58. Jahrestagung des Leibniz-Instituts für Deutsche Sprache vom 15. bis 17. März 2022
(als Online-Konferenz)

KORPORA IN DER GERMANISTISCHEN SPRACHWISSENSCHAFT – MÜNDLICH, SCHRIFTLICH, MULTIMEDIAL

METHODENMESSE

Mittwoch, 16. März 2022, 15:45 Uhr bis 17:45 Uhr

Das Berlin Dialogue Corpus (BeDiaCo) und das Corpus of Non-Native Addressee Register (CoNNAR) – themen- und aufgabenorientierte spontansprachliche Dialoge in direkter und videobasierter Kommunikation

Malte Belz, Bianca Sell, Robert Lange, Megumi Terada, Christine Mooshammer, Anke Lüdeling

1 Forschungsgegenstand

Das *Berlin Dialogue Corpus* enthält spontane gesprochensprachliche Dialoge in verschiedenen Situationen und dient der Grundlagenforschung in Bezug auf phonetisch-phonologische, morphologische, syntaktische, prosodische, pragmatische und interaktionslinguistische Phänomene spontaner gesprochener Sprache. Auch Forschung zu intra- und interindividueller Variation und registerspezifischen Fragestellungen sind möglich.
Das *Corpus of non-native addressee register* (CoNNAR) wurde mit der Fragestellung entwickelt, ob Sprecher*innen sich in ihrem Sprechstil an die tatsächlichen oder angenommenen Bedürfnisse ihrer Gesprächspartner*innen anpassen (Bell, 1984), wenn sie mit erwachsenen Lerner*innen kommunizieren (andere Begriffe sind *foreigner talk*, *foreigner-directed speech* oder Xenolekt, vgl. Bradlow and Bent, 2002; Roche, 1998). Auch dieses Korpus kann für weitere Fragestellungen herangezogen werden.

2 Korpus BeDiaCo

BeDiaCo v.2 (Belz et al. 2021a) enthält 150 000 Token von 36 Sprecher*innen in zwei Subkorpora mit insgesamt 11,2 h spontansprachlicher Dialoge (reine Artikulationszeit) und vorgelesener Wortlisten. Das Subkorpus BeDiaCo-main (BeDiaCo$_m$) enthält 46 000 Token von 16 SprecherInnen (10 männlich, 6 weiblich) zwischen 18 und 31 Jahren ($\bar{x}$ = 24,1, s = 4,2) mit 3 h Artikulationszeit, wobei sich die GesprächspartnerInnen vorher nicht kannten und in acht *face-to-face*-Dialogen miteinander sprechen. Das Subkorpus BeDiaCo-videocall (BeDiaCo$_v$) enthält 104 000 Token von 20 SprecherInnen (10 männlich, 10 weiblich) zwischen 19 und 32 Jahren ($\bar{x}$ = 25,7, s = 3,8) mit 8 h Artikulationszeit, die einander kennen (Mitbewohner*innen, Geschwister, Partner*innen/Eheleute). Sie wurden in zwei Bedingungen durchgeführt: face-to-face und mit dem Videokonferenztool Zoom.
In beiden Subkorpora wird der freie Dialog themenorientiert mit einer Frage des Experimentleiters nach der Qualität des Mensaessens (BeDiaCo$_m$), oder dem Thema Berlin bzw. Traumreise (BeDiaCo$_v$) begonnen und dauert zwischen 10 und 15 min, wobei die Versuchspersonen schnell dazu übergehen, über selbstgesetzte Themen zu sprechen. Die aufgabenbasierten Dialoge nutzen ins Deutsche übersetzte Diapix-Aufgaben (Baker & Hazan 2011, Van Engen et al. 2010) mit zwei fast identischen Bildern, in welchen die Versuchspersonen Unterschiede finden müssen. Zusätzlich sind Wortlisten für alle 36 Versuchspersonen enthalten, in denen ein Trägersatz *Sage X bitte* ein zweisilbiges Wort X enthält, welches auf der ersten Silbe betont ist. Die Wörter enthalten alle Monophthonge des Deutschen und die beiden Reduktionsvokale

[ə ɐ] in der zweiten unbetonten Silbe. Das Experiment ist folgendermaßen aufgebaut: Wortliste, Diapix, freier Dialog, Diapix, Wortliste. Die Versuchspersonen saßen in einer schallisolierten Kabine (mit Hörsprechgarnitur *beyerdynamics Opus 54*) bzw. in der Zoombedingung (mit Kopfhörern und Stabmikrofonen) zusätzlich in einem Büro, welches an das Labor angrenzt. Das Experiment dauerte maximal eine Stunde und wurde mit 10 bzw. 11 Euro kompensiert.

