Check for updates

DATA NOTE

REVISED Curating and extending data for language comparison in Concepticon and NoRaRe

[version 3; peer review: 2 approved]

Annika Tjuka[1], Robert Forkel[1], Johann-Mattis List[1,2]

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Saxony, 04103, Germany
[2]Chair for Multilingual Computational Linguistics, University of Passau, Passau, Bavaria, 94032, Germany

v3 **First published:** 16 Dec 2022, **2**:141
https://doi.org/10.12688/openreseurope.15380.1
Second version: 28 Apr 2023, **2**:141
https://doi.org/10.12688/openreseurope.15380.2
Latest published: 24 May 2023, **2**:141
https://doi.org/10.12688/openreseurope.15380.3

Abstract

Language comparison requires user-friendly tools that facilitate the standardization of linguistic data. We present two resources built on the basis of a standardized cross-linguistic format and show how the data is curated and extended. The first resource, the Concepticon, is a reference catalog for standardized concepts from linguistic research. 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

Open Peer Review

Approval Status ✓ ✓

	1	2
version 3 (revision) 24 May 2023	✓ view	
version 2 (revision) 28 Apr 2023	✓ view	✓ view
version 1 16 Dec 2022	✗ view	✓ view

1. **Gemma Boleda**, Universitat Pompeu Fabra, Barcelona, Spain

2. **Thanasis Georgakopoulos**, Aristotle University of Thessaloniki, Thessaloniki, Greece

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Languages and Literature gateway.



This article is included in the Horizon Europe gateway.



This article is included in the European Research Council (ERC) gateway.



This article is included in the Horizon 2020 gateway.



This article is included in the Linguistic Diversity collection.

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) as well as the Horizon Europe programme (grant agreement No. 101044282). 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: © 2023 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 3; peer review: 2 approved]** Open Research Europe 2023, **2**:141 https://doi.org/10.12688/openreseurope.15380.3

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

REVISED **Amendments from Version 2**

In this version, we implemented minor changes based on the reviewer's suggestions. The changes include two additions: details on the ontological categories and a definition of the term "typed data".

Any further responses from the reviewers can be found at the end of the article

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 comparison. Therefore, we use a set of tools that facilitate the standardization of data so that they become comparable. By implementing 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 Version 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. Therefore, it is necessary to create a standardized format that applies to all languages, and at the same time, to provide the tools for effortless standardization. A community-led initiative has recognized this need and 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. We apply a test-driven data approach, i.e., specific tests are carried out that match the formal requirements of the data automatically with the specifications developed in CLDF.

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 "comparative 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 refinements to the accompanying Python packages, the publication of Concepticon and NoRaRe as CLDF datasets and the publication of NoRaRe in a web application built on Cross-Linguistic Linked Data (CLLD, clld.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. Due to the scope of the data note, we cannot present the entire background information on both resources (for detailed overviews, see List *et al.*, 2016a; Tjuka *et al.*, 2022a). The interested reader may find additional information on technical details and tutorials on how to use the tools presented here on our blog Computer-Assisted Language Comparison in Practice (calc.hypotheses.org). Table 1 contains a glossary of relevant terminology.

Table 1. Glossary of relevant terms and abbreviations occurring in the text. The table is adapted from Tjuka *et al.* (2022a).

Term	Definition
elicitation gloss	In linguistic fieldwork, an "elicitation gloss" is used to denote a concept in a metalanguage. For example, when English is used as a metalanguage, a researcher would use the elicitation gloss *tree* to elicit the expression for the concept TREE in another language, while a Spanish researcher would use the elicitation gloss *árbol* to elicit the same concept.
concept	We define "concept" as a non-linguistic psychological representation of an object in the world. The standardised concepts in Concepticon are built on the theoretical framework of "comparative concepts" proposed by Haspelmath (2010). These comparative concepts are defined by the researcher as a tool for language comparison. In psychology, there is a longstanding debate about what a "concept" constitutes (cf. Barsalou, 2017; Malt & Majid (2013); Murphy, 2002). In comparison, the discussions in linguistics focus mainly on the term "word meaning" (cf. Bolognesi, 2020; Riemer, 2014). We do not propose a solution for these theoretical discussions.
concept identifier	A "concept identifier" is a unique number that is used to identify a concept set.
concept label	A "concept label" is a unique word or phrase describing a concept set.
concept set	The concept identifier and the concept label together form a "concept set", for example, 906 TREE. Concept sets in Concepticon are described with additional metadata: a definition, a semantic field, and an ontological category. The concept sets are defined to standardize the concepts which are used for language description and comparison.
concept list	The term "concept list" refers to a compilation of concepts in the form of elicitation glosses. They are used by linguists who want to elicit a concept in a particular language. In contrast to dictionaries, the lists are based on questionnaires and are compiled for language comparison or documentation (List, 2018). The list is usually a table of elicitation glosses such as *tree*, *you*, *what*, and *bring* which represent the concepts TREE, YOU, WHAT, and BRING.
Abbreviations	
CLLD	Cross-Linguistic Linked Data (clld.org): The overarching project structure under which all of the data are published.
CLDF	Cross-Linguistic Data Formats (Forkel *et al.*, 2018): The format used to standardize the information provided in linguistic research.
CSVW	CSV on the Web (CSVW): A standard format to represent metadata (csvw.org).
JSON	JSON (JavaScript Object Notation): A text format which allows for import and export of tabular data (json.org/json-en.html).
SQL	Structured Query Language (SQL): A programming language to query and manage relational databases.

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 surface, 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 concrete, 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 concept lists in Concepticon are accounted for by giving them tags which make it more straightforward to search for a particular kind of data (List, 2018). Table 2 shows the 22 tags we use in the Concepticon (List *et al.*, 2022a). For example, the 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).

Table 2. Tags for the Concepticon concept lists (List *et al.*, 2022a). 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	Description	Count	Example
acquisition	Concept lists related to studies on language acquisition.	6	(Łuniewska *et al.*, 2019)
annotated	Concept lists that contain additional annotations, for example, on semantic change or colexification degrees. These lists provide extra information beyond the regular content of a concept list.	12	(Urban, 2011)
areal	Concept lists designed to elicit concepts in a specific geographic area.	127	(Walworth & Shimelman, 2018)
basic	Concept lists which are supposed to represent the basic vocabulary.	190	(Sagart *et al.*, 2019)
body parts	Concept lists which concentrate on body parts.	9	(Majid & van Staden, 2015)
colors	A list that is used to elicit color terms.	8	(Haynie & Bowern, 2016)
documentation	Concept lists which serve to document one language or one language family.	5	(Krisadawan, 2000)
hihi	A list of highly reconstructable and highly retentive items (term adopted from McMahon & McMahon, 2005).	3	(McMahon & McMahon, 2005)
lolo	A list of less stable basic items, with low reconstructability and low retentiveness (term adopted from McMahon & McMahon, 2005).	4	(Galucio *et al.*, 2015)
historical	A list which is historically interesting, mostly referring to lists published before the 20th century.	13	(Dunn *et al.*, 2017)
naming test	A list designed for a naming test in neurology or psycholinguistics to assess the linguistic capability of children and adults.	5	(Ardila, 2007)
proto-language	A list illustrating the concepts in a proto-language that can be reconstructed with high certainty.	5	(Bodt & List, 2019)
questionnaire	A questionnaire for linguistic fieldwork.	51	(Buck, 1949)
ranked	A list that shows items in a ranked order and has one column reflecting the rank. The ranks are based on, for example, phylogenetic analyses or analysis of borrowing frequencies.	21	(Pagel *et al.*, 2007)
sign language	A list which was designed to investigate sign languages.	3	(Woll *et al.*, 2010)
specific	A list that we consider specific because it is not easily comparable with other lists in the Concepticon. Often the lists include a specific phenomenon studied in linguistics, such as binominal lexemes, pronouns, or semantic groupings.	30	(Pepper, 2020)
stable	A list that is supposed to represent the stable part of a larger list. Usually, the stable part has an unstable counterpart.	8	(Matisoff, 2009)
ultra-stable	A usually very short list of the supposedly most stable concepts.	15	(Pagel *et al.*, 2013)
unstable	A list that is supposed to represent the unstable part of a larger list. Usually has a stable counterpart.	5	(Daniel *et al.*, 2021)
norms	A word list that includes data measured by taking samples from a set, such as word frequency.	6	(Cai & Brysbaert, 2010)
ratings	A word list comprised of ratings of a word by human participants on a scale or other measures, for example, whether a word is positive or negative.	34	(Monnier & Syssau, 2014)
relations	A word list containing various types of semantic relations (e.g., lexical semantic similarity).	11	(Vulić *et al.*, 2020)

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 2.

