

# Einführung in die Computerlinguistik

## Übung 3

Jan Hendrik Dithmar - Matrikelnummer 2031259

## Aufgabe 3.1

|           |                                       |     |
|-----------|---------------------------------------|-----|
|           | Möglichkeiten für Formulierungen<br>→ |     |
| Wörter    | (a)                                   | (b) |
| Wortarten | (a)                                   | (b) |
| Syntax    | (b)                                   | (a) |
| Semantik  | (b)                                   | (a) |
| Pragmatik | (a)                                   | (b) |

Die folgenden Erklärungen beziehen sich grundsätzlich auf die im Diagramm angegebenen (ansteigenden) Möglichkeiten für Formulierungen:

- **Wörter:** Der Armchair-Linguist beschäftigt sich weniger mit dem Nachdenken über Wörter, sondern über Eigenheiten der Sprache. Der Korpus-Linguist beschäftigt sich eher mit Wörtern, auf Grund des vorhandenen Korpus.
- **Wortarten:** Auch hier wird der Armchair-Linguist eher weniger aktiv. Er denkt nicht über neue Wortarten nach, sondern überlässt dem Korpus-Linguisten die Suche danach in seinem Korpus bzw. seiner Korpora.
- **Syntax:** Der Korpus-Linguist ist bestrebt, mögliche syntaktische Formulierungen in seinen Korpora zu finden. Seine Bestrebung wird auch in gewissem Maße erfüllt, jedoch ist es eher dem Armchair-Linguisten zugedacht, mögliche syntaktische Formen durch Nachdenken zu erkennen und aufzuschreiben.
- **Semantik:** Gleiches wie für die Syntax gilt für die Semantik, wobei natürlich das Wort syntaktisch durch semantisch ersetzt werden muss.
- **Pragmatik:** Bei der Pragmatik sieht es wiederum anders aus. Hier ist es so, dass der Armchair-Linguist teilweise pragmatische Formulierungen findet, jedoch ab einem gewissen Punkt nicht mehr weiterkommt und die weiterführende Suche nach möglichen pragmatischen Formulierungen dem Korpus-Linguisten überlassen muss.

## Aufgabe 3.2

- **Unaufmerksamkeit:** Wenn der Annotator Vorschläge bekommt, wird er unaufmerksam. Man neigt in diesem Fall leicht dazu, einfach den Vorschlag zu bestätigen.
- **mögliche hohe Fehlerquote:** Das System, welches die Vorschläge macht, kann sich irren. Eine falsche Annotation kann weitreichende Folgen haben, da – falls ähnliche Annotationen ab diesem Zeitpunkt automatisch gemacht werden – diese durchgängig falsch annotiert werden.

## Aufgabe 3.3

Je detaillierter die Annotation ist, desto häufiger können Zweifelsfälle auftreten. Angenommen, wir haben das Wort „Rock“. Klar ist, dass dies ein Nomen ist. Doch die Einteilung in detailliertere Formen kann schwierig werden, da es sich z.B. um einen Eigennamen (Stadt) handelt. Lässt man zu viele Wahlmöglichkeiten, so können Annotationen sehr unterschiedlich ausfallen.

Was die Nutzung betrifft, so ist ein Korpus mit detaillierterer Annotation besser zu gebrauchen, da es wesentlich mehr Informationen enthält. Korpora mit nicht sehr detaillierter Annotation beinhalten weniger Informationen und sind somit nicht ganz so nützlich.

## Aufgabe 3.4

Korpora mit Annotationen auf „tiefen“ linguistischen Ebenen sind deshalb deutlich kleiner, da es u.a. schwierig wird, bei tiefgehenden Annotationen die Qualität zu sichern. Solche Annotationen bringen mit sich, dass es wahrscheinlich mehrere Möglichkeiten gibt, etwas zu annotieren. Diese Möglichkeiten müssen ausdiskutiert werden. Bei Korpora mit „flachen“ Annotationen ist es einfacher, die Qualität zu sichern, da es bei der Annotation wesentlich weniger Zweifelsfälle gibt. Um den Überblick zu behalten, werden Korpora mit Annotationen auf „tiefen“ linguistischen Ebenen entsprechend klein gehalten.

Die empirische Modellierung tieferer linguistischer Ebenen fällt wesentlich schwieriger aus, da es auf eine Frage möglicherweise nicht nur eine empirische Antwort gibt.

## Aufgabe 3.5

1. Bei Telefongesprächen hat man selten vollständig korrekte Sätze. Viele Sprecher unterbrechen sich selbst, bauen ihre Sätze um oder produzieren unvollständige Sätze. Oft identifiziert man auch Sprechpausen, dargestellt z.B. durch „mh“ oder „äh“.
2.
  - Switchboard wurde komplett ohne menschliche Eingriffe erstellt. Falls der Computer nun etwas falsch annotiert hat, könnte dies dem Parser Probleme bereiten.
  - Wenn beispielsweise zwei Personen A und B sich unterhalten und Person B Person A ins Wort fällt, könnte der Parser darüber stolpern, da er erkennen muss, dass das von B Gesagte nicht zu dem Satz von A gehört.