# Linking Norms, Ratings, and Relations of Words and Concepts Across Multiple Language Varieties

Annika Tjuka[1], Robert Forkel[1], and Johann-Mattis List[1]

[1]Max Planck Institute for the Science of Human History, Jena, Germany

Corresponding author: Annika Tjuka (tjuka@shh.mpg.de)

**Abstract**

Psychologists and linguists have collected a great diversity of data for word and concept properties. In psychology, many studies accumulate norms and ratings such as word frequencies or age-of-acquisition often for a large number of words. Linguistics, on the other hand, provides valuable insights into relations of word meanings. We present a collection of those data sets for norms, ratings, and relations that cover different languages: 'NoRaRe.' To enable a comparison between the diverse data types, we established workflows that facilitate the expansion of the database. A web application allows convenient access to the data (https://digling.org/norare/). Furthermore, a software API ensures consistent data curation by providing tests to validate the data sets. The NoRaRe collection is linked to the database curated by the Concepticon project (https://concepticon.clld.org) which offers a reference catalog of unified concept sets. The link between words in the data sets and the Concepticon concept sets makes a cross-linguistic comparison possible. In three case studies, we test the validity of our approach, the accuracy of our workflow, and the applicability of our database. The results indicate that the NoRaRe database can be applied for the study of word properties across multiple languages. The data can be used by psychologists and linguists to benefit from the knowledge rooted in both research disciplines.

## 1 Introduction

Words are connected to entities in the world around us. Hence, they contain information about different aspects of the objects which we perceive in our everyday life. If native speakers of English see the word *tree*, they can instantly answer several questions regarding the nature of the concept represented by *tree*. For example, whether they perceive the word as concrete. In addition, English speakers can estimate how many times they came across the word *tree* in the last week or whether *tree* is related to the word *wood*. Yet, when native speakers of Spanish perceive the word *árbol*, or Chinese speakers hear *shù* 树, will we get the same or similar answers? While it is likely that the answers coincide, it will be difficult to substantiate this assumption with evidence. One could start with a thorough study of the available literature that provide empirical studies on norms, ratings, and relations of *tree* in the three language varieties. If the data are available in a machine-readable format, the information could be assembled and unified. Furthermore, the words would be translated into a meta-language to compare the data sets. In the end, one could assess the similarity of the answers across the different languages, i.e., English, Spanish, and Chinese.

Given the increased interest in cross-linguistic (multilingual) studies in the field of psychology (e.g., Gibson et al., 2017; Jackson et al., 2019; Jackson, Watts, List, Drabble, &

Lindquist, 2020), it would be desirable to have a catalog that unites the wealth of data on norms, ratings, and relations of words and concepts, which have been published across a variety of languages. Recent approaches offer bibliographies that list information on norm databases (Buchanan, Valentine, & Maxwell, 2019b; Winter, Wedel, & List, 2017), or unify information on concepts across languages (Speer, Chin, & Havasi, 2017). However, no resource is available that includes a sufficient amount of cross-linguistically comparable data on *word and concept properties* from *norm* and *rating* studies in psychology. In addition, we do not find a database that relates norms and ratings from psychology with data on word relations from linguistic fields, such as historical linguistics and linguistic typology.

Linguistic data sets include various rankings of concepts regarding linguistic constructs such as *stability* (the robustness of the connection between word form and word meaning over time), *borrowability* (the likelihood that a word is transferred from one language to another), or *polysemy* (the degree by which a word form expresses multiple concepts). These relations between words and concepts are usually derived from the comparison of multiple languages, whereas psychological norms and ratings are generally collected for one particular language. The linguistic data could inspire psychologist to embrace a cross-linguistic perspective. At the same time, linguists would benefit from having access to norms and ratings collected in large studies. For example, ratings for valence or arousal as well as norms for frequency facilitate the prediction of certain linguistic constructs (Calude & Pagel, 2011; Jackson et al., 2019).

In linguistics, the comparison of concepts across languages is crucial in many regards, for instance, the study of semantic change. The first attempts to compare concepts cross-linguistically were enabled by the establishment of the Concepticon project (List, Cysouw, & Forkel, 2016).[1] The Concepticon links concepts in more than 300 concept lists to more than 3,000 *concept sets*. Each concept set receives a unique identifier, an English gloss (for convenience), and a definition. The primary intention of the Concepticon project was to provide stable identifiers for concepts used in the linguistic literature in order to ease the aggregation of data sets from different sources. The first Concepticon version (List et al., 2016) already contained data sets that have been compiled for applications in psychology as well as data sets offering *conceptual metadata*, such as frequency norms (Brysbaert, Warriner, & Kuperman, 2014) and links to WordNet (Fellbaum, 1998).

The Concepticon project has been growing steadily over the past five years. It is based on a computer-assisted data curation workflow that is well-established. The data is versionized and regularly released with at least one update per year. Therefore, the Concepticon is a perfect starting point for the task of linking norms, ratings, and relations of words and concepts across multiple language varieties. Based on our experience with the Concepticon project, we created a new collection containing 71 data sets with additional information on word and concept properties. For linking the data sets, we use a transparent, computer-assisted, and replicable workflow. The collection is accessible through a software API (written in Python) that allows to test the data for internal consistency and at the same time, offers quick access to the data. Furthermore, we provide a web-based front-end so that other researchers can easily examine the data.

In Section 2, we give an overview of existing data sets in psychology and introduce the challenges we face when trying to link different data sets. In Section 3, we describe our computer-assisted data curation workflow. To validate our approach, we present basic statistics of the data we have assembled so far, review our experience with our current data curation workflow, and show in three case studies, how the data can be put to concrete use (Sect. 4). Finally, we discuss the implications of the newly assembled data in Section 5 and point to

[1] The database curated by the Concepticon project can be accessed via a web application under the following link: `https://concepticon.clld.org`

future plans regarding the extension of the data collection.

## 2 Combing Forests of Data

If we step out into a forest, we see a variety of trees. Some are as tall as a four-story building others are thin like a pencil. The wealth of data for word properties is no less diverse. For our purposes, we divided the different types of data into three groups: *norms*, *ratings*, and *relations*.

*Norms* are determined by taking samples from a total quantity, for example, counts of word occurrences in a corpus. They are collected and applied predominantly in the field of psychology.[2] The norms we found in the literature include data on word frequencies in subtitles for various languages, for instance, English (Brysbaert & New, 2009), Spanish (Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2011), Chinese (Cai & Brysbaert, 2010), and Dutch (Keuleers, Brysbaert, & New, 2010). Additionally, we classified reaction time studies (e.g., Ferrand et al., 2010; Tsang et al., 2018) as norms. These studies are comprised of a large number of words (often more than 60,000 words). Norm data are commonly based on a broad text or word basis. They are rarely compiled for smaller languages due to the lack of available sources.

*Ratings* are conducted with participants who judge a given word either on a scale or on other measures, for instance, the age at which a word was acquired. We found various ratings on age-of-acquisition using different scales (e.g., Alonso, Fernandez, & Díez, 2015; González-Nosti, Barbón, Rodríguez-Ferreiro, & Cuetos, 2014; Kuperman, Stadthagen-González, & Brysbaert, 2012). Other studies inquire a diverse set of psychological aspects which are evoked while reading a word: concreteness, imageability, arousal, valence, discrete emotions (happiness, sadness, anger, fear, disgust), sensory modality (auditory, gustatory, haptic, olfactory, visual) and so on. The number of words included in the rating studies varies from small to large. For example, Maciejewski and Klepousniotou (2016) collected meaning frequencies for 100 homonyms, whereas Stadthagen-González, Imbault, Pérez-Sánchez, and Brysbaert (2017) offer ratings for valence and arousal for approximately 14,000 words, and Lynott, Connell, Brysbaert, Brand, and Carney (2020) provide ratings on perceptual and action strength for 40,000 words. Most studies are conducted with speakers of well-documented languages which is typical for psychological research. The over-representation of data from a Western, educated, industrialized, rich, and democratic (WEIRD) population (Jones, 2010) is striking. In recent years, linguistic diversity has been increasing, as shown by the publication of new ratings for Turkish (Kapucu, Kılıç, Özkılıç, & Sarıbaz, 2018) or a diverse set of languages from Afrikaans to Western Armenian (Łuniewska et al., 2016, 2019).

*Relations* include a variety of data types such as *rankings*, *semantic field categorization*, and *semantic networks*. Data sets of this type provide information on the *relations* between words and concepts. Data on relations are typically collected in the field of comparative linguistics which deals with various questions related to the evolution of languages (historical linguistics) and the general properties of the world's languages (linguistic typology). But we also find data for relations in the field of Natural Language Processing and other data-driven fields investigating language and semantics.

Typical examples for relations are lists (of words and concepts) in which items are *ranked*, *tagged*, or directly *associated* with other items in the same list. In *ranked lists*, words and concepts are ordered by cross-linguistic categories, such as *borrowability* (referring to the likelihood by which words expressing a given meaning tend to be borrowed, see Tadmor, 2009) and *stability* (referring to the language-family-specific or general resistance of words

[2]Note that ratings are also described as norms in the literature (e.g., Scott, Keitel, Becirspahic, Yao, & Sereno, 2019).

expressing a given concept to change their meaning over time, see Calude & Pagel, 2011). In *tagged lists*, a given word or concept is described by a certain tag or a set of tags, and different words and concepts can be compared by means of the tags they share (the list of headwords and senses by Starostin (2000) is a classic example for such a data set). Lists providing concept *associations* are most typically represented by the *WordNet* ontology (Fellbaum, 1998). But association data sets, such as the Edinburgh Associative Thesaurus (Kiss, Armstrong, & Milroy, 1973), would also fall under this category as would the recently proposed data sets of cross-linguistic colexifications[3] (Rzymski et al., 2020).