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 subsequent 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 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. 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[1] (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/norare-data/tree/v1.0.1/examples. A blog post described the use of the scripts and how NoRaRe datasets are compared (Tjuka, 2021c). For example, we compared word frequencies across English, German, and Chinese, and the results showed that word frequencies were similar (Tjuka, 2020b). Another study compared sensory modality across English, Dutch, and Italian showing subtle differences in each sensory modality (Tjuka, 2022). Thus, the database can be used to compare word properties and reveal both cross-linguistic similarities and differences.

The result of our data curation efforts is 413 concept lists across 56 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. Table 3 provides an overview of the descriptive statistics.

Manual and automated mapping

Detailed descriptions of our workflows are provided in the articles introducing Concepticon (List *et al.*, 2016b) and NoRaRe (Tjuka *et al.*, 2022a) as well as in the tutorials on our blog (Tjuka, 2020a; Tresoldi, 2019a; Tresoldi, 2019b). Here, we summarize the basic steps of the manual and automated mapping workflows.

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 4 shows a small subset of the 3,914 Concepticon concept sets. Each concept set has a unique identifier (ID) and a Concepticon gloss. In addition, the semantic field, a description, and the ontological category are provided. The ontological categories are based on those used by the World Loanword Database (Haspelmath & Tadmor, 2009). However, they have been modified with regard to the specific terms used, because we are dealing with concepts rather than words that can be divided into parts of speech (List *et al.*, 2016a: 2394). An algorithm based on previous mappings generates a list of pre-selected concept sets which are possible for a particular elicitation gloss (examples of

[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.

Table 3. Descriptive statistics of Concepticon 3.0 and NoRaRe 1.0.

	Concepticon 3.0		NoRaRe 1.0
Lists	413	**Datasets**	113
Concepts sets	3914	**Word properties**	75
Languages	56	**Languages**	39
Language families	14	**Language families**	8

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 1284 BANK or 3463 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. For example, if a list contains the elicitation gloss *smoke* without any further information about whether the verb or noun meaning is intended, we do not map it to one of the concept sets 778 SMOKE (EXHAUST) or 1689 SMOKE (INHALE). Elicitation glosses which are not assigned to a concept set are automatically linked to the concept set 0 NA. 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 1,000 items, whereas the basis for lists in NoRaRe can be thousands of words. The datasets 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 (e.g., *tree, árbol*) and compares them with the word in a given list (e.g., *tree, forest, wood*). 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. NoRaRe also includes mappings of typed data, for example, entries in Wikidata (wikidata.org) and BabelNet (babelnet.org), which is an added benefit compared to the data in Concepticon. The information on the websites is retrieved via an API and then stored in a JSON file with the corresponding metadata on data types and relations in the original dataset, i.e., typed data.

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, for example, `create_metadata` which automatically creates the metadata JSON file. Guides on how to use the functionalities can be found on our blog (Tjuka, 2020a; Tresoldi, 2019a; Tresoldi, 2019b).

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, incorrect numbering of the concept sets, and many more 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`

Similar 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`. Guides on how to use the functionalities can be found on our blog (Tjuka, 2021b; Tjuka, 2021c).

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

Table 4. Examples of Concepticon concept sets across the six ontological categories.

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

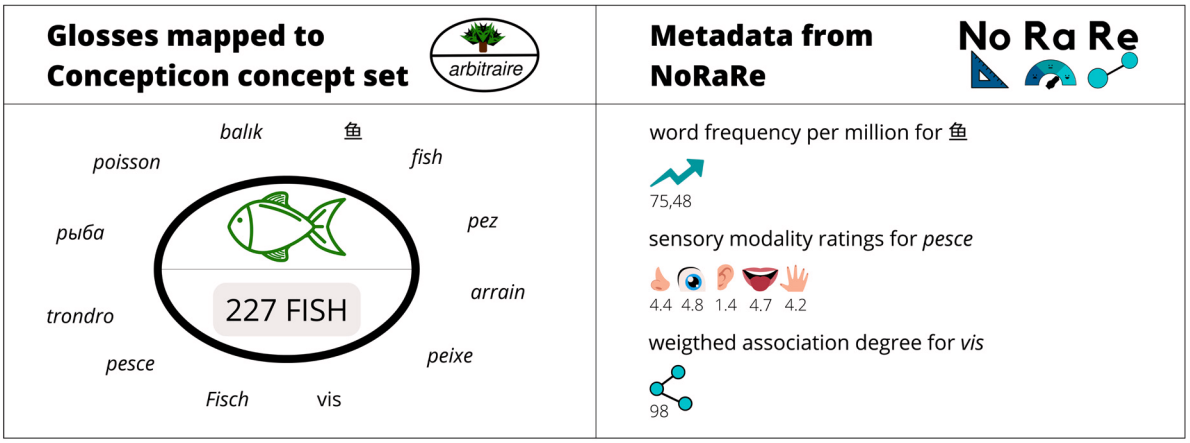


Figure 1. Illustration of the content of the Concepticon (List *et al.*, 2016a) on the left and NoRaRe (Tjuka *et al.*, 2022a) on the right.

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 datasets links for each concept set.

Increasing reusability through Cross-Linguistic Data Formats

Cross-Linguistic Data Formats (CLDF, Forkel *et al.*, 2018) – a 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 NoRaRe data to increase interoperability and reusability[2]. 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 that 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. CLDF datasets follow the data model of relational databases, where tables can link to each other using foreign keys. This means that CLDF datasets can automatically be converted to relational databases to facilitate efficient data access. It also means that CLDF datasets follow the recommendations for "tidy" data (Wickham, 2014; Wilson *et al.*, 2017), in particular by providing 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

[2] We refer readers interested in the major design principles behind CLDF and the contrasts with other standardization efforts to the original publication introducing CLDF by Forkel *et al.* (2018).

categorical variables, with string values from a controlled vocabulary. The corresponding CSVW datatype descriptions then serve as documentation of valid assumptions for data reuse, but also as specifications for data consistency checks, which are built-in to CLDF validation.

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). While the raw data is updated and improved consistently in the respective GitHub repositories, the CLDF datasets allow for the reuse and exploration of the data in a different way. The data in the CLDF datasets are represented in tabular format and the 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. By releasing the data in Concepticon and NoRaRe as CLDF datasets, tools such as `csvkit` (pypi.org/project/csvkit/1.1.1) can be used to process and analyse the data conveniently. 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/doc.

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 1262 BROTHER, i.e., 559 BROTHER (OF MAN), 560 BROTHER (OF WOMAN), 1759 OLDER BROTHER, 1760 YOUNGER BROTHER, etc. The relations *broader* versus *narrower* are used to indicate concept sets that refer to specific parts of a more general concept. For example, we established the broader concept set 3626 KNOW to relate it to the narrower concept sets 1410 KNOW (SOMETHING) and 2248 KNOW (SOMEONE) because we found that several concept lists did not provide a clear distinction between the two concepts. It is important to note that the Concepticon does not provide a full ontology but rather that the relations are established bottom-up and mainly serve as an exploration of the data instead of providing a basis for inferences. 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 datasets, 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

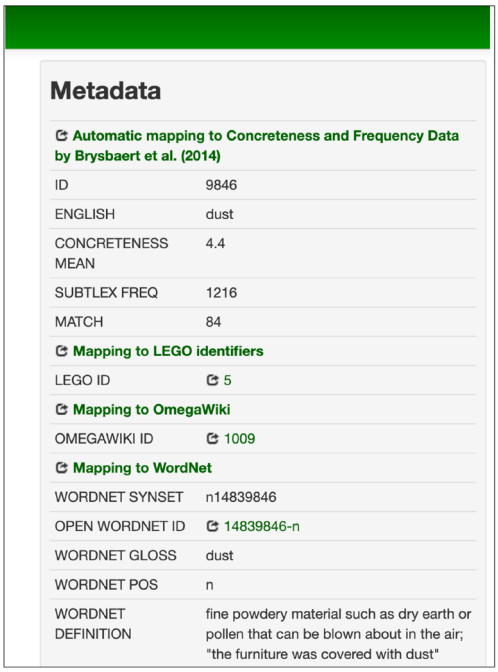
The CLDF datasets described above are the input for the `clld` applications developed in the Cross-Linguistic Linked Data (CLLD) project (clld.org and documentation under clld.readthedocs.io/en/latest). The CLLD project allows the curation and development of lexical and grammatical databases. The `clld` toolkit (pypi.org/project/clld/10.0.0) 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 word cloud 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

Before



After

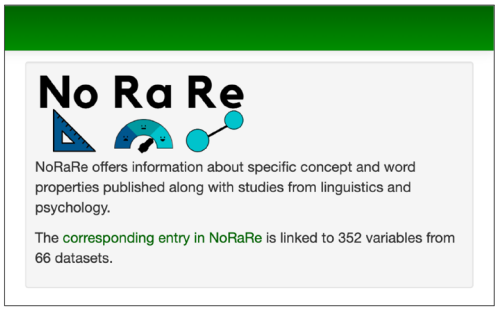


Figure 2. **Replacement of the metadata box with a link to NoRaRe for each concept set.**

