**Supervisor:** Leonie Weißweiler

**Examiner:** Hinrich Schütze

**BSc, MSc, Open:** BSc

**Title:** Creation of a Multilingual Gold Standard for Case Marker Extraction

**Summary:**

Case marker extraction is the task of inducing the set of case markers for a given language from unstructured text. A case marker is a prefix, suffix or infix that marks the case of a given noun, for example in German, “-s” and “-es” would both be considered case markers for the Genitive. Case markers can take many different shapes in diverse languages. Their form may depend on the noun that they are attached to, or other factors in the sentence.

If the task of case marker extraction were to be solved, this would open many possibilities for interesting linguistic discoveries through the automated analysis of case in large corpora. Particularly, as case systems differ wildly across languages, from very coarse- to very fine-grained, automatically marking case in highly parallel corpora would be a big step towards the automated marking of deep cases, which are very fine-grained semantic categories related to the role of a noun in a sentence, akin to Semantic Role Labelling.

This thesis will work with highly parallel corpora and standard descriptive works about the Syntax of diverse languages to compile a gold standard for this task, i.e., a complete set of case markers for every language, so that automated methods can be tested against it. This will mean developing a both linguistically and computationally sensible definition of case marker that works for morphologically and syntactically diverse languages, so that the resulting gold standard is as universal and as comparable as possible.

**Prerequisites (if any):** some background in linguistics would be helpful