

**Natural Language Processing  
Final Exam**

June 19th, 2026

1. **[2 points]** Introduce Herdan/Heaps law and Zipf/Mandelbrot law, and discuss their relevance for natural language processing.
2. **[5 points]** An English corpus  $T$  has been tokenized based on white spaces. The resulting dictionary and word frequencies are reported in the following table

| word | play | replay | replayed | stay | stayer |
|------|------|--------|----------|------|--------|
| freq | 17   | 11     | 7        | 19   | 3      |

Apply the byte-pair encoding (BPE) algorithm to derive subword tokens for  $T$ , using the special character ‘\_’ to mark the end of each word. Report and comment each of the first twelve iterations (merge operations) in a run of the algorithm, including the frequency updates.

3. **[2 points]** In the context of the estimation of the probabilities of  $N$ -gram models, explain what is meant for bias/variance tradeoff.
4. **[6 points]** In the context of neural language modeling (NLM), answer the following questions.
  - (a) Why  $N$ -gram language models have been largely supplanted by NLM?
  - (b) Draw and discuss a schematic representation of the general architecture for NLM, as presented in class and reported in the on-line course ‘NLP Course — For You’ by Elena Voita.
  - (c) Introduce the model for feedforward NLM and report the equations for all of the layers of the model along with your own explanation.

(see next page)

5. **[2 points]** In the context of decoder-only models, introduce the notion of key-value cache, motivate its need, and specify how it works in practice.
6. **[5 points]** In the context of pre-training for large language models, answer the following questions.
  - (a) Introduce the notion of self-supervised learning and provide the definition of the associated objective function, along with some discussion of how it works in practice.
  - (b) Explain the notion of teacher forcing and discuss its limitations.
  - (c) Provide information about the most popular datasets used in pre-training of large language models.
7. **[5 points]** In the context of post-training for large language models, answer the following questions.
  - (a) Introduce and motivate the approach of preference-based learning.
  - (b) Define and discuss the Bradley-Terry model as presented in the textbook.
8. **[6 points]** In the context of machine translation, answer the following questions.
  - (a) Introduce the encoder-decoder model based on RNN, dynamic context representation, and attention, providing the model equations and explaining how the model works.
  - (b) Define and discuss the two types of attention that have been presented in the lectures: dot-product attention, and bilinear attention.