Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Handling Non-Standard Datasets in NoRaRe: A Practical Guide

Mira Ahmedović
Department of Linguistic and Cultural Evolution
Max Planck Institute for Evolutionary Anthropology

NoRaRe, the Database of Cross-Linguistic Norms, Ratings, and Relations, is a resource that curates multiple datasets containing information on various properties of words and concepts. When researchers contribute their data, the format and structure can vary widely, presenting challenges for seamless integration. Here, I offer practical guidance for addressing common issues such as data being placed in different sheets, headers in unexpected rows, or datasets contained within `zip`-files. The strategies shared here offer a foundational approach to understanding and adapting NoRaRe's flexibility to accommodate the idiosyncrasy of each dataset.

1 Introduction

NoRaRe – the Database of Cross-Linguistic Norms, Ratings, and Relations – is a compilation of datasets that offer information on various properties of words and concepts, such as emotional ratings, concreteness, frequency, and more (Tjuka et al. 2022a, Tjuka et al. 2023). The current version, Version 1.1, can be found here: https://norare.clld.org/. The datasets are gathered from studies in linguistics and psychology. By linking datasets to the standardized concept sets of the Concepticon (List et al. 2016; Tjuka et al. 2023), NoRaRe allows researchers to explore cross-linguistic similarities and differences while adding context-specific data from published studies. The database is open-access and all researchers are welcome and encouraged to add data that are relevant. A previous contribution contains a general guide to adding datasets to NoRaRe (Tjuka 2021) and the data are publicly available on Github and Zenodo.

When adding data, it quickly becomes evident that no two datasets are exactly alike in terms of their format and structure. Some researchers put their data on the second

Computer-Assisted Language Comparison in Practice
Volume 8, Number 1
URL: https://calc.hypotheses.org/8331 DOI: 10.15475/calcip.2025.1.3

Published on 12/03/2025

sheet of a spread sheet, others have headers on the second or third line while yet others upload their electronic supplementary material in `.zip` files for ease of accessibility of many files at once. Van Zantwijk (2023) provides guidance for preparing better datasets. The following offers a resource for handling special cases – and combinations of them – effectively.

2 Accessing Data from Different Sheets

When a dataset's main data is not on the first sheet of an Excel file, running the command `norare map name-of-dataset` will fail. By default, NoRaRe assumes the primary data resides in the first sheet. This issue can be solved as follows.

In the `norare.py` file:

- Adjust the file name in the `sheet_list` object.
- Adjust the script to specify the correct sheet index that contains the dataset.
- Adjust the `dataset.extract_data` function and add `sheet` as the first argument.

The following example is a chunk of code from the Coso-2023-Emotions dataset (Ćoso et al. 2023).

```
def map(dataset, concepticon, mappings):
    sheet_list = dataset.get_excel(
        'CROWD-5e.xlsx', 1, dicts=False)
    sheet = [
        dict(
            zip(sheet_list[0], row)
        ) for row in sheet_list[1:]]

    dataset.extract_data(
        sheet,
        concepticon,
        mappings,
        gloss='ENGLISH',
        language='en'
    )
```

Here, the `sheet_list` object has an argument after the name of the file, namely `1`, which tells the program where to start extracting the data from – recall that Python uses zero-based indexing so 1 specifies the second sheet. The `sheet` argument tells the program where in the second sheet to look for the header, with `sheet_list[0], row` indicating the first row of the sheet. The second part of the `sheet` argument, namely `for row in sheet_list[1:]` tells the program to extract the actual data from every row after the header.

3 Unusual Headers and Values

Another common occurrence is that some researchers place column headers in lines other than the first because they either split the data into sub data types or add metadata in the first rows of the sheet. This issue will also become apparent once running `norare map name-of-dataset`. NoRaRe assumes that headers are placed in the first line. Furthermore, unexpected null values in the original dataset make correct treatment of the data impossible for the program. A workaround, which mainly consists of adjusting the `norare.py` file, is shown below.

