Erschienen in: Schneider, Roman/Storrer, Angelika/Mehler, Alexander (Hrsg.): Webkorpora in Computerlinguistik und Sprachforschung. Web Corpora for Computational Linguistics and Linguistic Research. Regensburg: GSCL, 2013. S. III-IV. (Journal for Language Technology and Computational Linguistics Jg. 28 (2013) H. 2)

Editorial

"Webkorpora in Computerlinguistik und Sprachforschung" war das Thema eines Workshops, der von den beiden GSCL-Arbeitskreisen „Hypermedia" und „Korpuslinguistik" am Institut für Deutsche Sprache (IDS) in Mannheim veranstaltet wurde, und zu dem sich am 27.09. und 28.09.2012 Experten aus universitären und außeruniversitären Forschungseinrichtungen zu Vorträgen und Diskussionen zusammenfanden. Der facettenreiche Workshop thematisierte Fragen der Gewinnung, der Aufbereitung und der Analyse von Webkorpora für computerlinguistische Anwendungen und sprachwissenschaftliche Forschung. Einen Schwerpunkt bildeten dabei die speziellen Anforderungen, die sich gerade im Hinblick auf deutschsprachige Ressourcen ergeben. Im Fokus stand weiterhin die Nutzung von Webkorpora für die empirisch gestützte Sprachforschung, beispielsweise als Basis für sprachstatistische Analysen, für Untersuchungen zur Sprachlichkeit in der internetbasierten Kommunikation oder für die korpusgestützte Lexikographie. Zusätzlich gab es eine Poster-/Demosession, in der wissenschaftliche und kommerzielle Projekte ihre Forschungswerkzeuge und Methoden vorstellen konnten. Eine Übersicht über das Gesamtprogramm sowie Abstracts und Folien der Workshopvorträge sind online unter *http://hypermedia.ids-mannheim.de/gscl-ak/workshop12.html* einsehbar.

Ausgewählte Beiträge des Workshops finden sich nun – zum Teil als Ergebnis von im Anschluss an das Treffen angebahnten wissenschaftlichen Kooperationen – im vorliegenden Themenheft des Journal for Language Technology and Computational Linguistics (JLCL). Dabei wird ein breites Spektrum aktueller Forschungsfragen rund um die Thematik „Webkorpora" abgedeckt:

- Michael Beißwenger und Lothar Lemnitzer geben einen Überblick über Motivation, Konzeption und laufende Arbeiten im Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation" (DeRiK), in dem ein deutschsprachiges Korpus zu Genres der internetbasierten Kommunikation aufgebaut wird. Das Korpus ist als eine Zusatzkomponente zu den Korpora im BBAW-Projekt „Digitales Wörterbuch der deutschen Sprache" (DWDS) konzipiert. Zunächst werden grundsätzliche Probleme der Repräsentation struktureller und linguistischer Besonderheiten von IBK-Korpora auf der Basis der Repräsentationsformate der Text Encoding Initiative (TEI) angesprochen, dann folgt eine Skizze möglicher Anwendungsszenarien.
- Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski und Torsten Zesch zeigen, wie sehr große Korpora für linguistische bzw. computerlinguistische Anwendungen auf Basis von Webinhalten – und deren Charakteristika – kompiliert werden können. Ihr Hauptaugenmerk liegt dabei auf den Bereichen Crawling, Vor- bzw. Weiterverarbeitung sowie Qualitätskontrolle. Weiterhin geben die Autoren einen Einblick in die Nutzung dieser Korpora für NLP-/CL-relevante Forschungsfragen.
- Bryan Jurish und Kay-Michael Würzner präsentieren eine neuartige Methode für die Segmentierung von Text in Sätze bzw. Token. Dabei nutzen sie Hidden Markov Modelle und berechnen Modellparameter unter Heranziehung der Segmentierung in etablierten Korpora bzw. Baumbanken. Die Verfahren werden an verschiedenen Korpora evaluiert, u.a. einem deutschen Korpus zur internetbasierten Kommunikation.
- Sabine Schulte im Walde und Stefan Müller präsentieren zwei Fallstudien, in denen die Eignung von Webkorpora für die automatische Akquisition lexikalisch-

semantischen Wissens untersucht wird. Hierfür vergleichen sie zwei unterschiedlich gut aufbereitete deutsche Webkorpora mit einem deutschen Wikipedia-Korpus und einem Zeitungskorpus. Die Fallstudien zeigen, wie sich verschiedene Parameter (z.B. Korpusgröße, Aufbereitung/Filterung, Domänenspezifik) auf die Qualität der Ergebnisse auswirken.

Wir danken allen Gutachtern und den JLCL-Herausgebern für die tatkräftige Unterstützung.
Wir hoffen, dass die Leser dieses JLCL-Themenhefts die darin enthaltenen Beiträge ebenso interessant und inspirierend finden wie wir.

Dezember 2013

Roman Schneider, Angelika Storrer, Alexander Mehler