

Das KOMMA-Korpus und seine vielseitige Nutzbarkeit: Die deutsche Standardvarietät in Südtirol

Mara Maya Victoria Leonardi (Bozen/Bolzano)

Abstract

This paper outlines the construction of the KOMMA corpus and the presentation and analysis of selected case studies. The paper contains written and oral data from young adults attending the final year of a German-language high school in South Tyrol. In addition to describing the data collection and the methods used, the paper will also go into more detail about the data processing (transcription, normalisation, lemmatisation and POS tagging). Finally, some research results illustrate the usability of the corpus, with selected case studies from the field of word formation and lexis being presented.

1 Einleitung

Die deutsche Sprache ist äußerst vielfältig und variantenreich. Neben den unterschiedlichen Nonstandardvarietäten und regionalen Varietäten gibt es auch in der deutschen Standardsprache¹ regionale Unterschiede und Varianten (cf. Berend 2009). Diese regionalen Unterschiede betreffen nicht nur die Aussprache, sondern auch die Lexik, Grammatik, Syntax, Morphologie und Pragmatik (cf. Ammon et al. 2004; Ammon/Bickel/Lenz 2016). Anhand eines Korpus (siehe Abschnitt 3), das in Südtirol erhoben wurde, sollen im vorliegenden Beitrag der Gebrauch und die Frequenz mancher Varianten, die den Bereich der Wortbildung und Lexik betreffen, untersucht werden (siehe Abschnitt 4). Es handelt sich hierbei um sprachliches Material, das die Grundlage für unterschiedliche qualitativ und quantitativ orientierte linguistische Fragestellungen bildet. Neben lexikalischen und grammatikalischen Analysen des Deutschen in Südtirol, erlaubt das Korpus zudem Untersuchungen zu Sprachkontakt und -variation sowie zu Spracheinstellungen und -gebrauch (cf. Glück 2018; Glück/Leonardi 2019; Leonardi 2020a, 2020b, 2021, 2022).

Im folgenden Abschnitt 2 wird kurz die Situation in Südtirol skizziert. Abschnitt 3 beschreibt das KOMMA-Korpus, berichtet über die Datenaufbereitung und stellt die schriftlichen und mündlichen Sprachdaten vor. Anschließend werden in Abschnitt 4 einige ausgewählte Fallbeispiele und Ergebnisse präsentiert.

¹ Als „Standardsprache“ wird hier im Sinne von Elspaß (2005: 294) jene Sprachform verstanden, „[...] die in einer Sprachbevölkerung überregional und über alle gesellschaftlichen Schichten hinweg akzeptiert ist“. Für Südtirol findet sich eine Übersicht der einzelnen Sprachformen (regionaler Standard, Umgangssprache, Dialekte) in Glück/Leonardi/Riehl (2019: 256–259).

2 Südtirol

Neben dem Italienischen sowie dem Ladinischen stellt Deutsch in Südtirol, einer Autonomen Provinz im Nordosten Italiens, eine ko-offizielle Amtssprache dar. Seit 1961 wird bei der Volkszählung auch die Sprachgruppenzugehörigkeit (deutsch/italienisch/ladinisch) abgefragt.² Seit 1991 wiederum wird die Sprachgruppenzählung getrennt von der eigentlichen offiziellen Volks- und Wohnungszählung durchgeführt. Dadurch sind, außer dem Wohnsitz, keine weiteren Informationen (Geschlecht, Alter, Beruf etc.) zur Person bekannt. Zusätzlich zu den drei vorgesehenen Sprachgruppen (deutsch/italienisch/ladinisch) gibt es für all jene, die sich keiner oder nicht nur einer der drei Gruppen zugehörig erklären wollen, auch die Möglichkeit, „Andere“ anzugeben. Im Rahmen der letzten Befragung im Jahr 2011 war dies bei knapp 1,7% der gültigen Erklärungen der Fall (cf. ASTAT 2021: 15). Bei der letzten Sprachgruppenzugehörigkeitserhebung, die im Jahr 2011 im Rahmen der letzten Volkszählung durchgeführt wurde, erklärten sich 69,6% der deutschen, 25,8% der italienischen und 4,6% der ladinischen Sprachgruppe zugehörig (cf. ASTAT 2021: 15). Somit bildet die deutsche Sprachgruppe zwar auf nationaler Ebene eine Minderheit, innerhalb der Provinz jedoch die Mehrheit. Personen aus bilingualen Familien³ sowie Personen, die mit einer anderen Sprache als Deutsch, Italienisch oder Ladinisch sozialisiert worden sind, scheinen in dieser Statistik nicht auf. Innerhalb Südtirols sind die drei Sprachgruppen nicht gleichmäßig verteilt. Die italienische Sprachgruppe konzentriert sich überwiegend auf die Hauptstadt Bozen (73,8%), auf Leifers (71,5%), Branzoll (62,0%), Salurn (61,9%) und Meran (49,1%), während sich die ladinische Sprachgruppe v. a. auf Ortschaften im Gadertal und in Gröden⁴ beschränkt (cf. ASTAT 2021: 16–18).

Aufgrund der politischen und schulischen Situation in Südtirol ist der Großteil der deutschsprachigen Bevölkerung nicht nur durch eine innere, sondern auch durch eine äußere Mehrsprachigkeit gekennzeichnet. Die innere Mehrsprachigkeit bezeichnet dabei die Beherrschung und Verwendung verschiedener Varietäten innerhalb derselben Sprache, während das Konzept der äußeren Mehrsprachigkeit den Gebrauch und die Kenntnis verschiedener Einzelsprachen bezeichnet. Für den Südtiroler Kontext bedeutet dies, dass die meisten deutschsprachigen Sprecher:innen über mindestens zwei Varietäten der deutschen Sprache, die Standardsprache und einen südbairischen Dialekt, und mindestens zwei Einzelsprachen, das Deutsche und das Italienische, verfügen (cf. Glück/Leonardi/Riehl 2019: 254–261; Lanthaler 2012: 140–143, 149–151). Im Hinblick auf die Varietätenverwendung im familiären und privaten Bereich kann festgestellt werden, dass die deutschsprachige Bevölkerung überwiegend den Dialekt verwendet (cf. ASTAT 2015: 142–151; Leonardi 2020a). Zur wissenschaftlichen Erforschung der deut-

² Kontrolliert wird die Aussage über die Sprachgruppenzugehörigkeit in der Regel nicht. Die Ergebnisse der Erhebung haben allerdings konkrete Folgen, da sie die Grundlage für den ethnischen Proporz darstellen. Dies bedeutet, dass jede Sprachgruppe den entsprechenden Anteil an Stellen im öffentlichen Dienst erhält, wie etwa Gesundheitswesen, Landesverwaltung, Schule oder Gemeinden. Außerdem kann jemand nur gegen das Vorweisen der Erklärung eine öffentliche Stelle antreten oder gewisse Leistungen in Anspruch nehmen.

³ Zur soziolinguistischen Situation und zum Sprachgebrauch in bilingualen Familien in Südtirol, cf. Egger (1994), Leonardi (2020b), Veronesi (2009) und Weber Egli (1992).

⁴ Abtei (94,07%), Corvara (89,70%), Enneberg (92,09%), St. Martin in Thurn (96,71%) und Wengen (97,66%) gehören zu den mehrheitlich ladinischsprachigen Gemeinden des Gadertals. St. Christina in Gröden (91,40%), St. Ulrich (84,19%) und Wolkenstein in Gröden (89,74%) gehören zu den mehrheitlich ladinischsprachigen Gemeinden in Gröden (cf. ASTAT 2021: 16–18).

schen Varietäten in Südtirol haben Ciccolone/Franceschini (2015: 459) festgestellt, dass „[...] es nicht einfach sein könnte, von Basis- und Ortsdialekten über Ausgleichsformen bis hin zu standardnahe Sprechen und (regionalem) Standard das Varietätenrepertoire in Südtirol genau zu bestimmen“. Als Halbzentrum der deutschen Sprache verfügt Südtirol zudem über kein eigenes autorisiertes Nachschlagewerk (cf. Ammon/Bickel/Lenz 2016). Nichtsdestotrotz gibt es in Südtirol mittlerweile mehrere Untersuchungen als auch Nachschlagewerke, die die sprachlichen Besonderheiten der deutschen Varietäten in Südtirol untersuchen und dokumentieren, insbesondere was die Lexik anbelangt (cf. Abel 2018; Abfalterer 2007; Ammon et al. 2004; Ammon/Bickel/Lenz 2016). Sprachliche Charakteristika der deutschen Standardsprache, wie etwa phonetische, morphosyntaktische und lexikalische Merkmale, die sie von anderen deutschen Standardvarietäten unterscheiden, finden sich in Glück/Leonardi/Riehl (2019: 259).

Neben der individuellen Mehrsprachigkeit gibt es in Südtirol ein nach Sprachen getrenntes Schulsystem. An den deutsch- bzw. italienischsprachigen Schulen wird Deutsch bzw. Italienisch als Unterrichtssprache verwendet, wobei die jeweils andere Landessprache als Zweitsprache unterrichtet wird. In den paritätischen Schulen mit ladinischer Unterrichtssprache werden alle drei Sprachen, also Deutsch, Italienisch und Ladinisch, verwendet (cf. Glück/Leonardi/Riehl 2019: 251f.).

