Die datengeleitete Ermittlung des gemeinsamen sprachlichen Inventars der Geisteswissenschaften

,
cordula.meissner@uni-leipzig.de
Universität Leipzig, Deutschland

,
f.wallner@rz.uni-leipzig.de
Universität Leipzig, Deutschland

Hintergrund

Sprache ist in der Wissenschaft nicht nur ein Instrument, um Sachverhalte zu vermitteln, sondern spielt für das wissenschaftliche Denken eine konstitutive Rolle. Dies gilt insbesondere für die geisteswissenschaftlichen Disziplinen, da hier selbst die Gegenstände der Forschung größtenteils sprachlich verfasst sind (vgl. Kretzenbacher 2010). Die nicht-terminologische, disziplinenübergreifend verwendete Wissenschaftssprache spiegelt dabei in besonderem Maße die in Sprache niedergelegten Erkenntnisprozesse wider und ist somit von wesentlicher wissenschaftsmethodologischer Bedeutung. Zu ihr gehören beispielsweise Ausdrucksmittel des Voraussetzens, des Begründens, des Folgerns, des Einschränkens, des Übertragens und Vergleichens. Für diesen Bereich, der unter dem Begriff der allgemeinen oder auch alltäglichen Wissenschaftssprache zusammengefasst wird (Schepping 1976; Ehlich 1999), steht eine systematische lexikographische Erschließung und Beschreibung jedoch bislang noch aus. Der einzige vorliegende Ansatz zu einer lexikografischen Erfassung der allgemeinen Wissenschaftssprache nimmt das gesamte Spektrum akademischer Fächer in den Blick und erlaubt so eine nur geringe Beschreibungsdetailliertheit (Erk 1972, 1975, 1982, 1985).

Das Projekt GeSIG (Das gemeinsame sprachliche Inventar der Geisteswissenschaften) setzt sich daher zum Ziel, erstmals das Inventar der allgemeinen Wissenschaftssprache der Geisteswissenschaften auf empirischer Grundlage zu bestimmen und damit den Grundstein für seine umfassende Erschließung zu legen. Ein auf diese Weise bestimmtes Inventar stellt eine wertvolle Grundlage für die Dokumentation und Erforschung der Sprache der Geisteswissenschaften dar. Das Projekt ist als Pilotprojekt angelegt und soll Vorarbeiten liefern für den Aufbau einer umfassenden elektronischen lexikographischen Ressource dieses Sprachbereichs.

Der Beitrag stellt das Projekt GeSIG vor. Im ersten Teil wird die datengeleitete Ermittlung des gemeinsamen sprachlichen Inventars der Geisteswissenschaften beschrieben. Während diese auf einer sehr feindifferenzierten Einteilung geisteswissenschaftlicher Disziplinen basiert, ist für die lexikografische Bearbeitung eine Bündelung notwendig. Der zweite Teil geht daher der Frage nach, welche Großbereiche aus lexikografischer Perspektive unterschieden werden sollten. Es wird eine empirische Studie vorgestellt, die datengeleitet versucht, diese Frage zu beantworten.

Die Ermittlung des gemeinsamen sprachlichen Inventars der Geisteswissenschaften

Die Datenbasis für die Ermittlung des Inventars bilden Korpora verschiedener geisteswissenschaftlicher Fachbereiche. Zur Operationalisierung der „Geisteswissenschaften“ wurde dabei die Umfangsbestimmung des Wissenschaftsrates (2010) zugrunde gelegt, der sich an die Systematik des statistischen Bundesamtes anlehnt (vgl. Statistisches Bundesamt 2013). In dieser werden 19 geisteswissenschaftliche Disziplinengruppen unterschieden (wie etwa Geschichte, Romanistik, Philosophie, Musikwissenschaften u.a.). Diese Einteilung bildete die Basis für die Erstellung von Teilkorpora. Es wurden für jeden Bereich mindestens 10 Dissertationen und mindestens 1 Mio. Token erhoben. Die Analysegrundlage setzt sich insgesamt aus 197 Dissertationen mit einem Gesamtumfang von ca. 19 Mio. Token zusammen.

Um einen systematischen Zugriff auf den Wortschatzbestand der allgemeinen Wissenschaftssprache der Geisteswissenschaften zu ermöglichen, wurde eine datengeleitete Vorgehensweise gewählt. Hierfür war zunächst eine Bereinigung der Sprachdaten erforderlich. Anschließend wurden die Texte mit Hilfe des TreeTaggers (Schmid 1995) und unter Anwendung der Richtlinien des STTS (Schiller et al. 1999) annotiert sowie lemmatisiert, um eine systematische Auswertung auf Lemmaebene und im Hinblick auf Wortarten durchführen zu können. Zusätzlich erfolgten weitere manuelle Nachbearbeitungsschritte zur Desambiguierung automatisch ermittelter Homonyme sowie zur Lemmatisierung der Partikelverben und unvollständiger Wortformen.

Auf der Grundlage der so aufbereiteten Teilkorpora wurde der allgemeinwissenschaftliche Wort¬schatz der Geisteswissenschaften ermittelt. Dieser wurde operationalisiert durch das disziplin¬übergreifende Vorkommen von Lemmata. Für jedes Teilkorpus wurde hierzu eine Lemmaliste erstellt und eine Schnittmenge

aus diesen 19 Listen gebildet. Die Schnittmenge enthält jene sprachlichen Mittel, die der Form nach in geisteswissenschaftlichen Disziplinen übergreifend gebraucht werden. Sie umfasst insgesamt 4.668 Lemmata (z.B. Nomen wie *Jahr*, *Form*, *Frage*, *Arbeit*, *Bild*, Verben wie *geben*, *zeigen*, *finden*, *sehen*, *darstellen* und Adjektive wie *gut*, *verschieden*, *deutlich*, *f olgend*).