In the `norare.py` file:

- Add a `def replace_null` function, which tells the program how to handle null values. Adjust it for the specific unexpected string in the original data, `#NULL!` in the example below.
- Adjust the file name in the `sheet_list` object.
- Adjust the script and define the `valid_fields` list object. List all columns that are in the dataset in the order that they appear in. The names do not have to be identical to the original file but their order is important to ensure correct extraction. Avoid duplicates.
- Add the first number of the first row of data to the loop that iterates over all rows (`for row in sheet_list`).
- Adjust the `sheet.append` function to replace null values correctly and to iterate over all columns specified in `valid_fields` appropriately.
- Add the `sheet` argument in the `dataset.extract_data` function.

In the `metadata.json` file:

- Use the names in the `valid_fields` list object from the `norare.py` file as the `titles` fields when setting up the metadata.

The following example is a chunk of code from the Bonin-2018-Concreteness dataset (Bonin et al. 2018).

```
def replace_null(v):
    if v == '#NULL!':
        return None
    else:
        return v
```

```
def map(dataset, concepticon, mappings):
    sheet_list = dataset.get_excel(
        '13428_2018_1014_MOESM3_ESM.xlsx', 0,
        dicts=False)
    sheet = []
    valid_fields = [
        'items', 'English translation', 'mean',
        'sd', 'Mean context availability',
        'SD context availability', 'Mean valence',
        'SD valence', 'Mean arousal', 'SD arousal'
    ]
    for row in sheet_list[2:]:
               sheet.append({
            k: replace_null(v)
            for k, v in zip(valid_fields, row[:10])})

    dataset.extract_data(
        sheet,
        concepticon,
        mappings,
        gloss='FRENCH',
        language='fr'
    )
```

4 Handling a Mix of Issues at Once

Cases that combine the previous two issues require the same approach previously discussed. An example of such a case is the Repetto-2023-Sensorimotor dataset (Repetto et al. 2023), which additionally also contains Part of Speech (POS) data. This information is added to the `metadata.json` file as a separate column.

In the `norare.py` file:

- Adjust the file name in the `sheet_list` object.
- Define the column names in the `valid_fields` list object, avoiding duplicates.
- Adjust the `for row in sheet_list` loop to iterate over the data after the header.
- Adjust the `sheet +=` statement to iterate over all `valid_fields`.
- Add the `sheet` argument to the `dataset.extract_data` function.
- Add the `pos` tag to the `dataset.extract_data` function. The `pos_mapper` dictionary contains the shorthand used in the original data on the left and the corresponding NoRaRe category of the POS on the right hand-side. Specify the column name of the POS as it appears in the `metadata.json` file.

In the `metadata.json` file:

- Use the names in the `valid_fields` list object from the `norare.py` file as the `titles` fields when setting up the metadata.
- Define the POS column.

The following chunk of code is an example from the Repetto-2023-Sensorimotor dataset (Repetto et al. 2023).

```python
def map(dataset, concepticon, mappings):
    sheet_list = dataset.get_excel(
        'sensorimotor dataset.xlsx', 1, dicts=False)

    valid_fields = [
        'Ita_Word', 'WordClass', 'Let_ITA', 'FreqColfis',
        'Ln_Colfis', 'FreqRepub', 'Ln_FreqRep',
        'N_OrtNeig', 'MeanFreq_Neig', 'N_Part_p',
        'M_Fam', 'SD_Fam', 'M_Ima', 'SD_Ima',
        'M_Con', 'SD_Con', 'N_Part_a', 'M_Val', 'SD_Val',
        'M_Aro', 'SD_Aro', 'M_Dom', 'SD_Dom',
        'N_Part1', 'M_head', 'SD_head', 'M_foot/leg',
        'SD_foot/leg', 'M_hand/arm', 'SD_hand/arm',
        'M_mouth/throat', 'SD_mouth/throat', 'M_torso',
        'SD_torso', 'N_Part2', 'M_int', 'SD_int',
        'M_taste', 'SD_taste', 'M_smell', 'SD_smell',
        'M_touch', 'SD_touch', 'M_audition',
        'SD_audition', 'M_vision', 'SD_vision',
        'Max_strength.action', 'Exclusivity.action',
        'Dominant.action', 'Max_strength.perceptual',
        'Exclusivity.perceptual', 'Dominant.perceptual',
        'Max_strength.sensorimotor',
        'Exclusivity.sensorimotor',
        'Dominant.sensorimotor'
    ]

    sheet = []
    for row in sheet_list[2:]:
        sheet += [dict(zip(valid_fields,
            row[:len(valid_fields)]))]

    dataset.extract_data(
        sheet, concepticon, mappings,
        gloss='ITALIAN',language='it', pos=True,
        pos_mapper = {
            'n': 'Person/Thing',
            'v': 'Action/Process'},
            pos_name = "ITALIAN_POS")
```

5 Data in a Zip-File

Some datasets are contained in a `.zip` file with additional information. By adjusting the `norare.py` file, we work around this issue.