Studies on word and concept relations often only include a much smaller number of items compared to norm and rating studies. However, the items are carefully selected and chosen based on their comparability across multiple languages. While norm and rating studies cover large numbers of words and concepts in individual languages, studies of relations typically cover small numbers of words and concepts across samples of diverse languages, including many languages that are notoriously underrepresented in cross-linguistic studies.

Psychologists provide platforms that include norms and ratings on several psycholinguistic criteria with the possibility to create balanced stimulus sets (for English, see Buchanan, Valentine, and Maxwell 2019a; Guasch, Boada, Ferré, and Sánchez-Casas 2013; Wilson 1988; for German, see Heister et al. 2011). However, data sets on different languages are not included in those databases which makes a cross-linguistic comparison difficult. If we want to examine the frequency of the word *tree* in English, Spanish, and Chinese, we would translate *tree* into *árbol* and *shù* 树. The Center for Reading Research at the Department of Experimental Psychology of Ghent University offers many resources on norms and ratings, also in different languages. To find our frequencies, we could access the data on English (Brysbaert & New, 2009), Spanish (Cuetos et al., 2011), and Chinese (Cai & Brysbaert, 2010) and search for the words *tree*, *árbol*, and 树. The advantage of those data sets is that they are in a rather uniform format and are stored on the same website[4]. Nevertheless, the homogeneous data provided by the Center for Reading Research is an exception rather than the norm.

In order to guarantee that data can be easily found, accessed, employed, and reused, Wilkinson et al. (2016) have proposed the FAIR guiding principles for scientific data management. They suggest that data should be *findable*, *accessible*, *interoperable*, and *reusable* (FAIR). Although many researchers might know about the principles, the data supplemented in research articles often do not qualify as FAIR in the sense of Wilkinson et al. (2016).

While it is becoming more and more common to add a section introducing the supplementary material of a given study, some journals obscure the access to the repositories in which data sets are stored. The fact that the data sets are archived on a journal's website is also problematic. Journals are not properly equipped for long-term archiving, licensing, and regular release updates for the data. The best practice for storing one's data is, therefore, scientific archiving services, for instance, Zenodo (`https://zenodo.org`), or the Open Science Framework (`https://osf.io`). These possibilities enjoy increasing popularity (for studies that store their data on one of the two archives see Kapucu et al., 2018; Lynott et al., 2020; Rzymski et al., 2020).

Even if data can be easily found and accessed, this does not necessarily mean that they can be used and reused. Most data sets presenting word and concept properties are available in the form of tabular data. In a spreadsheet, words or concepts are given in a row of the table, and properties are listed in additional columns. Information on the semantics of the columns, however, is often lacking. Other researchers who would like to apply the data have

[3]The term was first introduced by François (2008). It is a cover term for polysemy and homophony. Thus, colexification refers to those cases in which the same word in a given language is used to express two or more concepts, such as Russian *ruka*, Hausa *hannu*, or Vietnamese *tay* all denoting 'hand' and 'arm.'

[4]The data of the Center for Reading Research can be accessed online `http://crr.ugent.be/programs-data/subtitle-frequencies`

to guess the nature of the content based on the table headers. This issue is illustrated in Figure 1. Given that many data sets offer similar norms, ratings, and relations for words and concepts, it would be highly desirable to have *uniform exchange formats*. In addition, a clear licensing policy with open licenses should be provided to ensure that the data can be reused in other studies as well. While many data sets are published without a license, some data sets have a license that explicitly restricts to build upon the data or use them in other scientific studies.

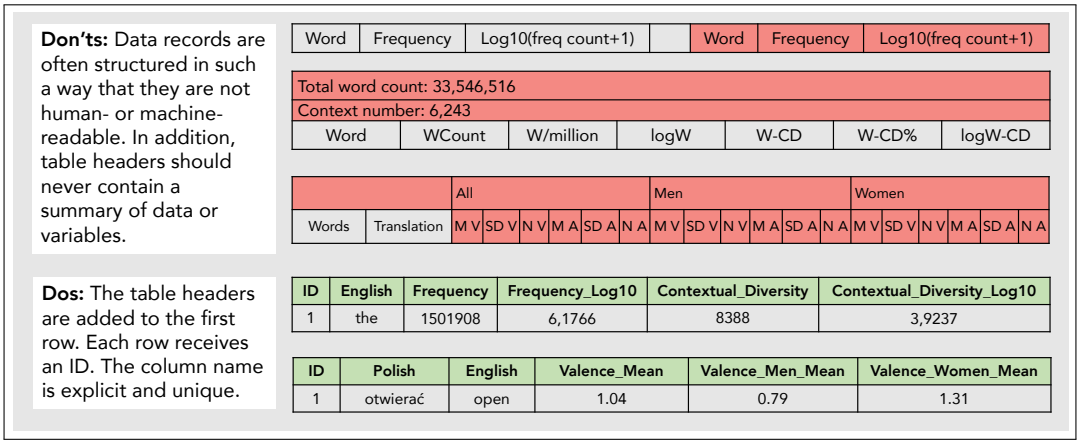


Figure 1: Best practice examples for structuring data sets. The data should be in a machine-readable form. Additionally, the information of the column content should be easily understandable by other researchers.

The advantage of having openly available data on a wide range of word and concept properties in one place is that we can compare, evaluate, and answer interrelated questions with a mixed bag of data. Furthermore, with FAIR data sets, studies can be carried out more rapidly and gaps as well as inconsistencies would become apparent. But the clearest benefit would be the possibility to link the data sets to other resources and make them cross-linguistically comparable.

# 3 Materials and Methods

## 3.1 Materials

We collected 71 data sets with 415 word properties (see Tab. 2 for an overview and the supplementary material for a complete list of the data sets). We made the diverse data sets comparable with each other by (1) normalizing the raw data, (2) linking the concepts and words to the Concepticon database, and (3) classifying and labeling the word properties provided by each data set (for details see Sec. 3.2). Since our collection includes norms, ratings, and relations for words and concepts across multiple languages, we call it *NoRaRe*. It is a new feature to the Concepticon resource that was previously only sporadically linked to metadata on word and concept properties (List et al., 2016).

In its current state (List et al., 2020), the Concepticon database offers identifiers, definitions, and short glosses for as many as 3,721 *concept sets*. The Concepticon identifiers are used to link individual multilingual elicitation glosses provided to describe concepts in 331 *concept lists* from four centuries of research in linguistics and beyond. Since its first publication in 2016 (List et al., 2016), the Concepticon project has been growing constantly.

### 3.2 Methods

Given their different origins, the data sets that we collected come in various formats and flavors. While data sets on norms and ratings are often large, listing more than 10,000 distinct words for a given language, some data sets in comparative linguistics only consist of 100 and at times even fewer items. The large data sets for norms and ratings and the small data sets in comparative linguistics are always discrete in size. In addition, we added a third category of data sets which are not available in discrete form, and can only be queried, for example, through a website. These data sets include word properties from databases such as Wikidata[5] or BabelNet[6].

The fact that we are dealing with three different types of data sets forced us to develop three different workflows to add them to our collection. For *small lists* of words and concepts, we relied on the well-established workflow to add concept lists to the database curated by the Concepticon project, which we will discuss in detail in Section 3.2.1. For *large (but discrete) lists*, we designed a new automated method that links the words and concepts in a given list to the Concepticon, which we present in Section 3.2.2. For databases so large that our mapping algorithm could not be used, we designed a semi-automated approach in which the automatically generated queries are manually checked (see Sec. 3.2.3).

All three workflows yield a unified output: either all or a certain part of the items (words or concepts) in the original data set are provided in tabular format along with the information on word and concept properties, and – where available – a link to the corresponding Concepticon identifier. The tabular format in which we provide the data is strictly standardized, following the recommendations of the W3C for tabular data on the web (Tennison, n.d.), also known as CSVW (for details about the use of CSVW for linguistic data, see Forkel et al., 2018). The core idea of CSVW is to increase the interoperability of tabular data by adding metadata in JSON format that conforms to specific recommendations. The CSVW Python package (Bank & Forkel, 2018) allows to automatically test for consistency as well as parse and manipulate data that is conforming to the CSVW recommendations.

After the data have been normalized and converted to a tabular format, all data sets are reviewed by colleagues who were not involved in preparing the data. They ensure that the data is modified when errors are spotted. To minimize errors in this stage, specific tests that check the formal requirements for the data are carried out, based on unit test facilities as they are typically used for the testing of code in software development. Once a data set has passed this *test-driven data curation* process, the word and concept properties provided by a particular data set are *classified* and *labeled* in order to make them comparable against other data sets. Since we use `git` for version control and `GitHub` for data curation, and `Zenodo` for data storage, all stages of the data curation workflow are transparently documented and can also be directly inspected by anybody interested in the details. Figure 2 (a) shows a schematic illustration of our test-driven data curation process. Figure 3 provides an example of the resulting cross-linguistic resources that offer data on norms, ratings, and relations across languages for the concept sets of the Concepticon database.