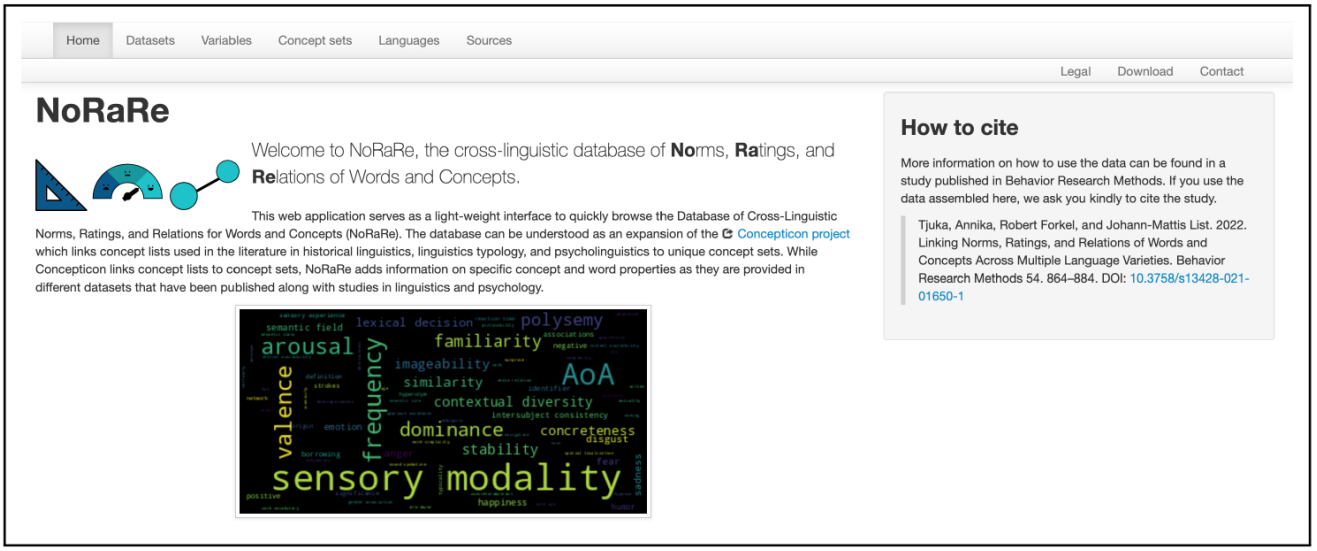


Figure 3. **Word cloud illustrating the tags used for each variable in NoRaRe (norare.clld.org).**

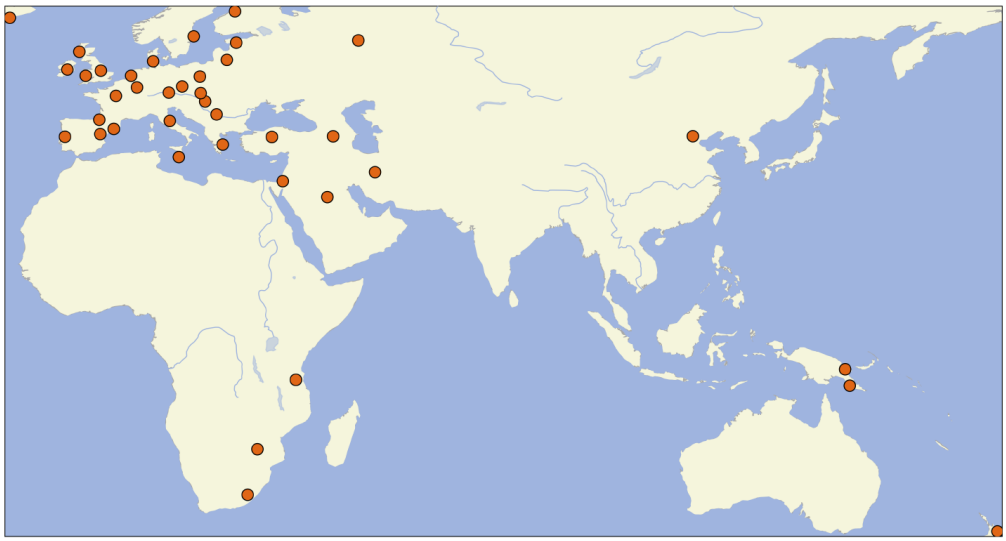


Figure 4. Distribution of languages in the NoRaRe datasets.

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 were published in CLDF format and for NoRaRe, a web application was created.

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 across 56 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 CC-BY license (doi.org/10.5281/zenodo.7298060) (Tjuka *et al.*, 2022b).

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.*, 2022c).

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.clld.org) and NoRaRe (norare.clld.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

Ardila A: **Toward the development of a cross-linguistic naming test.** *Arch Clin Neuropsychol.* 2007; **22**(3): 297–307.
PubMed Abstract | **Publisher Full Text**

Barsalou LW: **Cognitively plausible theories of concept composition**. In: James A. Hampton & Yoad Winter (eds.), *Compositionality and concepts in linguistics and psychology*. (Language, Cognition, and Mind), Basel: Springer International Publishing, 2017; 9–30.
Publisher Full Text

Bodt TA, List JM: **Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages.** *Pap Hist Phonol.* 2019; **4**(1): 22–44.
Publisher Full Text

Bolognesi M: **Where words get their meaning: Cognitive processing and distributional modelling of word meaning in first and second language.** (Converging Evidence in Language and Communication Research 23). Amsterdam: John Benjamins, 2020.
Publisher Full Text

Brysbaert M, New B: **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.** *Behav Res Methods.* 2009; **41**(4): 977–990.
PubMed Abstract | **Publisher Full Text**

Buck CD: **A dictionary of selected synonyms in the principal Indo-European languages: A contribution to the history of ideas**. University of Chicago Press, Chicago, Illinois, 1949.
Reference Source

Cai Q, Brysbaert M: **SUBTLEX-CH: Chinese word and character frequencies based on film subtitles.** *PLoS One.* 2010; **5**(6): e10729.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Daniel M, Chechuro I, Verhees S, *et al.*: **Lingua francas as lexical donors: Evidence from Daghestan.** *Language.* 2021; **97**(3): 520–560.
Publisher Full Text

Dunn M, Dewey TK, Arnett C, *et al.*: **Dative sickness: A phylogenetic analysis of argument structure evolution in Germanic.** *Language.* 2017; **93**(1): 1–22.
Publisher Full Text

Forkel R, List JM, Greenhill SJ, *et al.*: **Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics.** *Sci Data.* 2018; **5**: 180205.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Forkel R, List JM: **CLDFBench: Give your cross-linguistic data a lift**. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference.* Marseille, France, European Language Resources Association, 2020; 6995–7002.
Reference Source

Foundation OmegaWiki: **OmegaWiki. A dictionary in all languages**. 2020.
Reference Source

Galucio AV, Meira S, Birchall J, *et al.*: **Genealogical relations and lexical distances within the Tupian linguistic family.** *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas.* 2015; **10**(2): 229–274.
Publisher Full Text

Haspelmath M: **Comparative concepts and descriptive categories in crosslinguistic studies.** *Language.* 2010; **86**(3): 663–687.
Reference Source

Haspelmath M, Tadmor U: **The Loanword Typology project and the World Loanword Database.** In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages*. Berlin: De Gruyter Mouton. 2009.

