Gesprochene Muttersprache vs. Lernersprache

Aufbau und Auswertung eines Korpus

Simon Sauer, Linda Giesel, Myriam Klapi, Daisy Krüger, Isabelle Nunberger und Oxana Rasskazova

Im Tutorium „Gesprochene Muttersprache vs. Lernersprache – Aufbau und Auswertung eines Korpus“ am Institut für deutsche Sprache und Linguistik wurden gemeinsam mit den Studierenden Sprachdaten systematisch erhoben und aufbereitet. Diese Daten wurden in einem Korpus nachhaltig und für Forschung und Lehre frei verfügbar zugänglich gemacht. Dies ermöglicht empirisch-fundierte kontrastive Analysen zur Sprache von MuttersprachlerInnen und LernerInnen des Deutschen, welche für Forschung und Fremdspracherwerb von großer Bedeutung sind.

Motivation und inhaltliche Fragestellungen

Im Rahmen der linguistischen Forschung und insbesondere im Bereich Deutsch als Fremdsprache nimmt die Untersuchung von Lernersprache eine zentrale Rolle ein. Viele Forschungsbeiträge beschäftigen sich jedoch ausschließlich mit geschriebener Lernersprache, da Ressourcen für gesprochene Sprache rar sind. Um solche Daten zugänglich und systematisch durchsuchbar zu machen, wird in der Linguistik mit Korpora gearbeitet. Es handelt sich hierbei um kontrolliert erhobene und zusammengestellte Sammlungen sprachlicher Daten. Die meisten vorhandenen Korpora beinhalten geschriebene ‚Standardsprache‘ (z.B. Zeitungssprache). Für solche Korpora gibt es bereits viele frei verfügbare computerlinguistische und korpuslinguistische Werkzeuge zur Anreicherung mit linguistischen Informationen (Annotation). Allerdings deckt geschriebene Standardsprache nur einen geringen Teil der sprachlichen Wirklichkeit ab. In den letzten Jahren gab es daher ein wachsendes Interesse an systematisch erhobenen und tief annotierten Daten zu „Nichtstandardvarietäten“, da die Strukturen in diesen Daten teilweise stark vom Standard abweichen. Dazu zählen unter anderem spontan gesprochene Sprache und Lernerdaten. Korpora dieser sogenannten Nichtstandardvarietäten sind bisher oft noch klein – außerdem fehlen Werkzeuge, die diese Daten bearbeiten können. Für das Deutsche gibt es bisher nur drei sehr kleine öffentlich zugängliche Korpora gesprochener Lernersprache, keines dieser Korpora kann weiter bearbeitet werden. Eines dieser Korpora ist das Hamburg Map Task Corpus (HAMATAC; Schmidt et al. 2010). Unter gesprochener Lernersprache werden hier kontrolliert erhobene Gespräche von fortgeschrittenen LernerInnen des Deutschen als Fremdsprache verstanden. HAMATAC besteht aus Gesprächen, in denen jeweils zwei Deutschlernende eine zielorientierte Aufgabe erfüllen. Die Aufgabenstellung besteht darin, dass sich zwei ProbandInnen gegenseitig jeweils eine Route auf einer Karte erklären, die dann gezeichnet werden soll (vgl. Anderson et al. 1991; s. Abb. 1). Damit wird eine spontansprachliche Gesprächssituation erzeugt, welche jedoch thematisch klar abgegrenzt ist, was Vergleichbarkeit und Generalisierungen erleichtert.