#### 3.2.1 Manual Concept Mapping

Given that the Concepticon resource already links to all kinds of concept lists of different sizes, purposes, and languages, it is straightforward to use the well-established data curation workflow to link small to moderately large data sets providing norms, ratings, and relations. While most of the concept lists released with the first version of the Concepticon (List et al., 2016) were linked manually, the growing body of elicitation glosses from different languages

[5]The Wikidata project is available at the following link: `https://www.wikidata.org/`
[6]BabelNet is available online: `https://babelnet.org/`

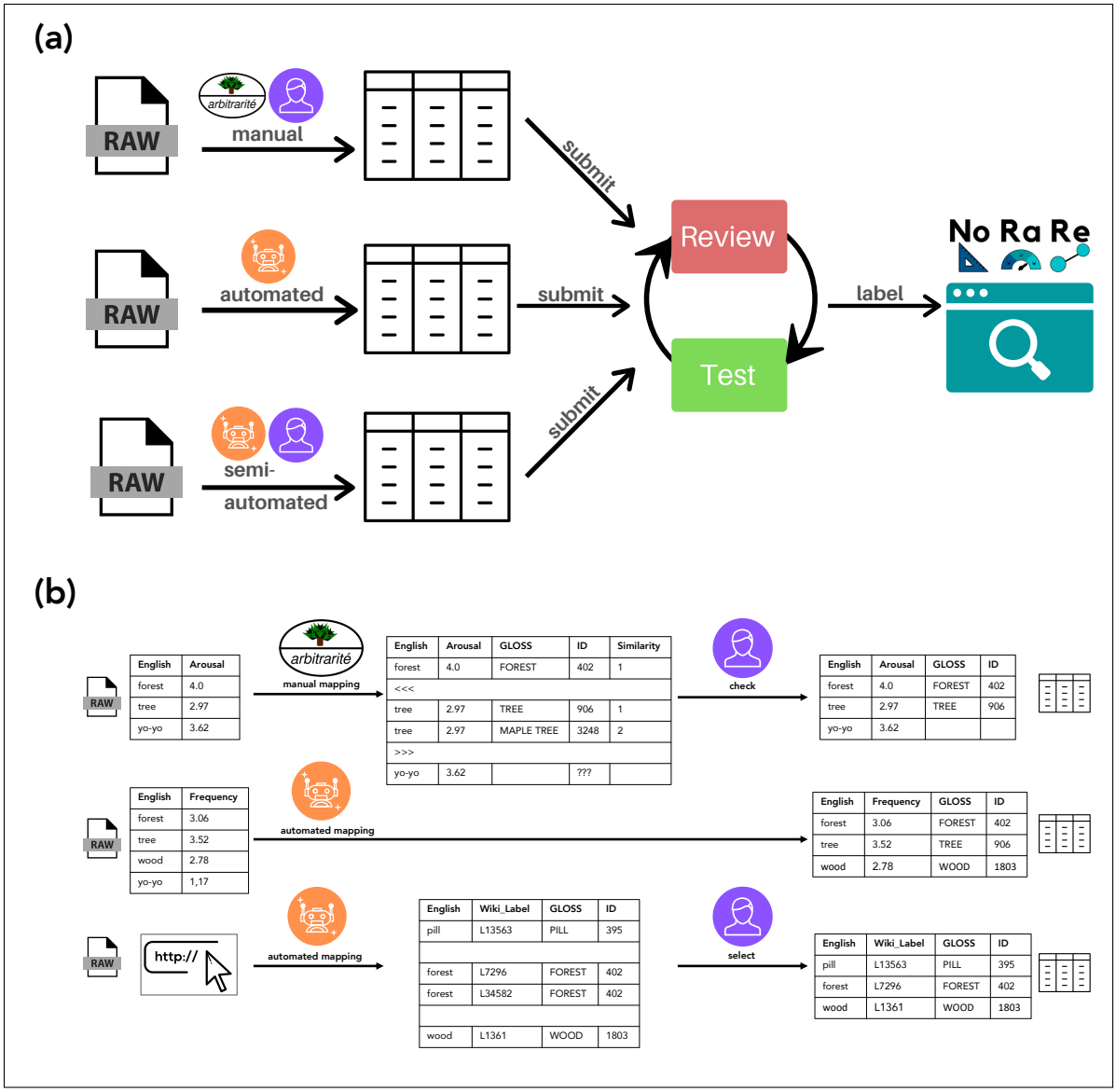


Figure 2: Workflows for data curation. Fig. (a) shows how raw data is converted to unified tabular data formats and consecutively labeled. Fig. (b) provides details for the individual steps involved in the linking of the different data sets to Concepticon.

made it possible to add an automated mapping algorithm in later versions of the Concepticon. This algorithm, which checks a given elicitation gloss against previous manual mappings, works surprisingly well, can currently be carried out in 30 languages, and is provided along with the `pyconcepticon` Python package which also allows to test the data for internal consistency (Forkel, Rzymski, & List, 2019). For individual concepts, users can consult a web-based lookup tool which offers a slightly simplified mapping algorithm that supports currently seven languages (List et al., 2018). In addition, users who want to contribute can consult tutorials for different levels of expertise (Tjuka, 2020; Tresoldi, 2019).

As should be clear from its name, the Concepticon deals primarily with *concept lists* which need to be distinguished from *word lists*. In a typical concept list, scholars try to assemble different concepts by means of *elicitation glosses* in a certain language in order to express the meaning of the concept they want to list. Since concept elicitation has never been standardized (and the Concepticon can be seen as one project that tries to help in this

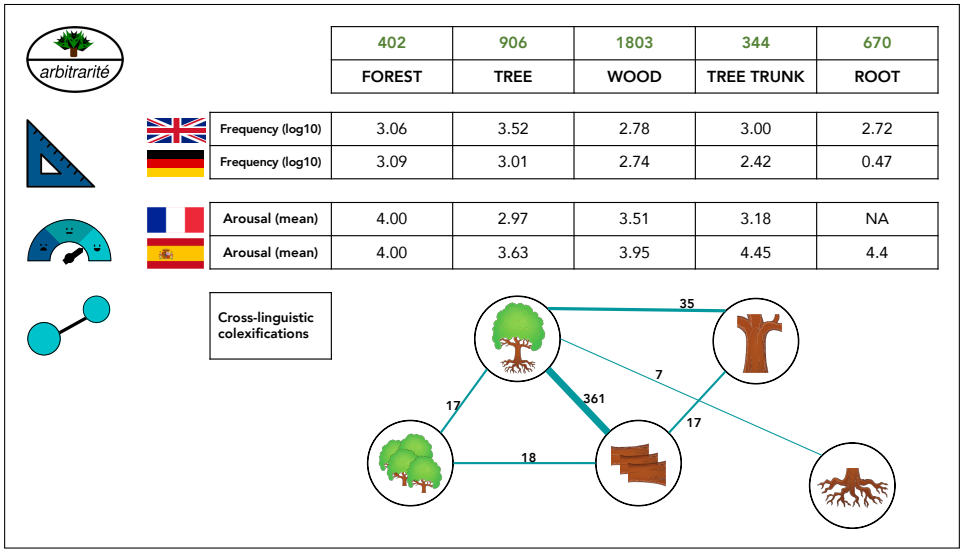


Figure 3: Comparing different kinds of data on word and concept properties as they have been proposed in the literature. (a) The Concepticon on top of the figure offers stable identifiers for more than 3,000 concepts. (b) The SUBTLEX data sets offer frequency counts for words across different languages based on subtitles (Brysbaert et al., 2011; Brysbaert & New, 2009). (c) User-rated collections of psychological categories, such as arousal, have been published for different languages (Riegel et al., 2015). (d) The CLICS database allows to estimate the semantic closeness of concepts by measuring how often they are colexified in the languages of the world (Rzymski et al., 2020).

regard by providing concept sets with stable definitions and identifiers), it is at times difficult to decide how to interpret the intended meaning of a specific elicitation gloss. This becomes even more difficult when dealing with proper word lists, where no attempt was undertaken to distinguish between the different meanings a word can have. As a result, an elicitation gloss or word like *wave* can be interpreted as the verb form 'waving motion of one's hand' or the noun form 'a sequence of ridges created by the wind across the sea.' The Concepticon provides a description of each concept which distinguishes both forms in separate concepts, namely WAVE (VERB) (ID: 3544) and WAVE (ID: 978). This distinction guarantees that the correct concept set is mapped to the word form. To avoid errors when adding new concept lists to the Concepticon, each new list is accompanied by an extensive data review that is conducted independently by colleagues. Since most data sets compiled for studies in psychology do not provide any information on potentially intended meanings of word forms, we decided to leave ambiguous cases unmapped instead of mapping them incorrectly to a specific concept. Figure 2 (b) contrasts the manual mapping procedure with the automated and semi-automated mapping procedure.

### 3.2.2 Automated Concept Mapping

The detailed manual mapping procedure, including extensive review by colleagues, is not feasible for data sets with more than 2,000 items. In order to make it possible to have access to the specific word properties offered by these data sets, we decided to set up a highly schematic workflow for automated concept mapping which is implemented in Python. It is dedicated to the data curation for large data sets. The basic idea of the mapping algorithm is to employ all previous mappings available from the Concepticon and order them by priority to check for direct matches against a specific data set.

