The exam is a paper on a research question of your choice, and using corpus data of your choice, in which you apply one (or more) of the (families of) methods introduced in this course to your corpus data for the purpose of addressing your research question. The research question can, but need not be linguistic.

The corpus data you use can either be one (or several) of the corpora we used in the course, or it could be another set of documents of your choosing. In principle, any collection of text files is acceptable, provided that the documents are a logical source of empirical evidence, given your research question . If you fear that your corpus is too small for the type of analysis you choose to conduct, you can always check with me.

The methods that can be chosen as the 'main method' of the paper include:

collocation and/or keyword analysis

the comparison of variants using (mixed-effects) logistic regression analysis and/or conditional inference tree analysis

the comparison of varieties using correspondence analysis

factor analysis

It is definitely allowed to compare the results from different methods in the paper, thus making the paper (in part) a methodological study, and in such comparisons you are free to also include methods not covered in the course.

Depending on which method you choose as your 'main study', your paper will typically fit in one of the scenarios listed below.

Scenario 1: collocations and/or keywords（I chose this one collostructional analysis）

In scenario 1, you base your paper on an elaborate keyword analysis, collocation analysis or collostructional analysis. In this scenario you should include (and compare) at least three different association measures in your study.

The requirements include at least 3 association measures, which I have selected according to my research questions: LLR, PMI and **Chi-square**

In this type of analysis, which typically has an exploratory nature, the interpretation of the results should cover the careful interpretation of semantic, stylistic or regional (etc) patterns in the (groups of) keywords or collocates as well as the comparison of results across the different measures under scrutiny.

Layout and size of the paper

A rough indication of the size of the paper is about 10 to 15 pages (tables, figures and references included). You can organize the paper any way you want, but the following items should be present:

a detailed description of the procedure for corpus selection (as well as its characteristics), data retrieval and data tagging [technical detail may be put in an appendix]

the hypotheses that you formulated before you conducted the case study, and that you wanted to test by means of the case study

the relevant references to the literature (if any)

analyses based on association scores (scenario 2), a logistic regression analysis (scenario 2), or a dimension reduction technique such as correspondence analysis or factor analysis (scenario 3)

interpretation and discussion of the results

The rules for the size of the paper are not strict, so 10 to 15 pages should be seen as a very rough guideline. Also, it is allowed to add appendices, e.g. for additional tables or figures. There is not stylesheet or set of guidelines for the layout of the paper. For bibliographic references, you can use any stylesheet.

The paper will be written using Quarto, with an output of choice, and submitted along the rest of the code in an R Project. The corpus need not be submitted if it's too big (discuss it with me).

Submission procedure

The deadline for submitting the paper is January 15 (midnight) [or otherwise August 27 (midnight)]. You should submit your full R project either as a zip file via the Assignments tab in Toledo, or by pushing your changes in the Github directory provided at the beginning of the semester.

If you used relatively small corpora (less than twenty million tokens), you should send them to me in a zip-file. However, if you used a large corpus for your case study (more than twenty million tokens), then simply inform me on where I can find that corpus for the purpose of reproducing your analyses. If you used one of the corpora provided in mcl.zip, then there's no need to send it to me.

Evaluation criteria

The following are the main criteria that are used for evaluating the paper.

5 points: clear formulation of research questions / research hypotheses (and appropriate references to the literature if applicable)

5 points: accurate choice of data, design and statistical method, and correct interpretation of the results

5 points: technical mastery of the techniques (e.g. queries, scripts, conversion issues, R instructions) and clear documentation of steps in the R script(s) - if you show code in the paper, do it sparingly (or with code-folding).

5 points: clarity and structure of report and argumentation

In general, depending on your background, you can choose to put more emphasis on the scientific/theoretical underpinning, the computational, or the statistical aspects of your case study, which is perfectly fine, provided that none of the other aspects suffer from this choice. In other words, all should in any case receive sufficient attention for the result to be solid. "Negative results" are also valid results, as long as you show the effort in obtaining meaningful conclusions.

Regarding the easy reproduction of the study: it should be possible for me to pull the changes from your GitHub repository (or unzip the zip-file you submit), and then run your R Script(s) and report without error without making any adjustments to the code. Make sure to clearly indicate at the top of your script(s) where the corpus is stored (so I place my copy in the right location) and which R packages you have used (in case I need to install any).