Haynie HJ, Bowern C: **Phylogenetic approach to the evolution of color term systems.** *Proc Natl Acad Sci U S A.* 2016; **113**(48): 13666–13671.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Hill F, Reichart R, Korhonen A: **SimLex-999: Evaluating semantic models with (genuine) similarity estimation.** *Comput Linguist.* 2015; **41**(4): 665–695.
Publisher Full Text

Krisadawan H: **Rgyalthang Tibetan lexicon and an apraisal of a Southeast Asian wordlist.** *Mon-Khmer Studies.* 2000; **30**: 83–94.
Reference Source

Kuperman V, Stadthagen-González H, Brysbaert M: **Age-of-acquisition ratings for 30,000 English words.** *Behav Res Methods.* 2012; **44**(4): 978–990.
PubMed Abstract | **Publisher Full Text**

List JM, Cysouw M, Forkel R: **Concepticon: A resource for the linking of concept lists**. In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, European Language Resources Association. 2016a; 2393–2400.
Reference Source

List JM, Cysouw M, Forkel R: **Concepticon. A resource for the linking of concept lists (Version 1.0)**. Max Planck Institute for the Science of Human History, Jena, Germany, 2016b.
Publisher Full Text

List JM, Greenhill SJ, Rzymski C, *et al.*: **Concepticon. A resource for the linking of concept lists (Version 2.0)**. Max Planck Institute for the Science of Human History, Jena, Germany, 2019.
Publisher Full Text

List JM: **Towards a history of concept list compilation in historical linguistics.** *History and Philosophy of the Language Sciences.* 2018.
Reference Source

List JM, Tjuka A, Rzymski C, *et al.*: **Concepticon. A resource for the linking of concept lists (Version 3.0)**. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 2022a.
http://www.doi.org/10.5281/zenodo.7296458

List JM, Tjuka A, Rzymski C, *et al.*: **CLLD Concepticon 3.0.0 as CLDF dataset**. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 2022b.
http://www.doi.org/10.5281/zenodo.7298023

Łuniewska M, Wodniecka Z, Miller CA, *et al.*: **Age of acquisition of 299 words in seven languages: American English, Czech, Gaelic, Lebanese Arabic, Malay, Persian and Western Armenian.** *PLoS One.* 2019; **14**(8): e0220611.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lynott D, Connell L, Brysbaert M, *et al.*: **The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words.** *Behav Res Methods.* 2020; **52**(3): 1271–1291.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Malt BC, Majid A: **How thought is mapped into words.** *Wiley Interdiscip Rev Cogn Sci.* 2013; **4**(6): 583–597.
PubMed Abstract | **Publisher Full Text**

Majid A, van Staden M: **Can nomenclature for the body be explained by embodiment theories?** *Top Cogn Sci.* 2015; **7**(4): 570–594.
PubMed Abstract | **Publisher Full Text**

Matisoff JA: **Stable roots in Sino-Tibetan/Tibeto-Burman**. In: Yasuhiko Nagano, editor, *Issues in Tibeto-Burman historical linguistics.* in Senri Ethnological Studies, National Museum of Ethnology, Osaka, 2009; **75**: 291–318.
Publisher Full Text

McMahon A, McMahon R: **Language classification by numbers**. Oxford University Press, Oxford, UK, 2005.
Reference Source

Miller GA, Beckwith R, Fellbaum C, *et al.*: **Introduction to WordNet: An on-line lexical database.** *Int J Lexicogr.* 1990; **3**(4): 235–244.
Publisher Full Text

Monnier C, Syssau A: **Affective norms for French words (FAN).** *Behav Res Methods.* 2014; **46**(4): 1128–1137.
PubMed Abstract | **Publisher Full Text**

Murphy GL: **The big book of concepts**. Cambridge: MIT Press, 2002.
Reference Source

Navigli R, Ponzetto SP: **BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network.** *Artif Intell.* 2012; **193**: 217–250.
Publisher Full Text

Pagel MD, Atkinson QD, Calude AS, *et al.*: **Ultraconserved words point to deep language ancestry across Eurasia.** *Proc Natl Acad Sci U S A.* 2013; **110**(21): 8471–8476.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Pagel MD, Atkinson QD, Meade A: **Frequency of word-use predicts rates of lexical evolution throughout Indo-European history.** *Nature.* 2007; **449**(7163): 717–720.
PubMed Abstract | **Publisher Full Text**

Pepper S: **The typology and semantics of binominal lexemes**. PhD thesis, University of Oslo, Oslo, 2020.
Reference Source

Riemer N: **Word meanings**. In: John R Taylor (ed.), *The Oxford handbook of the word*. Oxford: Oxford University Press, 2014; 305–319.
Publisher Full Text

Sagart L, Jacques G, Lai Y, *et al.*: **Dated language phylogenies shed light on the ancestry of Sino-Tibetan.** *Proc Natl Acad Sci U S A.* 2019; **116**(21): 10317–10322.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Swadesh M: **Towards greater accuracy in lexicostatistic dating.** *Int J Am Linguist.* 1955; **21**(2): 121–137.
Reference Source

Tjuka A: **Adding concept lists to Concepticon: A guide for beginners.** *Comput Assist Lang Learn Practice.* 2020a; **3**(1): 5–10.
Reference Source

Tjuka A: **General patterns and language variation: Word frequencies across English, German, and Chinese**. In: Michael Zock, Emmanuele Chersoni, Alessandro Lenci, and Enrico Santus, editors, *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon.* Barcelona (Online), Association for Computational Linguistics, 2020b; 23–32.
Reference Source

Tjuka A, Forkel R, List JM: **NoRaRe. A database of cross-linguistic norms, ratings, and relations for words and concepts (Version 0.2)**. Max Planck Institute for the Science of Human History, Jena, Germany, 2021.
Publisher Full Text

Tjuka A: **How to review concept lists in collaboration (How to do X in linguistics 6).** *Comput Assist Lang Learn Practice.* 2021a; **4**(3): 1–5.
Reference Source

Tjuka A: **Adding data sets to NoRaRe: A guide for beginners.** *Comput Assist Lang Learn Practice.* 2021b; **4**(8): 1–5.
Reference Source

Tjuka A: **Comparing NoRaRe data sets: Calculation of correlations and creation of plots in R.** *Comput Assist Lang Learn Practice.* 2021c; **4**(11): 1–5.
Reference Source

Tjuka A: **Comparing sensory properties of words between English, Dutch, and Italian**. The Biannual Conference of the German Cognitive Science Society, Freiburg, Germany, 2022.
Reference Source

Tjuka A, Forkel R, List JM: **Linking norms, ratings, and relations of words and concepts across multiple language varieties.** *Behav Res Methods.* 2022a; **54**(2): 864–884.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Tjuka A, Forkel R, List JM: **NoRaRe. A database of cross-linguistic norms, ratings, and relations for words and concepts (Version 0.2)**. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 2022b.
http://www.doi.org/10.5281/zenodo.4647878

Tjuka A, Forkel R, List JM: **Database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts as CLDF dataset (Version 1.0.0)**. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 2022c.
http://www.doi.org/10.5281/zenodo.7312927

Tresoldi T: **Using pyconcepticon to map concept lists.** *Comput Assist Lang Learn Practice.* 2019a; **2**(4): 33–43.
Reference Source

Tresoldi T: **Using pyconcepticon to map concept lists (II).** *Comput Assist Lang Learn Practice.* 2019b; **2**(4): 44–56.
Reference Source

Urban M: **Asymmetries in overt marking and directionality in semantic change.** *J Hist Linguist.* 2011; **1**(1): 3–47.
Publisher Full Text

Vulić I, Baker S, Ponti EM, *et al.*: **Multi-SimLex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity.** *Comput Linguist.* 2020; **46**(4): 1–51.
Publisher Full Text

Walworth M, Shimelman A: **Vanuatu basic vocabulary list.** *Zenodo.* 2018.
Publisher Full Text

Wickham H: **Tidy data.** *J Stat Softw.* 2014; **59**(10): 1–23.
Publisher Full Text

Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wilson G, Bryan J, Cranston K, *et al.*: **Good enough practices in scientific computing.** *PLoS Comput Biol.* 2017; **13**(6): e1005510.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Woll B, Crasborn O, van der Kooj E, *et al.*: **Extended Swadesh list for sign languages**. Max Planck Institute for the History of Science, Berlin, 2010.

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 3

Reviewer Report 25 May 2023

https://doi.org/10.21956/openreseurope.17368.r32031

© **2023 Boleda G.** This is an open access peer review report 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.

✔ **Gemma Boleda**
Department of Translation and Language Sciences, ICREA, Catalan Institution for Research and Advanced Studies, Universitat Pompeu Fabra, Barcelona, Spain

No further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Semantics, lexicon, linguistics, computational linguistics, cognitive science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 15 May 2023

https://doi.org/10.21956/openreseurope.17268.r31333

© **2023 Boleda G.** This is an open access peer review report 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.

✔ **Gemma Boleda**
[1] Department of Translation and Language Sciences, ICREA, Catalan Institution for Research and Advanced Studies, Universitat Pompeu Fabra, Barcelona, Spain
[2] Department of Translation and Language Sciences, ICREA, Catalan Institution for Research and Advanced Studies, Universitat Pompeu Fabra, Barcelona, Spain

I congratulate the authors for the revision. There are only two minor remaining changes I would like to see in the article:

- add this clarification that you gave in the response to the body of the article: "The ontological categories are based on the categories used by the World Loanword Database (Haspelmath and Tadmor 2009). However, they were modified regarding the concrete terms used, in order to make clear that we do not talk about concrete parts of speech since we are dealing with concepts and not with words (see List et al. 2016: 2394)."

- It is still not clear what "typed data" are; can you provide a small explanation or paraphrase? (This concerns "NoRaRe also includes mappings of typed data, for example, entries in Wikidata (www.wikidata.org) and BabelNet (www.babelnet.org), which is an added benefit compared to the data in Concepticon.")