The algorithm consists of four steps. First, all Concepticon mappings for a given language are assembled and ranked according to their frequency of occurrence throughout the concept lists linked to the Concepticon. In a second step, the algorithm iterates over each item in the target data set and checks if the item can be found in the list of assembled mappings. If this is the case, the item will be appended to the list of *potential mappings* for a given Concepticon concept set. In a third stage, the algorithm iterates over all concept sets for which a mapping was identified and selects one, according to the priority rank.

As an example, consider again the English word *wave* which occurs as elicitation gloss linked to two Concepticon concept sets, namely `918` WAVE and `3544` WAVE (VERB). While *wave* occurs as an elicitation gloss as many as 19 times in the Concepticon data, it has been linked 18 times to the noun reading (`918`) and only once to the verbal reading (`3544`). The verbal reading, in this case, is justified since the database in which the reading occurs explicitly deals with verbal meanings (Kibrik, 2012). Given that *wave* refers to the concept WAVE in the overwhelming majority of cases, the algorithm will ignore the verbal reading and link the word to the concept set `918` WAVE. To further increase the precision of this procedure, it is possible to add part-of-speech information, when available, to give preference in matches for the same part-of-speech.

While the concept mapping procedure provided by the Concepticon resource can be directly invoked from the command line, we figured that the automated workflow needs to be more neatly integrated into a data curation workflow since the results are no longer checked manually. For this reason, we established a new workflow that is based on Python scripts that can be invoked with the help of a new Python package: `pynorare` (List & Forkel, 2020). The package also automatizes the download of the data from dedicated URLs and the conversion of individual formats to the standardized tabular format that we employ for all data sets. With this workflow, each data set receives a custom Python script that can be called from the `pynorare` Python library. The script downloads the data set, unpacks it (if needed), pre-processes the data (if needed), and maps it automatically to the Concepticon.

By offering users to download the data themselves with the help of our Python library, we contribute to the *reusability* of the data. Since not all data sets have permissive licenses and some even explicitly prohibit creating derivative data sets of the original data, our approach gives users full access to these data sets. The data we link to the Concepticon often only consists of a small part of the original data and can, therefore, be freely redistributed under the assumption of fair use. Figure 2 (b) contrasts the automated mapping procedure with the manual and the semi-automated mapping procedure.

### 3.2.3 Semi-Automated Concept Mapping

There are a certain number of data sets that cannot be easily downloaded and treated with the mapping procedure described in the previous sections. Typical obstacles are their size (while most data sets come in the order of megabytes, dumps of big databases may amount to gigabytes and even more), their availability (apart from web-services), or their structure. While the former two are technical obstacles, the problem of the structure may pose a direct issue. Data sets on word frequencies list each word only one time, whereas some data sets for relations offer many candidates for the same concept. A search for the item *foot* on OmegaWiki[7], for example, gives us three possible senses, namely 'The part of a human's body below the ankle [...]', 'A unit of measurement equal to twelve inches [...]', and 'The lowest support of a structure.'

When linking the Concepticon concept set `1301` FOOT to OmegaWiki by hand, it is obvious that we would select the first over the second and the third option. It is perfectly

[7]The OmegaWiki project can be accessed with the following link: `https://omegawiki.org`

possible to search big databases like OmegaWiki manually in order to identify matches for the Concepticon concept sets, but we figured it would be easier to create a semi-automated approach in which we use software APIs provided by individual databases. This gives us the possibility to query the data and later manually decide which of the three or more possible matches should be the preferred one. This procedure notably differs from the normal data curation workflow used in the Concepticon project, since we do not identify the best mappings for a discrete concept list but rather find the closest counterparts for our Concepticon concept sets in large semantic databases. Figure 2 (b) contrasts the semi-automated mapping procedure with the automated and the manual mapping procedure.

### 3.2.4 Labeling Word and Concept Properties

Normalizing and linking data sets alone does not guarantee that word and concept properties can be easily compared across different data sets. To make a comparison between the data more convenient, we classified and labeled the word and concept properties in all data sets. Each column that includes information on a specific variable received multiple tags.

Our NoRaRe collection offers tags for several categories. The most general label is the 'NoRaRe' tag which indicates whether a data point belongs to the *norms*, *ratings*, or *relations* category. Furthermore, we specified the type of each column in that it was tagged with a keyword for properties such as *frequency*, *AoA* (age-of-acquisition), *semantic field*. Additionally, we tagged whether the column provides data on a subset, for instance, female/male participants, young/old participants. As of yet, 60 different labels for the various data sets in NoRaRe are available. The values were further grouped into structural categories such as *mean*, *logarithmic*, *percentage*. Along with the type and structural labels, each column received a language tag and a note with a detailed description of the column content, for example, which rating scale was used by the authors of the study. The distinct labels provide additional semantics for the different values in a data set. They are available for every data set in the NoRaRe collection (including data sets with a NoRaRe label in Concepticon).

The labeling allows for an easy retrieval of a specific data point connected to a Concepticon concept set. The information is conveniently accessed through a web application (`https://digling.org/norare/`). Here, one can search for a word and select an appropriate Concepticon concept set. Figure 4 illustrates a subset of the values for the concept set `906` TREE across three different data sets shown in the NoRaRe web application.

**Norms**, **Ratings**, and **Relations** for 906/TREE

| Dataset | Language | Structure | Type | Other | Value |
|---|---|---|---|---|---|
| Bond 2013 OMW | | numeric | semantic | in degree | 1 |
| Bond 2013 OMW | | numeric | semantic | out degree | 180 |
| Alonso 2015 AoA | | mean | AoA | | 2.62 |
| Alonso 2015 AoA | | numeric | AoA | minimum | 1 |
| Alonso 2015 AoA | | numeric | AoA | maximum | 6 |
| Brysbaert 2009 Frequency | | tokens | frequency | | 3315 |
| Brysbaert 2009 Frequency | | tokens | contextual diversity | | 1622 |

Figure 4: A screenshot of the NoRaRe web application (`https://digling.org/norare/`) illustrating the values for the Concepticon concept set `906` TREE across three different data sets (Alonso et al., 2015; Bond & Foster, 2013; Brysbaert & New, 2009).

# 4 Validation

In the previous sections, we introduced our NoRaRe collection for words and concepts that covers different languages. We also presented three workflows that can be used to expand the data collection. Additionally, we showed that the data can be conveniently accessed via a web application. The NoRaRe collection enables psychologists and linguists to compare and access a wealth of data from both research disciplines and for many languages.

To illustrate the current scope of our NoRaRe database, we report descriptive statistics of the data (Sect. 4.1), evaluate the efficiency of the data curation workflow (Sect. 4.2), and test the applicability of our data in three case studies (Sect. 4.3). The results of the studies replicate findings with existing data sets (Case study 1), prove the accuracy of the automated concept mapping (Case study 2), and show that genealogical different languages have similar word frequencies (Case study 3).

## 4.1 Descriptive Statistics of NoRaRe

With the help of our three workflows for test-driven data curation, we have assembled as many as 415 word and concept properties derived from 71 different data sets. Thirteen out of 71 reflect *norms* in the notion defined above, 38 reflect *ratings*, and 20 belong to our category of *relations*. Table 2 provides an overview of a small part of the data and also shows how many concepts we managed to link to our Concepticon concept sets.

The distribution of the concept identifiers across the 71 data sets is illustrated in Figure 5. The graph shows that most Concepticon identifiers occur in only a few data sets. Nevertheless, a large group of Concepticon identifiers is linked to 15 to 25 data sets. The most frequently occurring concept sets that are mapped to 52 up to 55 data sets are given in Table 1. Interestingly, almost half of them belong also to the most frequent concept sets in the concept lists curated in the Concepticon database, namely BONE, STAR, DOG, BIRD, SUN, and EYE.

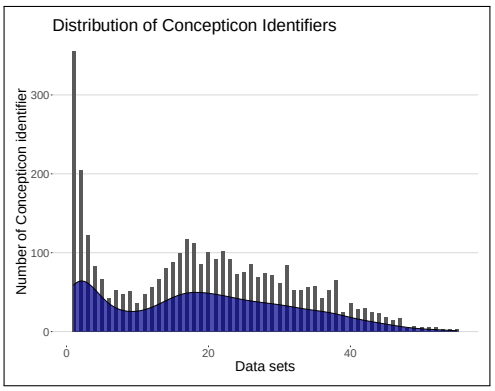


Figure 5: Distribution of Concepticon identifiers across the NoRaRe data sets. The x-axis gives the number of data sets in which the identifier occurs. The y-axis provides the number of Concepticon identifiers.

The numbers reported in this section illustrate that the NoRaRe collection offers a wide range of data sets across multiple structural types such as numeric, categorical, and relational data. Although we restrict the number of words in a given data set to the number of Concepticon concept sets, we provide the basis for a cross-linguistic comparison. Our data collection allows an in-depth study of word and concept properties across a variety of languages.

