Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

# Comparing NoRaRe data sets: Calculation of correlations and creation of plots in R

Annika Tjuka
Department of Linguistic and Cultural Evolution
Max Planck Institute for the Science of Human History

In a recent blog post, I introduced the Database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (NoRaRe) and demonstrated how to add new data sets (Tjuka 2021). The database currently includes 65 unique word and concept properties based on 98 different data sets across 40 languages (NoRaRe v0.2, Tjuka et al. 2021a) and can easily be expanded further. But what can we do with the data? The article presenting the NoRaRe database already included three case studies that illustrate the application of the database (Tjuka et al. 2021b). This blog post, therefore, provides a tutorial on how to compare NoRaRe data sets in R by conducting a new case study that correlates ratings on arousal in English and Dutch.

## How to find comparable data sets

The first step is to find out which data sets are available in NoRaRe. Since the NoRaRe database is constantly updated, it is best to retrieve the latest version either from the web interface (https://digling.org/norare/) or GitHub (https://github.com/concepticon/norare-data). The web interface will show you an overview of the available data and the GitHub repository includes the data in its entirety. The current version of NoRaRe (v0.2, Tjuka et al. 2021a) contains 16 data sets of the type *norms*, 54 reflect *ratings*, and 34 belong to the data type *relations* (sometimes several data types are included in one data set). Data such as word occurrence counts in a corpus (i.e., word frequency) are categorized as norms. The data type ratings includes studies based on participant judgments, for example, age-of-acquisition or sensory modality. Belonging to the data type relations are semantic field categorization, semantic networks, among others.

*Computer-Assisted Language Comparison in Practice*
Volume 4, Number 11
https://calc.hypotheses.org/3109/

Published on 24/11/2021

The easiest way to find comparable data sets is the file `norare.tsv` on GitHub: https://github.com/concepticon/norare-data/tree/v0.2/norare.tsv. This file includes all NoRaRe data sets that are available for comparison. It can be opened in any text editor, Excel, or similar. Each row in the file is a specific data type found in a given data set and the columns offer several categories to filter the data. The NORARE column indicates whether the data is tagged as *norms*, *ratings*, or *relations* and the TYPE column offers a more fine-grained tag of the data, for instance, *frequency*, *sensory modality*, or *semantic field*. The LANGUAGE column includes ISO language names. Thus, with a simple filter function, one can search for data sets that collected sensory modality ratings for Italian and would find the studies by Morucci et al. (2019) and Vergallito et al. (2020). Importantly, the NOTE column gives a detailed description of how the data was collected. Many rating studies on the same data type such as concreteness or imageability use different scales (e.g., 5-, 7-, or 9-point scales) and are therefore not suitable for direct comparison.

For the present case study, I chose ratings on arousal collected for English and Dutch words on a 9-point scale. Data were taken from studies by Scott et al. (2019) for English and Moors et al. (2013) for Dutch.

## Correlation and plots in R

In the following, I provide a description of how to set up a comparison with NoRaRe data sets. The full script can be found on GitHub: https://github.com/concepticon/norare-data/blob/master/examples/correlation-arousal-valence.R (more scripts are available in the "examples" folder). For this tutorial, I assume that one has downloaded the GitHub repositories for NoRaRe and Conception (List et al. 2016, 2021) as described in Tjuka (2021).

First, the data sets need to be imported. Since all word lists are stored in `.tsv` file format, they can be easily imported with `library(readr)` and `read_delim()` and the file path to the NoRaRe GitHub repository:

```
English_Scott_2019 <- read_delim("PATH/TO/concepticon/norare-
                                 data/concept_set_meta/Scott-2019-Ratings/
                                 Scott-2019-Ratings.tsv", "\t", escape_double = FALSE,
                                 col_types = cols(CONCEPTICON_ID = col_integer()),
                                 trim_ws = TRUE)
```

Note that the path to the data might differ. Data sets stored in Concepticon have the identifier format "Author-Year-Number of items" (e.g., "Lynott-2013-400"), whereas the identifier for NoRaRe data sets is structured differently: "Author-Year-Main property" (e.g., "Lynott-2020- Sensorimotor"). For loading data from Concepticon, use the path to the Concepticon GitHub repository:

```
Lynott_2013_400 <- read_delim("PATH/TO/concepticon/concepticon-
                              data/concepticondata/conceptlists/Lynott-2013-
                              400.tsv", "\t", escape_double = FALSE,
                              col_types = cols(CONCEPTICON_ID = col_integer()),
                              trim_ws = TRUE)
```

Once you have loaded two data sets that you would like to compare, they can be merged with `merge()` to see how many concepts they have in common:

```
overlap_English_Dutch <- merge(English_Scott_2019, Dutch_Moors_2013, by =
                               "CONCEPTICON_ID")

nrow(overlap_English_Dutch)
```