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Semantics, lexicon, linguistics, computational linguistics, cognitive science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 16 May 2023

Annika Tjuka

We thank the reviewer for the positive assessment of our revised manuscript. We implemented the two suggestions and uploaded a final version of the manuscript.

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 May 2023

https://doi.org/10.21956/openreseurope.17268.r31332

© 2023 Georgakopoulos T. This is an open access peer review report 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.

Thanasis Georgakopoulos
[1] School of English, Department of Theoretical and Applied Linguistics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2] School of English, Department of Theoretical and Applied Linguistics, Aristotle University of Thessaloniki, Thessaloniki, Greece

No further comments to make.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Historical Linguistics; Lexical Semantics; Lexical Typology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 22 February 2023

https://doi.org/10.21956/openreseurope.16625.r30509

© 2023 Georgakopoulos T. This is an open access peer review report 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.

Thanasis Georgakopoulos
[1] School of English, Department of Theoretical and Applied Linguistics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2] School of English, Department of Theoretical and Applied Linguistics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[3] School of English, Department of Theoretical and Applied Linguistics, Aristotle University of Thessaloniki, Thessaloniki, Greece

1. General overview and assessment

The publication of this data note marks the release of two important resources, namely Concepticon (Version 3.0) and NoRaRe (Version 1.0), which come with new data and various improvements over their predecessors. The joint presentation of Concepticon and NoRaRe is noteworthy as they share several common features but also exhibit certain differences. Specifically, Concepticon consists of data from over 400 concept lists, which are collections of concepts put together by different scholars at various times. These lists contain concept labels that refer to the same concept, which are referred to as concept sets. The current version of Concepticon contains more than 3,900 such concept sets, which allow comparison between multiple languages. NoRaRe is considered an extension of Concepticon with a wider scope, designed to foster the dialogue between different disciplines, mainly psychology and linguistics. It consists of three different types of data, namely norms (No), ratings (Ra), and relations (Re), which can be used to compare different concept/word properties across languages.

The protocols used for the creation of the resources are deemed appropriate, and the work carried out by the editors is technically sound. Both resources tackle the difficult task of standardizing cross-linguistic datasets, which is widely recognized as a challenging undertaking. Curators employ a combination of manual, automated, and semi-automated workflows to maintain and add new data. Both Concepticon and NoRaRe are curated using the GitHub platform

and are archived under the CC-BY license on Zenodo. In addition, they adhere to the Cross-Linguistic Data Formats Initiative (CLDF; see Forkel et al., 2018; https://cldf.clld.org), which aims at representing lexical data in straightforward tabular formats. As a result, the datasets are presented in a useable and accessible format. Overall, all datasets comply with the FAIR principles, according to which the data should be *findable*, *accessible, interoperable*, and *reusable* (Wilkinson et al., 2016). Finally, these CLDF datasets are also available on GitHub and serve as input for the CLLD web applications for Concepticon (concepticon.clld.org) and NoRaRe (norare.clld.org).

The authors provide adequate information about their methods and materials, which enables researchers to reproduce the results of studies that utilize the data provided by these resources. Additionally, this information allows for the addition of new data to the resources by interested researchers. It is worth mentioning that several technical issues are discussed in blog posts authored by the editors, which may further encourage the use of these resources (and which can be seen as complementary to this data note).

In light of the growing need for standardized cross-linguistic datasets and their importance in enabling reproducible research in linguistics, the newer versions of Concepticon and NoRaRe are a timely and valuable contribution. Therefore, I recommend accepting this data note (with minor changes; see my comments below).

2. Specific comments (on pdf pages)

- p. 5: *The authors write that "Concepticon currently has 3,914 concept sets with an average of 231.76 concept sets mapped in a given list."*
  - It would be informative to learn which lists have the highest and lowest coverage of concept sets. A table featuring the top 5 and bottom 5 lists would be sufficient.
- p. 5: *In the paper, the term 'elicitation gloss' is used.*
  - The authors can provide further clarification on the term (and the rationale behind its use). My understanding is that the term "elicitation gloss" is equivalent to "word" in NoRaRe, whereas in Concepticon, it is equivalent to "concept." This difference could potentially be confusing for researchers who intend to use both resources. As was done in Tjuka et al. (2022a), it would be helpful if the authors could elaborate on the distinction between concepts and word forms and the reasons why one could use them interchangeably.
- p. 6: *The authors write that "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."*
  - Providing an example where the described limitation made it impossible to create a mapping would be helpful for the reader.
- p. 7: *The authors write that "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*.
  - Could you offer additional clarification on the term "foreign key constraints" as used in this context? As far as I understand, it pertains to a set of rules that ensure that the values in one table correspond to the values in another table. Nevertheless, it may be beneficial to provide further explanation here.
- p. 7: The URL for the Python package pyconcepticon (pypi.org/project/pyconcepticon/3.0.0) leads to the Concepticon collection, whereas the URL for the Python package pynorare (pypi.org/project/pynorare/1.0.1) does not provide direct access to the NoRaRe collection.

- p. 7: The authors make reference to csvkit (pypi.org/project/csvkit/1.0.7), a tool that enables the processing and detailed analysis of the data in Concepticon and NoRaRe. It is worth noting that a newer version (1.1.0) of this tool is available.
- P 7: For readers who are not acquainted with the Concepticon, the term *"narrow concept set"* may be unfamiliar. Could you please provide further explanation of the distinction between broad and narrow concepts (even though the content becomes obvious with the examples that follow)?
- p. 7: *The authors write: "1,262 BROTHER, i.e., 559 BROTHER (OF MAN), 560 BROTHER (OF WOMAN), 1,759 OLDER BROTHER, 1760 YOUNGER"*
  - Please remove the commas in the numbers. The correct text should read as follows: "1262 BROTHER, i.e., 559 BROTHER (OF MAN), 560 BROTHER (OF WOMAN), 1759 OLDER BROTHER, 1760 YOUNGER BROTHER."
- p. 8: The authors may refer to the newer version of the clld toolkit (clld 10.0.0)

References

1. Forkel R, List J, Greenhill S, Rzymski C, et al.: Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*. 2018; **5** (1). Publisher Full Text
2. Tjuka A, Forkel R, List J: Linking norms, ratings, and relations of words and concepts across multiple language varieties. *Behavior Research Methods*. 2021; **54** (2): 864-884 Publisher Full Text
3. Wilkinson M, Dumontier M, Aalbersberg I, Appleton G, et al.: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016; **3** (1). Publisher Full Text
4. List JM, Tjuka A, Rzymski C, Greenhill S, et al.: Concepticon 2.6.0. Max Planck Institute for Evolutionary Anthropology. 2022. Reference Source

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Historical Linguistics; Lexical Semantics; Lexical Typology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 27 Feb 2023

Annika Tjuka

We thank the reviewer for this detailed assessment of our manuscript. We will now start preparing a new version of the manuscript, addressing the points raised in the two reviews.

Competing Interests: No competing interests were disclosed.

Author Response 21 Apr 2023

Annika Tjuka

We would like to thank the reviewer for the extensive and detailed summary of our data note. It shows that the manuscript was carefully evaluated in light of the larger context of our study. The point regarding our blog encouraged us to include a reference to it in the introduction. We implemented the suggested changes and respond to the reviewer's comments below (order: reviewer comment, our response, implemented change).

- p. 5: The authors write that "Concepticon currently has 3,914 concept sets with an average of 231.76 concept sets mapped in a given list." It would be informative to learn which lists have the highest and lowest coverage of concept sets. A table featuring the top 5 and bottom 5 lists would be sufficient.

We thank the reviewer for this suggestion. In light of the comments from reviewer 1, who suggested a table with a different set of descriptive statistics, we decided to include only one additional table (Table 3). However, we would like to refer the reviewer to the Concepticon GitHub repository, where additional statistics can be found, which are updated with each release. Furthermore, the "Representation" tab on the Concepticon website shows how many concept lists are mapped to a specific concept set. The list can be sorted to show the highest and lowest number of coverage. The relevant links are listed below.

- https://github.com/concepticon/concepticon-data/tree/v3.0.0/concepticondat
- https://concepticon.clld.org/parameters

Table 3: Concepticon 3.0. NoRaRe 1.0. Lists 413 Datasets 113 Concepts sets 3914 Word properties 75 Languages 56 Languages 39 Language families 14 Language families 8

- p. 5: In the paper, the term 'elicitation gloss' is used. The authors can provide further clarification on the term (and the rationale behind its use). My understanding is that the term "elicitation gloss" is equivalent to "word" in NoRaRe, whereas in Concepticon, it is equivalent to "concept." This difference could potentially be confusing for researchers who intend to use both resources. As was done in Tjuka et al. (2022a), it would be helpful if the authors could elaborate on the distinction between concepts and word forms and the reasons why one could use them interchangeably.