3 Korpus

Eine ganze Reihe deutschsprachiger Korpora basieren auf der Schriftsprache, wie etwa das ZDL-Regionalkorpus, das Artikel aus Lokal- und Regionalressorts deutscher Tageszeitungen enthält (cf. Nolda/Barbaresi/Geyken 2022) oder das Deutsche Referenzkorpus DeReKo (cf. Kupietz/Lüngen/Diewald 2022), das am Leibniz-Institut für Deutsche Sprache (IDS) aufgebaut wird. Das Projektkorpus *Varietäntengrammatik* besteht aus Zeitungsartikeln verschiedener Regionen des deutschsprachigen Raums, mit dem Ziel, areale grammatische Variation im Standarddeutschen zu ermitteln und zu dokumentieren (cf. Elspaß 2022). In Südtirol dokumentiert das Korpus *Südtirol*, das inhaltlich in vier Kategorien eingeteilt werden kann (Belletristik, Sachtexte, journalistische Prosa und Gebrauchstexte) die geschriebene deutsche Standardsprache in Südtirol (cf. Abel/Anstein/Petrakis 2009; Anstein/Oberhammer/Petrakis 2011). Es gibt aber nicht nur Korpora der geschriebenen Sprache, sondern auch der gesprochenen. Am IDS wurde das Forschungs- und Lehrkorpus *Gesprochenes Deutsch (FOLK)* aufgebaut. Ziel des FOLK-Korpus ist die Untersuchung der gesprochenen Sprache in „natürlicher“ Interaktion (cf. Reineke/Deppermann/Schmidt 2022).

3.1 Untersuchungsgebiet und Datenerhebung

Im Folgenden werden zunächst die wichtigsten Eckpunkte des Projekts KOMMA („Sprachkompetenzen von Maturantinnen und Maturanten: Schulsprache Deutsch und Kontaktphänomene im mehrsprachigen Kontext“) dargestellt, das aus einem schriftlichen und mündlichen Datensatz besteht.⁵ Die Erhebungsorte umfassen das gesamte Südtirol: Schlanders im Vinschgau, Meran im Burggrafenamt, die Landeshauptstadt Bozen, Brixen im Eisacktal und Bruneck im Pustertal. An der Erhebung, die zwischen 2010 und 2015 stattgefunden hat, nahmen

⁵ Projektleitung: Rita Franceschini.

insgesamt 13 deutschsprachige Oberschulen teil: drei in Schlanders, drei in Meran, drei in Bozen, drei in Brixen und eine in Bruneck. Zum Zeitpunkt der Erhebung befanden sich die Schüler:innen im fünften – und somit im letzten – Oberschuljahr, dem sogenannten „Maturajahr“.⁶

Bei den schriftlichen Texten handelt es sich um authentische Aufsätze, die im Schulkontext in verschiedenen Schuljahren (2009/2010, 2010/2011, 2011/2012, 2012/2013 und 2014/2015) produziert und von Seiten der Schulen dem Projekt anonym⁷ zur Verfügung gestellt wurden (cf. Glück/Leonardi 2019: 448–451).

Die mündlichen Datenerhebungen fanden in den jeweiligen Schulgebäuden der Maturant:innen statt. Die Auswahl der Proband:innen erfolgte von Seiten der Schule. Vorgabe für die Schulen war eine zufällige Auswahl von Schüler:innen, die Deutsch als Erstsprache sprechen, wobei allerdings im Informationsblatt hinzugefügt wurde, dass eine weitere Erstsprache kein Problem darstellen würde. Vor dem Interview wurden einige soziodemographische Grunddaten über einen schriftlichen Fragebogen, den die Proband:innen selbst ausfüllen mussten, eingeholt (Abbildung 1). Anschließend wurden alle Interviews auf Standarddeutsch geführt. An der mündlichen Datenerhebung nahmen insgesamt sieben Interviewer:innen teil, die – mit Ausnahme von einer Interviewerin, Sprecherin der lokalen Standardvarietät, – alle aus Deutschland stammten.

- Name
 - Geburtsjahr
 - Deine Muttersprache
 - Muttersprache Mutter
 - Muttersprache Vater
 - Welche Sprachen sprichst Du?
 - Mit wie vielen Deiner Freunde sprichst Du Deutsch, mit wie vielen Italienisch (ca. in Prozent)?

Abbildung 1: Schriftlicher Erhebungsbogen

Der mündliche Datensatz umfasst Sprachbiographien, Nacherzählungen, Erklärungen und Fokusgruppendifkussionen, die im Folgenden genauer erläutert werden.

Sprachbiographie

Als Grundlage für das Gespräch diente ein Interviewleitfaden (Abbildung 2), wodurch zum einen besser vergleichbare Daten vorliegen. Zum anderen konnten somit alle forschungsrelevanten Themenkomplexe in die Erzählungen der Schüler:innen inkludiert werden. Es handelt sich hierbei um subjektive Sprachdaten, die einen Einblick in die Sprachgeschichte des Erzählers geben. Die Fragen des leitfadengesteuerten Interviews lassen sich grob in vier Themenbereiche gruppieren:

⁶ Auf die fünfjährige Grundschule folgt die dreijährige Mittelschule. Die Oberschule umfasst anschließend weitere fünf Jahre.

⁷ Da die meisten schriftlichen Texte anonym zur Verfügung gestellt wurden, kann zum größten Teil des schriftlichen Datensatzes keine Auskunft über den Produzenten gemacht werden. Allerdings gibt es einige Ausnahmen, siehe Abschnitt 3.2.

- soziodemographische Grunddaten,
- Angaben zu den Sprachkenntnissen,
- Einstellungen gegenüber Mehrsprachigkeit,
- Sprachkompetenzen und literale Aspekte (Schreib- und Lesegewohnheiten).

Bei den soziodemographischen Grunddaten wurde nach der Sprachverwendung im familiären Umfeld und im Alltag (mit Freunden, Nachbarn) gefragt. Die Fragen im zweiten Themenbereich beziehen sich auf den Spracherwerb und die Sprachverwendung im schulischen Kontext sowie im Alltag. Im dritten Bereich wird nach den Einstellungen gegenüber Sprachen im Allgemeinen sowie gegenüber der Mehrsprachigkeit in Südtirol gefragt. Der vierte Themenbereich bezieht sich auf die Kompetenzeinschätzungen der Proband:innen in ihren Sprachen und ihr individuelles Schreib- und Leseverhalten.

Einstiegsfragen
<ul style="list-style-type: none"> - Erzähl doch mal, wie Du zu Deinen Sprachen gekommen bist bzw. wie Du sprachlich aufgewachsen bist.
<p>Familie</p> <ul style="list-style-type: none"> - Wo sind Dein Vater, Deine Mutter und Deine Großeltern aufgewachsen? - Welche Sprachen wurden bei Euch zu Hause gesprochen? - Wie sprechen die Eltern untereinander? Wie sprechen Deine Geschwister untereinander? - Beherrschen Deine Geschwister die Sprachen gleich gut wie Du? - Gab/Gibt es eine Sprache, die Du zur Besprechung z. B. eines Problems bevorzugst?
<p>Nachbarn/unmittelbares Umfeld</p> <ul style="list-style-type: none"> - Hattest Du in der Kindheit Freunde, die mehrsprachig waren? - Gab es sonst noch Sprachen, die in Deiner Nachbarschaft gesprochen wurden?
<p>Schule</p> <ul style="list-style-type: none"> - Welche Sprachen hast Du in der Schule gelernt und wann? - Warum hast Du Dich für das Erlernen einer bestimmten Sprache entschieden? - Gibt es außerdem noch Sprachen, die Du gerne lernen würdest? - Bist Du gerne in den Sprachunterricht gegangen? Was hat Dir Freude oder Mühe bereitet? - Was war die Unterrichtssprache? Wie war das Gefühl, wenn Du Standarddeutsch gesprochen hast? - Welche Stellung hatten Kinder aus anderen Ländern in der Schule?
<p>Sprachspezifische Fragen</p> <ul style="list-style-type: none"> - Hast Du einen bestimmten Akzent in einer Sprache? - Gibt es Situationen, in denen Du eine Sprache besser sprichst als die anderen? - Früheste Erinnerungen in einer Sprache (z. B. erstes Wort). - Wie lernst Du eine Sprache am liebsten? - Hat sich nach dem Erlernen irgendeiner Sprache Deine Haltung ihr gegenüber geändert? - Ist das Erlernen von Sprachen für Dich wichtig? - Ist Mehrsprachigkeit für Dich wichtig? - Schreibst Du viel/gerne? Liest Du viel/gerne?
<p>Heutige Kompetenzen</p> <ul style="list-style-type: none"> - Wie schätzt Du heute Deine Kompetenzen in den von Dir gelernten Sprachen ein? - Wie bewusst sind Dir beim Sprechen in den einzelnen Sprachen die grammatikalischen Regeln?

