# Introduction to Computational Linguistics (WS 23/24)
# Assignment 1

November 9, 2023

1. Describe different types of writing systems and how they vary in their use of phonetic and semantic properties. Include at least two different writing systems and limit your answer to 300 words. [3 points]

2. How do different encoding systems, such as ASCII and Unicode, address the challenge of storing and representing characters in a computer's memory, considering factors like the number of bits used and the potential for misidentification? Limit your answer to 300 words. [3 points]

3. Convert the following binary numbers to hexadecimal and vice versa. Show all the steps involved in the conversion. Clearly show the conversion process for each binary-to-hexadecimal and hexadecimal-to-binary conversion. [4 points]

   - 1100110010101101
   - 1010101101110011
   - 1A3F
   - B7E9

---

**Important:** Please upload your answers in a PDF on Moodle
**before Wednesday, Nov 22, 2023 16:00.** Keep the Zero-Points-Sheet in mind..

# Introduction to Computational Linguistics (WS 23/24)
# Assignment 2

November 13, 2023

1. Describe the difficulties of real-word errors. Limit your answer to 100 words.[1 point]

2. Examine the text below and list at least four different errors. For each, provide an error type, a likely cause, and a sequence of edit operations that correct the misspelling. [3 points]

   *"City of Glass" is a story about a lonely man named Quinn, whose a writer of detective stories. In the middle of the night he recieves a strange phonecall and some one on the other end asks him for help in a criminal vase. The person who called believes that Quinn is a detectiv named Paul Auster and wants to meet him. After the third call, Quinn decides to take Auster's identitz and goes to the meeting as a man he isn't. And and so he gets involved into a case that can't be more confusing and puzling.*
   (adapted from https://schulzeug.at/englisch/summary/paul-auster-city-of-glass/)

3. List all word uni-grams and bi-grams together with their frequencies for the following text [3 points]
   *a cat and a dog went to th beach, and a beautiful day it was!*

4. Search the web for a linguistic corpus that you find interesting. Next, describe your choice in one or two paragraphs. Limit your answer to 350 words and be sure to include the following information: [3 points]
   - Type of linguistic data (e.g., language, written or spoken, speech acts, pronunciation, etc.)
   - Intended purpose of corpus
   - Annotation types (global metadata, linguistic annotations)
   - Method used for collection
   - Data serialization (e.g., file formats)
   - Licensing information (e.g., freely available, or subscription required)

---

**Important:** Please upload your answers in a PDF on Moodle
**before Wednesday, Dec 6, 2023 16:00.** Keep the Zero-Points-Sheet in mind..

---

# Introduction to Computational Linguistics (WS 23/24)
# Assignment 3

November 29, 2023

1. Find and explain 7 potential challenges for tokenization and sentence segmentation in the following text. [7 points, 1 point for each instance]

   *Regarding the contest plans, I've just had a really good idea about what to offer as a prize: we could fly the winners overseas to a really first-rate restaurant. There's one I like a lot in San Francisco, close to 29th Ave. and Rochester... Of course that would mean we'd need enough budget to send them to the good ol' U.S.A. Have a look at their website (it's https://fullbelliesinc.com). Yeah, it's called Full Bellies Inc. - what a ridiculous name! :). I just had a look at their menu. There are some nice choices, e.g. filet mignon and baked potatoes. Have a look and tell me what you think. Of course, we'd have to think about scheduling; we can't just say "OK, you're flying to the States tomorrow", can we?*

2. From Dickinson et al.: You are the owner of a spam filtering service: Currently, your server gets 2000 spams per hour and only 500 good messages. The filter classifies 95% of the spams correctly, and misclassifies the other 5%. It classifies 99% of the good mes- sages correctly, and misclassifies the other 1%. Tabulate the re- sults, How many false positives and how many false negatives do you expect to see each hour? Calculate the precision, recall, and TNR of the spam filter and include the calculation steps. [3 points]

---

**Important:** Please upload your answers in a PDF on Moodle
**before Wednesday, Dec 20, 2023 16:00.** Keep the Zero-Points-Sheet in mind..