We thank the reviewer for the suggestion. To clarify the use of the terms, we included a glossary at the beginning of the manuscript (Table 1). Table 1: Term Definition elicitation gloss In linguistic fieldwork, an "elicitation gloss" is used to denote a concept in a metalanguage. For example, when English is used as a metalanguage, a researcher would use the elicitation gloss tree to elicit the expression for the concept TREE in another language, while a Spanish researcher would use the elicitation gloss árbol to elicit the same concept.

- p. 6: The authors write that "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."
Providing an example where the described limitation made it impossible to create a mapping would be helpful for the reader.

We integrated the suggestion of the reviewer by adding the following sentence. "For example, if a list contains the gloss *smoke* without any further information about whether the verb or noun meaning is intended, we do not map it to one of the concept sets 778 SMOKE (EXHAUST) or 1689 SMOKE (INHALE)."

- p. 7: The authors write that "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.
Could you offer additional clarification on the term "foreign key constraints" as used in this context? As far as I understand, it pertains to a set of rules that ensure that the values in one table correspond to the values in another table. Nevertheless, it may be beneficial to provide further explanation here.

Reviewer 1 had similar concerns about the sentence so we paraphrased it to make the terminology clearer. "CLDF datasets follow the data model of relational databases, where tables can link to each other using foreign keys. This means that CLDF datasets can automatically be converted to relational databases to facilitate efficient data access. It also means that CLDF datasets follow the recommendations for "tidy" data (Wickham 2014; Wilson et al., 2017), in particular by providing unique identifiers for each row of a table."

- p. 7: The URL for the Python package pyconcepticon (pypi.org/project/pyconcepticon/3.0.0) leads to the Concepticon collection, whereas the URL for the Python package pynorare (pypi.org/project/pynorare/1.0.1) does not provide direct access to the NoRaRe collection.

We thank the reviewer for the thorough check of our link collection. We decided to reference a link to the PyPI page rather than GitHub for the Python packages. Now that data access works via CLDF, pynorare and pyconcepticon are only implementation details. We have stopped archiving releases on Zenodo for Python packages since PyPI is the authoritative archive. That is why we no longer make releases on GitHub (which only served to move the Python package to Zenodo).

- p. 7: The authors make reference to csvkit (pypi.org/project/csvkit/1.0.7), a tool that enables the processing and detailed analysis of the data in Concepticon and NoRaRe. It is worth noting that a newer version (1.1.0) of this tool is available.

We updated the link to the newest version of csvkit as suggested by the reviewer. "Tools such as csvkit (pypi.org/project/csvkit/1.1.1) allow the processing and in-depth study of the available data in Concepticon and NoRaRe."

- P 7: For readers who are not acquainted with the Concepticon, the term "narrow concept set" may be unfamiliar. Could you please provide further explanation of the distinction between broad and narrow concepts (even though the content becomes obvious with the examples that follow)?

We added further information on the broader/narrower concept distinction. "The relations broader versus narrower are used to indicate concept sets that refer to specific parts of a more general concept. For example, we established the broader concept set 3626 KNOW to relate it to the narrower concept sets 1410 KNOW (SOMETHING) and 2248 KNOW (SOMEONE) because we found that several concept lists did not provide a clear distinction between the two concepts. It is important to note that the Concepticon does not provide a

full ontology but rather that the relations are established bottom-up and mainly serve as an exploration of the data instead of providing a basis for inferences."

- p. 7: The authors write: “1,262 BROTHER, i.e., 559 BROTHER (OF MAN), 560 BROTHER (OF WOMAN), 1,759 OLDER BROTHER, 1760 YOUNGER” Please remove the commas in the numbers. The correct text should read as follows: "1262 BROTHER, i.e., 559 BROTHER (OF MAN), 560 BROTHER (OF WOMAN), 1759 OLDER BROTHER, 1760 YOUNGER BROTHER."

The commas were included in the typesetting of the manuscript. We thank the reviewer for noticing them. We deleted all additional commas in the identifiers of concept sets in the revision. "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 1262 BROTHER, i.e., 559 BROTHER (OF MAN), 560 BROTHER (OF WOMAN), 1759 OLDER BROTHER, 1760 YOUNGER BROTHER, etc."

- p. 8: The authors may refer to the newer version of the clld toolkit (clld 10.0.0)

We updated the link to the newest version of clld toolkit as suggested by the reviewer. "The clld toolkit (pypi.org/project/clld/10.0.0) is a Python package that integrates functions for building and maintaining CLLD web applications."

Competing Interests: No competing interests were disclosed.

Reviewer Report 20 February 2023

https://doi.org/10.21956/openreseurope.16625.r30510

© 2023 Boleda G. This is an open access peer review report 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.

Gemma Boleda
[1] Department of Translation and Language Sciences, ICREA, Catalan Institution for Research and Advanced Studies, Universitat Pompeu Fabra, Barcelona, Spain
[2] Department of Translation and Language Sciences, ICREA, Catalan Institution for Research and Advanced Studies, Universitat Pompeu Fabra, Barcelona, Spain
[3] Department of Translation and Language Sciences, ICREA, Catalan Institution for Research and Advanced Studies, Universitat Pompeu Fabra, Barcelona, Spain

SUMMARY

This article presents major releases of two large-scale datasets (and accompanying code and web apps) for research in cross-linguistic phenomena related to the lexicon: Concepticon Version 3.0 and NoRaRe Version 1.0. Concepticon provides a standardized set of concepts that are linked to words/concepts in 413 pre-existing datasets; and NoRaRe links the corresponding words (and their concepts from Concepticon) to word properties available in 113 pre-existing datasets.

Both datasets represent huge compilation and standardization efforts that are very useful for the community and allow for much larger-scale research than was possible before. Furthermore, the way the authors have made the data public follows FAIR principles: the data are easily findable, accessible, interoperable, and reusable. I particularly like it that they have built web apps that make browsing the datasets easy. Another great aspect is that the authors have worked towards a workflow that allows external people to contribute to the datasets. This is an aspect that merits more discussion in the paper.

While the work itself seems to be excellent (as far as I can tell; see below for missing information, which I need to be able to properly evaluate it) and is certainly relevant for the research community, the reporting in the article needs quite a bit of work. In particular, it should provide more information and less technical detail, as well as structure the information better, as follows.

ASPECT TO IMPROVE #1: MORE INFORMATION

The article should be self-contained and provide sufficient information about the dataset for the reader to be able to understand both how it was constructed and the information it contains. The following needs to be addressed:

- Please include a table with relevant descriptive statistics of the datasets in their 3.0 and 1.0 releases (minimally n. of datasets, concepts, words, languages, language families).

- The main notions around which the data are organized should be more clearly explained, with examples. The notion of "concept set" was particularly hard for me to understand; I had to check the websites, and even there it took some digging: The example on the homepage of the CLLD web app (https://concepticon.clld.org/) talks about the concept set SIBLING, so it looks like the other concepts in the network shown (OLDER SIBLING, YOUNGER BROTHER, ...) are members of this set. However, if one queries the interface, OLDER SIBLING and YOUNGER BROTHER are themselves concept sets. It turns out that "concept set" does not refer to sets of concepts in Concepticon, but to sets of concepts defined in the primary datasets. Please explain and illustrate how you use the following terms: "concept", "concept set", "concept label", "concept list". Also explain how these terms relate to words as commonly defined by linguists, and what the place of words is in the Concepticon (including improving the claim that "concept lists look like a list of words. However, words represent concepts in the mind" -- check e.g. G. Murphy's chapter on word meaning in The Big Book of Concepts).

- Also provide information about the ontological categories: how was the set of categories defined, and how were they assigned to specific concepts?

- The mapping between the primary datasets and Concepticon includes an automatic step: "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. (...) If there is a match, the word is mapped." Please explain how the checking and comparing is made, and what the criteria are to decide whether there is a match, at an adequate level of abstraction for the present report. Also, explain "elicitation glosses" -- I thought they were definitions but Figure 1 suggests they are words. Maybe it depends on the dataset?

- Also please provide information on how many datapoints from the primary sources are left

unlinked at present (after the manual and automatic mappings you have done so far).

- Please explain the data curation workflow better, clearly specifying the different steps before moving to a description of each. CLDF is presented in different places and I am still not sure I understand how it is used for the present datasets -- again, examples would help. Also please distinguish clearly in the text between the format and the functionalities it affords.

- CLDF, CSVW, CLLD, ... it gets confusing, please help the reader.

- Please also clarify why you release datasets in two formats (CLDF, another -- which?).

- Also please include a brief note on how CLDF relates to other standardization efforts such as TEI.

ASPECT TO IMPROVE #2: LESS TECHNICAL DETAIL

- Please remove technical details that belong in the documentation, not the article. Two examples:
  - 1) In the description of the Python packages, details such as "Typing *concepticon -h* will give a list of the functionalities." Important: Make sure that the information that is relevant to understand the functionalities of the packages remains.