Das im Tutorium aufgebaute Berlin Map Task Corpus (BeMaTaC) dient als muttersprachliches Vergleichskorpus zu HAMATAC. Den Studierenden wurde hierbei die Gelegenheit geboten,

selbstbestimmt sowie praxisorientiert zu arbeiten und am Aufbau eines Korpus beteiligt zu sein. Die Motivation für den Aufbau von BeMaTaC entstand in Anlehnung und als Reaktion auf HAMATAC. Das ursprüngliche Grundgerüst des Korpusdesigns von BeMaTaC orientiert sich daher an den Strukturen von HAMATAC, welche allerdings in vielerlei Hinsicht, insbesondere für automatisierte Abfragen, unzureichend sind. Durch ein vergleichbares Korpusdesign wird mit BeMaTaC eine Ressource geschaffen, die für die Spracherwerbsforschung wesentliche kontrastive Analysen ermöglicht, ohne die beispielsweise die Vermeidung von bestimmten Konstruktionen nicht erkennbar ist. Der korpusbasierte Vergleich von Muttersprache und Lernersprache ist für eine Reihe von Forschungsthemen interessant, die im Tutorium diskutiert worden sind.

In dem folgenden Beispiel, in dem ein Proband seinem Gesprächspartner einen Weg auf einer Karte erklärt (siehe unten), sieht man, dass LernerInnen manchmal ungrammatische Wörter oder Formen (unter das Motorrad, Nageln) benutzen, dass sie viele ungefüllte (durch Doppelklammern mit der Längenangabe angezeigt) und gefüllte Pausen (äh) machen, dass sie manche Ausdrücke abbrechen und wieder neu anfangen (gehs/ fährst).

gehs/ fährst du unter das Motorrad ((0,8s)) durch ((1,2s)) nach links ((1,0)) dann ((1,1s)) unter den äh

((0,8s)) Nageln

Einige dieser Phänomene können Hinweise auf den Spracherwerbsverlauf und Interferenzen aus der Muttersprache der LernerInnen geben. Andere, wie die gefüllten Pausen, sind aber typische Phänomene der gesprochenen Spontansprache, die auch in muttersprachlicher Spontansprache häufig sind. Forschungsfragen, die im Tutorium behandelt wurden, waren unter anderem:

- Welche Wörter, Formen und Strukturen finden wir in spontansprachlichen Dialogen bei LernerInnen des Deutschen als Fremdsprache? Welche davon sind typisch spontansprachliche Phänomene und welche sind typische Lernerphänomene?
- Wie kann man ein multimodales Korpus aufbauen? Welche Formate sind zu beachten? Wie kann Nachhaltigkeit sichergestellt werden? Welche Werkzeuge gibt es?

Die Studierenden hatten die Aufgabe, eine selbst gewählte Forschungsfrage zu bearbeiten und somit erste Untersuchungen mit BeMaTaC vorzustellen. Dabei gab es kontrastive Untersuchungen zu HAMATAC und BeMaTaC, wie auch Forschungsbeiträge, die sich nur auf BeMaTaC bezogen, zum Beispiel:

- Gibt es unterschiedliche Strategien zur Lösung von Missverständnissen oder Irrtümern zwischen MuttersprachlerInnen und LernerInnen?
- Nähern sich die Sprechgeschwindigkeiten zwischen den GesprächspartnerInnen im Laufe des Gesprächs an?
- Lässt sich bei MuttersprachlerInnen in subordinierenden Sätzen (Nebensätzen) ein systematisches Autreten von Verbzweitstellung beobachten?

Erhebung der Daten und ihre technische Aufbereitung

Die Aufnahmen der ProbandInnen wurden mit Hilfe der TutorInnen und der jeweiligen studentischen Arbeitsgruppe in den schallisolierten Laboren des Lehrstuhls für Phonetik am Institut für deutsche Sprache und Linguistik der Humboldt-Universität zu Berlin mit den dafür vorgesehenen Video- und Audioaufnahmegeräten Olympus LS-11 und LS-20 erhoben. Die dabei verwendeten Karten stammen

aus einem Forschungsprojekt des IDS Mannheim (Brinckmann et al. 2008). Von allen ProbandInnen wurden systematisch linguistisch relevante Metadaten durch Fragebögen erhoben.

