Curating and extending data for language comparison in Concepticon and NoRaRe [version 1; peer review: awaiting peer review]

Annika Tjuka, Robert Forkel, Johann-Mattis List

Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Saxony, 04103, Germany

Abstract
Over the past decade, there have been several attempts to standardize cross-linguistic datasets. Since language comparison is a notoriously difficult endeavor, it requires tools that facilitate standardization and are convenient to use. The Concepticon is based on a toolkit provided for cross-linguistic comparison and offers a reference catalog for comparable concepts that appear in concept lists. While curating the Concepticon, we found that a variety of studies in distinct research fields collected information on word properties. However, until recently, no resource existed that contained these data to enable the comparison of the different word properties across languages. This gap was filled by the Database of Norms, Ratings, and Relations (NoRaRe), which is an extension of the Concepticon. Here, we present the major release of both resources - Concepticon Version 3.0 and NoRaRe Version 1.0 - which represents an important step in our data development. We show that extending and adapting the data curation workflow in Concepticon to NoRaRe is useful for the standardization of cross-linguistic datasets. In addition, combining datasets from different research fields enables studies grounded in language comparison. Concepticon and NoRaRe include lexical data for various languages, tools for test-driven data curation, and the possibility for data reuse. The first major release of NoRaRe is also accompanied by a new web application that allows convenient access to the data.

Keywords
Cross-linguistic database, Test-driven data curation, Lexical data, Word properties, Language comparison
Corresponding authors: Annika Tjuka (annika_tjuka@eva.mpg.de), Johann-Mattis List (mattis_list@eva.mpg.de)

Author roles: Tjuka A: Conceptualization, Data Curation, Investigation, Project Administration, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Forkel R: Conceptualization, Data Curation, Methodology, Software, Validation, Writing – Review & Editing; List JM: Conceptualization, Data Curation, Funding Acquisition, Software, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 715618). AT was additionally supported by a fellowship 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.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Tjuka A et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Tjuka A, Forkel R and List JM. Curating and extending data for language comparison in Concepticon and NoRaRe [version 1; peer review: awaiting peer review] Open Research Europe 2022, 2:141
https://doi.org/10.12688/openreseurope.15380.1

First published: 16 Dec 2022, 2:141 https://doi.org/10.12688/openreseurope.15380.1
Plain language summary

There are more than 7,000 spoken and signed languages in the world today. They vary in their vocabulary and structure their lexicon differently to some degree. Standardizing this diversity is a challenge for researchers interested in language comparisons. Therefore, we use a set of tools that facilitate the standardization of data so that they become comparable. Using these tools, we created a database with a list of words that appear in studies by historical linguists who want to find out how languages are related to each other. Our goal was to create a catalog of standardized words and links to the lists in which they appear. To this end, we reformatted the available data into a simple text file format to make them comparable across languages. This database is called Concepticon. When we added metadata that described various properties of the words to the Concepticon, we noticed that a wealth of additional data existed. These data provide information about when a particular word was acquired, how often it changes its meaning over time, or which words it is associated with. For this reason, we decided to create a separate resource for these types of data: the Database of Norms, Ratings, and Relations (NoRaRe). NoRaRe is an extension of Concepticon and uses a customized workflow to add new data. Here, we introduce the most recent release of Concepticon Version 3.0 and NoRaRe 1.0. We illustrate the content of the two resources, our approach to adding more data, and present a web application that allows for convenient exploration of the data.

Introduction

The comparison of languages is made possible by standardizing data from various sources. To facilitate this standardization, we need tools to help systematically unify the data and provide them in a FAIR format, i.e., the data need to be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). Striving for this standard is especially difficult when dealing with linguistic data since languages vary greatly and language scientists choose to structure their data differently. For this reason, a community-led initiative developed the Cross-Linguistic Data Formats (CLDF, Forkel et al., 2018) which provide specifications on how to format a given dataset to comply with the FAIR principles (Wilkinson et al., 2016). The CLDF format specifically targets interoperability and reusability while the storage of the data on Zenodo (zenodo.org) accounts for findability and accessibility. The advantage of this framework is that cross-linguistic data from diverse languages becomes comparable by converting them into a standardized tabular format and thus, adding new data becomes straightforward. The data curation workflows established through CLDF also allow the curation and extension of data for language comparison.

