EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

PHILOSOPHISCHE
FAKULTÄT
Seminar für Sprachwissenschaft

Data Science for Linguists
Summer 2025
Johannes Dellert

Assignment 04:
Analysing Morphological Data in a UD Corpus

handed out: 14 May, 14:00
to be submitted by: 21 May, 12:00

In this sheet, you will use the data preparation and cleanup capabilities of Pandas in order to extract data about the forms of the Basque auxiliary paradigm from a UD corpus in CoNLL-U format (`eu_bdt-ud-dev.conllu`).

Task 1: Loading a CoNLL-U file into Pandas

Read up on the CoNLL-U format for dependency treebanks in case you are not familiar with it. Create a new Jupyter notebook and import `pandas`. Load the development set of the Basque UD corpus from the file provided into a `DataFrame` object, filtering out lines which are empty or start with `#`. Specify lowercase versions of the ten CoNLL-U field names (`id`, `form`, `lemma`, `upos`, ...) as the column index.

Task 2: Handling Missing Data and Cleaning Out Variables

a) Convert all values consisting of an underscore into an appropriate missing data type.
b) Replace all the null values in the column with the morphological features by the empty string.
c) Remove all columns where more than 80% of the values are empty. Which columns have disappeared?

Task 3: Reduction to Datapoints of Interest

a) Derive a reduced dataset which only covers the forms of the lemmas *izan* (“to be”) and *ukan* (“to have”), no matter whether they are tagged as verb (`VERB`) or as auxiliary (`AUX`).
b) Apply vectorised string matching with a regular expression to extract an inventory of all the morphological features which are used on these forms. (Hint: you should see number and person agreement features for three arguments, a unique feature of Basque!).
c) Discard all forms with a feature that occurs less than 50 times in the dataset (these indicate uncommon forms for which we will not enough instances to say anything meaningful).

Task 4: Ingesting Data From a Complex Text Format

a) Apply vectorised string operations which involve regular expressions in order to convert the remaining morphological features into a more useful format (one new column per feature, feature values in each row; example: column `Number[abs]` with values `Sing` and `Plur`).
b) Conventionally, number and person agreement features are not treated separately, but joint together into labels such as `1sg` “first person singular”. Join the columns for each argument (absolutive, ergative, and dative agreement) into columns called `Abs`, `Erg`, and `Dat` with values in this format.
c) Convert all columns where that seems useful into categorical format.

Task 5: Frequencies of Forms and Features

a) How often does each form of *izan/ukan* occur in this development set? Create a new dataframe consisting of the form, the columns with the morphological features, and these counts.
b) Visualise for each feature which share of the tokens is specified for it. As an example, your plot should make it easy to see which percentage of tokens has `Dat` agreement specified compared to, say, `Abs`.