Geparste Korpora für alle!

Eigene Korpora syntaktisch parsen und abfragen mit dem Erlanger treebank.info-Projekt

Auch wenn viele Forscher heute für Ihre Vorhaben wortartannotierte Korpora verwenden, so blieb bislang die Nutzung syntaktisch annotierter Korpora weitestgehend Computerlinguisten oder anderen computeraffinen Spezialisten vorbehalten. Das Erlanger treebank.info-Projekt hat sich zum Ziel gesetzt, diesen Umstand zu ändern, indem es Linguisten ohne computerlinguistische Kenntnisse die Möglichkeit gibt, geparste Korpora über eine Web-Oberfläche zu erstellen und zu nutzen. Somit gelingt es, mehrere Ursachen für die bisherige Zurückhaltung der Forschung bei der Nutzung geparster Korpora zu beseitigen und Benutzern ohne Spezialwissen Zugriff auf derlei Ressourcen zu gewähren. Die wichtigsten Vorteile sind:

1. **Der Parser kann mit wenigen Klicks gestartet werden.** Die meisten Parser sind immer noch nur umständlich per Kommandozeile nutzbar, was zu gewissen Berührungsängsten bei Computeranwendern führt, die eine grafische Benutzeroberfläche gewohnt sind. Beim treebank.info-Projekt führt der Upload-Vorgang im Web-Interface direkt zum Start des Parsers, der bereits mit sinnvollen Voreinstellungen versehen ist.
2. **Das Parsen großer Mengen Text kann unabhängig von der vorhandenen Rechenkraft und Speicherausstattung des eigenen PCs vorgenommen werden.** Parsing kann sehr zeitaufwändig sein und stellt hohe Anforderungen an die verwendete Hardware. Je nach verwendetem Parser und je nach Grammatikmodell kann es auf einem gewöhnlichen Bürocomputer gut 60 Tage dauern, 100 Millionen Wortformen zu parsen (und auch das nur, wenn der Rechner über ausreichend Arbeitsspeicher verfügt). Das treebank.info-Projekt nutzt die Hochleistungsrechner des Regionalen Rechenzentrums Erlangen und kann bei Bedarf mehr als 1000 Prozessorkerne gleichzeitig verwenden.
3. **Eine intuitiv zu bedienende grafische Benutzeroberfläche im Web-Browser wird zur Abfrage der Korpora verwendet.** Selbst wenn geparste Korpora zur Verfügung stehen, wird eine sinnvolle Auswertung oft dadurch erschwert, dass zur Abfrage zwar mächtige, aber meist hochkomplexe Anfragesprachen zur Verfügung stehen, die nicht nur hohen Einarbeitungsaufwand erfordern, sondern auch fehlerträchtig und somit oft frustrierend sind. Außerdem bedarf das Einpflegen geparster Korpora in geeignete Abfragesysteme oft spezialisierter Computerkenntnisse. Durch die Integration von Parser und Abfragesystem muss der Benutzer keine weiteren Schritte (Konvertierung, Änderung der Zeichencodierung, etc.) durchführen, sondern findet sein frisch geparstes Korpus automatisch im Web-Interface.

Der Workshop gibt eine Einführung in die Benutzung der Erlanger treebank.info-Oberfläche und einen Überblick über die verwendeten Grammatikmodelle. Die Teilnehmer können eigene Korpora hochladen (auch wenn diese unter Umständen nicht zwischen Beginn und Ende des Workshops fertig geparst werden können) und anhand von Übungen erste Erfahrungen darin sammeln, Korpora syntaktisch abzufragen. Sowohl deutsche als auch englische Korpora werden behandelt. Es sind **keine Vorkenntnisse** erforderlich. Die Veranstaltung ist auf 30 Teilnehmer beschränkt.

Peter Uhrig & Thomas Proisl
Interdisziplinäres Zentrum für Lexikografie, Valenz- und Kollokationsforschung
FAU Erlangen-Nürnberg