy, and polysemy. The database comprises colexifications for almost 3,000 Concepticon concept sets across more than 2,000 language varieties. The colexifications are computed on the basis of the information in the concept lists by identifying whether a given elicitation gloss is mapped to multiple Concepticon concept sets. The database also offers colexification weights between concept sets. For example, the concept set 1209 I colexifies with 1212 WE in 31 language varieties compared to the colexification with the concept set 1405 NAME in 3 language varieties.[7]

All three resources are accessible online and the data can be easily retrieved. In addition, the data sets are presented in a standardized format. The workflows for the creation of each database rely on the standardization efforts of the Cross-Linguistic Data Formats initiative (CLDF) (Forkel et al., 2018).[8] The data is converted into a tabular format with an additional metadata file. This allows to instantly compare the data sets and reuse them. The Concepticon concept sets as a reference provide the further possibility for cross-linguistic comparison. The present study uses word frequencies in the SUBTLEX

[2] A web application of the Concepticon is available at `https://concepticon.clld.org/`

[3] The data is openly accessible on GitHub: `https://github.com/concepticon/concepticon-data`

[4] A web application of the NoRaRe database is available at `https://digling.org/norare/`

[5] The data is curated on GitHub: `https://github.com/concepticon/norare-data`

[6] A web application of the CLICS database is available at `https://clics.clld.org/`

[7] The data is available on GitHub: `https://github.com/clics/clics3`

[8] Wilkinson et al. (2016) proposed that data should be *findable*, *accessible*, *interoperable*, and *reusable* (FAIR). The CLDF initiative builds on this principle and offers standards for multiple data types.

data sets for English (Brysbaert and New, 2009), German (Brysbaert et al., 2011), and Chinese (Cai and Brysbaert, 2010) in the NoRaRe database. For information on colexifications in each of the languages, data included in CLICS from Key and Comrie (2016) as well as Haspelmath and Tadmor (2009) was selected.

The study presented in this article aims to test two 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.

The hypotheses are examined on the basis of two comparisons: English–German and English–Chinese. The first language pair (English–German) was chosen because the languages represent closely related languages (both belong to the Germanic sub-branch of the Indo-European language family) while English–Chinese is the other side of the extreme, as the languages do not have a common ancestral language and therefore, are not related. To my knowledge, no study has tested either of the hypotheses with data on word frequencies before. Therefore, I assume that the correlation between word frequencies in English and German is higher than between English and Chinese. In addition, greater differences in frequency values for individual concepts between English and Chinese compared to English and German are expected. The results of the study are discussed in the next section.

3 Results

3.1 Correlations of Frequencies

The links between the Concepticon concept sets and the data in English (Brysbaert and New, 2009), German (Brysbaert et al., 2011), and Chinese (Cai and Brysbaert, 2010) are already provided in the NoRaRe database. Each data set consisted of more than 1,000 Concepticon concept sets with the respective values for word frequencies in subtitles of films and TV-series in English (2,329 concept sets), German (1,291 concept sets), and Chinese (1,644 concept sets). The language pair English–German had an overlap of 1,149 concept sets. In the language pair English–Chinese, the overlap amounted to 1,313 concept sets.[9] The shared concept sets were the basis for the correlation between each language pair.

To test the hypothesis that related languages have more similar frequencies across a set of shared concepts than non-related languages, two correlations were performed. First, the $\log_{10}$ frequencies of the 1,149 concept sets in English and German were correlated. The Pearson coefficient was 0.67 with a statistically highly significant p-value of $p < .001$. The distribution of the word frequencies is illustrated in Figure 1. Second, the $\log_{10}$ frequencies of the 1,313 concept sets in English and Chinese were compared. The Pearson coefficient was 0.55 with a statistically highly significant p-value of $p < .001$ (for the distribution see Fig. 1).

The correlation coefficients for both language pairs were not particularly high. However, there seems to be a slight difference between the data in the related languages English and German compared to English and Chinese. The next section investigates the differences between the data in more detail.

3.2 Cases of Language Variation

The mapping of the word frequency data sets to the Concepticon makes a qualitative cross-linguistic comparison possible. Tables 1 and 2 show the 15 most frequent concept sets in the two language pair data sets English–German and English–Chinese.

