Exercise Sheet 7
Due: Tuesday, 19. December 2006

Exercise 7.1

a.) Part-of-speech-tagging is the task of assigning (grammatical) word categories or "tags" to the individual words in a sentence. Manually assign tags to the words in the sentence "I saw a man with a telescope". Use the following set of tags:

- N - a noun.
- Vi - an intransitive verb.
- Vt - a transitive verb.
- D - a determiner (like "the","a").
- PR - a preposition.

b.) Parsing a sentence means recovering its internal grammatical structure in a so-called parse tree. Find two different possible parse trees for the sentence "I saw a man with a telescope" to show the syntactic ambiguity in this sentence. Use the tags given above and the following non-terminals in the tree:

- S - the whole sentence.
- NP - a noun phrase.
- VP - a verb phrase.
- PP - a prepositional phrase, which is a preposition followed by a noun phrase. Prepositional phrases can occur both in noun phrases and verb phrases.

Exercise 7.2

Definite clause grammars are an extension to context free grammars that include arguments in non-terminal symbols and unification.

- Give the most general unification (substitution and result of unification) for the following pairs of non-terminals, or say why they cannot be unified:

  (a) $PN(Number,Case)$ and $PN(singular,Case)$
  (b) $NP(N)$ and $VP(singular)$
  (c) $VP(Any, accusative)$ and $VP(Number, accusative)$
  (d) $A(s(N))$ and $A(s(s(M)))$
• As explained in the lecture, DCGs are strictly more expressive than CFGs and can, for example, represent the language \( \{a^n b^n c^n \mid n \in \mathbb{N}\} \). Show how to derive the sentence \( aabbcc \) using the grammar presented in the lecture.

Exercise 7.3

Consider the following bigram statistics of a corpus over the vocabulary \( \{a, b, c, d\} \):

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>( N(w_x) )</th>
<th>( T(w_x) )</th>
<th>( Z(w_x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>60</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>400</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The word in row \( i \) is followed \( n \) times by the word in column \( j \) in the text, e.g. \( b \) is followed by \( c \) 60 times.

Assume we want to use Witten-Bell smoothing to compute the bigram probability estimates. Compute the necessary statistics \( N(w_x), T(w_x), Z(w_x) \) for \( w_x \in \{a, b, c\} \). From these, compute the bigram probabilities \( p^*(w_i|w_x) \) for \( w_i \in \{a, b, c, d\} \) and \( w_x \in \{a, b, c\} \) according to the formulas given in the lecture.