UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Exam in INF2820 Computational linguistics
Day of exam: 1 June 2017
Exam hours: 1430-1830
This examination paper consists of 4 page(s).
Appendices: 0
Permitted materials: none

Make sure that your copy of this examination paper
is complete before answering.

- **You may answer in English, Norwegian, Danish or Swedish.**
- **You should answer all the questions. The weights of the various questions are indicated.**
- **You should read through the whole set to see whether anything is unclear so that you can ask your questions to the teachers when they arrive.**
- **If you think some assumptions are missing, make your own and explain them!**

Exercise 1 (15% for the whole exercise)

Question 1.1 (5%)

Explain briefly what is meant by *word occurrence, word form*, *lexeme*, and *lemma* taking (1) as departure

1. *mouse, mice, rat, rats*

together with the text

2. *Rats aren't smarter than mice, and that's important. "Anything we could train a rat to do we could train a mouse to do as well", says Tony Zador. This finding is important because using mice in experiments instead of rats could open up all kinds of new research options.*

Question 1.2 (5%)

What is stemming and what is lemmatization? What are the differences between the two?

Question 1.3 (5%)

What is meant by tokenization in language technology? With the text (2) above as point of departure, discuss briefly choices that may arise when constructing a tokenizer.

Exercise 2 (15% for the whole exercise)

Let grammar G be the following:

```
S   -> NP VP
NP  -> ND
NP  -> DET NI
NP  -> NP PP
PP  -> P NP
DET -> NP G
VP  -> lo | smilte | danset
ND  -> faren | mora | broren | søstera | kona | mannen
NI  -> far | mor | kone | mann | søster | bror
P   -> til
G   -> sin
DET -> en | ei
NP  -> William | Kate | Elizabeth | Philip |
      Charles | Michael | Pippa | Camilla
```

Question 2.1 (8%)

Draw the two trees the grammar assigns to the word sequence:

3. *Faren til William sin kone danset*

Qestion 2.2 (7%)

How many different syntactic analyses does sequence (4) have?

4. *Faren til William sin kone sin søster danset*

Do all the syntactic analyses have different meanings? Explain which pairs that have the same meaning – if any – and which pairs that have different meanings – if any – and what the differences in meaning are.

Exercise 3 (20% for the whole exercise)

Question 3.1 (10%)

Convert grammar G into a grammar G1 in Chomsky normal form (CNF) and show how the CKY-algorithm will recognize sequence (4).

Question 3.2 (10%)

Show how a chart parser for grammar G can recognize sequence (5).

5. *William sin kone danset*

Exercise 4 (10%)

Is the language generated by grammar G regular? State reasons for your answer.

Exercise 5 (20%)

Possessive relations and similar relations, e.g., family relations, in Norwegian can be expressed by *sin* as we have seen so far. However, they may alternatively be expressed with a genitive *s*. For example, sentence (6) has as many analyses as sentence (4) with corresponding interpretations.

6. *Faren til Williams kones søster danset*

There are several options for describing such relations. Here we choose to consider *Williams* and the other word forms with genitive *s* as regular words and then modify the syntactic rules from grammar G to suit this choice. First, we add the genitive word forms to the lexicon.

```
ND  -> farens | moras | brorens | søsteras | konas | mannens
NI  -> fars | mors | kones | manns | søsters | brors
NP  -> Williams | Kates | Elizabeths | Philips |
      Charles' | Michaels | Pippas | Camillas
```

You should then extend and modify grammar G with rules such that the grammar generates sequence (6) with the correct number of analyses. The genitive construction may be repeated an arbitrary number of times (*søsteras brors mors fars* ...).

You should also equip the grammar rules with features such that the grammar does not overgenerate. In particular, it should not generate

7. ** Moras danset*
8. ** Mora søster danset*

In addition, you should use features to get correct agreement with the indefinite article:

ei mor, ei søster, ei kone
en mann, en far, en bror

We will not accept

** ei mann,* ei far,* ei bror*

But we will accept

en mor, en søster, en kone

Then you should draw the trees – with features – the grammar ascribes to the sequence.

9. *Faren til Williams kone danset*

Exercise 6 (20% for the whole exercise)

We will study classification and we will use the name set from NLTK. It contains 7944 names which each is ascribed one of the two classes *male* or *female*. We put aside 794 names for testing and use 7150 names for training. Of these 7150 names, 4500 names are classified as *female* and 2650 are classified as *male*.

Question 6.1 (10%)

We will make a Naive Bayes classifier. It will only consider one feature, the last letter of the name. For the letter *n,* there are 335 names in the training set classified as *female* and 425 names classified as *male*. Explain how the classifier will proceed to classify the name *John*. You don't have to smooth the values.

Question 6.2 (5%)

This classifier has to ascribe the same class to *John* and to *Marilyn*. To separate between the two, we add another feature. We consider the suffix consisting of the last two letters of the name. There are 70 names in the class *female* having *yn* as the last two letters and 10 names in the class *male* having *yn* as the last two letters. No *female* names has *hn* as the last two letters, while 3 *male* names have *hn* as the last letters. How will this classifier proceed to classify M*arilyn* and *John*? Will there be any changes in class assignment when we add this feature?

Question 6.3 (5%)

The feature *hn* is quite rare. The same applies e.g., to *wn* which occurs 3 times in the *female* class and once in the *male* class, and *ln* which occurs once in the *male* class and 0 times in the *female* class. Explain why this might be a problem. One way to correct for this is to use smoothing, where the simplest form is called "add-one" smoothing. How can we use this method for smoothing the values of the suffix feature?

THE END