KoLaS – Kommentiertes Lernendenkorpus akademisches Schreiben

Melanie Andresen und Dagmar Knorr (Juli 2017)

1 Entstehungsgeschichte

Im Institut für Interkulturelle Bildung der Fakultät für Erziehungswissenschaft an der Universität Hamburg wird seit mehreren Jahrzehnten der Zusammenhang zwischen Bildungserfolg und Bildungssprache erforscht. Es wurde beobachtet, dass Studierende mit Migrationshintergrund Probleme beim Verfassen akademischer Texte haben. Die – naheliegende – Hypothese war, dass dies an den sprachlichen Fähigkeiten der Studierenden liegt. Allerdings ist eine Überprüfung der Hypothese ohne entsprechendes empirisches Datenmaterial schwierig.

Die Fachliteratur zum deutschsprachigen akademischen Schreiben weist hier eine Lücke auf: Zugängliche Korpora von authentischen, deutschsprachigen akademischen Texten von Studierenden, die möglichst auch noch longitudinal untersucht werden können, existierten nicht. Mit dem Korpus KoLaS leistet die Schreibwerkstatt Mehrsprachigkeit[1] einen Beitrag zur Forschung, indem authentische Texte von Studierenden für Forschungszwecke zur Verfügung gestellt werden.

Die Korpusversion 1.0 wurde im Dezember 2015 veröffentlicht. Im Februar 2017 sind mit der Version 2.0 zahlreiche zusätzliche Texte ergänzt worden. Seit Juli 2017 ist die Version 2.1 verfügbar, in der nur kleine Korrekturen vorgenommen wurden. Da das Projekt „Schreibwerkstatt Mehrsprachigkeit“ beendet ist, ist die Korpuserstellung abgeschlossen.

2 Vorgehen bei der Erstellung des Korpus KoLaS

2.1 Datenerhebung

Das Korpus KoLaS besteht aus Texten, mit denen Studierende in die Schreibberatung der Schreibwerkstatt Mehrsprachigkeit gekommen sind und für die die Studierenden ihr Einverständnis zur Veröffentlichung und Nutzung im Rahmen wissenschaftlicher Forschung und Entwicklung gegeben haben. Die Studierenden werden in einem Gespräch über die Verwendung der Daten und die Form der Anonymisierung aufgeklärt. In diesem Gespräch werden zudem biographische Daten zum Bildungsverlauf, zum Studium sowie zum Sprachgebrauch erfasst (siehe unten).

Die Texte der Studierenden werden von den Schreibberaterinnen und Schreibberatern der Schreibwerkstatt Mehrsprachigkeit schriftlich kommentiert. Diese kommentierten Fassungen bilden die Grundlage für Beratungsgespräche, zu denen von den Schreibberaterinnen und -beratern Protokolle angefertigt werden und die dem Korpus als Metadaten beigefügt sind. Auf der Basis der Beratungsgesprächs und der schriftlich kommentierten Fassung nehmen die Ratsuchenden unter Umständen eine Textüberarbeitung bzw. -weiterführung der Textproduktion vor, die wiederum Eingang in das Korpus findet.

Auf diese Weise entsteht eine Schleife: „1. Textfassung – (schriftliches und mündliches) Feedback – Weiterbe-/Überarbeitung – Feedback – 2. Textfassung... – Endprodukt“.

Diese Schleife stellt den Idealfall in der Dokumentation einer Textgenese dar. Da es sich um ein natürliches und gewachsenes Korpus handelt, ist nicht in allen Fällen ein „Endprodukt“ im Sinne

[1] Die Schreibwerkstatt Mehrsprachigkeit hatte eine Laufzeit von 2011 bis 2016, stand unter der Leitung von Prof. Dr. Ursula Neumann und war am Arbeitsbereichs DivER der Fakultät für Erziehungswissenschaft der Universität Hamburg angesiedelt. Sie wurde zunächst von der ZEIT-Stiftung Ebelin und Gerd Bucerius finanziert. Von 2012–2016 war sie Teil des Universitätskollegs. Das Universitätskolleg wurde im Rahmen des Qualitätspakts Lehre I vom BMBF unter dem Förderkennzeichen 01PL12033 finanziert.