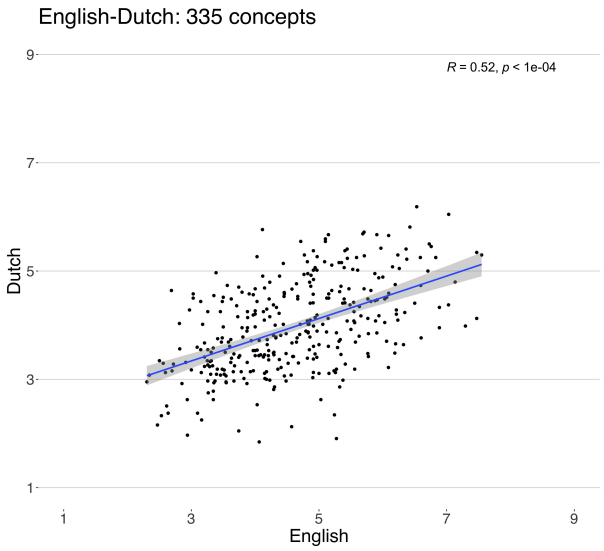
There are several methods for performing correlation analyses. For the present data, I chose a Pearson coefficient (Kirch 2008) analysis.

```
cor.test(overlap_English_Dutch$ENGLISH_AROUSAL_MEAN,
         overlap_English_Dutch$DUTCH_AROUSAL_MEAN, method="pearson")
```

The creation of the plot including a smoothing method can be achieved with `library(ggplot2)`, `library(ggthemes)`, and `library(ggpubr)`.

```
a_arousal <- ggplot(overlap_English_Dutch, aes(x=ENGLISH_AROUSAL_MEAN,
                    y=DUTCH_AROUSAL_MEAN)) + geom_point() +
                    scale_x_continuous(limits=c(1, 9),breaks=seq(1, 10, by = 2)) +
                    scale_y_continuous(limits=c(1, 9),breaks=seq(1, 10, by = 2)) +
                    geom_smooth(method = "gam", formula = y ~ x, se=TRUE,
                    fullrange=FALSE, level=0.95) +
                    labs(title = "English-Dutch: 335 concepts", y="Dutch", x= "English") +
                    stat_cor(method = "pearson", label.x = 7, label.y = 8.75, p.accuracy =
                    0.0001, size = 6) +
                    theme_hc(base_size = 24)
```

The result is the following plot:




## Summary

The NoRaRe database (Tjuka et al. 2021b) offers a variety of data from psychology and linguistics. Especially for cross-linguistic studies, the NoRaRe database is the perfect starting point and properties can be compared easily across languages. This blog post showed how to find comparable data sets, perform a correlation analysis of the data, and create a plot in R. While searching for data sets, the NoRaRe database can also be used to identify gaps, for example, languages for which we lack ratings on sensory modality or consistent age-of-acquisition ratings on the same scale. New data sets will continue to be added to the database in the future (Tjuka 2021) so that other researchers can use the available data for their cross-linguistic studies.

## References

Kirch, Wilhelm. 2008. Pearson's Correlation Coefficient. In: Kirch W. (eds) Encyclopedia of Public Health. Springer, Dordrecht. DOI: https://doi.org/10.1007/978-1-4020-5614-7_2569

List, Johann Mattis, Christoph Rzymski, Simon Greenhill, Nathanael Schweikhard, Kristina Pianykh, Annika Tjuka, Carolin Hundt, and Robert Forkel. 2021. Concepticon. A resource for the linking of concept lists (Version 2.5.0). Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.596412. https://concepticon.clld.org/.

List, Johann-Mattis, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2393–2400. Portorož, Slovenia: European Language Resources Association. https://aclanthology.org/L16-1379/.

Moors, Agnes, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. Behavior Research Methods, 45(1). 169-177. DOI: https://doi.org/10.3758/s13428-012-0243-8.

Morucci, Piermatteo, Roberto Bottini, and Davide Crepaldi. 2019. Augmented modality exclusivity norms for concrete and abstract Italian property words. Journal of Cognition, 2(1). DOI: https://doi.org/10.5334%2Fjoc.88.

Scott, Graham G., Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C. Sereno. 2019. The Glasgow Norms: Ratings of 5,500 words on nine scales. Behavior Research Methods, 51(3), 1258-1270. DOI: https://doi.org/10.3758/s13428-018-1099-3.

Tjuka, Annika. Adding data sets to NoRaRe: A guide for beginners, in Computer-Assisted Language Comparison in Practice, 11/08/2021, https://calc.hypotheses.org/2890.

Tjuka, Annika, Robert Forkel, and Johann-Mattis List. 2021a. NoRaRe. A database of cross-linguistic norms, ratings, and relations for words and concepts (Version 0.2). Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.3957681.

Tjuka, Annika, Robert Forkel, and Johann-Mattis List. 2021b. Linking norms, ratings, and relations of words and concepts across multiple language varieties. Behavior Research Methods. 1-21. DOI: https://doi.org/10.3758/s13428-021-01650-1.

Vergallito, Alessandra, Marco Alessandro Petilli, and Marco Marelli. 2020. Perceptual modality norms for 1,121 Italian words: A comparison with concreteness and imageability scores and an analysis of their impact in word processing tasks. Behavior Research Methods, 52(4). 1599-1616. DOI: https://doi.org/10.3758/s13428-019-01337-8.