In the `norare.py` file:

- Adjust the `def download(dataset)` function and change `dataset.download_file` to `dataset.download_zip`.
- Adjust the download link.
- Adjust the `.zip` file name.
- Adjust the name of the file that is supposed to be extracted from the `.zip` file. Use this file name throughout the rest if the `norare.py` file.

The following chunk of code is an example from the Syssau-2009-Valence dataset (Syssau and Monnier 2009).

```
def download(dataset):
    dataset.download_zip(
        'https://staticcontent.springer.com/esm/'
        'art%3A10.3758%2FBRM.41.1.213/MediaObjects/'
        '13428_2010_410100213_MOESM1_ESM.zip',
        'syssau_monnier_norms.zip',
        'syssau_monnier_norms.xls')
```

6 Conclusion

Some dataset come with their idiosyncrasies and not all of them fit perfectly into NoRaRe's default workflows. Whether it is dealing with data placed in different sheets, unconventional headers, `.zip` files, or even combinations of these deviations from the norm, there is always a work-around. By applying the solutions outlined here, accurate integration of datasets can be managed.

The examples presented above aim to illustrate a general approach: understanding the dataset structure, making targeted adjustments in the `norare.py` and `metadata.json` files and utilizing NoRaRe's flexibility to suit your individual dataset. With this in mind, tackling new and unexpected formats becomes a learning process, building upon the tools and techniques shared here.

References

Bonin, Patrick, Alain Méot & Aurélia Bugaiska. 2018. Concreteness norms for 1,659 French words: Relationships with other psycholinguistic variables and word recognition times. *Behavior Research Methods* 50(6). 2366–2387. https://doi.org/10.3758/s13428-018-1014-y

Ćoso, Bojana, Marc Guasch, Irena Bogunović, Pilar Ferré & José A. Hinojosa. 2023. CROWD-5e: A Croatian psycholinguistic database of affective norms for five discrete emotions. *Behavior Research Methods* 55(8). 4018–4034. https://doi.org/10.3758/s13428-022-02003-2

List, Johann-Mattis, Michael Cysouw & 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/

Repetto, Claudia, Claudia Rodella, Francesca Conca, Gaia C. Santi & Eleonora Catricalà. 2023. The Italian Sensorimotor Norms: Perception and action strength measures for 959 words. *Behavior Research Methods* 55(8). 4035–4047. https://doi.org/10.3758/s13428-022-02004-1

Syssau, Arielle & Catherine Monnier. 2009. Children's emotional norms for 600 French words. *Behavior Research Methods* 41(1). 213–219. https://doi.org/10.3758/BRM.41.1.213

Tjuka, Annika. Adding data sets to NoRaRe: A guide for beginners. Computer-Assisted Language Comparison in Practice 4(8). https://calc.hypotheses.org/2890

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2022a. Linking norms, ratings, and relations of words and concepts across multiple language varieties. *Behavior Research Methods* 54(2). 864–884. https://doi.org/10.3758/s13428-021-01650-1

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2022b. NoRaRe. A database of Cross-Linguistic Norms, Ratings, and Relations for Words and Concepts (Version 1.1). Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.14924483

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2023. Curating and extending data for language comparison in Concepticon and NoRaRe. *Open research Europe* 2(141): 1-13. https://doi.org/10.21956/openreseurope.17368.r32031

van Zantwijk, Mathilda. 2023. Five Recommendations for Creating Spreadsheets. Computer-Assisted Language Comparison in Practice 6(2): 93–96, https://doi.org/10.58079/m6m3

Supplementary Material

Data and code are curated on GitHub (https://github.com/concepticon/norare-data/tree/v1.1) and archived with Zenodo (https://doi.org/10.5281/zenodo.14924483).

Acknowledgements

I would like to thank Dr. Annika Tjuka for supervising the creation of this practical guide.