2.1 Annotation

Die Annotation ist in einem Mehrebenenmodell aufgebaut, mit dem akustischen Signal aligniert und ausführlich dokumentiert (Belz et al. 2021b). Enthalten sind manuell erstellte und automatisch alignierte Ebenen (z. B. die Transliteration mithilfe von WebMAUS, vgl. Kisler et al. 2017), automatisch erstellte Ebenen (z. B. Normalisierung, Lemmatisierung und Wortarten) sowie manuell erstellte und alignierte Ebenen (z. B. zu Füllpartikeln und Intonationsphrasen). Die Annotation erfolgt in Praat (Boersma & Weenink 2019).

2.2 Zugang und Wiederverwendung

Die pseudonymisierten Daten sind für wissenschaftliche Zwecke über das Medienrepositorium der Humboldt-Universität zu Berlin verfügbar. Das Korpus kann bspw. mit *emuR* (Winkelmann et al. 2018) in R analysiert werden. Eine solche vorerstellte Datenbank wird in einer neuen Version zur lokalen Verwendung bereitgestellt. Das Korpus wird für die Arbeit an neuen Forschungsfragen kontinuierlich verbessert und kollaborativ mit neuen Ebenen annotiert. Dies erfolgt über eine vom Computer- und Medienservice der Humboldt-Universität zu Berlin bereitgestellten Versionierungssoftware (GitLab). Weitere Annotationsebenen von Dritten können nach Rücksprache mit den HerausgeberInnen in das Korpus aufgenommen werden.

3 Korpus CoNNAR

CoNNAR besteht aus zwei Subkorpora, die insgesamt 40 Versuchspersonen sowie 16 instruierte Confederates enthalten. Die Versuchspersonen (L1 Deutsch, 20 männlich, 20 weiblich, zwischen 18 und 40 Jahren) durchlaufen das Experiment jeweils zweimal – einmal mit einem*r Nicht-Muttersprachler*in (L1 Englisch, Proficiency in Deutsch B1/B2 oder C1, Alter zwischen 18 und 40 Jahren) und einmal mit einem*r deutschen Muttersprachler*in, der*die dem*der anderen Confederate in Alter und Geschlecht entspricht. Das Experiment dauert 60–90 Minutenwird mit 11 Euro kompensiert.

Das Subkorpus CoNNAR-videocall (CoNNAR_{v}, ca. 110 000 Token) besteht aus 20 Versuchspersonen (je 10 weiblich und männlich, Alter 20–38, $\bar{x}$ = 26, s = 4,5) und 8 Confederates (je 4 weiblich und männlich, Alter 20–27, $\bar{x}$ = 22,9, s = 2, 4 Muttersprachler*innen und 4 Lerner*innen), die in getrennten Räumen (die VP in einer schallisolierten Kabine, der Confederate im angrenzenden Büro) sitzen und über das Videokonferenztool Zoom verbunden sind. Zoom wurde nur zur Kommunikation verwendet, die Audioaufnahmen erfolgte über Stabmikrofone. Das Experiment dauert jeweils etwa eine Stunde und wurde mit 11€ vergütet. Das Experimentdesign besteht aus Wortliste, freie Konversation, Diapix, Diapix, Wortliste, und ist damit ähnlich zu BeDiaCo aufgebaut, mit kleinen Unterschieden. In der freien Konversation unterhalten sich die Teilnehmer*innen 8 Minuten lang über ein Thema ihrer Wahl. In der Diapixaufgabe sollen die Teilnehmer*innen in 8 Minuten möglichst viele der 12 Unterschiede finden. Der Confederate bittet bei dem zweiten Diapix nach 3–4 gefundenen Unterschieden um eine klarere Aussprache.