Die Hauptaufgabe der Studierenden bestand darin, in Gruppen die Aufnahmen zu verschriften, das heißt nach genau festgelegten Kriterien zu transkribieren und weitere gemeinsam festgelegte Ebenen hinzuzufügen. Auf einer ersten Ebene der Transkription wurde die Aufnahme relativ nah am tatsächlich Gesprochenen verschriftet (dipl), während auf der zweiten normalisierten Ebene das Gesagte an die orthographische Norm (norm) angepasst und Wortgrenzen gesetzt wurden. Auf weiteren Ebenen haben die Studierenden in Äußerungseinheiten segmentiert (utt) und extralinguistische Ereignisse wie „lacht“ und „holt Luft“ (extra) annotiert. Pausen wurden auf einer weiteren dafür vorgesehenen Ebene (break) festgehalten.

Für diesen Arbeitsprozess haben die Studierenden mit der für die Linguistik sehr weit verbreiteten und frei verfügbaren Transkriptionssoftware Praat (Boersma 2010; s. Abb. 2) sowie EXMARaLDA (Schmidt/Wörner 2009) gearbeitet. Über das frei verfügbare und modular erweiterbare Konvertierungsframework SaltNPepper (Zipser/Romary 2010) wurde das Korpus in dem am Lehrstuhl für Korpuslinguistik und Morphologie entwickelten Such- und Visualisierungssystem ANNIS (Zeldes et al. 2009; s. Abb. 3) nachhaltig zugänglich und systematisch durchsuchbar gemacht.

Die Schritte der Transkription und Annotation wurden fortwährend begleitet und im Seminar gemeinsam mit den Studierenden diskutiert. Ergebnisse und Festlegungen dieses Austausches wurden in den Transkriptionsrichtlinien für BeMaTaC ständig erneuert und transparent gemacht.

Mithilfe engagierter studentischer Mitarbeit ist es uns nun möglich, ein Korpus zu präsentieren, das ein besonders detailliertes Annotationsschema für verschiedenste Phänomene gesprochener Sprache aufzeigt. Im Unterschied zu HAMATAC wurde eine gleichbleibende Formulierung der Aufgabenstellung gewährleistet, wurden neben Audio- auch Videoaufnahmen gemacht und es wird nicht nur auf Äußerungs- sondern auch auf Wortebene segmentiert. Zusätzlich werden jeweils in unterschiedlichen Ebenen Grundformen (lemma), Wortarten (pos), extralinguistische Ereignisse und akustische Pausen annotiert.

Erweiterung und Zukunft der Daten

In einem weiteren Tutorium im Sommersemester 2013 wird der Fokus auf Lernersprache gelegt. Zwar bleibt das grundsätzliche Setting der Map Task von HAMATACin BeMaTaC erhalten, doch unterscheiden sich die beiden Korpora in ihrer technischen Aufbereitung besonders in Bezug auf Multimodalität und Annotation mittlerweile so stark, dass eine neue Sammlung mit Lernerdaten sinnvoll erscheint. Ziel ist es, die neuen Daten mit identischem Korpusdesign und Transkriptionsrichtlinien aufzubereiten, um kontrastive Untersuchungen zu verschiedensten Fragestellungen zu ermöglichen.

Mehrere Studierende haben zudem vor, die im Tutorium erlernte Expertise sowie die gewonnen Forschungsdaten ihrer Präsentationen in Abschlussarbeiten anzuwenden. Einige Studierende, die erneut am Tutorium teilnehmen, erweitern das Korpus zudem mit weiteren Annotationsebenen und stehen den zukünftigen Teilnehmenden als MentorInnen zur Verfügung.

