

Natürliche Sprache

Computer *sprechen* in strukturierten Daten. Die Verarbeitung - auch riesig wirkender - Tabellen und anderer Datenstrukturen scheint kein Problem darzustellen, sofern genügend Speicher vorhanden ist. Menschen dagegen sprechen unstrukturiert: Aussagen sind weder immer gleich strukturiert noch immer gleich formuliert. Wir benötigen sogar häufig einen Kontext, um eine Aussage überhaupt korrekt deuten zu können. Zählen Sie beispielsweise kurz die Möglichkeiten, wie der folgende Satz zu interpretieren sei:

Bob sieht Alice mit einem Fernglas auf dem Berg.

Falls Sie Bob freundlich gesinnt sind, werden Sie vermuten, dass Alice im Besitz des Fernglases ist. Falls Bob Ihnen schon einmal negativ aufgefallen ist, werden Sie eventuell in eine andere Richtung mutmaßen. Und befinden sich nun eigentlich beide auf dem Berg?

Es ergeben sich nun also folgende Probleme beim Bearbeiten einer Aussage: 1. Mehrdeutigkeiten müssen aufgelöst werden. 2. Die für den Kontext zutreffende Bedeutung muss gewählt werden. 3. Die Absicht einer Aussage muss erkannt werden.

Die Computerlinguistik - im Englischen auch als *Natural Language Processing* (NLP) bezeichnet - beschäftigt sich daher mit der Frage, wie natürliche Sprache algorithmisch verarbeitet werden kann. Denn immerhin ist der überwiegende Anteil der Informationen, die uns umgeben, unstrukturiert und sollte dennoch nicht ungenutzt bleiben.

In diesem Kapitel betrachten wir daher die Grundlagen der Textanalyse, indem wir Texte zuerst in Sätze und anschließend in Wörter aufteilen, diese normalisieren und zuletzt einfache Analysen durchführen.