To compare lexical data across diverse languages, we use the CLDF format to curate resources such as the Concepticon (List et al., 2016a) and the Database of Norms, Ratings, and Relations (NoRaRe, Tjuka et al., 2022a). The goal of the Concepticon is to equip linguists with a reference catalog of “comparable concepts” (Haspelmath, 2010) through linking concept lists to standardized concept sets. As the Concepticon continues to develop, data curation workflows have proven useful in adding new data and improving existing data. From the beginning, the Concepticon contained a small number of metadata on word properties including age-of-acquisition ratings (e.g., Kuperman et al., 2012), naming tests (e.g., Ardila, 2007), and links to other databases such as BabelNet (Navigli & Ponzetto, 2012). However, the data were not continually enriched and a variety of different types of data seemed to be accumulating from different research fields, for example, psychology and natural language processing. We, therefore, decided to construct a new resource, building on the Concepticon but using a customized workflow for the available data. This led to the creation of NoRaRe (Tjuka et al., 2022a). The goal of NoRaRe is to facilitate exchange between different research fields in order to answer big-picture questions using cross-linguistic comparison. The data in Version 0.2 (Tjuka et al., 2021) included norms, ratings, and relations from studies in linguistics and psychology offering information on word properties such as word frequencies (e.g., Brysbaert & New, 2009), sensory modality ratings (e.g., Lynott et al., 2020), and similarity estimations (e.g., Hill et al., 2015), among others.

Here, we introduce the major release of Concepticon Version 3.0 (List et al., 2022a) and NoRaRe Version 1.0 (Tjuka et al., 2022b). Apart from new data and improvements, the releases include the publication of Concepticon and No-RaRe as CLDF datasets, refinements to the accompanying Python packages, and the publication of NoRaRe in a web application built on Cross-Linguistic Linked Data (CLLD, cld3.org). The improvements represent an important step in the development of both resources, and we illustrate the data curation workflows implemented in Concepticon Version 3.0 and NoRaRe Version 1.0 below.

Materials and methods

Concept and word lists

The Concepticon began with the collection of concept lists from studies in historical linguistics using cross-linguistic comparisons to create language family trees. These concept lists include basic vocabulary and cross-linguistically comparable concepts such as HAND, TREE, YOU, or GIVE. Historical linguists have used different versions of these lists to elicit the glosses for the concepts across languages and determine cognates indicating language relatedness. At the time one of the most commonly used lists of concepts was created (Swadesh, 1955), there was a lack of standardization efforts, so subsequent studies expanded or adapted the original list as they saw fit (List, 2018). The Concepticon was the first resource to include various concept lists and make them comparable. It enables researchers to find and use the available concept lists for their studies.

On the preface, concept lists look like a list of words. However, words represent concepts in the mind, and in the case of language comparison, there may not be a translational equivalent for a given concept in a language. Similarly, word lists that are used in psychology elicit word properties of concepts
to determine whether they are perceived as abstract or con-
crete, positive or negative, etc. These studies offer additional
data on a given word or concept in an individual language and
are also integrated into the Concepticon (List et al., 2016a).
While studies in linguistics are comprised of a small set of
items, with the rise of large-scale data collection, psychologists
publish word lists that include thousands of words. In order
to incorporate these data, the Concepticon was extended by
establishing the NoRaRe database (Tjuka et al., 2022a).
The different contents of the lists are accounted for by giv-
ing them tags which make it more straightforward to search
for a particular list (List, 2018). Table 1 shows the 22 tags we
use in the Concepticon (List et al., 2022a). For example, the