Abbildung 2: Leitfaden für das biographische Interview

Nacherzählung

Den Proband:innen wurde ein Ausschnitt aus dem Stummfilm *The Circus*⁸ von Charlie Chaplin gezeigt. Der Ausschnitt wurde gemeinsam mit den anderen Proband:innen und dem/der Forscher:in in einem ruhigen Raum der Schule unmittelbar vor der Datenerhebung angeschaut und sollte anschließend nacherzählt werden. Die Aufgabenstellung lautete: „Könntest Du mir bitte erzählen, worum es in diesem Filmausschnitt gegangen ist?“.

Erklärung

Die Erklärungsaufgabe bestand darin, das Spiel „Baseball“ auf der Grundlage eines im Vorfeld zugesandten schriftlichen Regelwerks zu erklären. Die Proband:innen sollten sich das Regelwerk alleine zu Hause durchlesen und bei der Datenerhebung folgende Frage beantworten: „Könntest Du mir erklären, was Baseball ist, wie man das spielt?“.

Fokusgruppendifkussion

Zwischen drei und fünf Proband:innen diskutierten über das Thema „Soziale Medien und deren Auswirkungen“. Auch hierzu erhielten die Proband:innen im Vorfeld Materialien zugeschickt, die sie zu Hause lesen sollten. Der/die Interviewer:in, der/die lediglich als Moderator:in funktionierte, leitete die Diskussion ein und regte sie mit Zwischenfragen oder kurzen Zusammenfassungen immer wieder an. Hierbei lautete die Fragestellung: „Ist es richtig, dass uns das reale soziale Leben zunehmend fremd wird, dass wir immer mehr virtuell leben? Kann der Mensch eigentlich dann nur noch in der virtuellen Welt glücklich sein?“.

3.2 Gewährspersonen

Zum Zeitpunkt der Erhebung befanden sich die Schüler:innen im fünften – und somit im letzten – Oberschuljahr, dem sogenannten „Maturajahr“. Das Durchschnittsalter der 65 Proband:innen betrug 19,21 Jahre, das Altersspektrum erstreckte sich von 19 bis 22 Jahren. Insgesamt nahmen 41 weibliche und 24 männliche Schüler:innen an der Erhebung teil. Dabei haben insgesamt 17 Proband:innen in Meran, 16 in Schlanders, jeweils 13 in Bozen und Brixen und sechs in Bruneck teilgenommen. Es ist anzumerken, dass der Schulort nicht zwingend mit dem Wohnort übereinstimmen muss, da einige Schüler:innen aus Dörfern des ländlichen Umlandes kommen. 39 besuchten ein Gymnasium, 18 eine Fachoberschule und vier eine Waldorfschule. Von vier Proband:innen liegen keine Angaben zum Schultyp vor (Tabelle 1).

⁸ In der Handlung wird ein Tramp, gespielt von Charlie Chaplin, fälschlicherweise des Taschendiebstahls verdächtigt. Er klärt das Missverständnis allerdings nicht auf, sondern flieht vor der Polizei, die von Passanten gerufen wurde. Unfreiwillig stolpert er bei seiner Flucht in eine Zirkusvorstellung. Auch andere Korpora bestehen aus Nacherzählungen von Stummfilmen. So basiert das Hamburg Modern Times Corpus auf Audioaufnahmen von Nacherzählungen eines Ausschnitts aus dem Film *Modern Times* von Charlie Chaplin. Allerdings wurden hierbei Erwachsene L2-Sprecher:innen des Deutschen mit unterschiedlichen L1 und unterschiedlichen Kompetenzen im Deutschen aufgenommen.

Mündlicher Datensatz			
Schulort	Erhebungsjahr	Proband:innen	Schultyp und Proband:innen
Bozen	2013	n=4	Waldorfschule (n=4)
	2015	n=9	Gymnasium (n=9)
Brixen	2012	n=4	Gymnasium (n=4)
	2015	n=9	Fachoberschule (n=5), keine Angabe (n=4)
Bruneck	2015	n=6	Gymnasium (n=6)
Meran	2012	n=17	Fachoberschule (n=9), Gymnasium (n=8)
Schlanders	2012	n=16	Fachoberschule (n=4), Gymnasium (n=12)

Tabelle 1: Mündliche Daten nach Erhebungsort, Erhebungsjahr, Proband:innenanzahl und Schultyp

Erstsprache(n)

Basierend auf dem Fragebogen (Abbildung 1), der von den Schüler:innen vor den mündlichen Interviews ausgefüllt wurde, gab die Mehrheit der Befragten (n=61) Deutsch⁹ als alleinige Erstsprache (L1) an, zwei der Befragten nannten Italienisch als L1 und zwei Schülerinnen gaben Deutsch/Italienisch bzw. Deutsch/Spanisch gleichermaßen als ihre Erstsprachen an. Die Angaben zu den L1 der Mutter und des Vaters stammen ebenfalls von den Proband:innen. Dabei nannten 54 Schüler:innen Deutsch als L1 beider Elternteile, während elf Schüler:innen zur Erstsprache der Eltern unterschiedliche Angaben machten (Tabelle 2).

Mündlicher Datensatz			
Proband:innen	Erstsprache(n) – Fragebogenangaben		
	L1 Proband:in	L1 Mutter	L1 Vater
n=54	Deutsch	Deutsch	Deutsch
n=3	Deutsch	Deutsch	Italienisch
n=2	Deutsch	Italienisch	Deutsch
n=1	Deutsch	Ladinisch	Deutsch
n=1	Deutsch	Deutsch	Deutsch/Italienisch
n=1	Italienisch	Deutsch	Italienisch
n=1	Italienisch	Italienisch	Deutsch
n=1	Deutsch/Italienisch	Italienisch	Deutsch
n=1	Deutsch/Spanisch	Spanisch	Deutsch/Italienisch

Tabelle 2: Erstsprache(n) der Proband:innen und der Eltern (Fragebogenangaben)

⁹ Im Fragebogen haben die Schüler:innen nicht zwischen den verschiedenen Varietäten des Deutschen unterschieden. In den Interviews wurde der Begriff *Deutsch* mit folgenden Ausdrücken spezifiziert: Südtiroler Dialekt, Österreichischer Dialekt, Österreichisches Standarddeutsch, Bayerischer Dialekt, Saarländischer Dialekt, Standarddeutsch oder Hochdeutsch.

In den sprachbiographischen Interviews hat sich anschließend gezeigt, dass insgesamt 13 der 65 Oberschüler:innen mehr als eine Sprache im familiären Kontext sprechen bzw. hören (cf. Leonardi 2020b: 172f.).

Sprachkenntnisse

Zusätzlich zur L1 wurde im Fragebogen auch nach den Sprachkenntnissen der Schüler:innen gefragt. Hierbei wurden neben Deutsch, Italienisch und Englisch folgende Sprachen aufgelistet: Französisch (von 13 Proband:innen), Latein (von fünf Proband:innen), Spanisch, Ladinisch und Griechisch (von jeweils einem Probanden/einer Probandin).

Sprachgebrauch mit Freunden

Was den Sprachgebrauch mit Freunden betrifft, so kann festgestellt werden, dass 56 der Schüler:innen im Umgang mit Freunden angaben, in 70% oder mehr Begegnungen ausschließlich Deutsch zu verwenden. Lediglich neun Schüler:innen gaben an, dass sie in mindestens 40% der Interaktionen mit Freunden Italienisch sprechen.

3.3 Datenaufbereitung und Annotation

Das kostenlose Transkriptionsprogramm ELAN¹⁰ ermöglicht die Transkription von Audio- und Videodateien, die anschließend bearbeitet, annotiert und durchsucht werden können. Wie wir in Abschnitt 4 sehen werden, ermöglicht eine einheitliche Datenaufbereitung die Untersuchung und den Vergleich von konkreten Phänomenen im geschriebenen und gesprochenen Südtiroler Standarddeutsch.

Für das POS-Tagging und die Lemmatisierung wurden der *TreeTagger* (cf. Schmid 1995) und das Stuttgart-Tübingen-TagSet (*STTS 2.0*) verwendet (cf. Schiller et al. 1999; Westpfahl 2014; Westpfahl/Schmidt 2013; Westpfahl et al. 2017). Im Folgenden sollen zuerst die Aufarbeitung und Annotation der schriftlichen Daten (3.3.1) und anschließend der mündlichen Daten (3.3.2) vorgestellt werden.

3.3.1 Schriftlicher Datensatz

3.3.1.1 Aufbereitung

Die schriftlichen Daten wurden zeichengetreu transliteriert, wobei Eigenkorrekturen (z. B. Streichungen, Einfügungen) der Proband:innen selbst in die Transkripte übernommen wurden. Durch Kopierverlust entstandene Lücken wurden teilweise von den Transkribent:innen rekonstruiert, wobei Fälle, in denen dies nicht möglich war, als unleserlich markiert wurden (Tabelle 3).¹¹

¹⁰ Das ELAN-Programm (ELAN) erlaubt die Erstellung verschiedener Tiers, die ein Part-of-Speech-Tagging (POS-Tagging) ermöglichen, sowie, bei Bedarf, auch das Einfügen einer Kommentarzeile, cf. auch Crasborn/Sloetjes (2008).

¹¹ Die einzelnen Belege werden aus einer Kürzelkombination aus Textart (S = schriftlicher Text) und Proband:innen-ID nachgewiesen.

