

Übung zur Vorlesung "Computerlinguistik II / Sprachtechnologie"

Sommersemester 2010, Prof. Dr. Udo Hahn, Erik Fäßler

Übungsblatt 7 vom 21.06.2010

Abgabe bis 22.06.2010, 23.59 Uhr; per Email (Standard Dateiformat: ps, pdf, oder doc) an erik.faessler@uni-jena.de

Aufgabe 1 : NLP Ressourcen

1. Wie in der Vorlesung gezeigt, besteht das PennTreebank-Corpus aus Klartextdateien. Sätze sind in einer Klamerschreibweise angegeben, wie sie beispielsweise in den folgenden zwei Sätzen aus dem Corpus demonstriert wird:

```
( (S (S (NP-SBJ An enormous turtle)
      (VP has
          (VP succeeded
              (SBAR-LOC (WHADVP-1 where)
                  (S (NP-SBJ the government)
                      (VP has
                          (VP failed
                              (ADVP-LOC *T*-1))))))))))
  :
  (S (NP-SBJ He)
      (VP has
          (VP made
              (S (S-NOM-SBJ (NP-SBJ *)
                  (VP speaking
                      (NP Filipino)))
                  (ADJP-PRD respectable))))))
  .))
```

Geben Sie für den Satz

The turtle Pong Pagong is a character who stars in children's television show "Batibot".

analog zu den obigen Beispielsätzen einen Parse in Klammer-Form an. Verwenden sie für die Phrasenbezeichnungen einfache Kategorien, wie sie zuvor in Vorlesung und Übung verwendet wurden, z.B. S, NP, VP etc. Falls Sie es für nötig erachten, können Sie auch weitere Kategorien verwenden.

2. Eine Treebank für die deutsche Sprache ist mit dem TiGER Corpus gegeben ¹. Auf das Corpus wird über die Applikation "TIGERSearch" zugegriffen, die von der TiGER Homepage heruntergeladen werden kann². Sehen Sie sich einige Beispiele aus dem Corpus mit TIGERSearch an³. Fallen Ihnen Unterschiede zu den PennTreebank Annotationen auf? Welche?

¹<http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/>

²Ignorieren Sie den Hinweis, dass TIGERSearch nicht stabil sei einfach.

³Oben links neben der Auswahl des Corpus gibt es einen Button mit einer kleinen Lupe darauf - über ihn können Sie durch das Corpus "blättern".

3. Begeben Sie sich per Dateibrowser in der Installationsverzeichnis von TIGERSearch. Betreten Sie das Verzeichnis "doc/examples". Sehen Sie sich die Datei "tigerstts.xml" mit einem Texteditor an. Beantworten Sie folgende Fragen:
- Was beinhaltet diese Datei?
 - Was sagt sie aus?
 - Welche Unterschiede gibt es zu der Form einer solchen Ressource, wie wir sie in der Vergangenheit kennen gelernt haben?