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Count</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>acquisition</td>
<td>Concept lists related to studies on language acquisition.</td>
<td>6</td>
<td>(Łuniewska et al., 2019)</td>
</tr>
<tr>
<td>annotated</td>
<td>Concept lists that contain further annotations which exceed the complexity</td>
<td>12</td>
<td>(Urban, 2011)</td>
</tr>
<tr>
<td>areal</td>
<td>Concept lists designed for a specific linguistic area.</td>
<td>127</td>
<td>(Walworth &amp; Shimelman, 2018)</td>
</tr>
<tr>
<td>basic</td>
<td>Concept lists which are supposed to represent the basic vocabulary.</td>
<td>190</td>
<td>(Sagart et al., 2019)</td>
</tr>
<tr>
<td>body parts</td>
<td>Concept lists which concentrate on body parts.</td>
<td>9</td>
<td>(Majid &amp; van Staden, 2015)</td>
</tr>
<tr>
<td>colors</td>
<td>A list that is used to elicit color terms.</td>
<td>8</td>
<td>(Haynie &amp; Bowern, 2016)</td>
</tr>
<tr>
<td>documentation</td>
<td>Concept lists which serve to document one language or one language family.</td>
<td>5</td>
<td>(Krisadawan, 2000)</td>
</tr>
<tr>
<td>hihi</td>
<td>A list of highly reconstructable and highly retentive items (term adopted</td>
<td>3</td>
<td>(McMahon &amp; McMahon, 2005)</td>
</tr>
<tr>
<td></td>
<td>from McMahon &amp; McMahon, 2005).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lolo</td>
<td>A list of less stable basic items, with low reconstructability and low</td>
<td>4</td>
<td>(Galucio et al., 2015)</td>
</tr>
<tr>
<td></td>
<td>retentiveness (term adopted from McMahon &amp; McMahon, 2005).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>historical</td>
<td>A list which is historically interesting, mostly referring to lists published</td>
<td>13</td>
<td>(Dunn et al., 2017)</td>
</tr>
<tr>
<td></td>
<td>before the 20th century.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>naming test</td>
<td>A list designed for a naming test in neurology or psycholinguistics to assess</td>
<td>5</td>
<td>(Ardila, 2007)</td>
</tr>
<tr>
<td></td>
<td>the linguistic capability of children and adults.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>proto-language</td>
<td>A list illustrating the concepts in a proto-language that can be</td>
<td>5</td>
<td>(Bodt &amp; List, 2019)</td>
</tr>
<tr>
<td></td>
<td>reconstructed with high certainty.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>questionnaire</td>
<td>A questionnaire for linguistic fieldwork.</td>
<td>51</td>
<td>(Buck, 1949)</td>
</tr>
<tr>
<td>ranked</td>
<td>A list that shows items in a ranked order and has one column reflecting the</td>
<td>21</td>
<td>(Pagel et al., 2007)</td>
</tr>
<tr>
<td></td>
<td>rank.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sign language</td>
<td>A list which was designed to investigate sign languages.</td>
<td>3</td>
<td>(Woll et al., 2010)</td>
</tr>
<tr>
<td>specific</td>
<td>A list that we deem specific since it is not easy to compare with other lists</td>
<td>30</td>
<td>(Pepper, 2020)</td>
</tr>
<tr>
<td></td>
<td>in our sample.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>stable</td>
<td>A list that is supposed to represent the stable part of a larger list. Usually,</td>
<td>8</td>
<td>(Matisoff, 2009)</td>
</tr>
<tr>
<td></td>
<td>the stable part has an unstable counterpart.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ultra-stable</td>
<td>A usually very short list of the supposedly most stable concepts.</td>
<td>15</td>
<td>(Pagel et al., 2013)</td>
</tr>
<tr>
<td>unstable</td>
<td>A list that is supposed to represent the unstable part of a larger list.</td>
<td>5</td>
<td>(Daniel et al., 2021)</td>
</tr>
<tr>
<td>norms</td>
<td>A word list that includes data measured by taking samples from a set, such as</td>
<td>6</td>
<td>(Cai &amp; Brysbaert, 2010)</td>
</tr>
<tr>
<td></td>
<td>word frequency.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ratings</td>
<td>A word list comprised of ratings of a word by human participants on a scale</td>
<td>34</td>
<td>(Monnier &amp; Syssau, 2014)</td>
</tr>
<tr>
<td></td>
<td>or other measures, for example, whether a word is positive or negative.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>relations</td>
<td>A word list containing various types of semantic relations (e.g., lexical</td>
<td>11</td>
<td>(Vulić et al., 2020)</td>
</tr>
<tr>
<td></td>
<td>semantic similarity).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Tags for the Concepticon concept lists (List et al., 2022a). Some parts of the table are repeated from a blog post (List, 2018). Some lists receive multiple tags, so the total number is higher than the number of lists in the Concepticon.
tag areal comprises lists that are used to elicit concepts in a particular geographic area such as Vanuatu (e.g., Walworth & Shimelman, 2018). The lists with the tag body parts include studies that elicit body part terms in various languages (e.g., Majid & van Staden, 2015). Lists that are tagged as ranked have a hierarchical order of the words indicating their rates of semantic change (e.g., Pagel et al., 2007). All lists represented in NoRaRe are given at least one of the tags norms, ratings, or relations. Since some lists include different contents, they receive multiple tags. More examples are given in Table 1.