einer Textfassung, die von den Studierenden abgegeben wird, im Korpus enthalten. Dies ist der Tatsache geschuldet, dass die Teilnahme an den Angeboten der Schreibwerkstatt Mehrsprachigkeit freiwillig erfolgt und somit die Ablieferung des Endprodukts nicht eingefordert werden kann.
Das Korpus KoLaS ist aufgrund seiner Genese nicht gleichförmig aufgebaut. Die Gründe hierfür sind folgende:

- Zeit: Der Zeitpunkt im Textproduktionsprozess variiert, zu dem die Schreibberatung von Studierenden aufgesucht wird. Einige kommen in einem frühen Stadium, andere erst später. Allein aufgrund der Zeitdauer bis zur Abgabe variiert die Anzahl der Texte pro Person.
- Kommentierungsbedarf: Einige Studierende lassen sich im Prozess der Ideenfindung und Strukturierung begleiten, andere haben konkrete Fragen. Dies führt zu unterschiedlichem Kommentierungsverhalten und Textbearbeitungszyklen.
- Freiwilligkeit: Die Studierenden bestimmen selbst, ob und in welchem Umfang sie sich in ihrer Textproduktion begleiten lassen möchten.
- Fachspezifik und Phase im Studium: Da die Schreibwerkstatt Mehrsprachigkeit ihre Angebote nicht auf eine bestimmte Studierendengruppe eingeschränkt, sind Texte von Studierenden aus verschiedenen Fachrichtungen und in unterschiedlichen Phasen des Studiums enthalten.
- Sprachliche Voraussetzungen: Die sprachlichen Voraussetzungen der Studierenden unterscheiden sich, da sich Texte von monolingual deutschsprachigen Studierenden ebenso im Korpus befinden wie Texte von Studierenden, die erst seit kurzem Deutsch lernen.

Dies führt dazu, dass die Vergleichbarkeit der Texte untereinander in sehr unterschiedlichem Maß gegeben ist. Aussagen über das Korpus als Ganzes sind nur eingeschränkt möglich, da zu viele Variablen die Gestalt der Texte beeinflussen. Andersherum hat diese Art der Korpuserstellung auch methodische Vorteile: Bei denen in das Korpus eingehende Texte handelt es sich ausschließlich um authentische Schreibprodukte, die das Ziel haben, eine Prüfungsleistung im Handlungsraum „Wissenschaft“ zu bestehen. Unserer Ansicht nach ist das Korpus sehr gut als Material für explorative Studien geeignet, die nicht unbedingt auf repräsentative Aussagen abzielen, sondern einen ersten Einblick gewinnen und Hypothesen generieren wollen. Dadurch, dass die Texte mit Metadaten verknüpft sind, haben alle Nutzer/innen des Korpus Einblick in die vielfältigen Entstehungsbedingungen der Texte und können ihre Eignung für eine gegebene Fragestellung einschätzen. Darüber hinaus bietet das Korpus einen Einblick in die Kommentierungspraxis von Peer-Tutorinnen und -Tutoren.

2.2 Datenaufbereitung

Wie oben beschrieben stammen die Texte aus dem Beratungsalltag der Schreibwerkstatt Mehrsprachigkeit, wo sie natürlich mit konkreten Personen in Verbindung gebracht werden müssen. Um eine Veröffentlichung der Daten möglich zu machen, war eine vollständige Anonymisierung aller Texte und Metadaten notwendig. Hierzu wurden zunächst die Namen aller Ratsuchenden durch einen sechsstelligen Nummerncode ersetzt und anstelle der Initialen der Schreibberater/innen werden Kürzel verwendet, die die Person als studentische Schreibberater/in erkennbar machen (SB01, SB02...). Kommentierungen von Dozierenden werden mit Kürzeln WB01, WB02... gekennzeichnet. Die Textdateien sind nach einem festen Schema benannt:
Beispiel: 12-02-05_HA-Bantusprache_03-19-57_SB03.doc

Elemente: 1 2 3 4

1. Datum im Format JJ-MM-TT
2. Eine Textbezeichnung, die Informationen zu Textart und Inhalt enthält
3. Nummerncode für die/den Ratsuchenden
4. Kürzel für den/die Schreibberater/in (wenn Kommentare vorhanden sind)