  - 2) You include links to many specific files in the repositories. In the article, you should provide only the main links and the information that is relevant to understand what they contain. For where to find which specific kind of information, you should rely on the documentation in the repositories.

- Are you sure Table 1 is needed? If so, expand the information you provide -- e.g. line 2 is not understandable ("Concept lists that contain further annotations which exceed the complexity of ranks"?).

ASPECT TO IMPROVE #3: STRUCTURE

The way the information is presented in the current version makes it difficult for the reader to understand the material. I list here what struck me the most, but the article needs to be thoroughly revised in this respect.

- The abstract mixes specific considerations about the historical development of the dataset that is being presented with general motivation and with relevant aspects of the dataset. The line "The Concepticon is based..." talks about the Concepticon as if it were already familiar to the reader.

- In the introduction, after a 3-sentence general motivation, the article jumps into details without giving the general picture: "For this reason, a community-led initiative developed the Cross-Linguistic Data Formats (CLDF, Forkel et al., 2018) [...] The CLDF format specifically targets interoperability and reusability..."

- "test-driven" is mentioned in the abstract and the keywords, but then it's not introduced explicitly in the article: the first mention is "This is the advantage of a test-driven data

curation workflow" -- please explain what a test-driven data curation workflow is beforehand.

- Please revise the structuring into sections: e.g., should section "Consistency and transparency through CLDF" really at the same level as Intro, Materials and Methods, etc.? Can you use a more transparent title? (This section is particularly unclear.) How about the subsections?

- More generally, please remove the many redundancies in the text (including two long links to the markdown file containing the list of contributors!).

CONCLUSION

While the work has merit, I cannot approve the article in its present form. I am looking forward to a revised version.

Typos/other:

- Figure 1 really helps! the "arbitrarité" part was confusing until I understood it was the logo for Concepticon. It's not the most transparent logo. :) (By the way, on the https://concepticon.clld.org/ website it says "arbitraire".)

- The text in Figure 2 is too small to be read.

- p. 3, "preface" ==> "surface".

- p. 4, "The different contents of the lists", not clear what "the lists" refers to here.

- p. 5, "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" ==> ?

- p. 7, "NoRaRe also allows the mapping of typed data such as networks, which is an added benefit compared to the data in Concepticon": what do you mean with "typed data such as networks"? Allows, or has been done already?

- p. 7, "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": please rephrase in non-technical terms and explain why this is good.

- p. 7, is "careless" really what you mean here?

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of methods and materials provided to allow replication by others?
No

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Semantics, lexicon, linguistics, computational linguistics, cognitive science

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 21 Feb 2023

Annika Tjuka

We would like to thank the reviewer for the thorough assessment of our manuscript. We will respond in detail to all points raised and prepare a new version as soon as all reviews are available.

Competing Interests: No competing interests were disclosed.

Author Response 21 Apr 2023

Annika Tjuka

We would like to thank the reviewer for the detailed comments on our manuscript. We have provided more information in the revision and modified the structure of the manuscript. However, we specifically choose the category “Data Note” to provide technical details on the data curation workflow and the relevant Python packages. We, therefore, kept the relevant passages in the new version of the manuscript. We address the reviewers' comments in detail below (order: reviewer comment, our response, implemented change) so that our decisions are transparent, and hope that they are okay with us not following all the suggestions. We try to take into account the specific article format specified by the journal, which we believe is partly in conflict with some of the suggestions in the review.

- Please include a table with relevant descriptive statistics of the datasets in their 3.0 and 1.0 releases (minimally n. of datasets, concepts, words, languages, language families).

We thank the reviewer for the suggestion. In addition to the descriptive statistics mentioned in the text, we added a table (Table 3) so that the data can be conveniently compared.

- The main notions around which the data are organized should be more clearly explained, with examples. The notion of "concept set" was particularly hard for me to understand; I had to check the websites, and even there it took some digging: The example on the homepage of the CLLD web app (https://concepticon.clld.org/) talks

about the concept set SIBLING, so it looks like the other concepts in the network shown (OLDER SIBLING, YOUNGER BROTHER, ...) are members of this set. However, if one queries the interface, OLDER SIBLING and YOUNGER BROTHER are themselves concept sets. It turns out that "concept set" does not refer to sets of concepts in Concepticon, but to sets of concepts defined in the primary datasets. Please explain and illustrate how you use the following terms: "concept", "concept set", "concept label", "concept list". Also explain how these terms relate to words as commonly defined by linguists, and what the place of words is in the Concepticon (including improving the claim that "concept lists look like a list of words. However, words represent concepts in the mind" -- check e.g. G. Murphy's chapter on word meaning in The Big Book of Concepts).

The reviewer is correct, and we noticed that the terminology used in linguistics does not always match the terminology used in psychology. Therefore, some translation work is required. To avoid confusion about the terminology and abbreviations used in our manuscript, we have included a glossary (Table 1) of relevant terms and their definitions.

- Also provide information about the ontological categories: how was the set of categories defined, and how were they assigned to specific concepts?

The ontological categories are based on the categories used by the World Loanword Database (Haspelmath and Tadmor 2009). However, they were modified regarding the concrete terms used, in order to make clear that we do not talk about concrete parts of speech since we are dealing with concepts and not with words (see List et al. 2016: 2394).

- The mapping between the primary datasets and Concepticon includes an automatic step: "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. (...) If there is a match, the word is mapped." Please explain how the checking and comparing is made, and what the criteria are to decide whether there is a match, at an adequate level of abstraction for the present report. Also, explain "elicitation glosses" -- I thought they were definitions but Figure 1 suggests they are words. Maybe it depends on the dataset?

We added a definition of “elicitation gloss” in the glossary (Table 1) at the beginning of the manuscript. We hope that the definition makes it clear that the automatic check compares word forms. The examples in the sentence should make the comparison more transparent.

"For this procedure, the algorithm checks the elicitation glosses that are mapped to a given concept set in Concepticon (e.g., tree, árbol) and compares them with the word in a given list (e.g., tree, forest, wood). "

- Also please provide information on how many datapoints from the primary sources are left unlinked at present (after the manual and automatic mappings you have done so far).

We added an example illustrating our decision to unlink a given concept. However, we would like to refrain from stating the number of unlinked data points. As the Concepticon and NoRaRe are constantly being developed, we add new links in each version and also update the links, for example, when a new concept set is introduced. We update our summary statistics with each release and the most recent numbers are provided here: https://github.com/concepticon/concepticon-data/tree/v3.0.0/concepticondata

Furthermore, data points which are not linked can also be inspected on our web page directly: when looking up the concept set “0 ”, one can see that we have 202 concept lists in which elicitation glosses are not linked to concept sets. These amount to more than 20,000 elicitation glosses (https://concepticon.clld.org/parameters/0). Given that Concepticon by now comprises 120,000 elicitation glosses in total, we succeed in linking more than 80% of the data.

"For example, if a list contains the elicitation gloss smoke without any further information about whether the verb or noun meaning is meant, we do not map it to one of the concept sets 778 SMOKE (EXHAUST) or 1689 SMOKE (INHALE). Elicitation glosses which are not assigned to a concept set are automatically linked to the concept set 0 NA."

- Please explain the data curation workflow better, clearly specifying the different steps before moving to a description of each. CLDF is presented in different places and I am still not sure I understand how it is used for the present datasets -- again, examples would help. Also please distinguish clearly in the text between the format and the functionalities it affords.

We recognize the reviewer's concern about the structure and presentation of our descriptions. We, therefore, moved the description of the Python packages pyconcepticon and pynorare below “Manual and automated mapping” in the “Materials and Methods” section. In this way, the information about the CLDF format is not interrupted by the description of functionalities. Due to the scope of the Data Note, we cannot provide an extensive description of the workflows but we added a note at the beginning of the section “Manual and automated mapping” where interested readers can find more details. Furthermore, we added references to our blog posts in which we provide tutorials on how to use the Python packages (see also comments below and the comment by Reviewer 2).

"Detailed descriptions of our workflows are provided in the articles introducing Concepticon (List et al., 2016b) and NoRaRe (Tjuka et al., 2022a) as well as in the tutorials on our blog (Tresoldi, 2019a, Tresoldi, 2019b; Tjuka, 2020a). Here, we summarize the basic steps of the manual and automated mapping workflows." "Typing concepticon -h will give a list of the functionalities, for example, create_metadata which automatically creates the metadata JSON file. Guides on how to use the functionalities can be found on our blog (Tresoldi, 2019a, Tresoldi, 2019b; Tjuka, 2020a)."

- CLDF, CSVW, CLLD, ... it gets confusing, please help the reader.

We added all abbreviations in the glossary (Table 1) at the beginning of the manuscript.