Data curation and reuse
The Concepticon was established in 2015 with its first major release in 2016 including 162 concept lists (Concepticon Version 1.0: List et al., 2016a; List et al., 2016b). Minor releases followed in the upcoming years and the last major release was in 2019 (Concepticon Version 2.0: List et al., 2019). The Concepticon then already contained 240 concept lists with additional links to metadata from studies collecting age-of-acquisition ratings (Kuperman et al., 2012) and other databases such as BabelNet (Navigli & Ponzetto, 2012) or OmegaWiki (OmegaWiki, 2020). At this point, the available metadata were slightly hidden, and we noticed that for words and concepts, there is a whole range of other data on norms, ratings, and relations to be found. Therefore, we decided to launch a satellite project that builds on the established workflows in Concepticon making the available data from linguistics and psychology FAIR (Wilkinson et al., 2016), especially interoperable and reusable. This is how NoRaRe (Tjuka et al., 2022a) got started. Since 2019, we are continuously adding new data to Concepticon and NoRaRe. This is the result of our data curation workflows, which are straightforward and can be managed by external and internal collaborators (e.g., student assistants). A list of all contributors to date can be found here: github.com/concepticon/concepticon-data/blob/v3.0.0/CONTRIBUTORS.md. The data curation of Concepticon is organized around the collaboration platform GitHub (concepticon-data repository: github.com/concepticon/concepticon-data). New lists or improvements start out as issues that can be triaged and addressed by the team of editors. If a list is selected for inclusion, the original data is transformed into a tabular format (if necessary), described by a metadata file, and additional metadata is added to the catalog. The concepts or words in the list are then mapped to the Concepticon concept sets (a detailed description of the concept mapping is given below). To facilitate the process for internal and external contributors, we offer documentation under github.com/concepticon/concepticon-data/blob/v3.0.0/CONTRIBUTING.md and blog posts that provide step-by-step tutorials (Tjuka, 2020a). All improvements and new concept lists are added by creating a pull request (PR) so that the changes can be reviewed by the Concepticon editors. The review process is described in detail in a blog post (see Tjuka, 2021a). The editors are a group of expert linguists who not only check the formal correctness of the contribution but also discuss questions of the content, for example, about concept mappings or additions of new concepts. Integrated into the PR are automated checks of the data that fail if a contribution is flawed (see the section “Testing and transparency through CLDF” below). These checks have proven extremely useful since accidental mistakes such as spelling errors or missing concept set identifiers tend to creep into the contribution workflow. This is the advantage of a test-driven data curation workflow and additionally, we have been able to identify mistakes in the original concept list with this process. The NoRaRe database is an extension of the Concepticon and thus, uses similar workflows. Established in 2020 with two minor releases (Tjuka et al., 2021), the first major release of NoRaRe Version 1.0 includes 113 datasets across 39 languages (Tjuka et al., 2022b). The NoRaRe database is also curated on GitHub (norare-data repository: github.com/concepticon/norare-data). To ensure the same quality of data as in Concepticon, improvements or new lists are added by creating a PR and are reviewed by one of the editors. Currently, the editor team of the NoRaRe database is small but will likely grow in the future. Documentation (see github.com/concepticon/norare-data/tree/v1.0.1) and a step-by-step guide in form of a blog post (Tjuka, 2021b) are provided as references for internal and external contributors. The data curation is also test-driven in that consistency checks are integrated to identify mistakes easily. In addition, the NoRaRe database offers predefined scripts that allow a quick correlation analysis of the data. The scripts are included in the folder examples here: github.com/concepticon/concepticon-data/blob/v3.0.0/examples. A blog post described the use of the scripts and how NoRaRe datasets are compared (Tjuka, 2021c). These correlation studies not only show interesting patterns arising from the data such as similarities in word frequencies across diverse languages (Tjuka, 2020b) but also illustrate that cross-linguistic comparable datasets for particular word properties such as sensory modality are not available yet (Tjuka, 2022). The database can therefore be used to identify gaps and improve cross-linguistic studies in psychology and linguistics.