Alle Metadaten zu den Ratsuchenden und den Beratungsereignissen werden in der Schreibwerkstatt Mehrsprachigkeit in einer FileMaker-Datenbank archiviert. Zum Zwecke der Veröffentlichung war es hier erstens notwendig, eine neue Version zu schaffen, die um zahlreiche persönliche Daten reduziert wurde (z.B. Kontaktdaten). Zweitens sollten die Daten mit der Veröffentlichung für jeden zugänglich und deshalb nicht an eine kostenpflichtige Software gebunden sein. In Absprache mit dem Hamburger Zentrum für Sprachkorpora (HZSK) haben wir uns für das Programm CoMa entschieden, das eine praktische Kombination der Metadaten zu den Personen einerseits und den Beratungsereignissen andererseits anbietet.

Diese neu entstandene Metadatendatei wird Nutzers des Korpus zusammen mit den Texten zur Verfügung gestellt.

Als aufwendig erwies sich die Anonymisierung der Texte selbst, die als Word- oder PDF-Dokumente vorliegen. Persönliche Daten finden sich hier insbesondere auf dem Deckblatt, wo neben Informationen zur Person auch solche zur dazugehörigen Lehrveranstaltung entfernt wurden. Namen werden außerdem häufig in den Kommentaren der Schreibberater/innen verwendet, die sich an ganz unterschiedlichen Stellen im Dokument befinden können. Überwiegend wird mit der Word-Kommentarfunktion gearbeitet, gelegentlich werden aber auch Abschlusskommentare am Ende des Dokumentes in den Fließtext eingefügt. Um hier zu vermeiden, dass Namen übersehen werden, wurde ein Python-Skript eingesetzt, dass nach den Namen sucht und Fundstellen ausgibt. Word-Dokumente enthalten außerdem Metadaten zur Autorin/zum Autor des Textes und den Namen möglicher Kommentatoren, die ebenfalls entfernt wurden. Personenbezogene Daten in Texten, von denen nur pdf-Dateien in das Korpus eingehen, wurden geschwärzt.

3 Korpusbeschreibung

KoLaS enthält 853 Texte aus dem Zeitraum September 2011 bis Dezember 2016. Die Texte stammen von insgesamt 122 unterschiedlichen Ratsuchenden. Die Anzahl der Texte oder Textversionen pro Ratsuchenden liegt im Durchschnitt bei 7 (schwarze Linie), ist aber großen Schwankungen unterworfen, wie die Abbildung „Anzahl Texte pro Ratsuchender/m“ verdeutlicht.

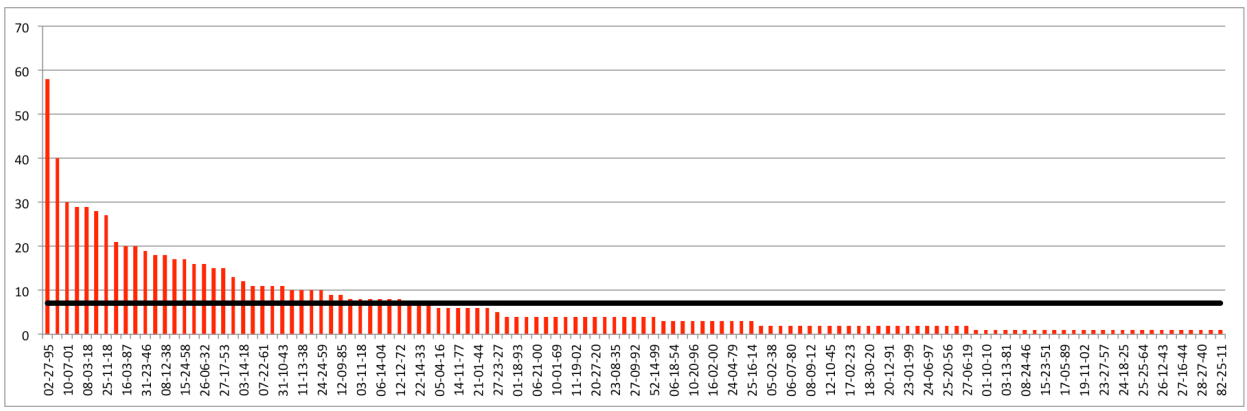

Abbildung 1: Anzahl Texte pro Ratsuchender/m

Von einer/m Ratsuchender/m liegen ganze 58 Texte vor, von 26 Personen nur ein einziger Text.

