Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Five Recommendations for Creating Spreadsheets

Mathilda van Zantwijk
Department for Linguistic and Cultural Evolution
Max Planck Institute for Evolutionary Anthropology

Through the rapid increase in digital data, the use of tabular formats for data has also increased notably. However, the reusability of data is still an issue due to the lack of transparency in the presentation of data in spreadsheets. In our work with the Concepticon, we sometimes encounter spreadsheets provided by researchers that contain information not transparent to an external audience. Therefore, I offer guidelines on how to format tables with data and provide five concrete recommendations.

1 Introduction

In the last 20 years, the amount of digital data has been increasing rapidly in linguistics. Data are often given in tabular formats or spreadsheets. It is crucial to format them well because it influences whether or not your data is usable or reusable. There is a variety of different formats in which researchers provide their data, which makes it almost impossible to compare and reuse the data.

The Concepticon is a reference catalog of standardized concept sets which are mapped to concept lists (List et al. 2016) and is one example where data are provided in tabular formats. When preparing data for the Concepticon, we encounter various files with different formatting. These differences begin with the type of file format and end with the information that is included. For the data to be reusable, for steps of data collection and for decisions on table header names to be transparent, it is crucial to have standard formatting of spreadsheets and use file formats consistently.

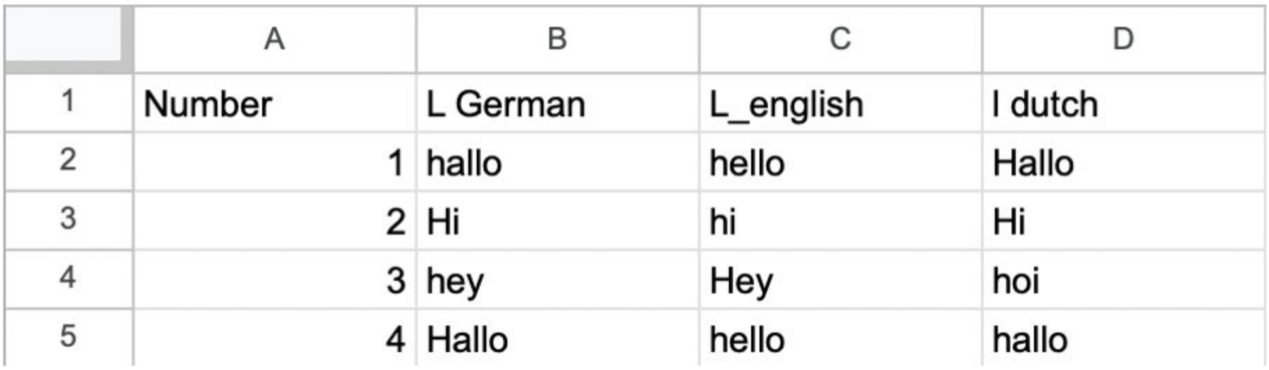
In the following, I begin by showing a number of common mistakes I have encountered during my work. After a short “What not to do” section, I continue by formulating some guidelines for working with spreadsheets and other tabular formats. In the end, I provide a list of further readings.

Computer-Assisted Language Comparison in Practice
Volume 6, Number 2, 93-96
URL: https://calc.hypotheses.org/6573/ DOI: 10.15475/calcip.2023.2.6

Published on 06/12/2023

2 What Not to Do When Working with Spreadsheets

To briefly summarize what not to do when working with a spreadsheet, one can say: be inconsistent. Using a consistent layout and way to format your data is one of the most important aspects to consider when using spreadsheets. The use of inconsistent capitalization, spelling, or inconsistent abbreviations and punctuation throughout the spreadsheet can lead to issues when analysing the data, but also when people want to filter and/or reuse the data given. The following graphic gives a constructed example of what not to do:



	A	B	C	D
1	Number	L German	L_english	l dutch
2	1	hallo	hello	Hallo
3	2	Hi	hi	Hi
4	3	hey	Hey	hoi
5	4	Hallo	hello	hallo