The result of our data curation efforts is 413 concept lists across 41 languages in Concepticon Version 3.0. (List et al., 2022a). Concepticon currently has 3,914 concept sets with an average of 231.76 concept sets mapped in a given list. 20,878 unique elicitation glosses are mapped to the Concepticon concept sets. NoRaRe Version 1.0. (Tjuka et al., 2022b) includes 113 datasets (25 norms, 65 ratings, and 38 relations) across 39 languages and data on 75 properties.

Manual and automated mapping
The first step in the mapping workflow in Concepticon is to generate a mapping to the Concepticon concept sets. The Concepticon concept sets include an identifier, a description, and elicitation glosses linked from each list. Table 2 shows a small subset of the 3,914 Concepticon concept sets. Each concept

---

1 Note that we counted Cantonese and Mandarin Chinese as two separate languages when introducing NoRaRe (Tjuka et al., 2022a). The data are now grouped under Chinese because some studies do not clearly distinguish the two dialects.
set has a unique identifier (ID) and a Concepticon gloss. In addition, the semantic field, a description, and the ontological category are provided. An algorithm based on previous mappings generates a list of pre-selected concept sets which are possible for a particular elicitation gloss (examples of elicitation glosses are illustrated in Figure 1). The contributor then checks whether the elicitation gloss represents the proposed Concepticon concept set. For example, depending on the information in the source, the elicitation gloss bank needs to be mapped either to the concept set 1,284 BANK or 3,463 RIVERBANK. It is important to note that we try to map as many elicitation glosses in a list as possible while at the same time, improving the mappings by not mapping an elicitation gloss to a concept set if the meaning cannot be disambiguated. The mappings found in the Concepticon are the basis for the automatic mapping workflow used in NoRaRe.

The automated mapping workflow in NoRaRe is used when it is not feasible to manually check each concept set mapping. Concept and word lists included in the Concepticon are small lists with about 100 to 1000 items, whereas the basis for lists in NoRaRe can be thousands of words. The data-sets in the NoRaRe database are generated by a Python script that automatically downloads the data from the source, i.e., an open repository or web page. Then the raw data are transformed into a tabular format and the words in the list are automatically mapped to the Concepticon concept sets.

For this procedure, the algorithm checks the elicitation glosses that are mapped to a given concept set in Concepticon and compares them with the word in a given list. Since the mapping is idempotent the output is unchanged even if the algorithm is run multiple times on the same data. If there is a match, the word is mapped. The result is a reduced list that includes only the words from the original list that were mapped to the Concepticon concept sets. This procedure decreases the file size and avoids consuming an unnecessary amount of data volume. We discussed more details about the manual and