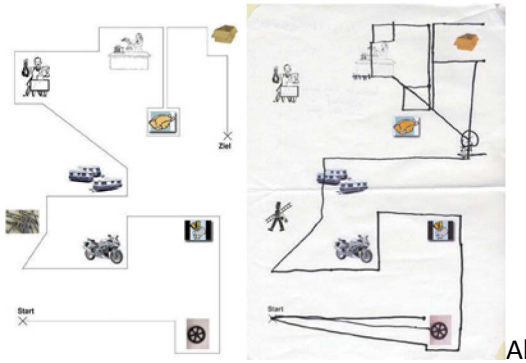

Abb.1

Bildrechte: Simon Sauer

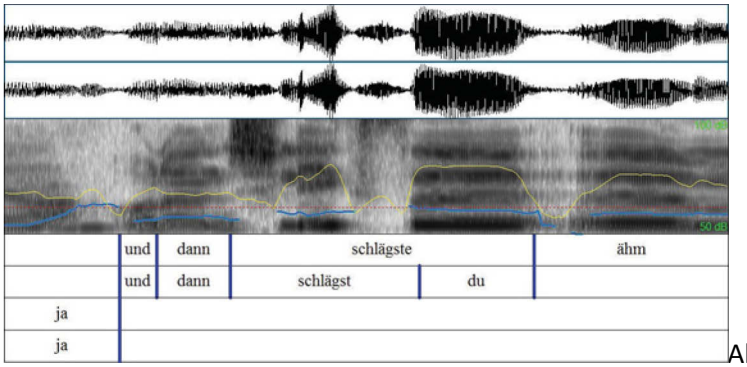

Abb.2

Bildrechte: Simon Sauer

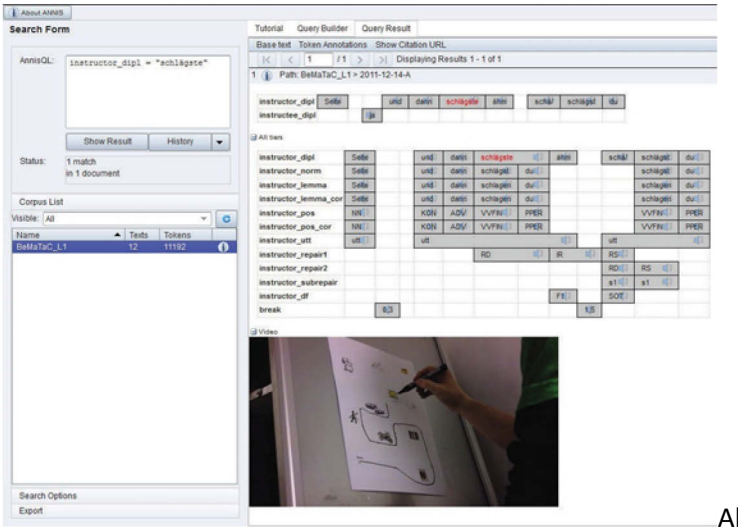

Abb.3

Bildrechte: Simon Sauer

Literaturverzeichnis

Anderson, Anne H./Bader, Miles/Gurman Bard, Ellen/Boyle, Elizabeth/ Doherty, Gwyneth/Garrod, Simon/Isard. Stephen/Kowtko, Jacqueline/ McAllister, Jan/Miller, Jim/Sotillo, Catherine/hompson, Henry/Weinert, Regina (1991): he HCRC Map Task Corpus. In: Language and Speech 34, S. 351-366.

BeMaTaC (2013): Webseite des Forschungsprojekts BeMaTaC. URL: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/ forschung/bematac [abgerufen am 17.04.2013].

Boersma, Paul (2010): Praat, a system for doing phonetics by computer. In: Glot International 5, 9-10, S. 341-345.

Brinckmann, Caren/Kleiner, Stefan/Knöbl, Ralf/Berend, Nina (2008): German Today: an areally extensive corpus of spoken Standard German. In: Proceedings 6th International Conference on Language Resources and Evaluation, LREC 2008.

Schmidt, Thomas/Wörner, Kai (2009): EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. In: Pragmatics, 19, 4, S. 565-582.

Schmidt, Thomas/Hedeland, Hanna/Lehmberg, Timm/Wörner, Kai (2010): HAMATAC – he Hamburg MapTask Corpus.

Zeldes, Amir/Ritz, Julia/Lüdeling, Anke/Chiarcos, Christian (2009): ANNIS: A Search Tool for Multi-Layer Annotated Corpora. In: Proceedings of Corpus Linguistics 2009, July, S. 20-23.

Zipser, Florian/Romary, Laurent (2010): A model oriented approach to the mapping of annotation formats using standards. In: Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010.