Winter Semester 2009/10
Prof. Dr. Detmar Meurers
Holger Wunsch

ISCL, Universität Tübingen
Computational Approaches
to Functional Elements

Lab Session November 4, 2009: A Majority Baseline for Predicting Determiners and Articles

1. The BNC Sampler Corpus

In this lab session, you will work with the BNC Sampler Corpus, which you can find at

`/afs/sfs/lehre/dm/ws-09-10-functional-elements/corpora/bnc-samp.tt`

2. Familiarize yourself with the POS tagset of the BNC Sampler

The POS tagset used in the BNC Sampler is the CLAWS7 tagset. It is documented at

`http://www.natcorp.ox.ac.uk/corpus/sampler/guide_C7.htm`.

Which tags are used for articles and which for prepositions?

3. Find out the frequency of prepositions and articles in the BNC Sampler

Write a program that uses the POS tags to extract all articles and prepositions in the corpus. You may use whatever programming language or tools that you like and you consider appropriate to solve that task.

Based on the output of your program, find answers to the following questions:

a) Which articles and prepositions occur in the BNC Sampler?

b) How often do they occur?

c) What are the ten most common prepositions in the BNC Sampler?

d) Are there tagging errors in the corpus?

4. Determine the majority baseline for predicting articles and prepositions

A majority baseline establishes the quantitative level of accuracy one can obtain simply by assuming that the most common preposition or determiner is chosen every time.

Write a program that computes the majority baselines of the determiners *the* and *a*, and the ten most frequent prepositions in the corpus. The program should

a) let you specify a list of articles or prepositions to be replaced

b) let you specify the article or preposition to be inserted at each occurrence

c) compute the number of times that the inserted article or preposition is correct

d) compute the number of times that the inserted article or preposition is **not** correct

e) print the accuracy of the prediction for articles and for prepositions.