The logarithmic word frequencies for the concept set 1209 I across the three languages is now apparent: English *I* has a $\log_{10}$ frequency of 6.31, German *ich* has a $\log_{10}$ frequency of 5.97, and Chinese *wǒ* 我

[9] The overlap of the concept sets in the language pair German–Chinese was only about 700 concepts. Thus, the comparison would have been based on a much smaller data set than in the other two language pairs. The differences in the size of the data sets would most likely blur the result. For this reason, the study focused on the comparison between English–German and English–Chinese.

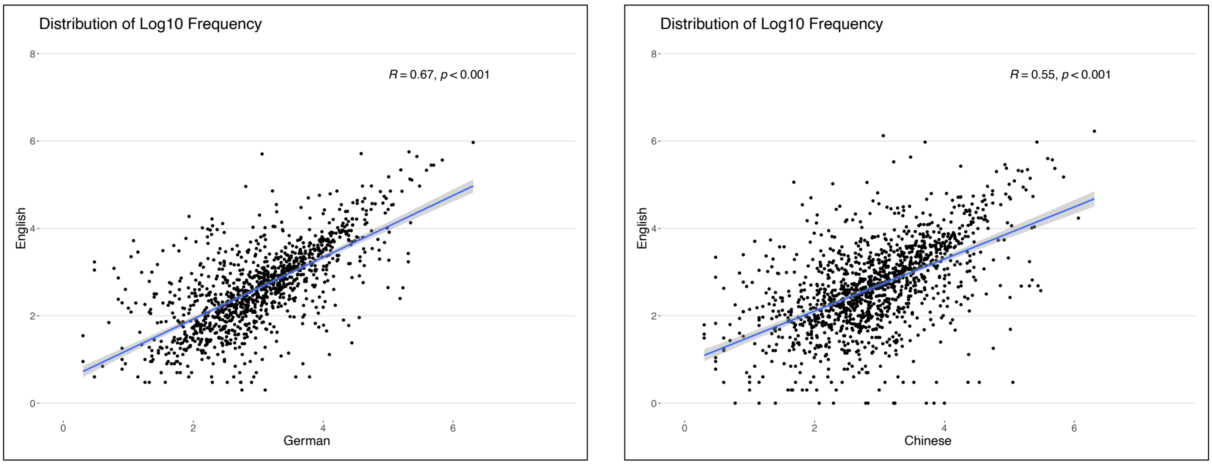

Figure 1: Distribution of the $\log_{10}$ word frequencies across the language pairs: English–German (left), English–Chinese (right). The data was taken from Brysbaert and New (2009), Brysbaert et al. (2011), and Cai and Brysbaert (2010) provided as subsets in the NoRaRe database (Tjuka et al., 2020b).

Table 1: The 15 most frequent concept sets in the overlapping data of English (Brysbaert and New, 2009) and German (Brysbaert et al., 2011) sorted by the English $\log_{10}$ word frequencies. The blue cell color indicates that the concept set does not appear in the English–Chinese language pair data set.

ID	**Label**	**English $\log_{10}$**	**German $\log_{10}$**
1209	I	6.31	5.97
1577	AND	5.83	5.57
1236	WHAT	5.70	5.45
1212	WE	5.66	5.45
1211	HE	5.59	5.33
1269	NO	5.48	4.97
1240	NOT	5.44	5.65
136	HERE	5.36	5.11
1937	THERE	5.35	4.13
2336	OF THIS KIND (SUCH)	5.34	5.13
817	THEY	5.32	5.75
1019	RIGHT	5.31	3.24
1117	LIKE	5.31	3.43
684	OWN	5.23	2.64
1376	NOW	5.21	4.85

has a $\log_{10}$ frequency of 6.23. The similar values indicate that the first-person pronoun occurred almost equally frequent in all three corpora. Other concept sets that are similarly common across all three data sets include 1577 AND, 1236 WHAT, 1212 WE, 1211 HE, 1937 THERE, and 817 THEY.

In contrast, some concept sets which have a high frequency in English appear to have lower frequencies in German and/ or Chinese. For example, the concept sets 1269 NO and 1240 NOT have relatively high $\log_{10}$ frequencies in English (5.48 and 5.44, respectively) and German (4.97 and 5.65, respectively), whereas the $\log_{10}$ frequencies in Chinese are considerably lower with 2.58 for the concept set 1269 NO and 2.69 for the concept set 1240 NOT. In the case of the concept set 1019 RIGHT, the English $\log_{10}$ frequency is higher (5.31) compared to German and Chinese which have the same lower $\log_{10}$ frequency of 3.24.