Schriftlicher Datensatz		
Tags	Beschreibung	TRSCR-Tier (Transkriptionstier)
=/ /=	Selbststreichung (durch den Schüler/die Schülerin)	Es kam zur Verfolgung der Juden und all der Menschen, die laut Nazis =/, /= nicht der Norm entsprachen oder die gegen Nazis arbeiteten. [S06a] Könnte diese =/r/= Szene zur Realität unserer Enkel werden? [S04a]
< >	Selbsteinfügung (durch den Schüler/die Schülerin)	Diese mussten in den Schlachten vorne die Kanonen bedienen <weil sie> wegen ihrer Größe dazu geeignet waren. [S59a]
{ }	Rekonstruktion (durch den Transkribenten)	35 Prozent geben an, häufig {P}atienten wegen fehlender Kapazitäten {a}bzulehnen. [S53b]
@@@	Unleserliches	Mit anderen Worten, oft =/@@@/= ist Reden Silber und Schweigen Gold. [S22a]

Tabelle 3: Konventionen für die Darstellung schriftlicher Daten

Nach der Transliteration wurden die schriftlichen Daten in ELAN importiert. Anschließend erfolgte auf dem Tier der Normalisierung (NORM) die automatische Tokenisierung (TOK, Wort-für-Wort-Unterteilung). Die einzelnen Tokens wurden wiederum automatisch in Lemmata (LEM) und Wortarten (POS, Part-of-Speech-Tagging) klassifiziert. Die geschwungene Klammer der rekonstruierbaren Lücken kommt ausschließlich im Transkriptionstier (TRSCR) vor (Abbildung 3).

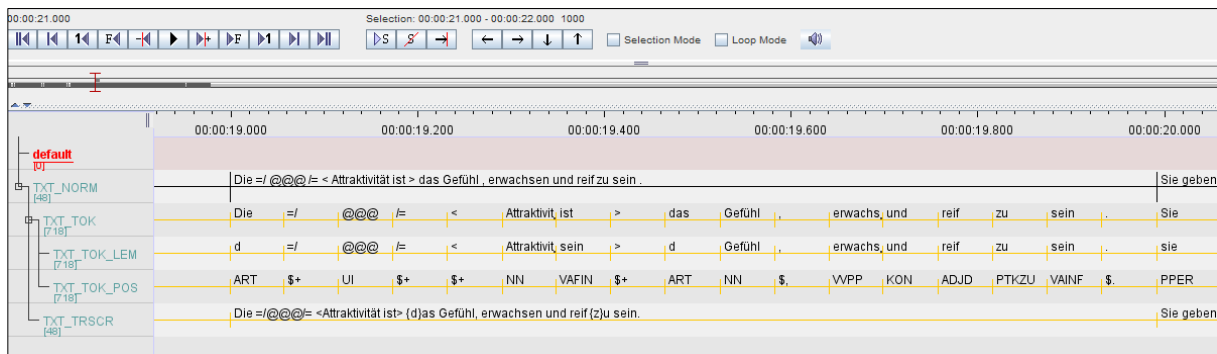


Abbildung 3: Transkription schriftlicher Daten in ELAN

3.3.1.2 Umfang

Der schriftliche Datensatz enthält insgesamt 64 Tags, von denen 55 reine POS-Tags sind und zusätzliche neun Tags entweder für Satzzeichen, unleserliche Äußerungen (UI), fremdsprachliches Material (FM), Buchstabiertes oder Buchstabe (SPELL), Abbrüche auf Wortebene (AB) oder Textlücken aufgrund von Kopierverlust (PAUSE) verwendet wurden. Da in Schiller et al. (1999) und Westpfahl et al. (2017) eine Beschreibung der einzelnen POS-Tags vorliegt, soll hier nicht näher darauf eingegangen werden.¹² Tabelle 4 beinhaltet eine Beschreibung der neun Sonder-Tags sowie Beispielbelege aus den Schülertexten.

Schriftlicher Datensatz			
POS-Tag	Beschreibung	TOK-Tier	TRSCR-Tier (Transkriptionstier)
\$.	Satzbeendende Interpunktion (durch den Schüler/die Schülerin)	. ? ! ; :	Oder doch etwas, wo ich anderen helfen kann? [S02a] Kinder brauchen Helden: [S03b]
\$,	Beistrich (durch den Schüler/die Schülerin)	,	Dabei sind sehr viele Leute, vor allem auch Zivilisten, ums Leben gekommen. [S06a]
\$(sonstige Satzzeichen (durch den Schüler/die Schülerin)	[] () / – “ ” ...	Von außen schaut das Szenario gar nicht bedrohlich aus, doch die meisten Zuschauer, die den “Tiefflieger” beobachten wissen nicht, dass es hunderten von Menschen das Leben kosten könnte. [S01b]
\$+	sonstige Satzzeichen (durch den Transkribenten)	< > = / !=	US-Filmmacher haben neulich eine kurze Szene gedreht, mit der sie <bei> den Leuten die Frage aufwerfen wollten, ob es wirklich erstrebenswert sei so sein zu wollen wie der lustige, übergewichtige Glatzkopf. [S04b] Solche Antihelden wären auch heute nicht schlecht, denn meines Erachtens =/ver/= würde so der Gesellschaftsdruck verschwinden. [S05b]
UI	Unleserliche Äußerung (durch den Transkribenten)	@@@	Columbus gelang es auch nicht alleine =/@@@/= Amerika zu entdecken. [S17b]
FM	Fremdsprachliches Material		“ Che emozioni ” würde der Italiener sagen. [S01a]
SPELL	Buchstabiertes oder Buchstabe		(A. Schnitzler in „Traumnovelle“). [S06b]
AB	Abbruch auf Wortebene		Trotzdem gibt es heute noch viel zu viele Menschen, die ihre Wert- und Normvorstellungen =/and/= anhand der Gesellschaft definieren. [S07a]
PAUSE	[[Textlücke]]		Gentechnik bei den Menschen ist ein [[Textlücke]] [S59b]

Tabelle 4: Konventionen für die Darstellung schriftlicher Daten

Das schriftliche Subkorpus umfasst insgesamt 430 schriftliche Texte. Bei den Texten handelt es sich um 196 Maturaaufsätze (45,6%), 42 Maturasimulationen (9,8%), 185 Schularbeiten (43,0%) und sechs Hausaufgaben (1,4%) sowie einen Text ohne genauere Angabe (0,2%). Es handelt sich hierbei um authentische Aufsätze, die im Schulkontext – entweder im regulären

¹² Da im Hinblick auf die geschriebene Sprache die Kategorisierung im STTS (cf. Schiller et al. 1999) nicht ausreichend war, wurde eine weitere Kategorie hinzugefügt (PAUSE) und der Tag der „sonstigen Satzzeichen“ (\$+) übernimmt zusätzliche Funktionen, wie etwa Streichung und Einfügung.

Schulunterricht bzw. während der Maturaprüfung (Maturaaufsätze, Maturasimulationen, Schularbeiten) oder zu Hause (Hausaufgaben) – in verschiedenen Schuljahren produziert wurden (cf. Glück/Leonardi 2019: 448–451). Da der größte Anteil der Texte (n=352) von Seiten der Schulen anonym und in Fotokopie zur Verfügung gestellt wurde, gibt es hierfür keine Informationen zum Produzenten. Von 40 Proband:innen, die ebenso an der mündlichen Datenerhebung teilgenommen haben, wodurch zusätzliche Metadaten vorliegen (siehe Abschnitt 3.2), gibt es allerdings mindestens je einen Text, der dem Projekt in diesem Fall ausnahmsweise nicht anonymisiert zur Verfügung gestellt wurde. Diese 78 Texte mit einem Umfang von 42.532 Tokens¹³ sollen im Folgenden genauer beschrieben werden. Tabelle 5 bietet einen Überblick der schriftlichen Daten nach Schulort, -jahr, Textanzahl und Produktionssituation.

Schriftlicher Datensatz			
Schulort	Schuljahr	Textanzahl	Produktionssituation und Textanzahl
Bozen	2014/2015	n=18	Hausaufgabe (n=5), Maturasimulation (n=5), Schularbeit (n=8)
Brixen	2011/2012	n=8	Maturasimulation (n=2), Schularbeit (n=6)
	2014/2015	n=10	Maturasimulation (n=4), Schularbeit (n=5), keine Angabe (n=1)
Bruneck	2014/2015	n=2	Maturasimulation (n=1), Schularbeit (n=1)
Meran	2011/2012	n=8	Schularbeit (n=8)
Schlanders	2011/2012	n=32	Maturasimulation (n=4), Schularbeit (n=28)

Tabelle 5: Schriftliche Daten nach Erhebungsort, Schuljahr, Textanzahl und Produktionssituation

Bei 15 der 78 Texte handelt es sich um „Essays“, 12 Texte sind „Zeitungsartikel“ und sieben Texte „Kommentare“. Auf die anderen Texttypen („Textanalyse“, „Erörterung“, „Bericht“, „Zeitungsbericht“) entfallen jeweils weniger als fünf Schülertexte.¹⁴ Aufgrund von Kopierverlusten kann zur Mehrheit der Schülertexte (n=32) keine Angabe gemacht werden.

