BeMaTaC, Falko, RIDGES

Linguistische Mehrebenenkorpora für Nichtstandard-Varietäten des Deutschen

Anke Lüdeling, Malte Belz, Hagen Hirschmann, Martin Klotz, Carolin Odebrecht, Laura Perlitz, Simon Sauer, Vivian Voigt

Einleitung

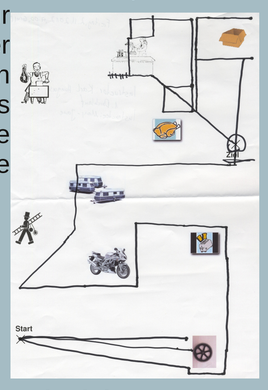
Mehrebenenkorpora ermöglichen die Analyse von Sprache auf verschiedenen Ebenen und sind Grundlage für die Beantwortung zahlreicher Forschungsfragen auch über die ursprüngliche Zielstellung hinaus.
Die hier vorgestellten Ressourcen decken unterschiedliche Nichtstandardvarietäten ab: gesprochene Sprache, historische Sprache und Lernersprache.

Für alle Korpora gilt:
- Durchsuchbarkeit im browserbasierten Such- und Visualisierungssystem ANNIS [1]
- freie Verfügbarkeit unter CC-BY 3.0 [2]
- konsistent erhobene Metadaten und umfangreiche Dokumentation
- automatische Lemmatisierung und Wortartentagging nach dem STTS [3] mit TreeTagger [4]

BeMaTaC

(Berlin Map Task Corpus) [5] besteht aus Dialogen fortgeschrittener Lerner von Deutsch als Fremdsprache sowie Dialogen deutscher Muttersprachler. Ziel des Korpus ist es, gesprochene Lernersprache systematisch untersuchbar zu machen. Da für gesprochene Sprache keine kanonische Form existiert, kann dies nur über einen Vergleich mit Muttersprache erfolgen. Die Verwendung einer konkreten Aufgabenstellung ermöglicht sowohl spontane konversationale Sprache als auch einen vergleichbaren kontrollierten Kontext. Die in BeMaTaC gestellte Aufgabe ist eine Map Task [6]: Eine Person erklärt der anderen eine Route

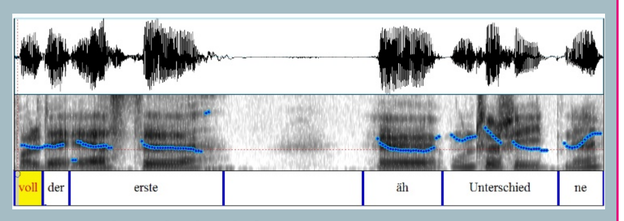
- Aktuelle Version: 2.1 / 2013-02.1
- Token: 19.046 (diplomatische Transkription, s.u.) in 17 Dialogen, 143 Minuten Audio- und Videomaterial
- Formate: EXMARaLDA-Partituren, Praat-TextGrids, Audio (WAVE, mp3), Video (QuickTime, WebM)
- Transkription und Annotation: Studierende aus 4 Tutorien und im Rahmen diverser Untersuchungen



Beispielroute

Korpusstruktur
- Diplomatische Transkription (dipl): enthält auch gefüllte Pausen, Abbrüche, Zusammenziehungen
- Normalisierung (norm): entspricht den amtlichen Regeln der deutschen Rechtschreibung (*haste* wird zu *hast du*)
- Äußerungen, Backchannelling, extralinguistische Ereignisse, stille Pausen
- Disfluencies: z.B. gefüllte Pausen, Wortdehnungen, Explicit editing terms
- Reparaturen: Reparandum, Interregnum, Reparans
- Reparatursubkategorisierungen: Wiederholungen, Ersetzungen, Einfügungen

Bisherige Forschung
- Kontrastive Analysen zu Disfluencies und Reparaturen bei Muttersprachlern und Lernern [7]
- Verhalten von Frauen und Männern beim Backchanneling [8]



Annotationsbeispiel

Falko

(fehlerannotiertes Lernerkorpus) [9][10] besteht aus Zusammenfassungen (FalkoSummary) sowie Essays (FalkoEssay, FalkoEssayWHIG) von Lernern und Muttersprachlern des Deutschen. Die Daten wurden unter konstanten Bedingungen erhoben (Prüfungssituation) und zu jedem verfassten Text liegen umfangreiche Metadaten (z.B. Alter, Geschlecht und Sprachbiografie des Autors) vor. Ziel des Projektes ist es, eine systematische Untersuchung von fortgeschrittener Lernersprache zu ermöglichen. Dies kann über einen Vergleich der Lerner-Subkorpora mit den korrespondierenden Muttersprachler-Subkorpora oder die Auswertung von Abweichungen von den annotierten Zielhypothesen (s.u.) erfolgen.