Some concept sets occurred only in one of the language pair data sets. On the one hand, the concept sets 2336 OF THIS KIND (SUCH), 1117 LIKE, 684 OWN, and 1376 NOW appeared in the data of the

Table 2: The 15 most frequent concept sets in the overlapping data of English (Brysbaert and New, 2009) and Chinese (Cai and Brysbaert, 2010) sorted by the English $\log_{10}$ word frequencies. The blue cell color indicates if a concept set does not appear in the English–German language pair data set.

ID	Label	English $\log_{10}$	Chinese $\log_{10}$
1209	I	6.31	6.23
2754	TOWARDS	6.06	4.24
1577	AND	5.83	5.18
1236	WHAT	5.70	5.37
1212	WE	5.66	5.57
1211	HE	5.59	5.60
1269	NO	5.48	2.58
1240	NOT	5.44	2.69
1579	BE	5.42	5.98
84	JUST	5.38	4.04
136	HERE	5.36	4.73
506	MAIZE	5.35	2.82
1937	THERE	5.35	4.01
817	THEY	5.32	5.15
1019	RIGHT	5.31	3.24

English–German language pair. The concept sets 2754 TOWARDS, 1579 BE, 84 JUST, and 506 MAIZE, on the other hand, occurred only in the English–Chinese data set.

The comparison of the 15 most frequent concept sets across the language pairs indicates that there are substantial differences in the data across the three languages. The second hypothesis of the present study was that fewer concepts have a large difference between frequencies in related languages than in non-related languages. To investigate this hypothesis, the differences in the $\log_{10}$ frequencies across the language pairs were compared. Tables 3 and 4 show the results of the comparisons for the English–German and English–Chinese data sets. Only concept sets that vary largely in their frequencies across the languages (difference greater than 3) were included for a qualitative comparison. These concept sets are extreme cases, but as discussed in the previous section, both language pair data sets share many concept sets that have similar $\log_{10}$ frequencies.

Table 3: 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

One obvious observation that becomes apparent in the comparison is the number of concept sets that have large differences between $\log_{10}$ frequencies. In the English–German data set only two concept sets vary greatly in their frequencies: 1301 FOOT and 492 THREE. Both concept sets occurred more often in English. The concept set 492 THREE refers to the natural number *three* in English and German (*drei*). The difference between the frequencies could be due to the fact that in German, the number word starts with a capital letter in some contexts, for instance, in the sentence *Sie hat eine Drei gewürfelt.* “She rolled a three.” The concept set 1301 FOOT refers to the human body part. In English and German, *foot* is used also in other contexts, for instance, *foot of the mountain* or *metrical foot*. However, in German, the word *Fuß* “foot” often occurs as a compound word, as in *Versfuß* “metrical foot.” This might explain the low frequency of the standalone word compared to English in which most compounds are written

Table 4: 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). The red row color indicates that the frequency of the concept set is higher in Chinese than in English.

ID	**Label**	**English $\log_{10}$**	**Chinese $\log_{10}$**	**Difference**
1235	WHO	5.05	0.48	4.58
1203	LONG	4.54	0.48	4.06
2483	COLD (OF WEATHER)	4.00	0.00	4.00
1417	KILL	4.36	0.48	3.89
702	CATCH	3.84	0.00	3.84
648	PAPER	3.72	0.00	3.72
1458	SAY	4.75	1.26	3.50
156	RED	3.88	0.48	3.40
705	GO UP (ASCEND)	1.68	5.06	3.38
1238	WHEN	5.02	1.69	3.33
1446	COME	4.37	1.11	3.26
1424	YELLOW	3.24	0.00	3.24
930	VILLAGE	3.23	0.00	3.23
711	TALL	3.22	0.00	3.22
1215	THOU	3.06	6.12	3.06
1208	CAT	3.53	0.48	3.05

with a space between the words, as in *three times* or *foot brake*.

