Aivars Glaznieks/Jennifer-Carmen Frey (Bozen)

Das DiDi-Korpus: Internetbasierte Kommunikation aus Südtirol

Südtirol ist eine mehrsprachige italienische Provinz, in der die Verwendung unterschiedlicher Sprachen, besonders Deutsch und Italienisch, sowie der lokalen deutschen Dialekte in der mündlichen Kommunikation in formalen wie informalen Sprechsituationen einen hohen gesellschaftlichen Stellenwert hat. Mit der Frage, welche Sprachen bzw. Varietäten in der schriftlichen Alltagskommunikation verwendet werden und welche soziolinguistischen Faktoren dabei eine Rolle spielen, hat sich das Projekt DiDi befasst, in dem die Sprach- und Varietätenverwendung in Facebook-Texten näher untersucht wurden.¹ Dabei stellte sich unter anderem heraus, dass das Schreiben im Dialekt besonders unter Jugendlichen weit verbreitet ist (Glaznieks/Frey 2018). Mithilfe des aus diesem Projekt entstandenen und für wissenschaftliche Nutzung frei zugänglichen Facebook-Korpus kann die Sprach- und Varietätenverwendung Südtiroler Facebooknutzer/innen in der internetbasierten Kommunikation aus unterschiedlichen linguistischen Perspektiven untersucht werden.

Das DiDi-Korpus (vgl. Frey/Glaznieks/Stemle 2016) besteht insgesamt aus ca. 40.000 Facebook-Texten (bzw. 600.000 Token). Das Korpus ist eine Sammlung von soziolinguistisch und computerlinguistisch aufbereiteten Facebook-Statusmeldungen, -Kommentaren und -Chatnachrichten. Alle Texte stammen von privaten Profilen von 133 Personen aus Südtirol. Die Texte stellen die gesamte Textproduktion dieser Personen eines Jahres dar und stammen aus dem Jahr 2013. Die Schreiberinnen und Schreiber der Texte waren zum Zeitpunkt der Texterstellung zwischen 15 und 76 Jahre alt und sind überwiegend deutsche Muttersprachler/innen. Auch die Texte sind überwiegend auf Deutsch verfasst, wobei das Korpus auch Texte in anderen Sprachen, allen voran auf Italienisch und Englisch, sowie gemischtsprachige Texte enthält. Eine weitere Besonderheit des Korpus stellt der relativ hohe Anteil an Texten dar, die in den lokalen Dialekten verfasst wurden. Alle dialektalen Texte wurden auf Tokenebene mit standardsprachlichen Token ergänzt, sodass Suchanfragen in der Standardsprache für alle Texte möglich sind. Zudem wurden alle Texte nach Wortart und Lemma automatisch und spezifische

¹ Das Projekt Digital Natives – Digital Immigrants (DiDi) wurde von 2013–2015 von der Wissenschaftsförderung der Provinz Bozen – Südtirol finanziell gefördert (Landesgesetz vom 13. Dezember 2006, Nr. 14 „Forschung und Innovation“). Weitere Informationen zum Projekt finden sich online unter: www.eurac.edu/didi (Stand: 7.8.2019).

Phänomene internetbasierter Kommunikation wie die Verwendung von Emoticons, Emojis, Hashtags u.a. halbautomatisch annotiert. Schließlich stehen diverse personenbezogene Metadaten wie Alter, Geschlecht, regionale Herkunft u.a. zur Verfügung, die das Korpus auch für soziolinguistische Untersuchungen interessant machen (vgl. Frey/Glaznieks 2018; Glaznieks/Frey 2018; Glaznieks/Glück 2019). Darüber hinaus kann das Korpus auch für didaktische Zwecke eingesetzt werden, um im schulischen und universitären Unterricht Besonderheiten von internetbasierter Kommunikation an authentischen Texten herauszuarbeiten und zu illustrieren.

Das Korpus kann über eine ANNIS-Suchoberfläche (<https://commul.eurac.edu/annis/didi>) direkt online durchsucht werden. Dies ermöglicht die Abfrage und Anzeige spezieller sprachlicher Phänomene in einer Keyword-in-Context-Ansicht, die auf mehrere Annotationsebenen zugreifen kann. Technisch versierte Benutzerinnen und Benutzer können das Korpus auch über die CLARIN-Infrastruktur (<https://clarin.eurac.edu/>; beide Stand: 7.8.2019) in Form von XML- oder JSON-Dateien herunterladen. Es steht somit auch für eine individuelle Weiterverarbeitung und Verwendung mit anderen Tools und Programmen zur Verfügung. Öffentlich zugänglich sind zurzeit nur die deutschsprachigen Texte; Texte in anderen Sprachen sollen jedoch zukünftig zum Korpus hinzugefügt werden.

Literatur

- Frey, Jennifer-Carmen/Glaznieks, Aivars (2018): The myth of the digital native: Analysing language use of different generations on Facebook. In: Vandekerckhove, Reinhild/Fišer, Darja/Hilte, Lisa (Hg.): Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-Corpora 2018). Antwerpen, S. 41–44.
- Frey, Jennifer-Carmen/Glaznieks, Aivars/Stemle, Egon W. (2016): The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In: Corazza, Anna/Montemagni, Simone/Semeraro, Giovanni (Hg.): Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016). Turin, S. 157–161.
- Glaznieks, Aivars/Frey, Jennifer-Carmen (2018): Dialekt als Norm? Zum Sprachgebrauch Südtiroler Jugendlicher auf Facebook. In: Ziegler, Arne (Hg.): Jugendsprachen: Aktuelle Perspektiven internationaler Forschung/Youth Languages: Current perspectives of international research. Berlin/Boston, S. 859–889.
- Glaznieks, Aivars/Glück, Alexander (2019): From the valleys to the World Wide Web: Non-standard spellings on social network sites. In: Stemle, Egon W./Wigham, Ciara R. (Hg.): Building computer-mediated communication corpora for sociolinguistic analysis. (= Cahiers du LRL 8). Clermont-Ferrand, S. 19–43.