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 VinKo-Korpus

Anne Kruijt, Stefan Rabanus, Marta Tagliani

Das VinKo-Korpus ist ein Parallelkorpus, das aus Tonaufnahmen deutscher und italienischer Dialekte und Minderheitensprachen besteht, die in den italienischen Regionen Trentino-Südtirol, Venetien und Friaul-Julisch Venetien gesprochen werden. Die Daten werden über die Crowdsourcing-Plattform des VinKo-Projekts ("Varietäten in/m Kontakt"; <https://www.vinko.it/index.php?lang=de>) gesammelt. Für die drei linguistischen Systemebenen Phonetik/Phonologie, Morphologie und Syntax sind Variablen so definiert, dass die Varianten sprachgruppenübergreifend (deutsch-italienisch) verglichen werden können. Bei der Erhebung wird eine Mischung aus traditionellen Übersetzungs- und Ausspracheaufgaben und innovativen Erhebungsmethoden in Form von Storyboard-Übersetzungsaufgaben und Aufgaben zur gesteuerten freien Sprachproduktion (Vervollständigung von konversationellen Paarsequenzen) verwendet. Der ursprüngliche, von der VinKo-Arbeitsgruppe verfolge Zweck der Datensammlung ist die Untersuchung von alter Mehrsprachigkeit, Sprachkontakt und sprachkontaktinduziertem oder -befördertem Sprachwandel im Untersuchungsgebiet. Im Mittelpunkt stehen die Tiroler Dialekte und germanische Minderheitensprachen wie Zimbrisch, Fersentalerisch, Sauranisch und Plodarisch einerseits, und romanische Varietäten wie das Ladinische und die Dialekte des Trentino und Venetiens andererseits.

Die VinKo-Arbeitsgruppe bekennt sich zum "open science"-Ansatz und zu den FAIR-Prinzipien für die Daten (*findable, accessible, interoperable, re-usable*). Aus diesem Grund werden die Daten der Forschung und der interessierten nicht-wissenschaftlichen Öffentlichkeit auf drei Arten zur Verfügung gestellt. (1) Eine Auswahl der Tonaufnahmen ist über eine interaktive Karte (GIS-System) und verschiedene Such- und Darstellungsfunktionen auf der VinKo-Webseite offen konsultierbar: <https://vinko.it/listen-explore.php?lang=de>. (2) Die Gesamtdaten sind erstens in einem passwortgeschützen Bereich der VinKo-Webseite zugänglich (Zugangsdaten können bei der VinKo-Arbeitsgruppe angefordert werden). (3) Zweitens kann das gesamte VinKo-Korpus (Rabanus et al. 2021) aus dem Repository des Eurac Research Clarin Centre (ERCC) in Bozen heruntergeladen werden (ERCC ist Teil der größeren europäischen CLARIN-Initiative [Common Language Resources and Technology Infrastructure]). Zum gegenwärtigen Zeitpunkt enthält der Datensatz im Repository alle VinKo-Daten aus dem Zeitraum von Juni 2017 bis Mai 2021 und umfasst damit insgesamt 37.806 Tonaufnahmen. Der Datensatz wird regelmäßig aktualisiert, um neue Daten aufzunehmen (im passwortgeschützten Bereich der VinKo-Webseite sind heute [08.02.2022] 90.000 Tonaufnahmen gespeichert). Das VinKo-Korpus besteht aus drei Hauptordnern: dem Audio-Ordner, der die Tonaufnahmen enthält; dem Metadaten-Ordner, der Tabellen mit relevanten Informationen zu den Stimuli (v.a. Explikation der linguistischen Variable und deutsche, italienische bzw. dialektale Ausgangsform für die Aufgabe) sowie soziolinguistische Informationen über die

Sprecher und damit auch die für die Analyse notwendigen Bezüge enthält; dem Bilder-Ordner, der die Bilder enthält, die als visueller Kontext für die Storyboard-Aufgaben verwendet werden. Alle Daten sind unter einer CreativeCommonsAttributionNon-CommercialShare-Alike-Lizenz lizenziert und können für nicht kommerzielle Zwecke frei verwendet werden.

Literatur:

Rabanus, Stefan, Alessandra Tomaselli, Andrea Padovan, Anne Kruijt, Birgit Alber, Patrizia Cordin, Roberto Zamparelli, and Barbara Maria Vogt. 'VinKo (Varieties in Contact) Corpus'. https://www.vinko.it, 2021. https://clarin.eurac.edu/repository/xmlui/handle/20.500.12124/32