- Aktuelle Version: 2.0
- Token: 381.447
- Formate: Excel

LT	Es	ist	diese	Gesetzte		die	wichtig
ZH1	Es	ist	dieses	Gesetz	,	das	wichtig
ZH1Diff			CHA	CHA	INS	CHA	

Annotationsbeispiel

Form	Kontinuum	Funktion
Minimale Zielhypothese (ZH1)		**Erweiterte Zielhypothese (ZH2)**
Minimale Veränderungen stellen eine morphosyntaktisch korrekte, grammatische Zielstruktur her.		Einbeziehung lexikalischer, semantischer und pragmatischer Information stellt eine funktional äquivalente Zielstruktur her.
+ gut operationalisierbar + hohe Konsistenz der Annotation zwischen Texten und Annotatoren + große strukturelle Nähe zur Lerneräußerung		+ bestmögliche Rekonstruktion der Lernerintention

Korpusstruktur
- Textebene mit durch die Zielhypothesenannotation bedingten Lücken (tok)
- Lernerreferenzebene:Textebene mit aufeinanderfolgenden Tokens (ctok)
- Zielhypothese 1: minimale Normalisierungsebene, korrigierte Orthografie und Morphosyntax (ZH1)
- Zielhypothese 2: weite Zielhypothese unter Berücksichtigung von Semantik, Pragmatik, Lexik (ZH2)
- POS und Lemma für ctok, ZH1 und ZH2
- Abweichungen zwischen ZH1-, ZH2- und ctok-Schicht (auch POS und Lemma)
- Dokument (TXTstructure, Annotation von Start und Ende) und Dokumentenstruktur (macro, strukturelle Annotation wie "title", "subtitle", ...)
- Satzspannen für FalkoEssay auf Basis von ctokpos
- Dependenz-Bäume auf ZH1 von FalkoEssayL2

Zielhypothesenannotation
Zunächst: Was ist ein Fehler?„a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterparts.“ [11] → Zielhypothesen als kanonische Entsprechungen von nicht-kanonischen Äußerungen oder Äußerungsteilen in Bezug auf bestimmte Regelsysteme [12]

Bisherige Forschung
- Komposition im fortgeschrittenen Lernerdeutsch [13]
- Over-Use- und Under-Use-Studien zu syntaktischen Kategorien [14]
- Parsing und Tagging [15]
- Erforschung von Fehlerkategorien bei Lernern und Muttersprachlern des Deutschen [16]
- Transferphänomene im Zweitspracherwerb [17]
- Untersuchungen zum Erwerb des Mittelfeldes [18]

RIDGES

[19] enthält 29 kräuterkundliche Textausschnitte aus dem 15. bis 19. Jahrhundert und deckt somit Frühneuhochdeutsch und modernes Deutsch ab. Ziel des Projektes ist es, die Entstehung und Entwicklung eines deutschen Wissenschaftsregisters auf verschiedenen sprachlichen Ebenen zu untersuchen.

- Aktuelle Version: 4.1
- Token: 154.267 (diplomatische Transkription, s.u.)
- Formate: ANNIS, Excel, PAULA
- Transkription und Annotation: HU-Studierende aus 3 Bachelor- und Masterseminaren [20]
- Aufarbeitung und Korrektur: im Rahmen des LAUDATIO-Projekts und unterschiedlichen Untersuchungen

Bisherige Forschung
- Komposita vs. Genitivkonstruktionen vom 15. bis 19. Jahrhundert [21]
- Optical Character Recognition frühneuzeitlicher Drucke [22]

Abbildung: Fuchs, Leonhardt (1543). New Kreüterbůch. Basel. Michael Isingrin. a1.

