Q U A T

Eine Plattform zur Vermittlung von quantitativen Analysen in den Geisteswissenschaften

QUAT

Für viele Fragestellungen in der Sprachwissenschaft braucht man authentische Sprachdaten: Welche Wörter kommen wie häufig in welchen Konstruktionen vor? In welchen Kontexten kommen bestimmte Konstruktionen vor? Wie verändert sich Sprache? Welche Fehler machen Sprachlerner? Für solche Fragestellungen werden Texte in Korpora gesammelt und mit zusätzlichen Informationen wie z.B. Wortarten, Lemmata (Grundformen) u.a. versehen. QUAT hat sich zum Ziel gesetzt, eine Webanwendung zur Vermittlung von Verfahren der qualitativen und quantitativen Auswertung von Korpora zu entwickeln.

QUAT erweitert das linguistische Datenbanksystem ANNIS. Dadurch wird die Suche in Korpora, die detaillierte Analyse einzelner Fundstellen und die quantitative Auswertung der Ergebnisse einer Suchanfrage ermöglicht. Suchanfragen werden mit der spezialisierten Anfragesprache AnnisQL formuliert. Das Ergebnis einer Anfrage ist im einfachsten Fall eine Konkordanz, d.h. eine Liste von Fundstellen mit zugehörigem Kontext. Zu jeder Fundstelle können je nach Bedarf alle verfügbaren Informationen angezeigt werden. Im Folgenden werden vier prototypische Anwendungsfälle der quantitativen Auswertung von Suchergebnissen illustriert, die in QUAT umgesetzt werden sollen.

Zur quantitativen Auswertung von Korpora werden die Ergebnisse einer Suchanfrage als Kontingenztabelle dargestellt. Eine Kontingenztabelle stellt die Auftretenshäufigkeit von ausgewählten Merkmalen bei den einzelnen Fundstellen zu einer Suchanfrage dar. Ein Beispiel für eine Kontingenztabelle ist in Abb. 1 illustriert. Die Abb. zeigt die Vorkommen von Nomen im Falko-Korpus, einem Korpus das Texte von Fremdsprachenlernern des Deutschen enthält. In der ersten Zeile werden Vorkommen des Nomens »Jahr« gezählt, in der zweiten Zeile Vorkommen anderer Nomen. In der linken Spalte werden Vorkommen in Texten von Lernerinnen, in der rechten Spalte Vorkommen in den Texten von Lernern gezählt. Diese Kontingenztafel ermöglicht es zu ermitteln, ob das Nomen »Jahr« von Lernerinnen häufiger verwendet wird als von den Lernern, indem die Auftretenshäufigkeit des Nomens »Jahr« relativ zur Häufigkeit von Nomen überhaupt betrachtet wird. Abb. 2 zeigt ein Ranking von Nomen, die bei weiblichen bzw. bei männlichen Lernern besonders häufig auftreten. Dieses Ranking basiert auf dem Log-Likelihood-Assoziationsmaß, das durch aufstellen einer Kontigenztabelle für jedes Nomen berechnet werden kann. Abb. 3 zeigt ein Beispiel für die graphische Auswertung einer Suchanfrage an das TIGER-Korpus, einem mit Satzstrukturen annotierten Korpus. Gesucht wurde nach Objektsätzen mit oder ohne einleitendem »dass« (»Hans sagt, dass er da sein wird.« vs. »Hans sagt, er wird da sein.«). Das Diagramm stellt für die vier häufigsten einbettenden Verben die relative Häufigkeit von Verben im Indikativ und im Konjunktiv Präsens in den gefundenen Objektsätzen dar. Beim Verb »erklären« treten deutlich häufiger Objektsätze im Indikativ auf als bei den anderen betrachteten Verben. Dies ist farblich hervorgehoben. Abb. 4 schließlich zeigt die Häufigkeit von Funktionswörtern im Falko-Korpus. Jede Zeile der Tabelle vergleicht die relative Häufigkeit eines Funktionswortes bei deutschen Muttersprachlern und bei Deutschlernern verschiedener Muttersprachen. Deutliche Unterschiede sind farblich hervorgehoben.

Ziel des Projektes ist es die hier vorgestellten Verfahren der quantitativen Auswertung linguistischer Korpora in einer Webanwendung umzusetzen. Besonders wichtig ist dabei die Integration von quantitativen Auswertungsverfahren mit den bekannten Mitteln der qualitativen Analyse einzelne Fundstellen. Dies geschieht in Form der direkten Verknüpfung von Überblicksansichten, wie sie auf diesem Poster gezeigt sind, mit den Fundstellen, die diesen Ansichten zugrunde liegen. Zur Vereinfachung der Vermittlung korpuslinguistischer Analyseverfahren, sollen beispielorientierte Einführungstexte zu diesen Verfahren direkt in die Webanwendung integriert werden. Solche Texte bestehen aus instruktiven Beispielen für Suchanfragen und deren Auswertung, die ebenfalls direkt mit der zu entwickelnden Webanwendung verknüpft sind. Auf diese Weise wird den Studierenden ein weitgehend selbstgesteuerter Zugang zur korpuslinguistischen Arbeit ermöglicht. Für ein vertiefendes Studium sind Referenzen zu relevanten Lehrmaterialien angegeben.

HU | Institut für deutsche Sprache und Linguistik
Emil Kroymann | emil.kroymann@gmail.com
http://www.linguistik.hu-berlin.de/korpuslinguistik/

multimedia projekt
HUMBOLDT-UNIVERSITÄT ZU BERLIN