Table 1: The fifteen most common Concepticon concept sets occurring in 52 up to 55 NoRaRe data sets.

| Rank | ID | Concept set | Data sets |
|---|---|---|---|
| 1. | 615 | HORSE | 55 |
| 2. | 1248 | EYE | 55 |
| 3. | 1489 | CLOUD | 55 |
| 4. | 937 | BIRD | 54 |
| 5. | 1343 | SUN | 54 |
| 6. | 1663 | BED | 54 |
| 7. | 227 | FISH | 53 |
| 8. | 1476 | CHAIR | 53 |
| 9. | 2009 | DOG | 53 |
| 10. | 978 | WAVE | 52 |
| 11. | 1223 | HEART | 52 |
| 12. | 1297 | LEG | 52 |
| 13. | 1352 | KNIFE | 52 |
| 14. | 1394 | BONE | 52 |
| 15. | 1430 | STAR | 52 |

### 4.2 Data Curation Workflow

The data curation workflows that we established to add data on norms, ratings, and relations to Concepticon proved to be very effective. With the pre-defined workflow for adding concept lists to Concepticon, we were able to add 15 new data sets with small numbers of words ($<$ 2,000 items) within seven months. The new data sets were included in the release of Concepticon Version 2.4.0-rc.1 in July, 2020 (List et al., 2020).[8] For the larger data sets ($>$ 2,000 items) and data from online resources like Wikidata, we created a new GitHub repository `concepticon/norare-data`. The first commit to this repository was on March 31$^{st}$, 2020 and since then we added 42 data sets with the automated and 5 data sets with the semi-automated workflow. The total number of 48 data sets uploaded within four months demonstrates that the data collection can be expanded in a short amount of time (in our case, approx. 10-15 data sets per month). Version 0.1 of the database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (NoRaRe) was released on July 23$^{rd}$, 2020 (Tjuka, Forkel, & List, 2020).

The Python package `pynorare` was established and developed parallel to the NoRaRe collection. It was expanded and adapted to account for the challenges of bringing completely different data set structures in a standardized format. The first release of `pynorare` v0.1.0 was on July 13$^{th}$, 2020. The next version update which included more tests for the data curation was uploaded on July 21$^{st}$ (List & Forkel, 2020).

The timeline of the releases for the NoRaRe database and the associated Python package shows that our workflows can be applied, constantly improved, and expanded. The longevity of the Concepticon project ensures regularly updated data. The Concepticon database allows for an advancement in cross-linguistic comparison and the development of features such as the NoRaRe collection. Therefore, it adds value to research disciplines like psychology by offering deliberately curated data on various aspects of word and concept properties.

[8]The Concepticon resource already included data on norms and relations in earlier versions. We added those nine data sets to the NoRaRe collection.

Table 2: Subset of the NoRaRe data. The table gives information on the language for which the data was collected, the data type, the original item number, and the number of matches to the Concepticon concept sets.

| | **Language** | **Types** | **Items** | **Matches** |
|---|---|---|---|---|
| **Norms** | | | | |
| Cai and Brysbaert (2010) | Chinese | frequency | 99,123 | 1,644 |
| Ferrand et al. (2010) | French | reaction time | 38,840 | 1,372 |
| Brysbaert et al. (2011) | German | frequency | 190,500 | 1,291 |
| Cuetos et al. (2011) | Spanish | frequency | 94,338 | 1,088 |
| Alonso, Fernandez, and Díez (2011) | Spanish | frequency | 67,979 | 1,016 |
| Tsang et al. (2018) | Chinese | reaction time | 25,156 | 827 |
| Keuleers et al. (2010) | Dutch | frequency | 437,503 | 640 |
| González-Nosti et al. (2014) | Spanish | reaction time | 2,765 | 554 |
| Mandera, Keuleers, Wodniecka, and Brysbaert (2015) | Polish | frequency | 377,843 | 215 |
| **Ratings** | | | | |
| Lynott et al. (2020) | English | sensorimotor | 40,000 | 2,437 |
| Brysbaert, Mandera, McCormick, and Keuleers (2019) | English | prevalence | 62,000 | 2,414 |
| Kuperman et al. (2012) | English | age-of-acquisition | 30,000 | 2,351 |
| Stadthagen-González et al. (2017) | Spanish | valence, arousal | 14,031 | 932 |
| Moors et al. (2013) | Dutch | age-of-acquisition, affective* | 4,300 | 444 |
| Łuniewska et al. (2019) | Diverse | age-of-acquisition | 299 | 284 |
| Verheyen, De Deyne, Linsen, and Storms (2020) | Dutch | age-of-acquisition, lexicosemantic, distributional, affective | 1,000 | 206 |
| Imbir (2016) | Polish | age-of-acquisition, affective | 4,900 | 159 |
| Kapucu et al. (2018) | Turkish | discrete emotions, affective | 2,031 | 75 |
| **Relations** | | | | |
| Wu, Nicolai, and Yarowsky (2020) | Global | core vocabulary | 10,000 | 2,460 |
| Matisoff (2015) | Sino-Tibetan (Global) | etymology | 6,431 | 2,159 |
| Starostin (2000) | Diverse | sense relation | 7,095 | 2,020 |
| Rzymski et al. (2020) | Global | polysemy | 1,624 | 1,624 |
| Bond and Foster (2013) | English | WordNet | 4,960 | 1,309 |
| Dellert and Buch (2018) | Eurasian | basicness, stability | 1,016 | 955 |
| Hill, Reichart, and Korhonen (2015) | English | semantic similarity | 999 | 524 |
| Calude and Pagel (2011) | Diverse | stability, frequency | 200 | 200 |
| Baroni and Lenci (2011) | English | semantic similarity | 200 | 140 |

*The term 'affective' summarizes different variables such as arousal, valence, dominance, concreteness, imageability.

### 4.3 Data Applicability

The NoRaRe database includes a broad range of norms, ratings, and relations. To test whether the data sets can be applied for investigations on words and concepts, we conducted three case studies. Each study uses a combination of different data sets to evaluate the validity of our database. In our GitHub repository, we provide a Python and several R scripts as examples to easily perform correlations between the data sets in NoRaRe.

#### 4.3.1 Case Study 1: Replication of existing Findings

In the first case study, we identified two similar data sets by using the column labels defined in the NoRaRe database. The data sets were chosen to replicate existing results. The Concepticon includes more than 3,000 concepts. In studies with more or different items than concept sets in Concepticon, a part of the data is not mapped and the number of items is reduced. Therefore, we computed the correlation of three variables across two data sets to see whether the results are still significant.

The NoRaRe collection was filtered by variables to find data sets with the same norms, ratings, and relations. We found several data sets that included ratings on arousal, valence, and dominance. To ensure that the data could be equally comparable, we identified the data sets with ratings in the same language and the same rating scale. The search was easily carried out because the column content of each data set is labeled within the NoRaRe workflow. We selected two data sets for our study that provide ratings of English words on a 9-point scale for arousal, valence, and dominance: Warriner, Kuperman, and Brysbaert (2013) and Scott et al. (2019). Both data sets were prepared with the automated mapping workflow.

The original data set in Warriner et al. (2013) consisted of 13,915 English words. The automated mapping algorithm found 2,067 matches between the words in Warriner et al. (2013) and the Concepticon identifiers. In the case of Scott et al. (2019), the original data set included 5,500 words and there were 1,459 matches with Concepticon identifiers. The overlap between both data sets in the NoRaRe database amounted to 1,397 concept sets (the overlap between the original data sets was 4,073 words). Table 3 shows the results of the correlations between the ratings for arousal, valence, and dominance in Warriner et al. (2013) and Scott et al. (2019). For each variable, the correlation (Pearson coefficients) was highly significant ($p < .00001$). The distribution of the ratings in Warriner et al. (2013) and Scott et al. (2019) across the 9-point scale for the 1,397 Concepticon concept sets is illustrated in Figure 6.

Table 3: Pearson coefficients for the variables arousal, valence, and dominance (see text). The values in parentheses indicate the original numbers, reported in Scott et al. (2019).

| **Overlap** | **Arousal** | **Valence** | **Dominance** |
|---|---|---|---|
| 1,397 (4,073) | 0.57 (0.62) | 0.92 (0.93) | 0.66 (0.69) |

The additional information for each data set in the NoRaRe database facilitates access to relevant content. The labels of data set columns provide the basis for an effortless comparison between the variety of data and allow a fast identification of compatible variables across different data sets. The results of the correlation between Warriner et al. (2013) and Scott et al. (2019) replicate the findings reported in Scott et al. (2019). Thus, the reduction of the items due to the restricted number of Concepticon concept sets still ensures the comparability of data sets. This result may not hold for all data sets in the NoRaRe database, but the Concepticon resource is growing steadily and more concept identifiers are added with each release.

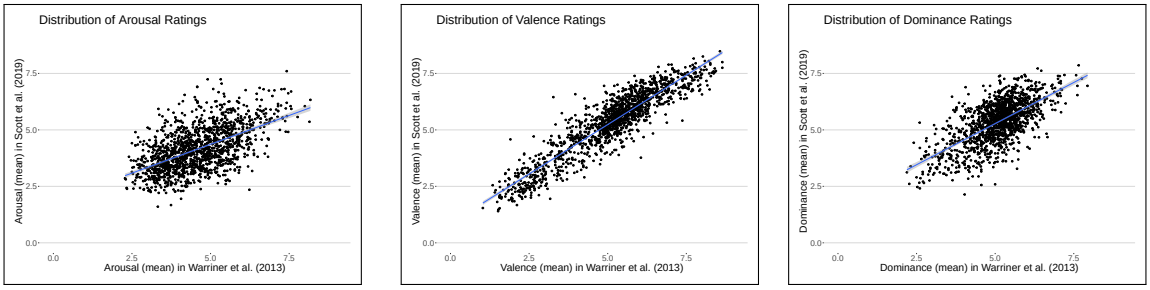


Figure 6: Distribution of the mean values for arousal (left), valence (middle), and dominance (right) in Warriner et al. (2013) and Scott et al. (2019) for 1,397 Concepticon concept sets.