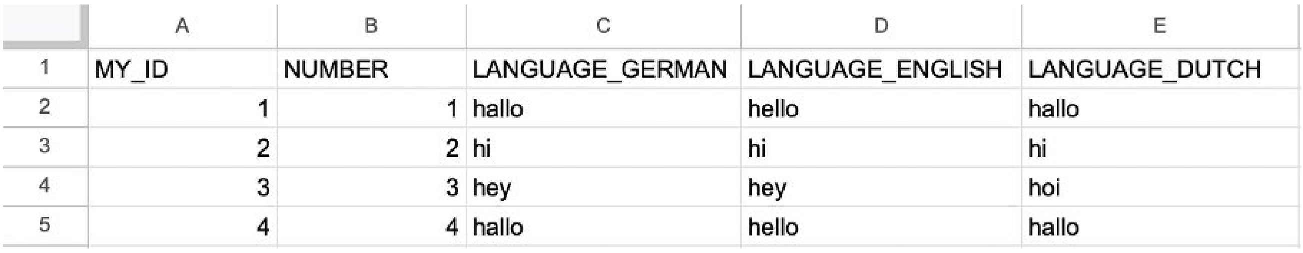
Figure 1: Constructed example of what not to do when creating a spreadsheet. This shows inconsistent use of formatting, capitalization, punctuation, and abbreviations. The data are not consistently formatted either.

3 Guidelines for Working with Tabular Data in Spreadsheets

As already established, using formatting consistently is crucial for making data useable and reusable (cf. Perkel 2022). This begins with the headers and categories chosen. For table headers (see also Tjuka et al. 2022), it is recommended to use capital letters. This allows the headers to be highlighted without the use of other fonts, bold or cursive letters, which get lost when converting the table to .CSV or .TSV formats. Additionally, it is recommended to avoid spaces and use an underscore instead of spaces. When it comes to abbreviations, it is crucial that they are comprehensible to people not familiar with the project. While abbreviations are important, it sometimes makes sense to use a longer term so that there are no obscure combinations of characters, which makes understanding the data from an outward perspective almost impossible.

Another aspect that should be considered is to add a column titled "MY_ID", or similar. When working with the data this is of great benefit since it makes steps you take traceable. For instance, if you remove a data point, try to link something, etc.

Last but not least, be consistent when entering your data. Inconsistent spelling or capitalization makes it virtually impossible to analyse and use the data - this point can therefore not be emphasized enough. On the same note, it is important to enter one piece of information in one cell. While these two tips may seem obvious, especially students encounter this problem when compiling their first datasets.



	A	B	C	D	E
1	MY_ID	NUMBER	LANGUAGE_GERMAN	LANGUAGE_ENGLISH	LANGUAGE_DUTCH
2	1	1	hallo	hello	hallo
3	2	2	hi	hi	hi
4	3	3	hey	hey	hoi
5	4	4	hallo	hello	hallo

Figure 2: Constructed example of what to do when creating a spreadsheet. This shows highlighted headers through capitalization, without opaque abbreviations and spaces. The data are formatted consistently and a MY_ID column was added to allow transparency of steps while working.

To conclude this section, I formulate the following five recommendations:

1. **Be consistent**: Be consistent in your formatting choices.
2. **Highlight headers:** Formulate your headers consistently and highlight them but avoid bold or cursive font and spaces.
3. **Capitalization and punctuation:** Avoid inconsistent capitalization and unnecessary or inconsistent punctuation.
4. **Intelligible abbreviations:** Use transparent abbreviations and add a description of each abbreviation in a separate file.
5. **Traceability:** Add a NUMBER and MY_ID column to make your steps retraceable for others.

4 Further Readings

As already mentioned, the article by Perkel (2022) on tips for better spreadsheets is a good place to start for further reading. There, Perkel formulates six recommendations ranging from consistency to general useability and machine-readability of data. In a more extensive compilation, Broman and Woo (2017) collect specific guidelines on, for instance, how to enter dates and in which format to save your data. An initiative that goes beyond the use of spreadsheets, but also focuses on data structure and reusability, is the CLDF initiative (Forkel et al. 2018). CLDF stands for Cross-Linguistic Data Formats and the initiative proposes new standards for data in historical linguistics and typological language comparison, as well as a framework to incorporate more data types and a software package. Overall, the importance of making data in linguistics and beyond comparable and reusable cannot be overstated.

References

Broman, Karl W. & Kara H. Woo. 2018. Data Organization in Spreadsheets. The American Statistician. Taylor & Francis 72(1). 2–10.https://doi.org/10.1080/00031305.2017.1375989.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarstrom, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics. Scientific Data 5(1). 1–10. https://doi.org/10.1038/sdata.2018.205.

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 10th International Conference on Language Resources and Evaluation, 2393–2400. Portorož, Slovenia: European Language Resources Association. https://aclanthology.org/L16-1379/. https://concepticon.clld.org/.

Perkel, Jeffrey M. 2022. Six Tips for Better Spreadsheets. Nature 608(7921). 229–230. https://doi.org/10.1038/d41586-022-02076-1.

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