---

# Introduction to Computational Linguistics (WS 23/24)
# Assignment 4

December 22, 2023

1. Explain in natural language what charaCter Sequences Are matched by the following regular expressions. [2 points each]

   (a) /Ha(ha)+!?/gi

   (b) /(\+49 |0)1615[0-9]{4}[5-9]{3}/g

2. Try the same search in several different search engines (Google, Bing, Baidu, DuckDuckGo, Ecosia, or others). What similarities or differences do you observe in the results and their ranking, as well as in the user experience, design, and advertisements? Give examples of the queries you searched for, and limit your answer to 2 - 3 paragraphs. [4 points]

3. Search the web for a Computer-Assisted Language Learning (CALL) application that uses Human Language Technology. Both commercial applications and applications from research projects are acceptable. Next, describe the application in two or three paragraphs, being sure to include the following information:

   - the name of the application you chose, and whether the system is a commercial application or a research project [1 point],
   - the intended target users (i.e., the kind of people expected to use the application) [1 point],
   - the type of language training offered by the app (e.g., listening comprehension, grammar practice, etc.), and what if any feedback is offered [2 points], and
   - whether or not the app has the task of processing well-formed (i.e., native) language, language produced by learners, or both. [1 point]

   Finally, speculate on which of the language technologies we have covered in class might be used in the CALL application's language processing pipeline. Examples of technology we have covered include dialogue systems, spell checking, language models, sentence segmentation, tokenization, POS tagging, parsing, and automatic speech recognition. [3 points]

---

**Important:** Please upload your answers in a plain text file on Moodle before January 12, 2024 16:00.

24 Jan Wed 16.00

# Introduction to Computational Linguistics (WS 23/24)
# Assignment 5

January 10, 2024

Limit all your answers to 150 words. The last question is a bonus question, meaning you can get up to four extra points in this assignment. Limit your answer to the bonus question to 300 words.

1. Define dialogue systems and differentiate between task-specific and task-independent dialogue systems. [1 points]

2. Describe two common implementation techniques for dialogue systems and discuss the pros and cons. [2 point]

3. Explain Grice's Maxims and their relevance to dialogue systems. [1 point]

4. Provide an example of a miscommunication in a dialogue system and explain how it might be rectified. [1 point]

5. Give the POS tags of all tokens in the following sentence including a short explanation. Explain if there are ambiguities and why. Use the Penn Tree Bank tagset. [2 points]
   *Time flies like an arrow. Fruit flies like bananas.*

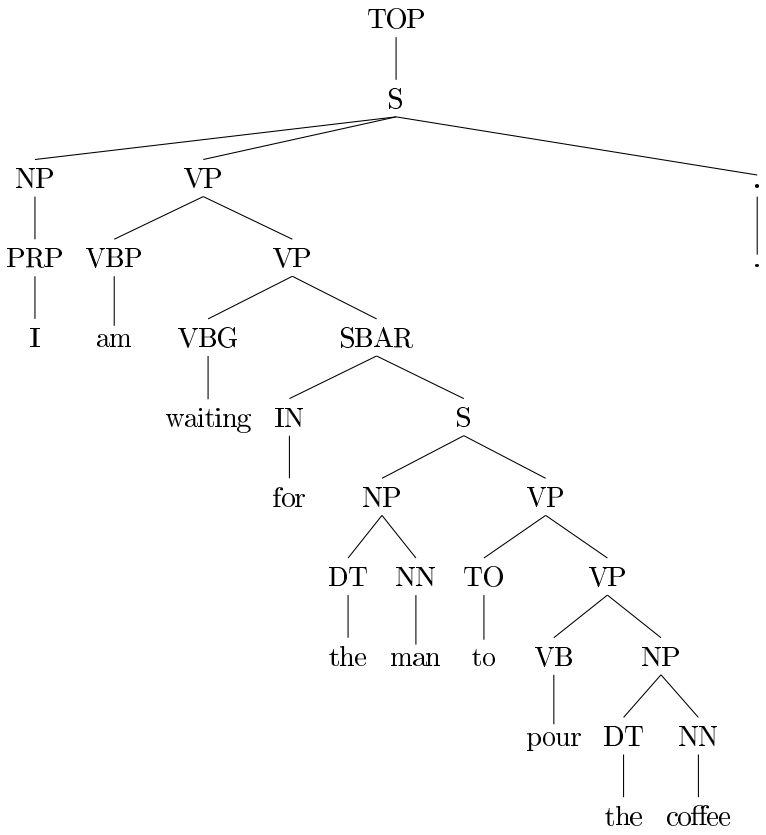
6. Provide a parentheses notation for the following syntactic parse tree. [3 points]

7. **Bonus**: Provide a short dialogue and analyze it in terms of speech acts, turn-taking, and adjacency pairs. You can provide a fictive dialogue that you come up with or generate a dialogue using generative AI. Do not use one of the dialogues discussed in the lecture. [4 points]

---

**Important:** Please upload your answers in a PDF on Moodle before January 24, 2024 16:00.

---


TOP
S
NP
VP
.
PRP
VBP
VP
.
I
am
VBG
SBAR
waiting
IN
S
for
NP
VP
DT
NN
TO
VP
the
man
to
VB
NP
pour
DT
NN
the
coffee