### 4.3.2 Case Study 2: Comparison of Concept Mappings

We used three workflows to link various data sets on norms, ratings, and relations to the Concepticon identifiers: *manual*, *automated*, and *semi-automated* mapping (for a detailed description of the workflows see Sect. 3.2). The main difference between the manual and both automated workflows lies in the check for accuracy of the mappings. In the manual workflow, the link between a given word and a Concepticon concept set is manually examined by a person who is familiar with the structure of Concepticon and a team of reviewers who discuss ambiguous cases. The automated workflow, on the other hand, uses an inherent rating system of the similarity between a word and the matches to the Concepticon identifiers without human intervention. In the second case study, we tested whether the mappings of both workflows are equal in their quality.

By searching the column labels provided in the NoRaRe database, we identified similar data sets that include information for the same variable and language. In addition, the identifiers for the data sets prepared with the manual versus automated workflow differ so that they can be easily distinguished. For the present study, we chose two data sets that offer ratings of English words on a 7-point scale for sensory modality (auditory, haptic, gustatory, olfactory, visual): Lynott and Connell (2013) and Lynott et al. (2020). The former was added to NoRaRe with the manual workflow, the latter with the automated workflow.

In the data of Lynott and Connell (2013), we linked 147 words to Concepticon identifiers from the original number of 400 items. The original data set in Lynott et al. (2020) comprised 40,000 English words and the algorithm detected 2,437 correspondences to Concepticon identifiers. The overlap between both data sets was 139 Concepticon concept sets. The results of the correlation between the five sensory modality ratings are shown in Table 4. The correlations (Pearson coefficients) were highly significant ($p < .00001$) across all five variables. Figure 7 illustrates the distribution of the ratings in Lynott and Connell (2013) and Lynott et al. (2020) across a 7-point scale for the 139 concept sets.

Table 4: Pearson coefficients for the sensorimotor variables auditory, gustatory, haptic, olfactory, and visual (see text). Abbreviations: AUD auditory; GUS gustatory; HAP haptic; OLF olfactory; VIS visual.

| **Overlap** | **AUD** | **GUS** | **HAP** | **OLF** | **VIS** |
|---|---|---|---|---|---|
| 139 | 0.87 | 0.95 | 0.88 | 0.91 | 0.86 |

The facilitated access to relevant information about the content of a given data set is provided by the labels in the NoRaRe database. We identified two data sets that were prepared with the manual and automated workflow, respectively. The accuracy of the automated mapping procedure seems to be as good as the manual mapping. The results indicate that both

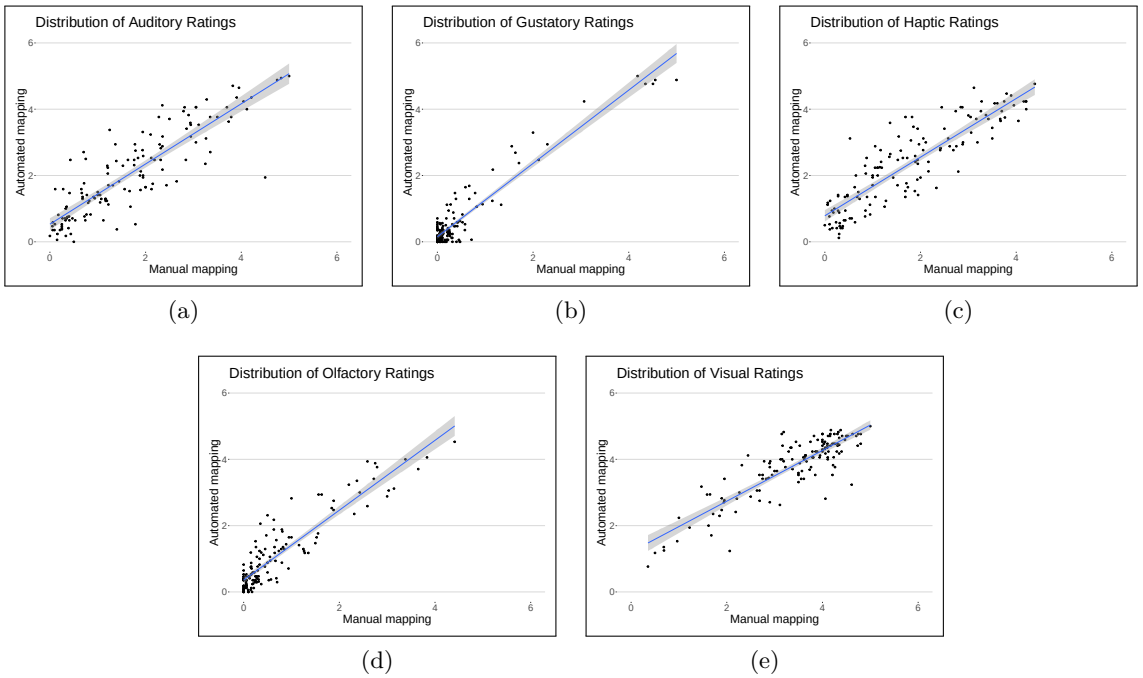


Figure 7: Distribution of the mean values for the five sensory modalities: auditory (a), gustatory (b), haptic (c), olfactory (d), and visual (e) in Lynott and Connell (2013) and Lynott et al. (2020) for 139 Concepticon concept sets.

workflows can be equally applied for the different data sets. Nevertheless, we will continue to add data sets with the established Concepticon workflow because it ensures the quality of the mapping algorithm although the automated workflow is faster. This is especially important because many data sets in psychology do not provide information on specific word meanings.

### 4.3.3 Case Study 3: Cross-Linguistic Comparison

One main goal of the NoRaRe database is to enable cross-linguistic studies of norms, ratings, and relations. Linking words and concepts in a data set to Concepticon concept sets enables us to compare the same variable across different languages. In the third case study, we present a cross-linguistic comparison of word frequencies in English, Spanish, and Chinese. The study examines the question of whether the frequencies of words in corpora of subtitles from movies and TV-series differ across those languages.

We searched for the respective SUBTLEX data sets for the three languages in the NoRaRe database and chose the logarithmic frequency norms to investigate our question. The data sets for English, Spanish, and Chinese were taken from SUBTLEX-US (Brysbaert & New, 2009), SUBTLEX-ESP (Cuetos et al., 2011), and SUBTLEX-CH (Cai & Brysbaert, 2010), respectively. All data sets were added to NoRaRe with the automated workflow.

The mappings of the SUBTLEX-US (English) data to the Concepticon identifiers resulted in 2,329 matches. In the SUBTLEX-ESP (Spanish) data, 1,088 Concepticon identifiers were mapped. The correspondence of the words in SUBTLEX-CH (Chinese) amounted to 1,644 Concepticon identifiers. The data sets were correlated with one another on the basis of the $\log_{10}$ word frequencies. The Pearson coefficients are shown in Table 5. The correlations between each language pair (English–Spanish, English–Chinese, Spanish–Chinese) were significant ($p < .0001$). The distribution of the word frequencies in each corpus is illustrated in Figure 8.

Table 5: Pearson coefficients for the variable $\log_{10}$ word frequency (see text).

| **Languages** | **Overlap** | **$\log_{10}$ Frequency** |
|---|---|---|
| English – Spanish | 995 | 0.64 |
| English – Chinese | 1313 | 0.55 |
| Spanish – Chinese | 722 | 0.53 |

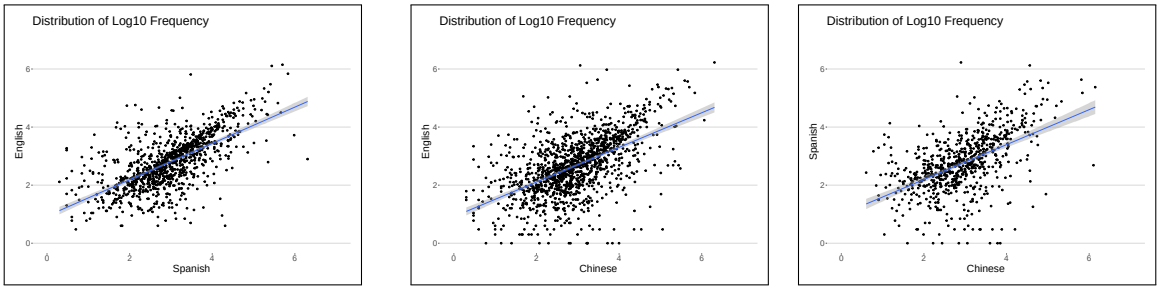


Figure 8: Distribution of the $\log_{10}$ word frequencies across the three languages: English–Spanish (left), English–Chinese (middle), and Spanish–Chinese (right). The data was taken from Brysbaert and New (2009), Cuetos et al. (2011), and Cai and Brysbaert (2010).