The comparison between the frequencies in English and Chinese resulted in 16 concept sets with a large difference (greater than 3) in their $\log_{10}$ frequencies (see Tab. 4). A closer look at some of the concepts revealed cases of language variation which could lead to the differences in the frequencies. For example, the concept set 1235 WHO is mapped to English *who*, but Chinese has two word-forms *shuí* 谁 and *shéi* 誰 to ask about one person or people. The former is written in the simplified Chinese script, whereas the latter uses the traditional Chinese characters. Because both of them occurred in the original data, but only one word is mapped to the concept, the data set includes the frequency for *shéi* 誰 instead of *shuí* 谁 which has a $\log_{10}$ frequency of 4.72. The zero frequency of the concept set 930 VILLAGE results from a choice between two words: the concept set was mapped to the compound *cūnzhài* 村寨 instead of the more frequent word *cūnzi* 村子 ($\log_{10}$ frequency: 2.66). The expression *cūnzhài* 村寨 is used to refer to an area in which specific cultural groups live. In contrast, *cūnzi* 村子 is a more general word that can be used for all villages. In the case of the concept set 2483 COLD (OF WEATHER), the word *liáng* 凉 was mapped instead of the more frequent compound *liángshuǎng* 凉爽 with a $\log_{10}$ frequency of 1.72. The term *liángshuǎng* 凉爽 would in fact be a more accurate word for the concept set 2483 COLD (OF WEATHER) since it relates to a state of weather with low temperature. Nevertheless, in English, the concept seems to appear more frequently than in Chinese. The reason could be the climate that Chinese speakers live in. The differences in the frequencies of the concept set 1238 WHEN (English $\log_{10}$ 5.02 and Chinese $\log_{10}$ 1.69) is due to Chinese having two distinct question pronouns: *shénme shíhòu* 什么时候 and *jǐshí* 几时 of which only the latter was included in the original data set. Note that the former seems to be the default option for the concept set 1238 WHEN in everyday language, whereas *jǐshí* 几时 is used by the older generation.

Interestingly, two concept sets – 705 GO UP (ASCEND) and 1215 THOU – appear to be more frequent in Chinese compared to English. Chinese uses *shàng* 上 to indicate an upward movement. It can also occur as a compound: *shàngqù* 上去 "go up." In English, however, there is a specific verb for moving from a lower to a higher position by walking or climbing: *ascend*. The difference in the frequencies of the concept set 1215 THOU, which describes a second-person pronoun singular form, can be explained by the fact that Chinese has two forms of second-person pronoun singular *nǐ* 你 and *nǐn* 您, which is the formal version. Similarly, German has *Du* and *Sie* (informal and formal, respectively). English used to have *thou* to indicate the second-person pronoun singular, but in common day English, *you* refers to both forms: second-person pronoun singular and plural. Note that the concept set 1215 THOU also has the highest diversity in glossing (List, 2018). The reference to one person or more was not distinguished in the SUBTLEX data and therefore, the frequencies cannot be separated. For the other concepts sets no conclusive explanation was apparent. The implications of the results are discussed in the next section.

4 Discussion

The present article set out to study the mental lexicon from a cross-linguistic perspective. The distribution of word frequencies across three languages, namely English, German, and Chinese was investigated. The advantage of the cross-linguistic database approach of the study is that it allowed a comparison of the same property across a set of diverse languages. The NoRaRe database (Tjuka et al., 2020b) was used to correlate data sets of frequencies in subtitles (Brysbaert and New, 2009; Brysbaert et al., 2011; Cai and Brysbaert, 2010) with one another and the CLICS database (Rzymski et al., 2020) was used to search for colexifications in the languages. Both databases are built upon the same reference catalog: Concepticon (List et al., 2016). This resource is based on a link between elicitation glosses in concept lists that comprise comparative concepts. The lists are provided by linguists and are used to compare basic meanings across languages. The Concepticon offers stable identifiers for those concepts and makes a direct comparison of concepts in multiple languages possible. The elicitation glosses are the basis for the word that can be mapped to a specific concept. Thus, the Concepticon can be used for an in-depth study of cross-linguistic lexical variation.

The goal of this study was to test whether related languages have more similar frequencies across a set of shared concepts than non-related languages. In addition, I examined the hypothesis that related languages share fewer concepts with a large difference in their frequencies than non-related languages. To test the hypothesis, correlations and qualitative analysis of individual concepts were carried out across two language pairs: English–German (related) and English–Chinese (non-related). Both hypotheses were supported by the findings in Section 3. The correlation of the frequencies between the language pair English–German was slightly higher than between English–Chinese. Furthermore, the comparison of the $\log_{10}$ frequencies of the concept sets shared in each language pair revealed language-specific variation. In the case of English–German, fewer concept sets with a large difference in their $\log_{10}$ frequencies were found (2 concept sets) compared to 16 concept sets in the English–Chinese data set.