Das Subkorpus CoNNAR-face-to-face ($\text{CoNNAR}_{\text{f2f}}$) befindet sich in der Erhebung. Geplant sind auch hier 20 Versuchspersonen und 8 Confederates, die gemeinsam in einer schallisolierten Kabine sitzen. Das Experimentdesign wird außerdem erweitert um eine Bildbeschreibungsaufgabe (die Versuchspersonen sehen ein Bild und lesen eine Bildbeschreibung vor bzw. beschreiben es selbst in einem Satz so, dass der*die Confederate das passende Bild aus 4 ähnlichen Bildern auswählen kann) und eine kognitive Belastung (vor den Diapixaufgaben sehen die Versuchspersonen jeweils eine 3x3-Matrix mit unterschiedlich komplex angeordneten Punktmustern, die nach der Diapixaufgabe reproduziert werden muss, vgl. De Neys and Schaeken, 2007).

3.1 Aufbereitung und Annotation

Die erhobenen Daten werden in Praat (Boersma and Weenink, 2019) transliteriert und in WEBMaus (Kisler et al., 2017) um segmentale Annotationen erweitert. Die Annotation ist in einem

Mehrebenenmodell aufgebaut und mit dem akustischen Signal aligniert. Mithilfe eines selbst erstellten Skripts werden Normalisierungen und POS-Tags ergänzt.

3.2 Zugang und Wiederverwendung

Die pseudonymisierten Daten werden nach Abschluss von Version 1 für wissenschaftliche Zwecke zur Verfügung gestellt. Das Korpus kann für die Arbeit an neuen Forschungsfragen kontinuierlich verbessert und kollaborativ mit neuen Ebenen annotiert werden (GitLab). Weitere Annotationsebenen von Dritten können nach Rücksprache mit den HerausgeberInnen in das Korpus aufgenommen werden.

Acknowledgements

Wir danken Melina Pfundstein, Alina Zöllner und Lea-Sophie Adam für die Hilfe bei der Erstellung und Annotation von BeDiaCo. BeDiaCo wurde gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1412, 416591334 und die Medienkommission des Akademischen Senats der Humboldt-Universität zu Berlin. CoNNAR wird wurde gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1412, 416591334.

Literatur:

Baker, Rachel & Valerie Hazan. 2011. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods* 43(3). 761–770. https://doi.org/10.3758/s13428-011-0075-y.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2).145–204.

Belz, Malte, Christine Mooshammer, Alina Zöllner & Lea-Sophie Adam. 2021a. *Berlin Dialogue Corpus (BeDiaCo): Version 2*. Medien-Repositorium. https://rs.cms.hu-berlin.de/phon.

Belz, Malte, Alina Zöllner, Megumi Terada, Robert Lange, Lea-Sophie Adam & Bianca Sell. 2021b. *Dokumentation und Annotationsrichtlinien für das Korpus BeDiaCo v2.* https://doi.org/10.5281/zenodo.4593351.

Boersma, Paul & David Weenink. 2019. *Praat: doing phonetics by computer [Computer program]*. http://www.praat.org/.

Bradlow, Ann R & Tessa Bent. 2002. The clear speech effect for non-nativelisteners. *J. Acoust. Soc. Am.* 112(1). 13.

De Neys, Wim & Walter Schaeken. 2007. When People Are More Logical Under Cognitive Load: Dual Task Impact on Scalar Implicature. *Experimental Psychology* 54(2). 128–133. https://doi.org/10.1027/1618-3169.54.2.

Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347. https://doi.org/10.1016/j.csl.2017.01.005.

Roche, Jörg. 1998. Variation in Xenolects (Foreigner Talk). *Sociolinguistica* 12(1). 117–139.

Van Engen, Kristin J., Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim & Ann R. Bradlow. 2010. The Wildcat Corpus of Native-and Foreign-accented English: Communicative Efficiency across Conversational Dyads with Varying Language Alignment Profiles. *Language and Speech* 53(4). 510–540. https://doi.org/10.1177/0023830910372495.

Winkelmann, Raphael, Klaus Jaensch, Steve Cassidy & Jonathan Harrington. 2018. *emuR: Main Package of the EMU Speech Database Management System*.