3.3.2 Mündlicher Datensatz

3.3.2.1 Aufbereitung

Für gesprächsanalytische Zwecke wurden die Daten in ELAN standard-orthographisch transkribiert, wobei keine phonetische Transkription vorgesehen war. Im Unterschied zum Transkriptionstier (TRSCR), die eine Kleinschreibung vorsieht, erfolgte bei der Normalisierung (NORM) der mündlichen Daten eine standardorthographische Normalisierung nach der Dudenrechtschreibung. Im Rahmen der Normalisierung wurden keine „Korrekturen“ von grammatischen Fehlern vorgenommen. Das heißt, dass in Bezug auf Kasus, Numerus, Genus etc. nicht normalisiert bzw. korrigiert wurde. Bei der Beurteilung wurde das Duden-Onlinewörterbuch konsultiert. Grammatische Fehler, also Normabweichung, wie etwa Kasusabweichungen (z. B. „oft begegnen **einen** ja auch **touristen**“ [B05]; „aber arbeiten gehen darf man erst nach neun **jahre**“ [B05]; „in **manche** geschäften“ [B23]) oder Verbstämme (z. B. „und und **wel** wird und geht dann wieder rüber und **fangt** an zu rennen und immer so weiter“ [E13]), wurden

¹³ Bei dieser Berechnung wurden die Satzzeichen nicht miteinberechnet.

¹⁴ Die Angaben stammen von den Schüler:innen selbst.

nicht korrigiert bzw. verändert. Alle Daten wurden wiederum automatisch tokenisiert, lemmatisiert und POS-getaggt (Abbildung 4). Diese Gliederung erfolgte bei jedem einzelnen Sprecher bzw. bei jeder einzelnen Sprecherin.

Track	Content
default	[317]
EXP_01_NORM	[317]
EXP_01_TOK	[2106]
EXP_01_TOK_LEM	[2106]
EXP_01_TOK_POS	[2106]
EXP_01_TRSCR	[317]
SBJ_01_NORM	[359]
SBJ_01_TOK	[3514]
SBJ_01_TOK_LEM	[3514]
SBJ_01_TOK_POS	[3514]
SBJ_01_TRSCR	[359]

Abbildung 4: Transkription mündlicher Daten in ELAN

Der mündliche Datensatz enthält insgesamt 61 Tags, von denen 53 reine POS-Tags sind und zusätzliche acht Tags entweder für Pausen (PAUSE), nonverbale Äußerungen (PARA), Abbrüche (AB), Unverständliches (UI), fremdsprachliches Material (FM), Buchstabiertes (SPELL) oder Nichtwörter¹⁵ (XY) verwendet werden (Tabelle 6). Im Rahmen der Kontrolle der Transkriptionen erfolgte auch die Anonymisierung der Transkripte. Um keine Rückschlüsse auf die Identität einer Person zu ermöglichen, wurden z. B. Personennamen oder Ortsnamen anonymisiert. Hierfür wurden folgende Arten von Siglen angelegt:

- Siglen für die Gewährspersonen (SBJ) im Korpus. Dabei kennzeichnet eine Zahl zwischen 01 und 65 den Sprecher bzw. die Sprecherin;
- Siglen für den Interviewer bzw. die Interviewerin (EXP), der/die die Interviews geführt hat. Als Identifikation wird eine Zahl zwischen 01 und 07 verwendet;
- Siglen für Personennamen (NAME);
- Siglen für Schulen oder Einrichtungen (SCHOOL, INSTITUTION);
- Siglen für Berufe oder Arbeitsstätten (PROFESSION);
- Siglen für Ortsnamen/Namen von Flüssen/Tälern (PLACE);
- Siglen für Sammelbezeichnung von Einwohnern einer Stadt/eines Dorfes (RESIDENTS).

Neben der Anonymisierung in den Transkripten wurden auch die Audio-Dateien an den zu anonymisierenden Stellen akustisch verfremdet.

¹⁵ Versprecher wurden in der Art verschriftet, wie sie geäußert wurden, und wurden anschließend als sog. „Nichtwörter“ getaggt.

Wie bereits erwähnt, handelt es sich bei den mündlichen Daten um die standardsprachliche Varietät des Deutschen in Südtirol, weshalb die verwendete Varietät nicht eigens in einem Tier markiert wurde. In den einzelnen Fällen, in denen eine andere Varietät/Sprache (z. B. Italienisch, Englisch oder Französisch) verwendet wurde, wurde dies auf dem POS-Tier markiert (FM). Im Gegensatz zu den gesprochenen lokalen Varietäten in weniger formellen Kontexten, wie etwa im Familien- oder Freundeskreis (cf. Dal Negro/Ciccolone 2018), lassen sich im gesprochenen Standarddeutsch auf lexikalischer Ebene nur sehr wenige Einflüsse aus anderen Sprachen beobachten (Abschnitt 4.2). Insgesamt wurden im KOMMA-Korpus 211 Tokens als FM markiert: davon verwenden 86,26% (n=182) die Proband:innen und 13,74% (n=29) die Interviewer:innen. Für die Darstellung von Parasprachlichem und Unverständlichem wurde teilweise auf die Konventionen von GAT 2 zurückgegriffen (cf. Selting et al. 2009). Die einzelnen Belege werden aus einer Kürzelkombination aus Textart (B=Sprachbiographie, D=Diskussion, N=Nacherzählung, E=Erklärung), Sprecher:in (SBJ=Proband:in, EXP=Interviewer:in) und Produzenten-ID nachgewiesen (Tabelle 6).

Mündlicher Datensatz		
Tags	Beschreibung	TRSCR-Tier (Transkriptionstier)
PAUSE	kurze (.) bzw. lange (-) Pause	SBJ_29: wenn der ball gleich sofort gefangen wird (.) bevor man [E29]
PARA	Parasprachliche/Nonverbale Handlung	SBJ_52: oder solche sachen aber halt ((lacht)) [B52]
AB	Wortabbruch	SBJ_05: eine schwes / eine ältere schwester und einen jüngeren bruder [B05]
UI	Unverständliche Passage	SBJ_41: und die handschuhe im dialekt () waren [B41]
FM	Fremdsprachliches Material	SBJ_22: ja waren halt io mi chiamo halt wie ich heiße wie alt ich bin das waren so die ersten wörter [B22]
SPELL	Buchstabiertes	SBJ_20: dass viele zu mir sagen dass ich das italienische q sag quando [B20]
XY	Nichtwörter	SBJ_23: das interget das internet vergisst nicht [D23_D24_D25]

Tabelle 6: Konventionen für die Darstellung mündlicher Daten

3.3.2.2 Umfang

Die Interviews wurden eigens für das Projekt KOMMA erhoben. Bei den Interviews handelt es sich also um elizitierte Gespräche in der deutschen Standardsprache. Die Daten wurden in den Jahren 2012, 2013 und 2015 von sieben Explorator:innen während der regulären Unterrichtszeit an den gleichen Oberschulen wie die schriftlichen Daten erhoben. Abgesehen von den Fokusgruppendifkussionen und in einzelnen Fällen, in denen es organisatorisch nicht anders möglich war, fanden die Gespräche einzeln zwischen einem Probanden bzw. einer Probandin und dem/der Interviewer:in im Standard statt. Tabelle 7 bietet einen Überblick der mündlichen Daten nach Erhebungsort, -jahr, Proband:innenanzahl und Gesprächstyp.

Mündlicher Datensatz			
Schulort	Erhebungsjahr	Proband:innen	Gesprächstyp und Proband:innen
Bozen	2013	n=4	Sprachbiographie (n=4), Erklärung (n=4), Nacherzählung (n=3), Diskussion (n=4)
	2015	n=9	Sprachbiographie (n=9), Erklärung (n=9), Nacherzählung (n=9), Diskussion (n=9)
Brixen	2012	n=4	Sprachbiographie (n=4), Erklärung (n=4), Nacherzählung (n=4), Diskussion (n=4)
	2015	n=9	Sprachbiographie (n=9), Erklärung (n=8), Nacherzählung (n=9), Diskussion (n=8)
Bruneck	2015	n=6	Sprachbiographie (n=6), Erklärung (n=6), Nacherzählung (n=6), Diskussion (n=3)
Meran	2012	n=17	Sprachbiographie (n=17), Erklärung (n=17), Nacherzählung (n=17), Diskussion (n=17)
Schlanders	2012	n=16	Sprachbiographie (n=16), Erklärung (n=16), Nacherzählung (n=16), Diskussion (n=16)

Tabelle 7: Mündliche Daten nach Erhebungsort, Erhebungsjahr, Proband:innenanzahl und Gesprächstyp

Die mündliche Datensammlung umfasst knapp 43 Stunden Audioaufnahmen und hat einen Umfang von 431.963 Tokens (Tabelle 8).