The findings of the study indicate that frequencies of the same concepts can differ greatly across languages. The detailed examination of the individual concepts showed that two processes may lead to the differences in frequencies. First, cultural diversity, for instance, different regional climates, drives the use of certain weather-related concepts such as COLD (OF WEATHER). Second, the use of two word-forms as in the case of *shuí* 谁 and *shéi* 誰 for the concept set 1235 WHO can result in varying cross-linguistic frequencies. The meaning-frequency law (Zipf, 1945) was not supported by the data. No influence of the number of colexifications for a concept on its frequency across languages was found.

When comparing frequencies across languages, it is challenging to consider the many differences that distinguish languages from one another. Nevertheless, researchers should not confine themselves to the study of single languages. The cultural differences that emerge from cross-linguistic studies offer valuable insights into the connections that languages draw between concepts in certain semantic domains, for example, emotions (Jackson et al., 2019). In addition, the study by Jackson et al. (2019) illustrated that general patterns of psycholinguistic measures, such as arousal and valency, exist independently of the family to which a given language belongs. Another advantage of the database approach used in this study is the possibility to explicitly look up comparative concepts and compare their properties across languages. Although some mappings might need refinement, the overall results prove the validity of the database. The comparison could be further improved by using frequencies in large parallel text corpora, but the data sets based on subtitles already account for a related context.

5 Conclusion

In recent years, a wealth of data for individual languages and data from cross-linguistic studies became available. The implementation of the diverse findings in databases makes a new field of exploration possible: a cross-linguistic comparison of variables such as word frequencies. Future studies can build on the hypotheses presented in this study and test other assumptions of general patterns or language variation in different areas of the mental lexicon. All data used in the present article are readily accessible and can be reused by other researchers.

Acknowledgements

Annika Tjuka was supported by a stipend from the International Max Planck Research School (IMPRS) at the Max Planck Institute for the Science of Human History and the Friedrich-Schiller-Universität Jena.

References

Jean Aitchison. 2012. *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons, 4 edition.

Francis Bond and Ryan Foster. 2013. Linking and extending an Open Multilingual WordNet. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics (ACL).

Francis Bond, Arkadiusz Janz, Marek Maziarz, and Ewa Rudnicka. 2019. Testing Zipf's meaning-frequency law with wordnets as sense inventories. In Christiane Fellbaum, Piek Vossen, Ewa Rudnicka, Marek Maziarz, and Maciej Piasecki, editors, *Proceedings of the Tenth Global Wordnet Conference*, pages 342–352, Wrocław, Poland. Oficyna Wydawnicza Politechniki Wrocławskiej.

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.

Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i Cancho, and Jaume Baixeries. 2019. Polysemy and brevity versus frequency in language. *Computer Speech & Language*, 58:19–50.

Christiane Fellbaum. 2012. WordNet. *The Encyclopedia of Applied Linguistics*.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(1):1–10.

Alexandre François. 2008. Semantic maps and the typology of colexification. In Martine Vanhove, editor, *From polysemy to semantic change: Towards a typology of lexical semantic associations*, volume 106, pages 163–215. John Benjamins Publishing.

Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T. Piantadosi, , and Bevil R. Conway. 2017. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences of the United States of America*, 114(40):10785–10790.

Martin Haspelmath and Uri Tadmor. 2009. *Loanwords in the world's languages. A comparative handbook.* de Gruyter, Berlin and New York.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Antoni Hernández-Fernández, Bernardino Casas, Ramon Ferrer-i Cancho, and Jaume Baixeries. 2016. Testing the robustness of laws of polysemy and brevity versus frequency. In Pavel Král and Carlos Martín-Vide, editors, *4th International Conference on Statistical Language and Speech Processing (SLSP)*, pages 19–29, Pilsen, Czech Republic. Springer, Cham.

Bahar Ilgen and Bahar Karaoglan. 2007. Investigation of Zipf's 'law-of-meaning' on Turkish corpora. In Ece G Schmidt, Ilkay Ulusoy, Nihan Ciçekli, and Ugur Halıcı, editors, *22nd International Symposium on Computer and Information Sciences*, pages 1–6, Ankara, Turkey. IEEE.

Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Peter J. Mucha, Robert Forkel, Simon J. Greenhill, and Kristen Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Mary Ritchie Key and Bernard Comrie. 2016. *The Intercontinental Dictionary Series*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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, Luxembourg. 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.

Johann-Mattis List. 2018. Towards a history of concept list compilation in historical linguistics. *History and Philosophy of the Language Sciences*, 5:1–14.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130.

Christoph Rzymski, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(13):1–12.

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2):121–137.

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.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, and Philip E. Bourne. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(160018):1–9.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256.