---

**Table 2. Examples of Concepticon concept sets across the six ontological categories.**

<table>
<thead>
<tr>
<th>ID</th>
<th>Gloss</th>
<th>Semantic field</th>
<th>Definition</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>GATHER</td>
<td>Spatial relations</td>
<td>To collect or gather (e.g. work, magazines, etc.).</td>
<td>Action/Process</td>
</tr>
<tr>
<td>153</td>
<td>GREY</td>
<td>Sense perception</td>
<td>Having a color between black and white, like ash or stone.</td>
<td>Property</td>
</tr>
<tr>
<td>1256</td>
<td>HEAD</td>
<td>The body</td>
<td>The part of the body of an animal or human which contains the brain, mouth, and main sense organs.</td>
<td>Person/Thing</td>
</tr>
<tr>
<td>2451</td>
<td>FOURTEEN</td>
<td>Quantity</td>
<td>The natural number fourteen (14).</td>
<td>Number</td>
</tr>
<tr>
<td>2166</td>
<td>PILLOW</td>
<td>(CLASSIFIER)</td>
<td>Classifier for pillows and other objects related to the bed.</td>
<td>Classifier</td>
</tr>
<tr>
<td>3843</td>
<td>BUT</td>
<td>Miscellaneous function words</td>
<td>A coordinating expression that signals a contrast to what was previously said, e.g. disagreement or a further point to consider.</td>
<td>Other</td>
</tr>
</tbody>
</table>

---

**Figure 1. Illustration of the content of the Concepticon (List et al., 2016a) on the left and NoRaRe (Tjuka et al., 2022a) on the right.**
automated mapping workflows and their outputs elsewhere (Tresoldi, 2019a, Tresoldi, 2019b; Tjuka et al., 2022a). NoRaRe also allows the mapping of typed data such as networks, which is an added benefit compared to the data in Concepticon.

Consistency and transparency through CLDF

CLDF (Forkel et al., 2018) – a package format for linguistic data that bundles a set of tables as CSV files with JSON metadata describing the relations between these tables – is used as a distribution format for the Concepticon and No-RaRe data to increase interoperability and reusability. To curate and package the data in Concepticon and NoRaRe as CLDF datasets, we use the tools in cldfbench (Forkel & List, 2020). CLDF’s extensibility makes it possible to add custom tables to such a package transparently while keeping the semantics of the fully standardized parts of the data, like the table of languages, intact. But the strict consistency promises CLDF makes and tools like pycldf (pypi.org/project/pycldf/1.29.0) enforce also serve as quality control during the data curation workflow to improve reusability. For example, the fact that CLDF datasets can be converted to relation databases turning relations between tables into foreign key constraints implies that valid CLDF datasets will have unique identifiers for each row of a table. In the context of NoRaRe another feature of CLDF, or rather the underlying CSVW specification (pypi.org/project/csvw/3.1.3), also helps with quality control. CSVW metadata can specify datatypes for data in CSV files, and thus augment the “raw” text data with well-defined conversions to typed data. For NoRaRe, this turns out to be particularly important, because NoRaRe variables provide many different types of data, ranging from continuous numbers from a limited range to categorical variables, with string values from a controlled vocabulary. The corresponding CSVW datatype descriptions then serve as documentation of valid assumptions for data re-use, but also as specifications for data consistency checks, which are built-in to CLDF validation.

Python package pyconcepticon

The Python package pyconcepticon (pypi.org/project/pyconcepticon/3.0.0) supports data curation in Concepticon. To use pyconcepticon, a copy of the Concepticon data must be locally accessible. When pyconcepticon is installed with pip, the integrated commands can be called through the command line. Typing concepticon -h will give a list of the functionalities.

The pyconcepticon package stores a number of tests that allow for consistency checks of the Concepticon data. Especially, if a new concept list or improvements on existing data are added, the tests that run with the command concepticon test can spot inconsistent mappings, missing files, the incorrect numbering of the concept sets, and many more careless mistakes. To check the consistency of an individual concept list, one can use the command concepticon check. The command concepticon validate inspects the availability of metadata for all concept lists. The pyconcepticon package includes several more commands that also simplify the addition of new lists as well as inspecting the available data.

