Learning on linguistics research data

Myra Spiliopoulou & Prof. Stavros Skopeteas (Univ Göttingen)

Topic 1. Modelling linguistic diversity in grammar

Data

WALS (World Atlas of Linguistic Structures) provides properties of grammars of different languages:

https://wals.info/

This database contains a list of questions about grammar (around 140 variables), e.g., how many cases does the language have (0 to >10) or does the language have tones (no, simple tone system, complex tone system). The variables are rarely binary, in most cases they have 4-6 values. Each language is characterized for these variables: the database contains more than 1000 languages, but there are (too) many missing values.

The data can be found in .csv files in the download area of the above website.

- RQ1: To what extend does the similarity between languages agree with our knowledge about the historical relations between languages? Historical relation means that the languages have either common origin (e.g., Spanish and Italian come from Latin) or have contact to each other (e.g., Basque and Spanish are in contact, but do not have common origin).
- RQ2: How can we incorporate interdependencies among the grammatical variables into our similarity function? The grammatical variables are not independent from each other (e.g., a language has case (nominative/genitive etc.) is more likely to have number (singular/plural) than a language that does not have case.
- RQ3: How to evaluate the similarity solution of RQ1 and RQ2?

The output of this student project is (1) a toolbox that encompasses (1a) similarity functions for RQ1 and RQ2 above, (1b) visualizations of the correlations between languages, (1c) a workflow for learning and (1d) evaluation functions, partitioning scheme and evaluation workflow, as well as (2) documentation and (3) literature discussion for the choice of similarity functions.

Topic 1xt: Augmenting linguistic diversity through lexicon information

Data

The ASJP (Automated Similarity Judgment Program) Database contains a 40-item word list (basic vocabulary such “woman”, “blood”, “egg”, etc.) in a huge amount of languages (around 10000 lists). This data is used to calculate similarity between languages (in the lexicon), which are used to infer historical relationships between languages.

https://asjp.clld.org/

The data can be obtained in the download area of this website. Distances between words are calculated with string comparison, in the simplest case Levenshtein distances, but there is a lot of discussion and proposals in this issue, see a summary article here:

List, Johann-Mattis, Simon J. Greenhill, Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12(1), e0170046.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0170046

- RQ4: How can we augment the similarity functions of RQ1/RQ2 with additional information about the languages? There is already a lot of work on obtaining inferences from this data about historical relationsships. Can we get better insights by taking into account demographical variables, for instance, by applying ‘gravity models’ (population size, distance) for the prediction of similarity?

This topic is not selfstanding. It is offered only as extension to Topic 1 for a large team of more than 7 students total.

The additional output of this topic consists of (1e) an augmentation solution on the basis of ‘gravity models’, (1f) an alternative augmentation solution, (1g) evaluation functions, (1h) an evaluation workflow that compares this solution to the solution of topic 1, as well as extensions to (2+) documentation and (3+) literature study of topic 1.