

Project I: Analysing language

This project consists of three parts and six deliverables. Submit the project report as a pdf and the code in a zip archive through Stud.IP. The file names should read

01_firstname.lastname-firstname_lastname-firstname._lastname.{pdf,zip}

with the first and last names of each group member. The reports are due on Monday evening (20:00 CET) of the **25th November**

If you run into any problems working on this project, describe the problem and/or give a partial analysis.

Part A: Language data


Deliverable 1: A description of the characteristics of two languages, including the number of speakers, geographic regions, language family, and grammatical structure. Language A should be familiar to at least one group member. We encourage you to extend your horizons beyond German and English. Language B should be unknown to every group member. At least one of the languages should belong to a language family outside the Indo-European branch.


For each selected language, ensure that you have access to a large corpus ([Wortschatz Leipzig](#) [↗](#), [Wikipedia's list of text corpora](#) [↗](#)) and a treebank with at least 10,000 sentences ([Universal Dependencies](#) [↗](#)). This will be important for Part B.

The description of each language should include:

- **Name(s) of the language.** Is the name the same in English and the language itself? Is there an older name that fell out of use?
- **Number of speakers.** How many people speak the language? How many use it as their L1?
- **Preservation status.** Is the language endangered? Check [Unesco WAL](#) [↗](#) for its status.
- **Geographic region.** Where is the language spoken? Is there a sizable diaspora? Is the current region the ancestral homeland of that language?
- **Language family.** To which major language family does your language belong (e.g., Indo-European, Afro-Asiatic, etc.)? What are its immediate parent families (e.g., Ger-

manic for Dutch or Sinitic for Mandarin)? Is it part of a Sprachbund? Which other languages are spoken in the same region (contact languages)?

- **Grammar.** What is the typical word order (e.g., SVO, SOV)? Is the language synthetic or analytic? Does it have an un-ergative or un-accusative system? What is the consonant-vowel ratio? Is it a tonal language? How many cases are present? Check [WALS](#)  for interesting features.
- **Orthography.** Which script is used for writing the language? Is it an alphabet, an abjad, a logographic system, or something else? When was the current form of this script established? Is the orthography phonemic?

Follow the [principles of good research practice](#)  and provide primary references for any information you cite.

Deliverable 2: A description of at least one corpus for each language following the datasheet structure described in Gebru et al. (2021).

Answer the questions in the sections on *Composition*, the *Collection process*, and *Preprocessing/cleaning/labeling* with a reasonable amount of research. Distinguish between ‘not findable’ and ‘not applicable’. You might find the examples in the appendix of Gebru et al. (2021) helpful.

Datasheets for datasets is intended to be filled in by the people curating a dataset. You might find certain questions difficult or impossible to answer. Not every question applies to text or language data. For example, when figuring out if information is missing, you might want to check if a profanity filter has been applied and how embedded images were treated in text crawled from the web.

Part B: Language analysis

Deliverable 3: A comparative analysis of word distributions of two corpora using concepts and techniques from the lecture.

Collect statistics of key concepts like

- **Number of sentences.**
- **Number of word forms** (tokens).
- **Number of distinct word forms** (types).
- **Sentence length** in characters and words.
- **Word form length.**

- **Word forms with frequency=1** (hapax legomena).
- **Most frequent words.**
- **Distribution of most frequent words.**
- **Most frequent bigrams, trigrams, etc.**

Compare the distribution of each concept between corpora. Do the languages have a different type–token ratio? Why? What are the averages for each language? How about the 10th and 90th percentile? Provide statistical indicators such as mean and standard deviation and discuss the extremes.

Select one of the following closed word classes and compare the relative frequency of members between corpora.

- Interrogative pronouns: why, what, where, when, etc.
- Articles: the, a, an, etc.
- Conjunctions: and, or, etc.
- Cardinal numerals (1–20): one, two, three, etc.
- Demonstratives: this, that, yonder, etc.

Alternatively, if you prefer to compare the languages on a semantic level, you can look up words and categories in the Swadesh list (Swadesh 1971). You can choose a semantic category and compare the frequency of its members across languages.

Do you think the differences between corpora are caused by the differences between languages, or is another factor in play?

Deliverable 4: A morphological analysis of two corpora using either a self-trained Unigram or BPE tokeniser, including a short description of the training process.

Train a sub-word tokeniser using Huggingface’s tokenizers library. Separate your corpus into a training and test split. Specify and justify which normalisation and pre-tokenization functions you used or did not use for your text. For example, accents (á, à, â, etc.) are not used consistently in the Tagalog corpus, so it might be a good idea to use the StripAccents normaliser.

Apply the tokeniser to your test corpus and provide basic statistics, including average token

length, token frequency distributions, number of tokens per word. Do these values differ for frequent words, infrequent words, and unseen words? Are affixes segmented as you would expect (e.g., plural marker -s in “student-s”)? Is that segmentation consistent or inconsistent across different instances of that feature (e.g., all plural forms)?

Discuss why the segmentation might not follow morphological intuitions and how/whether that might affect certain downstream tasks.

Deliverable 5: A comparative analysis of syntactic structures of two corpora using self-trained dependency parsers, including an evaluation of the parsers.

Split the treebank into a training and test split. Train a dependency parser on the training split for your languages. If your languages already have parsers available through SpaCy, you may use those. In any case, carefully describe the characteristics of the training data and the training parameters.

Follow Jurafsky & Martin (2024: Ch. 19.4) and report the following metrics as both macro- and micro-average:

- labeled attachment score (LAS),
- unlabeled attachment score (UAS),
- label accuracy score (LS).

Report the following statistics on the corpora from the previous tasks:

- Average tree depth.
- Distribution of degrees.
- Average distance of nouns, verbs, adjectives, determiners, conjunctions, ... to the root of the tree.
- Most common leaf node categories.
- Most common ancestors of nouns, verbs, ...
- Most common descendants of nouns, verbs, ...

Select a complex structure (labelled sub-tree of at least three nodes), and describe it with examples from the corpus. In which contexts does the structure appear? Are there outlier contexts in which it appears more often?

Part C: Reflection

Deliverable 6: **A discussion about your work experience.**

Write two to three paragraphs on how you approached the problem and how you split the work in your group. What challenges did you encounter? What surprised you? If you found a task too vague and had to make an interpretive decision, you can also discuss that decision process here.

Bibliography

- Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii & Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM* 64(12). 86–92.
- Jurafsky, Daniel & James H. Martin. 2024. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with language models*. 3rd. Online manuscript released August 20, 2024. <https://web.stanford.edu/~jurafsky/slp3>.
- Swadesh, Morris. 1971. *The origin and diversification of language: Edited post mortem by Joel Sherzer*. Chicago: Aldine.