Winter Semester 2009/10	ISCL, Universität Tübingen
Prof. Dr. Detmar Meurers	*Computational Approaches*
Holger Wunsch	*to Functional Elements*

Project starting December 9, 2009: Prediction of German articles and prepositions with a language model

Apply your system for predicting articles and prepositions, and evaluating these predictions, to German. Reuse as much as possible of your existing solution. Ideally, it should be sufficient to generate new training and test sets, and to replace the elements to be predicted with their German counterparts.

1. Generate training and test sets from the TüBa-D/Z 5.0

In this project, we will use the fifth version of the TüBa-D/Z, which has been released only a week ago. You can find it here:

`/afs/sfs/lehre/dm/ws-09-10-functional-elements/corpora/tuebadz-5.0.export`

The treebank is in NEGRA Export format. Convert the TüBa-D/Z to the column-based format that we have been using so far.

A sentence in Export format looks like this:

```
#BOS 1 3 1202391857 0 %% HEADLINE
Veruntreute          VVFIN   3sit          HD      500
die                  ART     nsf           -       504
AWO                  NN      nsf           -       501
Spendengeld          NN      asn           HD      502
?                    $.      --            --      0
#500                 VXFIN   --            HD      503
#501                 EN-ADD  --            HD      504
#502                 NX      --            OA      505
#503                 LK      --            -       506
#504                 NX      --            ON      505
#505                 MF      --            -       506
#506                 SIMPX   --            --      0
#EOS 1
```

Sentences start with the `#BOS` symbol, and end with `#EOS`. Tokens are located in the first column, and parts of speech in the second column.

The other information that is contained in the file is not relevant for our task.

> For completeness reasons: For words, the third column contains the morphological analysis of a token, and the fourth column the grammatical function label of the word. For each sentence, TüBa-D/Z contains a complete syntactic tree. Each node in the tree is assigned a unique number. Terminal nodes (i.e., words) are sequentially and implicitly numbered 1–499. Non-terminal nodes (i.e., inner nodes) start with `#500`. The fifth column contains the number of the parent node of the current node.

Inner nodes are encoded with their node number in the first column, followed by the category label in the second column. The third column is not used for inner nodes and therefore set to the value `--`. The fourth column contains the grammatical fuction, and finally the fifth column contains the number of the node's parent node.

Nodes that are "unattached", i.e., don't have a parent node in the tree, as well as the root node of the tree, are assigned the special parent node number 0.

The Export format contains additional tables delimited with `#BOT`/`#EOT` tags, which contain information about the origin of sentences, annotators, categories, and so on. Convert the TüBa-D/Z to the column-based format that we have been using so far.

2. Create training and test sets

Create ten-fold training and test sets using Kilian's `Preparer` tool. The tool needs to be adapted for use with German in the following places:

1. In `Constants.java`, update the word/POS combinations for articles and prepositions.
2. In `Preparer.java`, update the code that maps *a* and *an* to abstract *a/an* (lines 67 through 82).

Please have a look at task 3 below before you start with this. Both subtasks differ with respect to whether you prepare training and test material for 3(1) or 3(2).

3. Predict German articles

In a fashion similar to the second project 2 (November 17), predict German articles.

1. predict two **artifical** categories `__art-def__` and `__art_indef__`.
2. predict articles using all possible inflected surface forms.

(We will not predict articles in combination with morphological tags in this project).

4. Predict prepositions

Predict prepositions, using the set of ten most frequent prepositions in the TüBa-D/Z corpus.