Korpusstruktur
- Diplomatische Transkription (dipl): so nah am Original wie möglich
- 1. Normalisierung (clean): Ersetzung der historischen Sonderzeichen durch heutige Zeichen (ſ → s)
- 2. Normalisierung (norm): Anpassung an die moderne Rechtschreibung (darauf basieren Lemmata und Wortartentags) → so können alle historischen Schreib- und Flexionsvarianten ohne explizite Kenntnis darüber gefunden werden
- Lexikalisch: z.B. Krankheiten, Personennamen, Kräuterbezeichnungen
- Syntaktisch: z.B. Genitivattribute, Verbpositionen, Nebensatztypen
- Morphologisch: z.B. Komposita
- Graphisch: z.B. Zeilen- und Seitenumbrüche, Überschriften, Zitate

dipl	Von	Wermůt	.	Das	erſt	Capitel	.
clean	Von	Wermut	.	Das	erst	Capitel	.
norm	Von	Wermut	.	Das	erste	Kapitel	.

Annotationsbeispiel

Referenzen:
1. Amir Zeldes, Julia Ritz, Anke Lüdeling & Christian Chiarcos. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. Proceedings of Corpus Linguistics 2009, July, 20-23.
2. http://creativecommons.org/licenses/by/3.0
3. Anne Schiller, Simone Teufel, Christine Thielen. 1995. Guidelines fuer das Tagging deutscher Textkorpora mit STTS. Technical Report, IMS-CL, University Stuttgart.
4. Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing.
5. http://u.hu-berlin.de/bematac
6. Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson & Regina Weinert. 1991. The HCRC Map Task Corpus. Language and Speech 34, 351-366.
7. Malte Belz. 2013. Disfluencies und Reparaturen bei Muttersprachlern und Lernern – eine kontrastive Analyse. Masterarbeit. Humboldt-Universität zu Berlin, November 2013.
8. Clara Becker. 2013. Doing Backchanneling – Verhalten von Frauen und Männern beim Backchanneling im aufgabenorientierten Dialog. Bachelorarbeit. Humboldt-Universität zu Berlin, Juli 2013.
9. http://linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko
10. Anke Lüdeling, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt & Maik Walter (2008): Das Lernerkorpus Falko. In: Deutsch als Fremdsprache 2(2008), 67-73.
11. Paul Lennon. 1991. Error. Some Problems of Definition, Identification, and Distinction. Applied Linguistics 12 (2), 180-196.
12. Anke Lüdeling. 2008. Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora. Maik Walter, Patrick Grommes: Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung. DGfS. Linguistische Arbeiten; 520, 119-140.
13. Zeldes, Amir (2013): Komposition als Konstruktionsnetzwerk im fortgeschrittenen L2-Deutsch. Zeitschrift für germanistische Linguistik 41 (2). 240–276.
14. Hirschmann, Hagen; Lüdeling, Anke; Rehbein, Ines; Reznicek, Marc; Zeldes, Amir (2013): Underuse of Syntactic Categories in Falko. A Case Study on Modification. In: Granger, Sylviane; Gilquin, Gaëtanelle; Meunier, Fanny (Hrsg.): Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Proceedings of the First Learner Corpus Research Conference (LCR 2011). Louvain-la-Neuve: Presses universitaires de Louvain.
15. Rehbein, Ines; Hirschmann, Hagen; Lüdeling, Anke; Reznicek, Marc (2013): Better Tags give Better Trees–or do they? In: Linguistic Issues in Language Technology, 7 (10).
16. Anke Lüdeling (2008): Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Patrick Grommes & Maik Walter (Hrsg.) Fortgeschrittene Lernervarietäten, Niemeyer, Tübingen, 119-140.
17. Reznicek, Marc; Golcher, Felix: What Similarity Tells us about Transfer. Retrieving L1 from Learner Texts in Falko. Tübingen-Berlin-Meeting, Tübingen, 05.12.2011.
18. Reznicek, Marc: The German Learner Middlefield. Linearisation-Factors of Verbal Arguments in the Falko Advanced Learner Corpus. Tübingen-Berlin-Meeting, Tübingen, 06.12.2011.
19. http://korpling.german.hu-berlin.de/ridges
20. Malte Belz, Carolin Odebrecht, Laura Perlitz, Vivian Voigt. 2015. Annotationsrichtlinien zu Ridges Herbology Version 4.1. Humboldt-Universität zu Berlin.
21. Laura Perlitz. 2014. Konkurrenz zwischen Wortbildung und Syntax – historische Entwicklung von Benennung. Bachelorarbeit. Humboldt-Universität zu Berlin, August 2014).
22. Uwe Springmann, Anke Lüdeling, Felix Schremmer. 2015. Zur OCR frühneuzeitlicher Drucke am Beispiel des RIDGES-Korpus von Kräutertexten. Poster, DHd, Graz.