Versprecherdatenbank „schwierige Wörter“: Versprecherstatistiken und Versprecherklassifizierung

Sofiya Karnovska, Andrea Behrens, Anton Gadringer, Meryem Güvende, Astrid Kasztantowicz, Kalliopi Kyriakidis, Quoc Khanh Nguyen, Nicole Sommer, Christoph Draxler
Institut für Phonetik und Sprachverarbeitung, LMU München

Versprecherdaten sind eine wichtige Ressource, um Einblicke in die kognitiven Prozesse der Sprachproduktion zu gewinnen [1]. Ziel des Projekts *Schwierige Wörter* war es, im Rahmen eines Masterseminars ein Lesesprache-Versprecherkorpus als frei verfügbare Ressource für weitere und vergleichende Forschung zu erstellen.

Der Fokus bei der Datenerhebung lag auf Wörtern mit schwieriger lesesprachlicher Verarbeitung. Schwierige Wörter wurden verstanden als 1) Wörter mit mehreren Lesevarianten bei gleicher Orthografie, 2) Lehnwörter, 3) Komposita mit ambigen Teilwortgrenzen, 4) Wörter mit Silbendopplungen und 5) lange Wörter.

Das Korpus enthält sowohl schwierige Wörter als auch Kontrollwörter. Die Items stehen in unterschiedlichen Satzpositionen (medial, final sowie isoliert) und Satzkontexten (semantisch sinnvoller Satz, Trägersatz). Die Datenerhebung wurde online anhand visuell präsentierter Prompts über das Web-Aufnahmetool WikiSpeech [2] durchgeführt. Um Versprecher gezielt zu provozieren, wurden die Items in einem automatischen Ablauf unter Zeitdruck präsentiert und aufgenommen.

Zwischen dem 16.12.2021 und dem 19.01.2022 wurden 35 vollständige Sitzungen aufgenommen. Die Aufrufe zur Teilnahme erfolgten über lokale Verteiler, soziale Medien und die P&P Mailingliste.

Insgesamt wurden 6733 Äußerungen orthographisch transkribiert, annotiert und bezüglich der Existenz von Versprechern kategorisiert. Für die Transkription wurde OCTRA [3] verwendet. Die Transkription erfolgte in zwei Stufen: 1) Ersttranskription aller Aufnahmen, 2) zweite Transkription aller als kritisch oder unsicher markierten Ersttranskriptionen. Alle Transkripte wurden anschließend automatisch mit MAUS [4] segmentiert.

Um die Nutzbarkeit der Daten für künftige Forschungsvorhaben zu Versprechern sicherzustellen, wurde eine Analyse der Versprecherquote mithilfe deskriptiv-statistischer Methoden durchgeführt. In einem Datenausschnitt von 5593 Aufnahmen konnte eine Auftretenshäufigkeit von Versprechern von 19% ermittelt werden. Dabei machten die schwierigen Zielwörter einen Großteil (83%) der dokumentierten Versprecher aus, die restlichen Versprecher traten bei den Kontrollitems auf.

Darauf aufbauend erfolgte eine Versprecherklassifikation für den Zweck, einen Überblick über die vertretenen Versprecherarten zu schaffen. Eine Stichprobe von 222 Versprechern wurde hierfür manuell klassifiziert. Bei einem Großteil der Versprecher handelte es sich um Segmentationsfehler auf Silbenebene. Ebenfalls häufig vertreten waren Reparaturen und autoreflexive Kommentare, wie Lachen sowie auch antizipatorische Versprecher. Segmentationsfehler auf Lautebene wurden nur sehr selten beobachtet.

Aus den Versprecherstatistiken und -klassifikationen konnte abgeleitet werden, dass die erhobenen Daten eine gute Grundlage für die Untersuchung der Verarbeitung schwieriger Wörter im Deutschen darstellen, insbesondere in Bezug auf lexikalische Verarbeitung und Reparaturen bei Lesesprache unter Zeitdruck.

Es ist geplant, dieses Korpus als öffentliche Ressource im BAS Repository (https://clarin.phonetik.uni-muenchen.de/BASRepository/) abzulegen.

doi:10.11576/pundp2022-1020

[1] U. Schade, T. Berg, und U. Laubenstein, „Versprecher und ihre Reparaturen“, in *Psycholinguistik. Psycholinguistics: Ein internationales Handbuch*, G. Rickheit, T. Herrmann, und W. Deutsch, Hrsg. De Gruyter Mouton, 2008, S. 317–338. doi: doi:10.1515/9783110114249.3.317.

[2] C. Draxler und K. Jänsch, „Wikispeech - a content management system for speech databases“, in *Interspeech 2008*, Sep. 2008, S. 1646–1649. doi: 10.21437/Interspeech.2008-457.

[3] C. Draxler und J. Pömp, „OCTRA-A configurable browser-based editor for orthographic transcription“, in *Proceedings of Phonetik und Phonologie im deutschsprachigen Raum*, 2017, S. 145–148.

[4] F. Schiel, „MAUS Goes Iterative“, in: *Proceedings of the 4th Intl. conference on Language Resources and Evaluation*, Lisbon, 2004, S. 1015-1018

doi:10.11576/pundp2022-1020