Kompilation eines Diskursstruktur-annotierten deutschsprachigen Blogkorpus

Grumt Suárez, Holger
Holger.H.Grumt-Suarez@germanistik.uni-giessen.de
Justus-Liebig-Universität Gießen, Deutschland

Karlova-Bourbonus, Natali
Natali.Karlova-Bourbonus@germanistik.uni-giessen.de
Justus-Liebig-Universität Gießen, Deutschland

Lobin, Henning
Henning.Lobin@uni-giessen.de
Justus-Liebig-Universität Gießen, Deutschland

Das Poster „Interoperabilität bei der Erstellung eines deutschsprachigen Blogkorpus für die Repräsentation der Diskursstruktur“ informiert über das Vorgehen sowie die ersten Forschungszwischenergebnisse und die weiteren Ziele der Kompilierung und Annotation eines deutschsprachigen Blogkorpus. Gegenwärtig gibt es lediglich eine geringe Anzahl an öffentlich zugänglichen, umfangreichen Blogkorpora, wie zum Beispiel das englischsprachige Birmingham Blog Corpus der Birmingham Universität (vgl. WebCorp 2013) oder das bilinguale (deutsch-französische) Korpus d’apprentissage INFRAL (Interculturel Franco-Allemand en Ligne) (vgl. Abendroth-Timmer et al. 2014). Betrachtet man den großen Einfluss von Blogs für die Geschichte der Kommunikation im Internet, erscheint die geringe Anzahl überraschend.

Bislang existiert kein Standard für das Repräsentieren von sogenannten Computer-Mediated Communication-Daten (kurz CMC), allerdings arbeitet die Text Encoding Initiative CMC Special Interest Group (TEI SIG) (vgl. Beißwenger 2016) seit 2013 an einem Schema für die Repräsentation von CMC-Genres. Eine Standardisierung, wie sie die Text Encoding Initiative für CMC anstrebt, ist ein wichtiger Punkt, wenn es um ‚Digitale Nachhaltigkeit‘ geht. Unser Forschungsvorhaben leistet hierfür einen Beitrag.

Das Hauptziel des Vorhabens umfasst die semi-automatische Kompilation sowie die Repräsentation der Blogdiskursstruktur. Dabei sollen die Relationen zwischen den textuellen und multimodalen Elementen (Blogbeiträge, Kommentare, Hyperlinks, Bilder und Töne) und den verschiedenen Textproduzenten (Blogger, Kommentatoren) abgebildet werden. Das annotierte Blogkorpus soll am Ende als eine nachhaltige Ressource verfügbar gemacht werden. Hierfür stehen wir momentan in Kontakt mit der Redaktion von Spektrum der Wissenschaft Verlagsgesellschaft mbH, um die Form der Bereitstellung zu klären.

Die Grundlage des Korpus bildet das Wissenschaftsblogportal SciLogs – Tagebücher der Wissenschaft (SciLogs 2016) und deckt den Inhalt des Jahres 2015 vollständig ab. Die Daten wurden aus den vier SciLogs-Blogbereichen „WissensLogs“, „BrainLogs“, „KosmoLogs“ und „ChronoLogs“ erhoben. Das Korpus soll mit drei Informationstypen annotiert werden, die einerseits direkt, andererseits indirekt in den Blogdaten vorhanden sind oder anhand von statistischen Analysen und computerlinguistischen Tools sichtbar gemacht werden. Zum jetzigen Zeitpunkt beschränken sich die Annotationen des Korpus auf die direkt auslesbaren Informationen des Blogs wie beispielsweise der Titel des Blogbeitrags, der Name des Bloggers und das Einstelldatum des Blogeintrags. Ferner wird darauf geachtet, dass sämtliche Informationen, die die Inhalte der Scilogs-Website in Bezug auf die Bloginhalte liefern, ebenfalls annotiert werden. Wir sind der Meinung, dass auch auf den ersten Blick nicht für die Diskursstruktur relevante Informationen ausgezeichnet werden sollten. Im Fokus steht der Ansatz, dass das Blogkorpus aus CMC-Daten später für die Erforschung unterschiedlicher linguistischer Fragestellungen verwendet werden kann.

Zusammenfassend soll das Poster nicht nur unser Vorhaben vorstellen, sondern auch einen Einblick in unser grundsätzliches Vorgehen bei der Erstellung eines CMC-Korpus geben. Der Fokus für die DHd 2017 liegt unter anderem auf der Darstellung der Entscheidungsfindung innerhalb der Auszeichnungssprachen. Es soll erläutert werden, warum wir uns beispielsweise für die deskriptive Auszeichnungssprache TEI und nicht XML (Extensible Markup Language) entschieden haben. Des Weiteren möchten wir Einblicke in die semi-automatische TEI-Annotation geben und unsere Erkenntnisse mit dem vorläufigen, von der TEI SIG bereitgestellten, TEI-Schema teilen. Letztlich wollen wir auch das bisherige Korpus selbst vorstellen, das aus ca. 3.000.000 Tokens (ca. 1.200 Blogposts von 80 Bloggern und 15.000 Kommentaren von 1500 Kommentatoren) besteht.

Bibliographie

Abendroth-Timmer, Dagmar / Bechtel, Mark / Chanier Thierry / Ciekanski, Maud (2014): *Corpus d’apprentissage INFRAL (Interculturel Franco-Allemand en Ligne)*. Banque de corpus CoMeRe. Nancy: Ortolang.fr https://hdl.handle.net/11403/comere/cmr-infral [letzter Zugriff 1. Juli 2016].

Beißwenger, Michael (2016): *SIG: Computer-Mediated Communication* http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication [letzter Zugriff 1. Juli 2016].

SciLogs (2016): *SciLogs: Tagebücher der Wissenschaft*. Spektrum der Wissenschaft Verlagsgesellschaft mbH

http://www.scilogs.de/impressum/ [letzter Zugriff 1. Juli 2016].

WebCorp (2013): *Birmingham Blog Corpus. WebCorp: Linguist's Search Engine*. Birmingham City University http://wse1.webcorp.org.uk/cgi-bin/BLOG/index.cgi [letzter Zugriff 1. Juli 2016].