The possibility to search and identify data sets with the same content provided by the NoRaRe interface greatly facilitates the identification of comparable data. Furthermore, since Concepticon concept sets serve as multilingual comparative concepts, all data sets which have been linked to the Concepticon can be directly compared with each other, no matter if they reflect different languages or constructs in different disciplines.

The results of our study indicate that the words occurring in subtitles seem to have similar frequencies across three diverse languages, namely English, Spanish, and Chinese. The slightly higher Pearson coefficient of 0.64 for the comparison between English and Spanish might be due to a shared cultural background or due to the genetic closeness of the languages. The NoRaRe database offers manifold opportunities for future studies across diverse language varieties.

## 5 Discussion and Conclusion

We set out to establish a collection of norms, ratings, and relations for words and concepts across multiple languages. To achieve our goal, we implemented three workflows that unify diverse data sets and allow for a facilitated expansion and curation of the data collection. Additionally, we provide a web application that can be used to compare and access the data conveniently. The result is a cross-linguistic database of norms, ratings, and relations: NoRaRe.

Psychology and linguistics offer a diverse set of data on words and concepts. Although we provide workflows to standardize data sets, we hope that our description of the FAIR data principles (Wilkinson et al., 2016) inspires researchers to prepare their data sets in a more sufficient way. Other scholars are also invited to contribute their data sets to the NoRaRe database via GitHub. The comparability of data sets is especially important to fill gaps in the records because, by now, there is only a small number of languages for which norms and ratings have been consistently documented in psychology. In addition, our framework and tools facilitate the comparison of words and concepts across different properties that have

been investigated in the literature.

The workflows we established are based on our experience with the Concepticon project (List et al., 2016). The current team of Concepticon editors consists of nine linguists who ensure a high-quality resource for the study of comparative concepts. The manual workflow for adding new concept lists has already been proven valuable and many researchers have contributed to the Concepticon since its first release.[9] With the newly developed automated and semi-automated workflows, new data sets can easily be added and the test-driven data curation guarantees consistency of the data. The quality checks provided by the integrated tests in the Python libraries `pyconcepticon` and `pynorare` are supported by the review process on GitHub. We, therefore, offer a data curation workflow to constantly expand the data on norms, ratings, and relations. The regular updates of the Concepticon project are a huge advantage compared to other databases which are often set up at one point in time and are not curated afterward (e.g., Wilson, 1988; Winter et al., 2017)

The NoRaRe collection offers a wide range of data sets for word and concept properties. The large amount of data (71 data sets) distinguishes our collection from other databases that provide less data variety. Most databases focus on collecting norm data for one language alone (e.g., Baayen, Piepenbrock, & Gulikers, 1996; Heister et al., 2011) and they offer more content per data set (e.g., Buchanan et al., 2019a). The diversity of languages is a unique asset of the NoRaRe database. It allows researchers to investigate word and concept properties across diverse languages such as Chinese, Hebrew, Russian, or Icelandic. The broad comparison of different languages is facilitated by the link to the Concepticon concept sets, but is also limited to the number of available concepts. In contrast, databases like Buchanan et al. (2019a) offer a high number of words but without the possibility of a cross-linguistic comparison.

For a facilitated access to the vast amount of data in NoRaRe, each word or concept property documented in a given data set receives a unique label. These labels provide an additional semantic layer to the data structure. Each value in our database has a tag for the NoRaRe categories (norms, ratings, relations), the data type (i.e., frequency, AoA), the structural basis (i.e., logarithmic, mean), and the language. The information can be conveniently retrieved via the web application or searched in the spreadsheet underlying the application. The labels are valuable because they do not only guarantee that the data are comparable but also findable. In addition, researchers may well come across data that they were not aware of in advance. We hope that future users of the database will also point us to data sets that are still missing in our collection.

Another main advantage of our database is the link to the Concepticon concept sets. The Concepticon concept sets are carefully selected by linguists with the premise that the concepts are comparable across languages (Haspelmath, 2010). The innovation of our approach is that we link data sets on norms, ratings, and relations from multiple languages to the hand-curated concept sets in Concepticon. The concept sets are already linked to a variety of languages, which play a crucial role in language documentation, such as Spanish, Chinese, and Russian. The mappings between a given elicitation gloss and a Concepticon concept set is established by adding a concept list to Concepticon. The mappings allow a cross-linguistic comparison of multiple word and concept properties.

The validity of our workflows and applicability of our data was illustrated in three case studies. The results of the first case study showed that a correlation between two data sets can be replicated despite the reduced item number due to the limited Concepticon concept sets. In the second case study, we demonstrated that the accuracy of the automated workflow was comparable with the manual workflow. It is important that the automated workflow delivers similar results to manual mapping so that the data can be compared regardless of the processing method. The third case study demonstrated that the NoRaRe database can be

[9] For a list of Concepticon contributors see `https://concepticon.clld.org/contributors`.

used for a cross-linguistic comparison. Studies with the same variable in different languages offer valuable insights. We investigated word frequencies in subtitles across English, Spanish, and Chinese. The slight differences in the correlations between the three language pairs could be due to the diverse cultural backgrounds of the native speakers or the history of the languages (i.e., Chinese versus English and Spanish). The findings could also be a result of the diversity in the structure of the mental lexicon which is based on language-internal settings, especially word reuse. For example, emotion semantics vary across languages (Jackson et al., 2019) which indicates that we might find cross-linguistic correlations in other semantic fields as well.

The linking of norms, ratings, and relations of words and concepts across multiple languages comes with many challenges. Nevertheless, we present a large collection of data from different languages which are essentially comparable with each other. The NoRaRe database offers convenient access to the different data types we have assembled. We are optimistic that our data curation workflows are long-lasting and applicable to other researchers. The workflows and data are publicly curated on GitHub and regularly archived on Zenodo. Interested researchers are invited to test and use our database for studies on interrelated questions.

## Availability

The database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (NoRaRe) presented in this article is curated on GitHub (`https://github.com/concepticon/norare-data`) and archived with Zenodo (`http://doi.org/10.5281/zenodo.3957681`). The Concepticon database (List et al., 2020) is also curated on GitHub (`https://github.com/concepticon/concepticon-data`) and archived on Zenodo (`https://doi.org/10.5281/zenodo.3954155`).

R-Scripts that were used to produce the plots for the three case studies are available from the NoRaRe collection (on GitHub `concepticon/norare-data`, folder `examples`). The GitHub repository also provides detailed instructions on the installation of the curation software and the details of the data curation process. The Python package used for the data curation workflow can also be found on GitHub (`https://github.com/concepticon/pynorare`). The `pynorare` package is stored on Zenodo (`https://doi.org/10.5281/zenodo.3955051`) as well as PyPi (`https://pypi.org/project/pynorare/`).

For convenient access to the NoRaRe database, we offer a web application: `https://digling.org/norare/`

### Acknowledgments

AT and JML initiated the study, developed the specific data curation workflow, and wrote a first manuscript draft. RF and JML wrote the Python code to support the workflow. AT and JML prepared data for automated data curation. AT prepared data for manual and semi-automated data curation, labeled all data sets, created the figures, and conducted the analysis for the case studies. All authors revised the draft and agree with the final version of the manuscript. AT 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. JML was funded by the ERC Starting Grant 715618 Computer-Assisted Language Comparison (`https://digling.org/calc/`).

## Conflict of Interest

The authors declare that they have no conflict of interest.

## Open Practices Statement

The database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (NoRaRe) is available on GitHub (`https://github.com/concepticon/norare-data`) and archived with Zenodo (`http://doi.org/10.5281/zenodo.3957681`). The Python library `pynorare` submitted with this paper is also curated on GitHub (`https://github.com/concepticon/pynorare`) and archived with Zenodo (`https://doi.org/10.5281/zenodo.3955051`) as well as PyPi (`https://pypi.org/project/pynorare/`).

## Electronic supplementary material

The electronic supplementary material includes a list of data sets available in the NoRaRe database at the time of the submission of the article.

## References


Alonso, M. Á., Fernandez, A., & Díez, E. (2011). Oral frequency norms for 67,979 Spanish words. *Behavior Research Methods*, *43*(2), 449-458.

Alonso, M. Á., Fernandez, A., & Díez, E. (2015). Subjective age-of-acquisition norms for 7,039 Spanish words. *Behavior Research Methods*, *47*(1), 268-274. doi: https://doi.org/10.3758/s13428-014-0454-2

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *The CELEX lexical database.* Philadelphia: University of Pennsylvania.

Bank, S., & Forkel, R. (2018). *cldf/csvw: CSV on the Web.* Geneva: Zenodo. doi: https://doi.org/10.5281/zenodo.1325040

Baroni, M., & Lenci, A. (2011). *BLESS: Baroni & Lenci's evaluation of semantic similarity.* Retrieved from `https://sites.google.com/site/geometricalmodels/shared-evaluation`

Bond, F., & Foster, R. (2013). Linking and extending an open multilingual WordNet. In H. Schuetze, P. Fung, & M. Poesio (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (p. 1352-1362). Sofia, Bulgaria: Association for Computational Linguistics.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (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. doi: https://doi.org/10.1027/1618-3169/a000123

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467-479. doi: https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., & New, B. (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. doi: https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904-911. doi: https://doi.org/10.3758/s13428-013-0403-5

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019a). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*(4), 1849-1863.