Python package pynorare

Similarly to the pyconcepticon package, we created a Python package for the curation of the NoRaRe data collection, called pynorare (pypi.org/project/pynorare/1.0.1). The pynorare package can be installed with pip if a local copy of the NoRaRe data repository is downloaded. The command line is used to access the commands and a list of the functionalities can be retrieved by typing norare -h.

The consistency of the data in NoRaRe can be tested with the command norare check which checks the entries in the norare.tsv file (github.com/concepticon/norare-data/blob/v1.0.1/norare.tsv). The norare.tsv file includes information on all the variables in the NoRaRe datasets. It is important that these entries are consistent and that mistakes are immediately identified. Otherwise, the comparison across datasets would become unfeasible. Individual datasets can be checked for internal consistency by using norare validate. Furthermore, the command norare stats creates a summary statistic across all NoRaRe datasets and calculates the number of concept sets that have at least one link to a NoRaRe dataset as well as the available number of data sets links for each concept set.

Using and representing the data

CLDF datasets

With the major release of Concepticon Version 3.0 (List et al., 2022a) and NoRaRe Version 1.0 (Tjuka et al., 2022b), both resources were also made available as CLDF datasets (List et al., 2022b; Tjuka et al., 2022c). The data is represented in tabular format and corresponding metadata in JSON format. The bundling of the data makes it possible to represent relations between the tables. The data model is available here: github.com/concepticon/concepticon-cldf/blob/v3.0.0/cldf/README.md. Since the data is represented in text formats such as CSV and JSON, it can be conveniently explored. Tools such as csvkit (pypi.org/project/csvkit/1.0.7) allow the processing and in-depth study of the available data in Concepticon and NoRaRe. Examples of how to use the Concepticon data from the CLDF dataset can be found under github.com/concepticon/concepticon-cldf/tree/v3.0.0/doc and for NoRaRe under github.com/concepticon/norare-cldf/tree/v1.0.0.

An advantage of the CLDF datasets is that CLDF includes the information to load the data from Concepticon or NoRaRe into a relational database and perform various queries. The Python package pycldf (pypi.org/project/pycldf/1.29.0) can convert any CLDF dataset into an SQLite database that allows queries with SQL. This process replicates the construction of the web applications described below. For Concepticon, queries could include listing the Concepticon concept sets in a given concept list or showing concept set relations such as plotting all the narrower concept sets connected to 1,262 BROTHER, i.e., 559 BROTHER (OF MAN), 560 BROTHER (OF WOMAN), 1,759 OLDER BROTHER, 1760 YOUNGER
BROTHER, etc. For NoRaRe, one could query word frequencies for words expressing the same concept. Another possibility would be to compute correlations by assembling data from different data sets, for example, arousal ratings from different languages. The use of the Python packages pandas (pypi.org/project/pandas/1.5.1) and seaborn (pypi.org/project/seaborn/0.12.1) allows the creation of a dot plot for the correlation. Note that variables may have multiple values assigned to the same concept set because different words were mapped to the same concept set. Although these cases are rare, researchers need to inspect the mapped words before deciding whether or not they want to include them in a correlation study.

CLLD web applications

The CLDF datasets described above are the input for the clld applications developed in the Cross-Linguistic Linked Data (CLLD) project (clld.org and documentation under clld.readthedocs.io/en/latest). The CLLD project allows the curation and development of lexical and grammatical databases. The clld toolkit (pypi.org/project/clld/9.2.2) is a Python package that integrates functions for building and maintaining CLLD web applications. These web applications (short: web apps) can be conveniently accessed via a web browser. They are also a good way to check the consistency of data and are a form of data reuse. The CLDF datasets of Concepticon and NoRaRe include all the information to create clld web apps for each data collection. Each new data release is accompanied by an update of the web applications. The clld web application for Concepticon was already introduced with the first release of Concepticon Version 1.0 (List et al., 2016b) in 2016. The major release of Concepticon 3.0 (List et al., 2022a) and NoRaRe 1.0 (Tjuka et al., 2022b) brings a number of updates that affect the presentation of data. In the previous version of the Concepticon web application, different kinds of metadata on word frequency, concreteness ratings, links to WordNet (Miller et al., 1990), etc. were represented in a box beside the elicitation glosses linked to a given concept set. The most significant change for the Concepticon web app (concepticon.clld.org) apart from the data update is the integration of a link to the NoRaRe data and a summary statistic indicating the number of links to variables and datasets for each concept set that replace the metadata box (see Figure 2).