Mündlicher Datensatz				
Gesprächstyp	h:min:sek	Tokens	Aufnahmen	Proband:innen
Sprachbiographie	25:15:30 (Ø 24 min)	260.822 (60,38%)	n=63	n=65
Nacherzählung	04:35:20 (Ø 4 min)	42.911 (9,93%)	n=64	n=64
Erklärung	06:17:50 (Ø 6 min)	61.196 (14,17%)	n=64	n=64
Fokusgruppendifkussion	06:47:01 (Ø 27 min)	67.034 (15,52%)	n=15	n=61
Gesamt	42:55:41	431.963 (100%)	n=206	n=65¹⁶

Tabelle 8: Zusammenstellung des mündlichen Subkorpus

Wie aus Tabelle 8 ersichtlich ist, machen mit 60% der Tokens und über 25 Stunden Audioaufnahmen die Sprachbiographien den größten Anteil des mündlichen Subkorpus aus. Fokusgruppendifkussionen und Erklärungen sind mit 16% bzw. 14% vertreten, Nacherzählungen mit knapp 10%. Da fast alle Gewährspersonen an den verschiedenen Erhebungstypen teilgenommen haben, sind intra-individuelle Vergleiche möglich. Vervollständigt werden die Daten aus den sprachbiographischen Interviews durch die Abfrage von einigen Metadaten, wie etwa Erstsprache, Sprachkenntnisse und Sprachverwendung mit Freunden (Abschnitt 3.2).

¹⁶ Demnach ergibt sich die Gesamtzahl der Proband:innen in Tabelle 8 nicht aus der Summe an Proband:innen in den einzelnen Gesprächstypen.

4 Diskussion ausgewählte Analyseergebnisse

Einige Teilergebnisse zu Einzelphänomenen wurden bereits in einer Reihe von Aufsätzen dargestellt (cf. Glück 2018; Leonardi 2020a, 2020b, 2021, 2022). Im Folgenden sollen ausgewählte Ergebnisse aus den beiden Subkorpora, zum einen der schriftliche Datensatz bestehend aus 430 Texten und zum anderen der mündliche Datensatz bestehend aus Interviews mit 65 Oberschüler:innen und einer Gesamtzeit von ca. 43 Stunden, genauer beschrieben werden. Dieser Abschnitt zeigt die vielseitige Nutzbarkeit des KOMMA-Korpus anhand einiger ausgewählter Fallbeispiele im Bereich der Wortbildung (4.1) und der Lexik (4.2) (zum Kontakteinfluss im syntaktischen Bereich, cf. Glück/Leonardi 2019).

4.1 Wortbildung im geschriebenen und gesprochenen Standarddeutsch

In einer kleinen Fallstudie soll überprüft werden, inwieweit sich im Bereich der Wortbildung, mit Fokus auf die Variation bei der Adverbienbildung im Standarddeutschen, Unterschiede nachweisen lassen. Im Standarddeutschen können die Adverbien *durchweg/durchwegs*, *nochmal/nochmals*, *öfter/öfters* und *weiter/weiters* (in der Bedeutung ‚darüber hinaus‘) ohne Bedeutungsunterschied mit einem Nullmorphem oder durch das Anhängen eines -s gebildet werden (Duden 2006). Bereits in der *Variantengrammatik* der Deutschen (cf. Dürscheid/Elspaß 2015; Elspaß 2022) wurden areale Unterschiede innerhalb des Standarddeutschen im deutschsprachigen Gebiet festgestellt (cf. auch Ammon et al. 2004; Ammon/Bickel/Lenz 2016; Niehaus 2017).

Im gesamten KOMMA-Korpus sind insgesamt 289 Belege von *durchweg/durchwegs*, *nochmal/nochmals*, *öfter/öfters* und *weiter/weiters* belegt, wobei 40 davon auf den schriftlichen und 249 auf den mündlichen Datensatz fallen, wie aus Tabelle 9 ersichtlich wird.

Variante	KOMMA-Korpus		
	Schriftliche Daten	Mündliche Daten	
	Oberschüler:innen	Interviewer	Oberschüler:innen
<i>durchweg</i>	n=0	n=2	n=0
<i>durchwegs</i>	n=1	n=1	n=0
<i>nochmal</i>	n=2	n=132	n=57
<i>nochmals</i>	n=3	n=0	n=3
<i>öfter</i>	n=11	n=2	n=24
<i>öfters</i>	n=5	n=9	n=18
<i>weiter</i>	n=0	n=0	n=0
<i>weiters</i>	n=18	n=0	n=1
Gesamt	n=40	n=146	n=103
		n=249	

Tabelle 9: Verteilung der Varianten im KOMMA-Korpus

Wie man Tabelle 9 entnehmen kann, werden im mündlichen Datensatz 103 der hier untersuchten Adverbien mit und ohne Suffix von den Oberschüler:innen und 146 von den Interviewer:innen produziert. Werden die Belege der Interviewer:innen nicht berücksichtigt, so zeigen sich im schriftlichen Subkorpus für die Varianten mit Suffix 27 (67,5%) Belege und für die Varianten ohne Suffix 13 (32,5%) Belege. Im mündlichen Subkorpus sind es hingegen 22 (21,4%)

Belege für die Adverbien mit Suffix und 81 (78,6%) Belege für die Adverbien ohne Suffix. Diese Zahlen visualisieren die komplementären Häufigkeiten der Varianten mit und ohne -s.

Obwohl die absoluten Zahlen relativ klein sind, zeigt sich in Abhängigkeit vom Medium gesprochene vs. geschriebene Sprache ein Unterschied in der Verwendung der Adverbien im schriftlichen und mündlichen Datensatz (diamesische Variation): mit fast 70% der Belege überwiegt im schriftlichen Datensatz die Verwendung der Adverbien mit Suffix, während mit fast 80% der Belege im mündlichen Datensatz die Verwendung der Adverbien ohne Suffix überwiegt.¹⁷ Von einer Hyperkorrektur im Sinne einer Fehlanwendung zu sprechen, „[...] die dadurch entstehen [kann], dass jemand um jeden Preis Fehler vermeiden will und eine [...] korrekte Regel an der falschen Stelle anwendet“ (Löffler 2003: 2419), würde ich deshalb vermeiden, da die Adverbien *durchweg/durchwegs*, *nochmal/nochmals*, *öfter/öfters* und *weiter/weiters* laut dem Duden (2006) ohne Bedeutungsunterschied im Standarddeutschen gebildet werden können. Wie bereits erwähnt, wurden auch im *Variante(n)grammatik*-Korpus areale Unterschiede innerhalb des Standarddeutschen im deutschsprachigen Gebiet festgestellt (cf. auch Niehaus 2017), die wiederum durch die Daten aus dem KOMMA-Korpus bestätigt wurden. Laut Niehaus (2017: 77–79) wird *durchwegs* mehrheitlich in Österreich und Südtirol verwendet. In Deutschland wird hingegen fast ausschließlich die Variante *durchweg* benutzt. Obwohl beide Varianten, sowohl *durchweg* als auch *durchwegs*, als standardsprachlich einzustufen sind, zeigt das KOMMA-Korpus, dass sich die Südtiroler Oberschüler:innen im Schriftlichen vorwiegend an den nördlichen deutschen Sprachraum (Varianten ohne -s) orientieren, während sie sich im Mündlichen eher an den südlichen deutschsprachigen Raum (Varianten mit -s) orientieren (cf. Niehaus 2017). Eine mögliche Erklärung hierfür könnte der Mangel an Wissen der Südtiroler Schüler:innen über die Variation in der deutschen Standardsprache sein, der zur sprachlichen Unsicherheit führt. Wie bereits von Hofer (2017: 92) festgestellt wurde, könnte „[m]ehr Toleranz gegenüber standardsprachlicher Variation, insbesondere gegenüber Südtiroler Eigenheiten, [...] zu einer Stärkung des sprachlichen Selbstbewusstseins der Südtiroler Schüler_innen beitragen [...]“. Um bei den Sprecher:innen ein größeres sprachliches Sprachbewusstsein gegenüber der arealen Variation der deutschen Standardsprache schaffen zu können, ist eine Thematisierung unausweichlich: Wissen, Reflexion und Analyse sprachlichen Wissens erlauben einen angemessenen und selbstbewussten Sprachgebrauch (cf. Hofer 2017, 2020).

4.2 Lexikalische Einflüsse im gesprochenen Standarddeutsch

Dialekteinflüsse auf lexikalischer Ebene beschränken sich in den mündlichen Sprachdaten auf Einzelbelege, die im Folgenden anhand einiger ausgewählter Belege präsentiert werden sollen. Es handelt sich hierbei meist um Lexeme, die auch in der deutschen Standardvarietät existieren, allerdings in einer anderen Bedeutung: Im Dialekt übernimmt *fein* zusätzlich noch die Bedeutung von ‚angenehm‘ (Beispiel (1)–(3)), *ausstellen* hat noch die Bedeutung ‚aus dem Weg

¹⁷ An dieser Stelle soll angemerkt werden, dass die Zahlen in Tabelle 9 sowohl für eindeutige Ergebnisse zu klein als auch für intra-individuelle Vergleiche nicht aussagekräftig sind, wie korrekterweise von den anonymen Reviewer:innen angemerkt wurde. Nichtsdestotrotz geben sie einen Einblick in die Verwendung von ausgewählten Adverbien mit und ohne Suffix im mündlichen und schriftlichen Sprachgebrauch in Südtirol.

gehen‘ (Beispiel (4)–(5)). Diese semantische Bedeutung übertragen die Sprecher:innen in die deutsche Standardvarietät, wie die folgenden Beispiele zeigen.