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019b). LAB: Linguistic Annotated Bibliography – a searchable portal for normed database information. *Behavior Research Methods*, *51*(4), 1878-1888.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *Plos ONE*, *5*(6), 1-8. doi: https://doi.org/10.1371/journal.pone.0010729

Calude, A. S., & Pagel, M. D. (2011). How do we use language? Shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society B*(366), 1101-1107. doi: https://doi.org/10.1098/rstb.2010.0315

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, *32*(2), 133-143.

Dellert, J., & Buch, A. (2018). A new approach to concept basicness and stability as a window to the robustness of concept list rankings. *Language Dynamics and Change*, *8*(2), 157-181.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., … Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488-496. doi: https://doi.org/10.3758/BRM.42.2.488

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., … Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, *5*(1), 1-10.

Forkel, R., Rzymski, C., & List, J.-M. (2019). *concepticon/pyconcepticon: pyconcepticon 2.3.0*. Geneva: Zenodo. doi: https://doi.org/10.5281/zenodo.3516955

François, A. (2008). Semantic maps and the typology of colexification. In M. Vanhove (Ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations* (Vol. 106, p. 163). John Benjamins Publishing.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., … Conway, B. R. (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. doi: https://doi.org/10.1073/pnas.1619666114

González-Nosti, M., Barbón, A., Rodríguez-Ferreiro, J., & Cuetos, F. (2014). Effects of the psycholinguistic variables on the lexical decision task in Spanish: A study with 2,765 words. *Behavior Research Methods*, *46*(2), 517-525. doi: https://doi.org/10.3758/s13428-013-0383-5

Guasch, M., Boada, R., Ferré, P., & Sánchez-Casas, R. (2013). NIM: A web-based Swiss army knife to select stimuli for psycholinguistic studies. *Behavior Research Methods*, *45*(3), 765-771.

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

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). dlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*(1), 10-20.

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.

Imbir, K. K. (2016). Affective norms for 4900 Polish words reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability, and age of acquisition. *Frontiers in Psychology*, *7*, 1-18. doi: https://doi.org/10.3389/fpsyg.2016.01081

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Mucha, P. J., Forkel, R., … Lindquist, K. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517-1522. doi: https://doi.org/10.1126/science.aaw8160

Jackson, J. C., Watts, J., List, J.-M., Drabble, R., & Lindquist, K. (2020). From text to thought: How analyzing language can advance psychological science. *PsyArxiv*, *0*(0), 1-46. doi: https://doi.org/10.31234/osf.io/qat4r

Jones, D. (2010). A WEIRD view of human nature skews psychologists studies. *Science*, *328*(5986), 1627-1627. doi: https://doi.org/10.1126/science.328.5986.1627

Kapucu, A., Kılıç, A., Özkılıç, Y., & Sarıbaz, B. (2018). Turkish emotional word norms for arousal, valence, and discrete emotion categories. *Psychological Reports*, *0*(0), 1-22. doi: https://doi.org/10.1177/0033294118814722

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643-650. doi: https://doi.org/10.3758/BRM.42.3.643

Kibrik, A. A. (2012). Toward a typology of verbal lexical systems: A case study in Northern Athabaskan. *Linguistics*, *50*(3), 495–532. doi: https://doi.org/10.1515/ling-2012-0017

Kiss, G. R., Armstrong, C., & Milroy, R. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literary studies.* Edinburgh: Edinburgh University Press.

Kuperman, V., Stadthagen-González, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990. doi: https://doi.org/10.3758/s13428-012-0210-4

List, J.-M., Cysouw, M., & Forkel, R. (2016). Concepticon. A resource for the linking of concept lists. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (p. 2393-2400). Luxembourg: European Language Resources Association (ELRA). Retrieved from `http://www.lrec-conf.org/proceedings/lrec2016/summaries/127.html`

List, J.-M., & Forkel, R. (2020). *concepticon/pynorare: pynorare 0.2.0.* Geneva: Zenodo. doi: https://doi.org/10.5281/zenodo.3955051

List, J.-M., Greenhill, S. J., Anderson, C., Mayer, T., Tresoldi, T., & Forkel, R. (2018). CLICS². An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats. *Linguistic Typology*, *22*(2), 277-306. Retrieved from `http://clics.clld.org` doi: https://doi.org/10.1515/lingty-2018-0010

List, J.-M., Rzymski, C., Greenhill, S., Schweikhard, N., Pianykh, K., Tjuka, A., … Forkel, R. (2020). *Concepticon. A resource for the linking of concept lists (Version 2.4.0-rc.1).* Jena: Max Planck Institute for the Science of Human History. Retrieved from `https://concepticon.clld.org/` doi: https://doi.org/10.5281/zenodo.3954155

Łuniewska, M., Haman, E., Armon-Lotem, S., Etenkowski, B., Southwood, F., Anđelković, D., … Özlem Ünal Logacev (2016). Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order of words? *Behavior Research Methods*, *48*(3), 1154-1177. doi: https://doi.org/10.3758/s13428-015-0636-6

Łuniewska, M., Wodniecka, Z., Miller, C. A., Smolík, F., Butcher, M., Chondrogianni, V., … Haman, E. (2019). Age of acquisition of 299 words in seven languages: American English, Czech, Gaelic, Lebanese Arabic, Malay, Persian and Western Armenian. *Plos ONE*, *14*(8). doi: https://doi.org/10.1371/journal.pone.0220611

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods*, *45*(2), 516–526.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for

40,000 English words. *Behavior Research Methods*, *52*, 1271–1291. doi: https://doi.org/10.3758/s13428-019-01316-z

Maciejewski, G., & Klepousniotou, E. (2016). Relative meaning frequencies for 100 homonyms: British eDom norms. *Journal of Open Psychology Data*, *4*(e6), 1–5. doi: http://dx.doi.org/10.5334/jopd.28

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). SUBTLEX-PL: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, *47*(2), 471-483. doi: https://doi.org/10.3758/s13428-014-0489-4

Matisoff, J. A. (2015). *The Sino-Tibetan Etymological Dictionary and Thesaurus.* Department of Linguistics at the University of California, Berkeley. Retrieved from `https://stedt.berkeley.edu/`

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A.-L., … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*(1), 169-177. doi: https://doi.org/10.3758/s13428-012-0243-8

Riegel, M., Wierzba, M., Wypych, M., Żurawski, Ł., Jednoróg, K., Grabowska, A., & Marchewka, A. (2015). Nencki affective word list (NAWL): the cultural adaptation of the Berlin affective word list-reloaded (BAWL-R) for Polish. *Behavior Research Methods*, *47*(4), 1222-1236.

Rzymski, C., Tresoldi, T., J.Greenhill, S., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., … List, J.-M. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, *7*. Retrieved from `https://clics.clld.org/` doi: https://doi.org/10.1038/s41597-019-0341-x

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258-1270.

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In S. Singh & S. Markovitch (Eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligenze* (p. 4444-4451). Palo Alto: AAAI.

Stadthagen-González, H., Imbault, C., Pérez-Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, *49*(1), 111-123. doi: https://doi.org/10.3758/s13428-015-0700-2

Starostin, S. A. (2000). *The STARLING database program.* Moscow: RGGU. Retrieved from `http://starling.rinet.ru`

Tadmor, U. (2009). Loanwords in the world's languages. In M. Haspelmath & U. Tadmor (Eds.), (p. 55-75). Berlin and New York: Walter de Gruyter.

Tennison, J. (n.d.). *CSV on the Web: A primer. W3C Working Group Note 25 February 2016* (Tech. Rep.). W3C. Retrieved from `http://www.w3.org/TR/tabular-data-primer/`

Tjuka, A. (2020). Adding concept lists to Concepticon: A guide for beginners. *Computer-Assisted Language Comparison in Practice*, *3*(2). Retrieved from `https://calc.hypotheses.org/2225`

Tjuka, A., Forkel, R., & List, J.-M. (2020). *Database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (Version 0.1).* Jena: Max Planck Institute for the Science of Human History. doi: http://doi.org/10.5281/zenodo.3957681

Tresoldi, T. (2019). Using pyconcepticon to map concept lists. *Computer-Assisted Language Comparison in Practice*, *2*(4). Retrieved from `https://calc.hypotheses.org/1820`

Tsang, Y.-K., Huang, J., Lui, M., Xue, M., Chan, Y.-W. F., Wang, S., & Chen, H.-C. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods*, *50*(5), 1763-1777. doi: https://doi.org/10.3758/s13428-017-0944-0

Verheyen, S., De Deyne, S., Linsen, S., & Storms, G. (2020). Lexicosemantic, affective, and

distributional norms for 1,000 Dutch adjectives. *Behavior Research Methods*, *52*, 1108–1121. doi: https://doi.org/10.3758/s13428-019-01303-4

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191-1207.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(160018), 1-9. doi: https://doi.org/10.1038/sdata.2016.18

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6-10.

Winter, B., Wedel, A., & List, J.-M. (2017). *The Language Goldmine*. Jena: Max Planck Institute for the Science of Human History. Retrieved from `http://languagegoldmine.com/`

Wu, W., Nicolai, G., & Yarowsky, D. (2020). Multilingual dictionary based construction of core vocabulary. In N. Calzolari et al. (Eds.), *Proceedings of The 12th Language Resources and Evaluation Conference* (p. 4211-4217). Marseille: European Language Resources Association. Retrieved from `https://www.aclweb.org/anthology/2020.lrec-1.519`