Zur Frage der Fachbereichseinteilung im Hinblick eine lexikografische Bearbeitung des Inventars

Die quantitative Analyse des Inventars zeigt deutliche Frequenzunterschiede für einzelne Lemmata in bestimmten Disziplinen. Dies deutet darauf hin, dass einige der übergreifend gebrauchten Lexeme in den geisteswissenschaftlichen Disziplinen einen unterschiedlichen Stellenwert haben und möglicherweise fachterminologisch geprägt sind. Diese fachspezifische Prägung sollte auch bei der lexikografischen Bearbeitung des Inventars Berücksichtigung finden. Hierfür ist es erforderlich, die für die Ermittlung des Inventars zugrunde gelegten 19 geisteswissenschaftlichen Disziplinengruppen zu bündeln. Da in vorliegenden Fachbereichseinteilungen die Gruppierung geisteswissenschaftlicher Disziplinen uneinheitlich erfolgt, wurde mit Hilfe von Topic Modeling (Mallet, vgl. McCallum 2002) eine alternative, datengeleitete Fachbereichseinteilung vorgenommen. Die Grundlage hierfür bildeten die 197 Dissertationen, die auch zur Ermittlung des gemeinsamen sprachlichen Inventars der G eisteswissenschaften herangezogen wurden. Diese wurden mit Hilfe des Topic Modeling gruppiert, wobei sich die klarsten Ergebnisse zeigten, wenn der Berechnung sechs Topics zugrunde gelegt wurden. Aus der datengeleiteten Gruppierung der Dissertationen lassen sich die folgenden Bündelungen ablesen: 1. Dissertationen mit sprachwissenschaftlichem Schwerpunkt, 2. Dissertationen mit literaturwissenschaftlichem Schwerpunkt, 3. Dissertationen mit geschichtswissenschaftlichem Schwerpunkt, 4. Dissertationen mit philosophischem oder theologischem Schwerpunkt, 5. Dissertationen mit kunstwissenschaftlichem Schwerpunkt sowie 6. Dissertationen mit bibliothekswissenschaftlichem Schwerpunkt oder vorwiegend empirischer Ausrichtung. Im Beitrag werden die ermittelten Gruppen näher vorgestellt und im Vergleich zu vorliegenden Fachbereichseinteilungen diskutiert.

Insgesamt soll mit diesem Beitrag gezeigt werden, wie datengeleitete Verfahren nutzbar gemacht werden können, um einen Sprachverwendungsbereich lexikografisch zu erschließen.

Bibliographie

Ehlich, Konrad (1999): "Alltägliche Wissenschaftssprache", in: *Informationen Deutsch als Fremdsprache* 26: 3-24.

Erk, Heinrich (1972): *Zur Lexik wissenschaftlicher Fachtexte*. Verben, Frequenz und Verwendungsweise (= Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 4). München: Hueber.

Erk, Heinrich (1975): *Zur Lexik wissenschaftlicher Fachtexte*. Verben, Frequenz und Verwendungsweise (= Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 4). München: M. Hueber.

Erk, Heinrich (1982): *Zur Lexik wissenschaftlicher Fachtexte*. Verben, Frequenz und Verwendungsweise (= Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 4). München: M. Hueber.

Erk, Heinrich (1985): *Wortfamilien in wissenschaftlichen Texten*. Ein Häufigkeitsindex (= Schriften der Arbeitsstelle für wissenschaftliche Didaktik des Goethe-Instituts 9). München: M. Hueber.

Kretzenbacher, Heinz (2010): "Fach- und Wissenschaftssprachen in den Geistes- und Sozialwissenschaften", in: Krumm, Hans-Jürgen / Fandrych, Christian / Hufeisen, Britta / Riemer, Claudia (eds.): *Deutsch als Fremd- und Zweitsprache* (= Handbücher zur Sprach- und Kommunikationswissenschaft 35.1). Berlin, New York: de Gruyter 493-501.

McCallum, Andrew Kachites (2002): *MALLET: A Machine Learning for Language Toolkit.* http:// mallet.cs.umass.edu [letzter Zugriff 22. Februar 2016].

Schepping, Heinz (1976): "Bemerkungen zur Didaktik der Fachsprache im Bereich des Deutschen als Fremdsprache", in: Rall, Dietrich / Schepping, Heinz / Schleyer, Walter (eds.): *Didaktik der Fachsprache*. Beiträge zu einer Arbeitstagung der RWTH Aachen vom 30.9. bis 4.10.1974. Bonn-Bad Godesberg: DAAD 13-34.

Schmid, Helmut (1995): "Improvements In Part-of-Speech Tagging With An Application To German", in: *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland [letzter Zugriff 02. Oktober 2015].

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Technischer Bericht. Universitäten Stuttgart & Tübingen.

Statistisches Bundesamt (2013): *Bildung und Kultur. Studierende an Hochschulen - Fächersystematik.* https://www.destatis.de/DE/Methoden/ Klassifikationen/BildungKultur/ StudentenPruefungsstatistik.pdf [letzter Zugriff 16. Oktober 2014].

Wissenschaftsrat (2010): *Empfehlungen zur vergleichenden Forschungsbewertung in den Geisteswissenschaften. Drs. 10039-10.* http://

www.wissenschaftsrat.de/download/archiv/10039-10.pdf
[letzter Zugriff 30. Mai 2016].