Natural Language Processing 2025-2026
Homework 1: Words, Morphology and Spelling Variation

Deadline: 10 September (23:59).
Questions? Post them in the HW1 discussion on Canvas or send them to nlp-course@utwente.nl.

This assignment consists of a number of small exercises. Not much programming is required. The main thing to hand in is a report with the answers to the questions below. Please make your report not more than 8 pages long, excluding figures. (And font not smaller than 10pt.)

Please adhere to the Guidelines for using AI during your studies at UT and add an AI declaration to your assignment.

Exercise 1: Morphology (1.5 pt)

Have a look at the first page of Chapter 2 of the book by J&M (edition August 2025). On this page, try to find 3 examples (types, not tokens) of each of the following word formation categories (see lecture slides):

- Inflection
- Derivation
- Compounding

Don't include any that were already discussed in class! Some categories may occur less frequently, so if you can't find three examples of each category on the page, you can list fewer examples.

Explain for one example per category WHY it is an example of this type of word formation, and break this word down into its morphemes as in example (2.6) on page 5 of the book.

Exercise 2: Lemmatization (1,5 pt)

In this exercise you are going to lemmatize the words from the same page of the book. As input you use the file `sorted_types_HW1.txt` provided with this assignment on Canvas. It lists all the word types from the page in question, in alphabetical order. (We skip the step of *tokenization*, which normally comes before lemmatization.)

Stem the word list using the WordNet lemmatizer from NLTK (Natural Language Toolkit), a very useful NLP library. Here you can find a tutorial on how to do that:

`https://www.datacamp.com/community/tutorials/stemming-lemmatization-python`

Lemmatize the words. Based on the lemmatized word list, answer the following questions:

2a (0.5 pt) The original word list contained 238 word types. How many word types (= unique lemmas) are left after lemmatization?

2b (0,5 pt) Give 3 examples of lemmatization errors, if you can find them. Explain WHY they are errors.

2c (0,5 pt) Do you see any patterns, or anything else that's interesting about how some words, or categories of words, are lemmatized? Briefly discuss your observations about the lemmatization.

2d (optional, for 1 pt bonus) Lemmatize the words on the page using the NLP library SpaCy, which also includes a lemmatizer: `https://spacy.io/`. Briefly discuss how the results are different from the ones using the NLTK lemmatizer.

Exercise 3: Creative spelling analysis (1,5 pt)

In this exercise, 'creative spelling' means the intentional misspelling of existing words. For this you use the blog corpus that is provided on Canvas with this assignment. In the corpus, locate file F-train-189.txt. Read the

blog text and look at how the words are spelled. (The file contains the text of multiple blog posts from the same blogger, so don't expect a coherent story.)

3a (0.5 pt) Discuss the blogger's spelling in terms of the transformation categories from the paper of Mosquera & Moreda (2014). These categories are listed in Table 1 of their paper, and explained on the next page of the paper. What type of (intentional) misspellings does this blogger tend to make the most? Give some typical examples from the main categories used by the blogger.

3b (1 pt) Read Section 2.6 of the J&M book, which discusses linguistic variation and how this reflects, among other things, the demographic characteristics of the speaker. What do you think their language use says about the blogger? Briefly (in one or two paragraphs) mention anything that you found noticeable or remarkable.

Exercise 4: Vowel duplications (5,5 pt)

A common spelling variation used online is to duplicate characters for emphasis, for example, *reeeeaaallllyyyy?!* In this exercise we limit ourselves to duplication of vowels (a,e,i,o,u) and we use regular expressions to find words with such duplications in our blog corpus.

If you like working with Python, this has good functions for regular expressions, as explained in the book. If you prefer to use a specific tool, various tools exist that allow you to search a collection of text documents using regular expressions. An example tool for Windows that allows both search and replacement is PowerGrep, which offers a 15 days free evaluation trial: `https://www.powergrep.com/grep.html`.

Use Python or another tool of your own choice to find all cases of vowel duplication in the blog corpus (or at least as many as you can). Use the entire blog corpus for this exercise, ignoring the distinction between training and test files.

4a (1 pt) Provide the regular expression (or set of regular expressions) you used to get your results. Provide an explanation and motivation of the regular expression(s) you used. Also mention which tool you used.

4b (1,5 pt) For each vowel (a,e,i,o,u), provide the top 3 most frequent word types with duplications of that vowel, together with their frequency.

The frequency is the total number of instances in the corpus of that word type. The number of times the vowel gets duplicated does not matter. For example, if you find *coool* 2 times, *cooool* 6 times and *coooool* 3 times , then the frequency of word type *cool* with vowel duplication is 2 + 6 + 3 = 11. Use *case folding* so that tokens with and without capitalisation can be counted together.

4c (1,5 pt) Do the same as exercise 4b, but now separately for for male and female bloggers. Blogs of female bloggers start with F; blogs of male bloggers start with M. Show the results and discuss the differences you found between male and female bloggers.

4d (1,5 pt) Try to use regular expression *substitution*(see Section 2.7.7 from the book) to normalize the vowel duplications in the corpus. Explain how you did this and which (if any) problems you came across. You don't need to come up with a perfect solution: just give it a try, and describe your experiences / any problems you encountered in around half a page.

Handing in

Hand in the following things on Canvas (submission as a group):

- Your answers in a pdf document. Please include the name of both group members in the document!
- For exercise 2, also submit the lemmatized word list(s).