Lab Class NLP:I

Exercise 1 : Natural Language Processing

(a) Define the term *token*.

(b) What is the difference between a *token* and a *type* in natural language processing? Provide an example to illustrate the difference.

Exercise 2 : Morphology

(a) Explain the differences between the root, lemma, and stem of a word and how they relate to derivation and inflection.

(b) What affixes are in the word "reactors"? Identify whether each affix is *derivational* or *inflectional*.

(c) Using the word "reactors" explain the difference between *root* and *stem*.

(d) Consider the following sentence:

```
<s> the chair chaired the meeting </s>
```

(d1) Not counting the begin and end markers, how many tokens and how many types are in this sentence?

(d2) Identify which of the words (if any) have derivational morphology, and which (if any) have inflectional morphology.

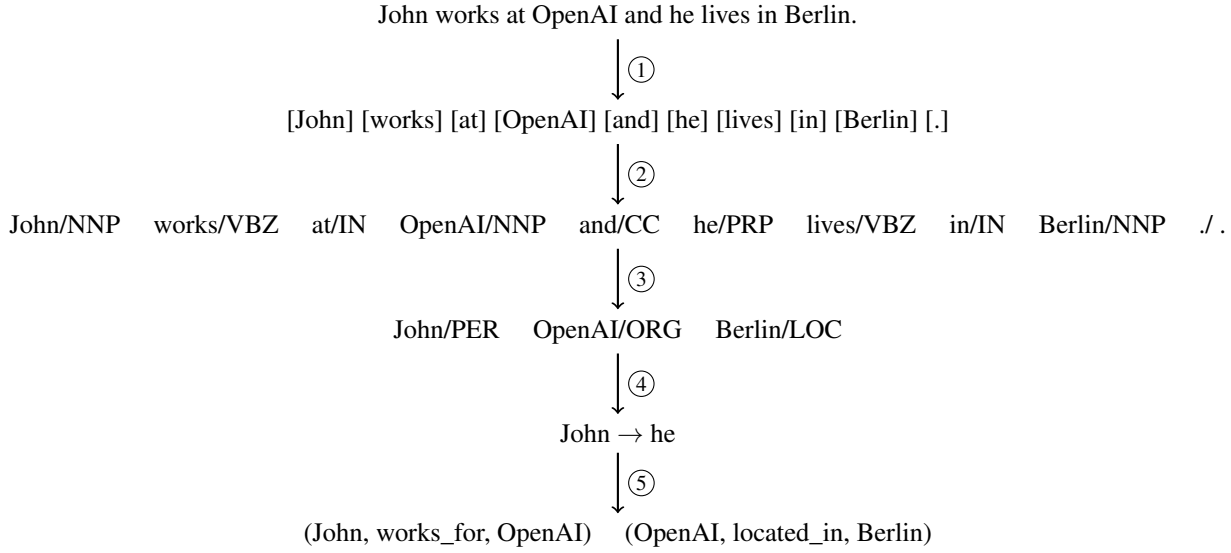
Exercise 3 : Morphology vs. Syntax

What is the difference between *morphology* and *syntax*?

Exercise 4 : Morphology

Each number indicates an NLP process. Assign the correct process name to each number.

© Gutekunst/Hagen 2026



Match the numbers to the processes:

- Entity Relation Extraction
- Tokenization
- Reference Resolution
- Named Entity Recognition (NER)
- POS tagging

Exercise 5 : Datasplits

When training and evaluating a ranking model, the dataset is usually separated into three "splits", **train-**, **test-**, and **validation split**.

(a) What is each of these splits used for?

(b) Why is the data split?

(c) Why do we need to separate evaluation splits? That is, why do we need separate test- and validation splits?

Exercise 6 : Effectiveness vs. Efficiency

Which of the following measures describe the effectiveness, and which describe the efficiency in NLP tasks? Give a short reason (up to 10 words)

(a) Throughput (e.g., tokens/s or texts/s) or latency

(b) BLEU score

(c) Recall

(d) Energy consumption

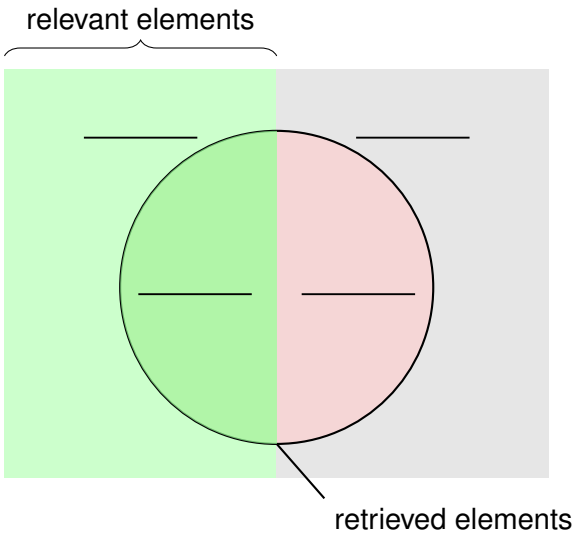
© Gutekunst/Hagen 2026

Exercise 7 : Effectiveness Measures in Analysis Tasks

The diagram below illustrates the relationship between “relevant” and “retrieved” elements and True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs) in an information retrieval task.

(a) Label the four regions in the diagram with TP, FP, FN, and TN.

(b) Complete the formulas for precision and recall in the fields provided.



Precision = _____

Recall = _____

Exercise 8 : Classification Effectiveness: Spam Detection (3 Points)

You developed a spam detector that classifies whether an email is “spam” (positive) or “ham” (negative). To test your spam detector, you evaluate it on the last 100 emails you got. Out of these, 10 were spam. Your detector classifies only 7 emails correctly as spam but manages to classify 92 emails correctly overall.

(a) Draw the confusion matrix.

(b) Compute the Accuracy, Precision, Recall, and F1 scores.

(c) Give one example of an NLP task that requires high precision and one that requires high recall.

Exercise 9 : Inter-Annotator Agreement

You conducted an annotation campaign on humor. Each annotator was given 6 texts and rated each on a 4-point scale from 1 (not funny) to 4 (hilarious). Since you have a fine grasp of the concept yourself, you also annotated each text as the “truth” label. The results of your campaign are as follows:

© Gutekunst/Hagen 2026

Text	**Annotator**						**Inference**		
	A	B	C	D	E	Truth	Majority	Mean	Median
1	1	2	1	1	1	1			
2	2	2	4	2	3	2			
3	2	3	1	2	2	2			
4	4	3	4	3	3	4			
5	1	1	1	1	2	1			
6	3	2	4	2	2	2			
Accuracy						–			
Micro-Precision						–			
Macro-Precision						–			
Micro-Recall						–			
Macro-Recall						–			

(a) Assess the performance of each annotator by calculating the accuracy, and the precision and recall (micro- and macro- averaged) of the annotations.

(b) Assume the annotator would assign labels randomly, what accuracy would you expect in such case?

(c) Infer the final annotation for each text from the votes of the five annotators (A-E) by calculating:

  (c1) the majority vote

  (c2) the mean of votes

  (c3) the median of votes

(d) Assess the performance of the different aggregation strategies (majority vote, mean, and median) by calculating the effectiveness metrics of the inferred annotations. Which one would you choose and why?

(e) Draw the precision-recall curve for the micro-averaged measures.

Exercise 10 : Regression vs. Classification (2 Points)

What is the difference between regression and classification tasks? Come up with 2 examples for each type of task and describe how they can be modelled as a regression/classification task. Include the input features and the output of the model in your examples.

© Gutekunst/Hagen 2026