The clld web app for NoRaRe (norare.clld.org) was introduced for the first time with the major release of NoRaRe Version 1.0 (Tjuka et al., 2022b). The NoRaRe web application has a similar structure to the Concepticon web application while at the same time, the features of NoRaRe including a list of all datasets and the variables for each concept set are highlighted. The wordcloud on the front page is automatically generated based on the tags used for each variable in NoRaRe (see Figure 3). The font size represents the frequency of the individual tags across all datasets in NoRaRe. Most NoRaRe datasets include multiple variables which become apparent by clicking on the link for a given dataset. The web application also shows which glosses are mapped to a Concepticon concept set. A map illustrates the distribution of languages associated with each value (see Figure 4). Since many datasets containing norms, ratings, and relations come from psychological studies, the bias toward Central European languages is obvious. However, once cross-linguistic data from linguistics is added, the distribution of languages extends to areas such as Africa and New Guinea.

Conclusion

The present article introduced the major release of the cross-linguistic databases Concepticon Version 3.0 (List et al., 2022a) and NoRaRe Version 1.0 (Tjuka et al., 2022b). We discussed the contents of both resources and their underlying data curation workflows. The Concepticon includes standardized concept sets that allow comparison across many languages. NoRaRe offers data on norms, ratings, and relations for words and concepts and is an extension of the Concepticon. With the major release, new data were added, the data was published in CLDF format and for NoRaRe, a web application was created.

Figure 2. Replacement of the metadata box with a link to NoRaRe for each concept set, for example, 2 DUST: concepticon.clld.org/parameters/2.
Our data curation workflows have proven applicable for the advancement of data for language comparison and we envision that other researchers use the two resources for their studies. The availability of the data as CLDF datasets and the Python packages make it possible to explore and test the data more conveniently. Web applications for Concepticon and NoRaRe offer an additional overview of the available data. At this point, Concepticon includes 413 concept lists with 41 mapping languages and 3,914 concept sets. NoRaRe contains 113 datasets with 75 word properties across 39 languages. We intend to further expand these data collections in the future.

**Ethical approval and consent**

Ethical approval and consent were not required.

**Data and software availability**

**Underlying data**

The Concepticon is curated on GitHub (github.com/concepticon/concepticon-data/tree/v3.0.0) and archived with Zenodo under CC-BY license (doi.org/10.5281/zenodo.7296458) (List et al., 2022a).

The CLDF dataset for Concepticon is published on GitHub (github.com/concepticon/concepticon-cldf/tree/v3.0.0) and Zenodo under CC-BY license (doi.org/10.5281/zenodo.7298023) (List et al., 2022b).

NoRaRe is also curated on GitHub (github.com/concepticon/norare-data/tree/v1.0.1) and archived with Zenodo under...
The CLDF dataset for NoRaRe is published on GitHub (github.com/concepticon/norare-cldf/tree/v1.0.0) and Zenodo under CC-BY license (doi.org/10.5281/zenodo.7312927) (Tjuka et al., 2022).

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Extended data

The Python packages used for the data curation workflows can be found on PyPi under Apache license: pyconcepticon (pypi.org/project/pyconcepticon/3.0.0) and pynorare (pypi.org/project/pynorare/1.0.1).

For convenient access, we offer a CLLD web app for Concepticon (concepticon.cldf.org) and NoRaRe (norare.cldf.org).

Acknowledgments

We would like to thank everyone who contributed to the Concepticon and NoRaRe by making their data available, adding new lists, pointing us to missing lists, and proposing improvements. A list of all contributors can be found for Concepticon and NoReRe here: github.com/concepticon/concepticon-data/blob/v3.0.0/CONTRIBUTORS.md and github.com/concepticon/norare-data/blob/v1.0.1/CONTRIBUTORS.md.

References