- (1) SBJ_07: aber ich versuche mich am wochenende (.) zu hause aufzuhalten einfach so auch (-) weil es **feiner** ist (.) wenn man die ganze woche in bozen ist dann ist es auch **fein** wenn man mal am wochenende so im dorf sein kann
(Art der Aufgabe: Sprachbiographie/Erhebungsort: Bozen/Geschlecht: weiblich)
- (2) SBJ_53: und da hab ich eben gemerkt dass es wichtig ist da italienisch zu sprechen weil da konnte ich mich (.) mit anderen kindern unterhalten und auch wenn ich (-) am strand (-) burgen baute war es **fein** wenn ich auch ihre sprache beherrschte
(Sprachbiographie/Schlanders/männlich)
- (3) SBJ_17: im (.) italienischen ist eigentlich die lehrperson relativ eine (-) **feine** und nette person
(Sprachbiographie/Brixen/männlich)
- (4) SBJ_46: kann man nicht **ausstellen** weil oft wenn wenn ich es dann auf dialekt sage versteht er mich nicht
(Sprachbiographie/Meran/weiblich)
- (5) SBJ_14 wenn ein spi/ äh spieler zur nächsten base läuft dann darf er nicht **ausstellen** er muss gerade laufen
(Erklärung/Brixen/männlich)

Lexikalische Einflüsse aus dem Italienischen zeigen sich im mündlichen KOMMA-Korpus in sehr geringem Maße. Eine Ausnahme bilden Diskursmarker, wie beispielsweise *mah* und *boh* in den Beispielen (6) und (7).

- (6) SBJ_35: das ist dann schon so irgendwie zwei verschiedene persönlichkeiten **mah** jetzt nicht so extrem aber ich denke schon dass da ein unterschied besteht
(Diskussion/Meran/weiblich)
- (7) SBJ_09: ich glaube ich spreche nur den dialekt in (.) ohne akzent oder wie man da sagt (-) weil (-) **boh** mir hat halt wenn man auch im ausland ist dann wird einem schon oft gesagt
(Sprachbiographie/Bozen/weiblich)

Bereits Dal Negro (2011) und Ciccolone/Dal Negro (2016) haben festgestellt, dass italienische Diskursmarker und Interjektionen auch von der deutschsprachigen Bevölkerung verwendet werden. Ihre empirischen Studien basieren allerdings auf einem mehrsprachigen Korpus von informellen Alltagsgesprächen (Familiengespräche, Interaktionen im Freundeskreis) in der lokalen deutschen Varietät (südbairischer Dialekt), der trentinischen Varietät und dem italienischen Standard (cf. dazu Dal Negro 2018). Insgesamt lassen sich im mündlichen KOMMA-Korpus 132 italienische Diskursmarker identifizieren: *boh* (n=68), *mah* (n=60), *ma dai* (n=2), *dai* (n=1) und *ostia* (n=1). Eine umfassende Diskussion zu den Vorkommenshäufigkeiten, Positionen und deren Funktion findet sich in Leonardi (2021). Wie bereits an anderer Stelle beschrieben (cf. Leonardi 2021: 52–54), könnte eine mögliche Erklärung der Verwendung der italienischen Diskursmarker in der deutschen Standardvarietät sein, dass sich die Proband:innen über deren Gebrauch nicht bewusst sind: *mah* und *boh* sind Teil des gesprochenen Standards in Südtirol (geworden).

5 Zusammenfassung und Ausblick

Im vorliegenden Beitrag wurde versucht, einen Überblick über das KOMMA-Korpus zu geben. Die systematische Aufbereitung in ELAN ermöglicht gezielte Abfragen, das Extrahieren von Belegen, die bestimmte Phänomene ausweisen können, sowie eine Vergleichbarkeit beider Subkorpora auf verschiedenen sprachlichen Ebenen. Dabei können sowohl in den Schülertexten als auch in den mündlichen Sprachdaten Phänomene aus den Bereichen der Lexik, Morphosyntax, Orthographie oder Pragmatik betrachtet werden. Sofern vom selben Produzenten bzw. derselben Produzentin auch ein mündlicher Diskurs vorliegt, kann auch ein Vergleich zwischen dem Geschriebenen und dem Gesprochenen angestrebt werden. Neben konkreten Sprachdaten bieten die sprachbiographischen Interviews subjektive Sprachdaten, die für eine große Bandbreite an Forschungsfragen genutzt werden können sowie zu einem besseren Verständnis des Zusammenspiels von Sprachkontakt, Spracheinstellungen und Sprachverwendung beitragen können (cf. Glück 2018; Glück/Leonardi 2019; Leonardi 2020a, 2020b, 2021, 2022). Weiters besteht die Möglichkeit, das Deutsche in Südtirol mit anderen deutschen Korpora zu vergleichen und zu kontrastieren, wie etwa dem Korpus *Südtirol* oder der *Varietengrammatik*.

Literaturverzeichnis

- Abel, Andrea (2018): „Von Bars, Oberschulen und weißen Stimmzetteln: zum Wortschatz des Standarddeutschen in Südtirol“. In: Rabanus, Stefan (ed.): *Deutsch als Minderheitensprache in Italien. Theorie und Empirie kontaktinduzierten Sprachwandels*. Hildesheim etc., Olms: 283–324.
- Abel, Andrea/Anstein, Stefanie/Petrakis, Stefanos (2009): „Die Initiative Korpus Südtirol“. *Linguistik Online* 38, 2/2009: 5–12. bop.unibe.ch/linguistik-online/article/view/502 [17.10.2023].
- Abfalterer, Heidemarie (2007): *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht. Lexikalisch-semantische Besonderheiten im Standarddeutsch Südtirols*. Innsbruck: Innsbruck University Press (= *Innsbrucker Beiträge zur Kulturwissenschaft. Germanistische Reihe* 72).
- Ammon, Ulrich et al. (2004): *Varietätenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin/New York: de Gruyter.
- Ammon, Ulrich/Bickel, Hans/Lenz, Alexandra N. (2016): *Varietätenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. 2., völlig neu bearbeitete und erweiterte Auflage. Berlin/Boston: de Gruyter.
- Anstein, Stefanie/Oberhammer, Margit/Petrakis, Stefanos (2011): „Korpus Südtirol – Aufbau und Abfrage“. In: Abel, Andrea/Zanin, Renata (eds.): *Korpora in Lehre und Forschung*. Bozen, University Press: 15–28.
- ASTAT. Landesinstitut für Statistik (2015): *Südtiroler Sprachbarometer. Sprachgebrauch und Sprachidentität in Südtirol – 2014*. Bozen: Autonome Provinz Bozen–Südtirol. astat.provinz.bz.it/de/aktuelles-publikationen-info.asp?news_action=4&news_article_id=516194 [17.10.2023].

- ASTAT. Landesinstitut für Statistik (2021): *Südtirol in Zahlen 2021*. Bozen: Autonome Provinz Bozen–Südtirol. astat.provinz.bz.it/de/suedtirol-in-zahlen.asp [17.10.2023].
- Berend, Nina (2009): „Standardsprache – Alltagssprache. Eine Aufnahmeaktion zur Untersuchung der Variation im gesprochenen Standarddeutsch“. In: Felder, Ekkehard (ed.): *Sprache*. Berlin/Heidelberg, Springer: 215–234.
- Ciccolone, Simone/Dal Negro, Silvia (2016): „Marcare il contrasto nel parlato bilingue. *Ma e obâr* in un corpus sudtirolese“. In: Bombi, Raffaella/Orioles, Vincenzo (eds.): *Lingue in contatto/Contact linguistics. Atti del XLVIII Congresso Internazionale di Studi della Società di Linguistica (SLI)*. Roma, Bulzoni: 97–113.
- Ciccolone, Simone/Franceschini, Rita (2015): „Südtirol zwischen Ortsdialekten und Sprachkontakt. DIAGRAMM und KONTATTO“. In: Kehrein, Roland/Lameli, Alfred/Rabanus, Stefan (eds.): *Regionale Variation des Deutschen. Projekte und Perspektiven*. Berlin/Boston, de Gruyter: 459–488.
- Crasborn, Onno/Sloetjes, Han (2008): „Enhanced ELAN functionality for sign language corpora“. *Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation*: 39–43.
- Dal Negro, Silvia (2011): „Tedesco di contatto in Italia“. In: Fazzini, Elisabetta (ed.): *Il tedesco superiore: Tradizione scritta e varietà parlate*. Alessandria, dell’Orso: 179–199.
- Dal Negro, Silvia (2018): „Finding patterns in bilingual speech“. *Lingue e Linguaggio* 17/1: 71–85.
- Dal Negro, Silvia/Ciccolone, Simone (2018): „Il parlato bilingue: italiano e tedesco a contatto in un corpus sudtirolese“. In: Bermejo Calleja, Felisa/Katelhön, Peggy (eds.): *Lingua parlata. Un confronto fra l’italiano e alcune lingue europee (= Kontrastive Linguistik/Linguistica Contrastiva 8)*. Berlin etc., Lang: 385–407.
- Duden (2006): *Die deutsche Rechtschreibung*. Band 1. 24., völlig neu bearbeitete und erweiterte Auflage. Mannheim etc.: Dudenverlag.
- Duden-Online Wörterbuch: *Duden – mehr als ein Wörterbuch*. duden.de/ [17.10.2023].
- Dürscheid, Christa/Stephan Elspaß (2015): „Variantengrammatik des Standarddeutschen“. In: Kehrein, Roland/Lameli, Alfred/Rabanus, Stefan (eds.): *Regionale Variation des Deutschen. Projekte und Perspektiven*. Berlin/Boston, de Gruyter: 563–584.
- Egger, Kurt (1994): *Die Sprachen unserer Kinder. Spracherwerb in einem mehrsprachigen Gebiet*. Meran: Alpha&Beta.
- ELAN: *Annotation tool for audio and video recordings*. tla.mpi.nl/tools/tla-tools/elan/ [17.10.2023].
- Elspaß, Stephan (2005): „Zum sprachpolitischen Umgang mit regionaler Variation der Standardsprache in der pluralistischen Sprachgesellschaft“. In: Kilian, Jörg (ed.): *Sprache und Politik. Deutsch im demokratischen Staat*. Mannheim, Dudenverlag: 294–313.
- Elspaß, Stephan (2022): „Die Variantengrammatik des Standarddeutschen als Brückenschlag zwischen Areallinguistik und Grammatikographie“. In: Brommer, Sarah/Roth, Kersten Sven/Spitzmüller, Jürgen (eds.): *Brückenschläge: Linguistik an den Schnittstellen*. Tübingen, Narr: 89–110. (= *Tübinger Beiträge zur Linguistik* 583).