Damit liegen von 96 Ratsuchenden mindestens zwei Texte oder Textversionen vor, was die Bearbeitung interessanter Forschungsfragen ermöglicht.
Die zeitliche Verteilung der Daten kann Tabelle 1 entnommen werden. Aus dem Jahr 2011 liegen nur wenige Dateien vor, da die Schreibwerkstatt Mehrsprachigkeit erst Ende des Jahres ihre Tätigkeit aufgenommen hat und die ersten Schreibberater/innen ausgebildet wurden. Die Diskrepanz entsteht dadurch, dass Beratungen maximal wöchentlich stattfinden, ggf. aber auch dazwischen neue Textversionen entstehen, die den Schreibberater/innen zur Verfügung gestellt werden.

Jahr	Texte	Beratungen
2011	28	20
2012	228	176
2013	195	167
2014	87	58
2015	170	126
2016	145	112
Summe	**853**	**659**

Tabelle 1: Zeitliche Verteilung der Korpustexte

Zu den Ratsuchenden (RS) liegen eine Reihe von Metadaten vor, die spezifischere Analysen ermöglichen. Hierzu gehören insbesondere Informationen zum Studium (Fach, angestrebter Abschluss) und zur Sprachbiografie (Sprachen, mit denen die RS aufgewachsen sind, die sie in der Schule gelernt haben, Erstsprachen der Eltern...). Die im Korpus erfassten Ratsuchenden haben im Feld „aufgewachsen mit“ 26 unterschiedliche Sprachen angegeben (Mehrfachnennungen möglich). Mit 43 bzw. 19 Ratsuchenden sind Deutsch und Russisch die am häufigsten genannten Sprachen, weiterhin sind folgende Sprachen vertreten (alphabetisch): Albanisch, Arabisch, Aserbainisch, Bassa, Chinesisch, Englisch, Fanti, Französisch, Georgisch, Griechisch, Hindi, Indonesisch, Italienisch, Japanisch, Kasachisch, Katalan, Kirgisisch, Kroatisch, Kurdisch, Latein, Multani/Kandahari, Persisch/Farsi, Polnisch, Portugiesisch, Schweizer Deutsch, Serbisch, Slowakisch, Spanisch, Sundonesisch, Tschechisch, Türkisch, Twi, Ukrainisch, Vietnamesisch und Yuruba. Diese große Sprachenvielfalt ist vorteilhaft, da viele Sprach- und Kulturregionen einbezogen werden. Allerdings liegen dadurch momentan nur wenige Texte zu einer einzelnen Sprache vor, sodass das Erstellen von Teilkorpora nach Sprache nur eingeschränkt möglich ist.
Die Ratsuchenden sind in ihrem Studium unterschiedlich weit vorangeschritten: Es sind Studierende vom ersten bis zum dreißigsten Semester vertreten, der Durchschnitt liegt bei 6 Semestern. In Bezug auf das Studienfach stammen die meisten Ratsuchenden aus der Fakultät für Erziehungswissenschaft sowie der für Sprach- und Kulturwissenschaften. Dieser Schwerpunkt hängt mit der ursprünglichen Zielgruppendefinition der Schreibwerkstatt Mehrsprachigkeit zusammen, die sich zunächst an „Lehramtsstudierende mit Migrationshintergrund“ richtete.
Die Textarten, mit denen die Studierenden in die Beratung kommen, sind ebenfalls sehr vielfältig. Die mit Abstand häufigste Textart ist die Hausarbeit, gefolgt von unterschiedlichen Typen von Abschlussarbeiten (BA-, MA-, Magister-, Diplom-Arbeiten). In geringerer Häufigkeit kommen Praktikumsberichte, Exposés, Protokolle, Essays oder auch Fragebögen vor.

4 Nutzung des Korpus KoLaS

Die Schreibwerkstatt Mehrsprachigkeit nutzte das Korpus in der Lehre (vor allem in der Ausbildung von studentischer Schreibberaterinnen und Schreibberater) sowie in der Forschung. Hier steht die Untersuchung der Textkommentierungen für Aus- und Weiterbildungszwecke von

Schreibberaterinnen und -beratern im Fokus. Da jedoch unsere Forschungskapazitäten sehr begrenzt sind, möchten wir mit KoLaS allen Studierenden sowie Wissenschaftlerinnen und Wissenschaftlern ermöglichen, Fragen rund um akademische Textproduktionsprozesse empirisch zu untersuchen.
Das Korpus steht daher kostenfrei über das Hamburger Zentrum für Sprachkorpora (HZSK) zur Verfügung. Hierfür muss lediglich ein Zugang für das Korpus über das HZSK beantragt werden.