- Please also clarify why you release datasets in two formats (CLDF, another -- which?).

We clarified the reasons for releasing CLDF versions of the data in the first paragraph of the section “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). While the raw data is updated and improved consistently in the respective GitHub repositories, the CLDF datasets allow for the reuse and exploration of the data in a different way. The data in the CLDF datasets are represented in tabular format and the 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. By releasing the data in Concepticon and NoRaRe as CLDF datasets, tools such as csvkit (pypi.org/project/csvkit/1.1.1) can be used to process and analyse the data conveniently. 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."

- Also please include a brief note on how CLDF relates to other standardization efforts such as TEI.

We added a footnote in which we refer readers interested in the major ideas behind CLDF and how they contrast with TEI to our original publication introducing CLDF from 2018 (Forkel et al. 2018). *Footnote: "We refer readers interested in the major design principles behind CLDF and the contrasts with other standardization efforts to the original publication introducing CLDF by Forkel et al. (2018)."*

- Please remove technical details that belong in the documentation, not the article.

We thank the reviewer for the suggestion but we decided to keep relevant technical details in the manuscript. The reason for the decision is that we consciously choose the category "Data Note" instead of "Research Article" so that we could provide a text that includes all necessary information in one place. While we also provide documentations on GitHub, we noticed that researchers tend to rely on published papers. Further details on how to use specific functionalities can also be found on our blog Computer-Assisted Language Comparison in Practice (https://calc.hypotheses.org/). The data note was intended to bring these different resources together and provide a reference point for researchers who are interested in using our resources. The links to the documentation in the GitHub repository are as follows:

1. General overview: https://github.com/concepticon/concepticon-data/blob/master/README.md
2. Getting started document including tutorials on the blog Computer-Assisted Language Comparison in Practice: https://github.com/concepticon/concepticon-data/blob/master/GettingStarted.md
3. Contributing procedure: https://github.com/concepticon/concepticon-data/blob/master/CONTRIBUTING.md
4. Release procedure:
    - https://github.com/concepticon/concepticon-data/blob/master/RELEASING.md
    - https://github.com/concepticon/norare-data/blob/master/RELEASING.md

- Two examples: 1) In the description of the Python packages, details such as "Typing concepticon -h will give a list of the functionalities." Important: Make sure that the information that is relevant to understand the functionalities of the packages remains.

We added an example of a functionality and further references to tutorials on our blog.

Typing concepticon -h will give a list of the functionalities, for example, create_metadata which automatically creates the metadata JSON file. Guides on how to use the functionalities can be found on our blog (Tresoldi, 2019a, Tresoldi, 2019b; Tjuka, 2020a)."

- 2) You include links to many specific files in the repositories. In the article, you should provide only the main links and the information that is relevant to understand what

they contain. For where to find which specific kind of information, you should rely on the documentation in the repositories.

We thank the reviewer for the suggestion but we decided to keep the links in the manuscript. As mentioned in the comment above, the Data Note is intended to provide a reference point so that researchers find all relevant information in one place. Linking to the various files in the GitHub repository and documentation is a way for us to bundle the information that already exists. We hope that researchers can use the Data Note as an entry point to explore our resources in more detail.

- Are you sure Table 1 is needed? If so, expand the information you provide -- e.g. line 2 is not understandable ("Concept lists that contain further annotations which exceed the complexity of ranks"?).

We included Table 1 (now Table 2) because the tags have been updated since the last publication in List (2018) and we wanted to incorporate the new tags to make it easier for researchers to find the data they are looking for. In addition, the table provides descriptive statistics and examples of lists available in Concepticon and NoRaRe. However, we recognise that some of the descriptions were not sufficient and improved them accordingly.

- The way the information is presented in the current version makes it difficult for the reader to understand the material. I list here what struck me the most, but the article needs to be thoroughly revised in this respect. The abstract mixes specific considerations about the historical development of the dataset that is being presented with general motivation and with relevant aspects of the dataset. The line "The Concepticon is based..." talks about the Concepticon as if it were already familiar to the reader.

We revised the abstract according to the reviewer's suggestion.

Abstract: "Language comparison requires user-friendly tools that facilitate the standardization of linguistic data. We present two resources built on the basis of a standardized cross-linguistic format and show how the data is curated and extended. The first resource, the Concepticon, is a reference catalog for standardized concepts from linguistic research. 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."

- In the introduction, after a 3-sentence general motivation, the article jumps into details without giving the general picture: "For this reason, a community-led initiative developed the Cross-Linguistic Data Formats (CLDF, Forkel et al., 2018) [...] The CLDF

format specifically targets interoperability and reusability..."

We included a sentence which facilitates the transition between the two sentences. *"Therefore, it is necessary to create a standardized format that applies to all languages, and at the same time, to provide the tools for effortless standardization. A community-led initiative has recognized this need and developed the Cross-Linguistic Data Formats (CLDF, Forkel et al., 2018)."*

- "test-driven" is mentioned in the abstract and the keywords, but then it's not introduced explicitly in the article: the first mention is "This is the advantage of a test-driven data curation workflow" -- please explain what a test-driven data curation workflow is beforehand.

We added a sentence in the introduction to clarify what “test-driven” entails. *"We apply a test-driven data approach, i.e., specific tests are carried out that match the formal requirements of the data automatically with the specifications developed in CLDF."*

- Please revise the structuring into sections: e.g., should section "Consistency and transparency through CLDF" really at the same level as Intro, Materials and Methods, etc.? Can you use a more transparent title? (This section is particularly unclear.) How about the subsections?

We revised the structuring into sections according to the reviewer’s comments above so that the descriptions of the Python packages are in the “Materials and methods” section. We used the section on CLDF to integrate the subsections on “Using and representing the data” under one umbrella. We also updated the heading of the section. *Heading: "Increasing reusability through Cross-Linguistic Data Formats"*

- More generally, please remove the many redundancies in the text (including two long links to the markdown file containing the list of contributors!).

We thank the reviewer for this comment, which emphasizes the relevance of readability. Due to the style guide of the Open Research Europe Journal, we have included all links in the text rather than in footnotes. Since we cannot assume that everyone will read the text from beginning to end, we decided to add the links in multiple places so that the data can be conveniently accessed. We also considered deleting the link to the contributors but decided to keep it. The data collection in Concepticon and NoRaRe is a collaborative effort and the link to contributors highlights how many people have contributed over the years.

- While the work has merit, I cannot approve the article in its present form. I am looking forward to a revised version.

We thank the reviewers for their valuable comments which contributed to improving the manuscript. We addressed all comments and uploaded a revised version of the manuscript.

- Figure 1 really helps! the "arbitrarité" part was confusing until I understood it was the logo for Concepticon. It's not the most transparent logo. :) (By the way, on the https://concepticon.clld.org/ website it says "arbitraire".)

We exchanged the logo in the figure with the logo from the Concepticon website.

- The text in Figure 2 is too small to be read.

We enlargened the part of the figure which was relevant to illustrate the difference on the web page.

- p. 3, "preface" ==> "surface".

We changed the typo. *"On the surface, concept lists look like a list of words."*

- p. 4, "The different contents of the lists", not clear what "the lists" refers to here.

We added further information to specify which lists were meant. *"The different contents of the concept lists in Concepticon are accounted for by giving them tags which make it more straightforward to search for a particular kind of data (List, 2018)."*

- p. 5, "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" ==> ?

We paraphrased the sentence. *"For example, we compared word frequencies across English, German, and Chinese, and the results showed that word frequencies were similar (Tjuka, 2020b). Another study compared sensory modality across English, Dutch, and Italian showing subtle differences in each sensory modality (Tjuka, 2022). Thus, the database can be used to compare word properties and reveal both cross-linguistic similarities and differences."*

- p. 7, "NoRaRe also allows the mapping of typed data such as networks, which is an added benefit compared to the data in Concepticon": what do you mean with "typed data such as networks"? Allows, or has been done already?

We specified the data type. *"NoRaRe also includes mappings of typed data, for example, entries in Wikidata (www.wikidata.org) and BabelNet (www.babelnet.org), which is an added benefit compared to the data in Concepticon."*

- p. 7, "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": please rephrase in non-technical terms and explain why this is good.

We rephrased the sentence with less technical terminology. *"CLDF datasets follow the data model of relational databases, where tables can link to each other using foreign keys. This means that CLDF datasets can automatically be converted to relational databases to facilitate efficient data access. It also means that CLDF datasets follow the recommendations for "tidy" data (Wickham 2014; Wilson et al., 2017), in particular by providing unique identifiers for each row of a table."*

- p. 7, is "careless" really what you mean here?

We deleted the adjective from the sentence. *"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, incorrect numbering of the concept sets, and many more mistakes."*

Competing Interests: No competing interests were disclosed.