- Glück, Alexander (2018): „(P)Artikelpositionen im regionalen Standarddeutschen in Südtirol“. In: Funk, Edith et al. (eds.): *Dialects are Forever. Die unbändige Lust an der Wortklauberei. Festschrift für Anthony R. Rowley*. Regensburg, edition vulpes: 51–66.
- Glück, Alexander/Leonardi, Mara Maya Victoria (2019): „Zur Verwendung von Präpositionen in Texten und Diskursen von Südtiroler Maturanten“. In: Kürschner, Sebastian/Habermann, Mechthild/Müller, Peter O. (eds.): *Methodik moderner Dialektforschung. Erhebung, Aufbereitung und Auswertung von Daten am Beispiel des Oberdeutschen*. Hildesheim/Zürich/New York, Olms: 445–470. (= *Germanistische Linguistik* 241–243/2019).
- Glück, Alexander/Leonardi, Mara Maya Victoria/Riehl, Claudia Maria (2019): „Südtirol“. In: Beyer, Rahel/Plewnia, Albrecht (eds.): *Handbuch des Deutschen in West- und Mitteleuropa: Sprachminderheiten und Mehrsprachigkeitskonstellationen*. Tübingen, Narr: 245–280.
- Hamburg Modern Times Corpus (HaMoTiC): *Hamburger Zentrum für Sprachkorpora* (HZSK). fdr.uni-hamburg.de/record/1483 [17.10.2023].
- Hofer, Silvia (2017): „Das Korrekturverhalten von Südtiroler Deutschlehrpersonen und ihre Rolle als Normautoritäten im sozialen Kräftefeld der deutschen Standardsprache in Südtirol“. *ÖDaF-Mitteilungen* 33/2: 79–95.
- Hofer, Silvia (2020): *Deutsch ist nicht gleich Deutsch: Zum Umgang mit der plurizentrischen Sprache Deutsch und standardsprachlicher Variation an Südtiroler Oberschulen*. Dissertation, Universität Wien.
- KOMMA-Korpus: koma.projects.unibz.it [05.11.2023].
- Kupietz, Marc/Lüngen, Harald/Diewald, Nils (2022): „Das Gesamtkonzept des Deutschen Referenzkorpus DeReKo“. In: Deppermann, Arnulf et al. (eds.): *Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial*. Berlin/Boston, de Gruyter: 1–28.
- Lanthaler, Franz (2012): „Innere und äußere Mehrsprachigkeit in Südtirol“. In: Drumbl, Hans/Sitta, Horst (eds.): *Franz Lanthaler. Texte zu Sprache und Schule in Südtirol (1974–2012)*. Meran, Alpha&Beta: 139–164.
- Leonardi, Mara Maya Victoria (2020a): „‘I hardly ever practice the real Standard German.’ Self-reported language use and language proficiency in South Tyrol (Italy)“. *Linguistik Online* 102, 2/20: 83–98. bop.unibe.ch/linguistik-online/article/view/6825 [17.10.2023].
- Leonardi, Mara Maya Victoria (2020b): „Famiglie plurilingui in Alto Adige. Pratiche linguistiche e appartenenza linguistica“. In: Marra, Antonietta/Dal Negro, Silvia (eds.): *Lingue minoritarie tra localismi e globalizzazione*. Milano, Studi AItLA 11: 167–181. officinaventuno.com/public/guest/AitLA/eBookAitLA_11.pdf [17.10.2023].
- Leonardi, Mara Maya Victoria (2021): „Una ricerca corpus-based sul tedesco standard parlato in Alto Adige: *mah* e *boh* in una situazione di contatto linguistico“. In: Favilla, Maria E./Machetti, Sabrina (eds.): *Lingue in contatto e linguistica applicata: Individui e società*. Milano, Officinaventuno: 41–57. (= *Collana Studi AItLA* 13) aitla.it/images/pdf/StudiAitLA13/003Leonardi.pdf [17.10.2023].
- Leonardi, Mara Maya Victoria (2022): „Varietätengebrauch und Spracheinstellungen. Eine qualitative Untersuchung sprachbiographischer Interviews mit MaturantInnen an deutschsprachigen Oberschulen in Südtirol“. In: Elspaß, Stephan/Vergeiner, Philip C./Wallner, Dominik (eds.): *Struktur von Variation zwischen Individuum und Gesellschaft. Akten der 14.*

- Bayerisch-Österreichischen Dialektologietagung, Salzburg, 2019*. Stuttgart, Franz Steiner: 247–260. (= *Zeitschrift für Dialektologie und Linguistik* – Band 189).
- Löffler, Heinrich (2003): „Hyperkorrekturen als Hilfe der Rekonstruktion von Sprachzuständen“. In: Besch, Werner et al. (eds.): *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*. 2., vollständig neu bearbeitete und erweiterte Auflage. 3. Teilband. Berlin/Boston, de Gruyter: 2419–2425.
- Niehaus, Konstantin (2017): „Die Begrenztheit plurizentrischer Grenzen: Grammatische Variation in der pluriarealen Sprache Deutsch“. In: Davies, Winifred V. et al. (eds.): *Standardsprache zwischen Norm und Praxis. Theoretische Betrachtungen, empirische Studien und sprachdidaktische Ausblicke*. Tübingen, Narr: 61–88.
- Nolda, Andreas/Barbaresi, Adrien/Geyken, Alexander (2022): „Korpora für die lexikographische Beschreibung diatopischer Variation in der deutschen Standardsprache. Das ZDL-Regionalkorpus und das Webmonitor-Korpus“. In: Deppermann, Arnulf et al. (eds.): *Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial*. Berlin/Boston, de Gruyter: 29–52.
- Reineke, Silke/Deppermann, Arnulf/Schmidt, Thomas (2022): „Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch (FOLK). Zum Nutzen eines großen annotierten Korpus gesprochener Sprache für interaktionslinguistische Fragestellungen“. In: Deppermann, Arnulf et al. (eds.): *Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial*. Berlin/Boston, de Gruyter: 71–102.
- Schiller, Anne et al. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*: 1–86. sfs.uni-tuebingen.de/resources/stts-1999.pdf [17.10.2023].
- Schmid, Helmut (1995): „Improvements in Part-of-Speech Tagging with an Application to German“. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland: 1–9. cis.lmu.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf [17.10.2023].
- Selting, Margret et al. (2009): „Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)“. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10: 353–402. gespraechsfor-schung-ozs.de/heft2009/px-gat2.pdf [17.10.2023].
- Variantengrammatik*: „Adverbien mit -s/ohne Suffix“. mediawiki.ids-mannheim.de/VarGra/index.php/Adverbien_mit_-s/_ohne_Suffix [17.10.2023].
- Veronesi, Daniela (2009): „‘Due lingue che sono entrambe mie’: biografia di parlanti ‘bilingui’ e ‘monolingui’ in un territorio di confine“. In: Riccioni, Ilaria (ed): *Multiculturalismi a confronto*. Bozen, University Press: 39–56.
- Weber Egli, Daniela (1992): *Gemischtsprachige Familien in Südtirol/Alto Adige. Zweisprachigkeit und soziale Kontakte. Ein Vergleich von Familien in Bozen und Meran*. Meran: Alpha&Beta.
- Westpfahl, Swantje (2014): „STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data“. In: Levin, Lori/Stede, Manfred (eds.): *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop*. Dublin, Association for Computational Linguistics and Dublin City University: 1–10. aclweb.org/anthology/W14-4901.pdf [17.10.2023].
- Westpfahl, Swantje/Schmidt, Thomas (2013): „POS für(s) FOLK – Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch“. *Journal for Language Technology and Computational Linguistics* 28/1: 139–153. ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/2223/file/Westpfahl_Schmidt_POS_2013.pdf [17.10.2023].

Westpfahl, Swantje et al. (2017): *STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS)*. Mannheim: Institut für Deutsche Sprache: 5–53. ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6063/file/Westpfahl_Schmidt_Jonietz_Borlinghaus_STTS_2_0_2017.pdf [17.10.2023].