Publikationen

- Andresen, Melanie/Knorr, Dagmar (2017). KoLaS – Ein Lernendenkorpus in der Schreibberatungsausbildung einsetzen. Zeitschrift Schreiben.
- Alagöz-Bakan, Özlem (2016): Textkommentierung unter Berücksichtigung der Nicht-Direktivität. In: Alagöz-Bakan, Özlem/Knorr, Dagmar/Krüsemann, Kerstin (Hrsg.): Akademisches Schreiben (Halbband 2). Sprache zum Schreiben – zum Denken – zum Beraten. Hamburg: Universität Hamburg [Universitätskolleg-Schrift; 14]
- Andresen, Melanie (2016): Im Theorie-Teil der Arbeit werden wir über Mehrsprachigkeit diskutieren – Sprechhandlungsverben in Wissenschafts- und Pressesprache. In: Zeitschrift für angewandte Linguistik 64/1. DOI: 10.1515/zfal-2016-0001

Präsentationen

- Andresen, Melanie: Using a Learner Corpus for Peer Tutor Training. Vortrag auf der Teaching and Language Corpora (TaLC) in Giessen vom 20.–23.07.2016.
- Andresen, Melanie/Knorr, Dagmar: KoLaS – Commented Learner Corpus of Academic German. Poster auf der Teaching and Language Corpora (TaLC) in Giessen vom 20.–23.07.2016. (pdf)
- Andresen, Melanie/Knorr, Dagmar: KoLaS – Kommentiertes Lernendenkorpus akademisches Schreiben. Poster auf Forum CA3 2016 in Hamburg vom 07.–08.06.2016.

Abgeschlossene, unveröffentlichte Projekte

- Tilmans, Anna (2013): „Einleitungen: Exemplarische Analysen studentischer Arbeiten“ (Bachelorarbeit).
- Andresen, Melanie (2014): „Im Theorie-Teil der Arbeit werden wir über Mehrsprachigkeit diskutieren – Verwendung von Sprechhandlungsverben in der deutschen Wissenschaftssprache“ (Masterarbeit, pdf)
- Stern, Claudia (2016): Ich-Verwendung in Lernendentexten (Hausarbeit)

Laufende Projekte

- Hansmeier, Judith: Fehlermuster im Artikelgebrauch russischsprachiger DaF-Lernender (Bachelorarbeit).

5 Perspektiven

Die nutzerfreundliche Aufbereitung der Daten ist unser Anliegen. Bisher liegen die Daten nur als Sammlung von Word- bzw. PDF-Dokumenten vor. Dies erschwert beispielsweise die gemeinsame Durchsuchbarkeit aller Dateien, außerdem sind die Nutzer/innen an eine spezifische Software gebunden. Geplant ist deshalb eine Integration in eine geeignete Software zur Korpusabfrage wie ANNIS oder cqp. Zusammen mit einer Annotation von Lemmata und Wortarten würde das die Analysemöglichkeiten vergrößern und vereinfachen.

Da sich die Rahmenbedingungen der Kooperation durch das Projektende der Schreibwerkstatt Mehrsprachigkeit geändert haben, sind wir dabei neue Wege der Finanzierung und Kooperation zu eruieren.

Kontakt

Dr. Dagmar Knorr
Leuphana Universität Lüneburg
ZeMoS – Zentraleinrichtung Moderne Sprachen
Scharnhorststraße 1
21335 Lüneburg
dagmar.knorr@leuphana.de
Tel. +49.4131.677-2651

Melanie Andresen M.A.
Universität Hamburg
Institut für Germanistik
Deutsche Sprache/Linguistik
Von-Melle-Park 6
20146 Hamburg
Tel. +49.40.42838-3227
Melanie.Andresen@uni-hamburg.de
https://www.slm.uni-hamburg.de/germanistik/personen/andresen.html