

Die Initiative Korpus Südtirol

Andrea Abel/Stefanie Anstein/Stefanos Petrakis (Bozen)

Abstract

Der Beitrag gibt einen Überblick über die Initiative Korpus Südtirol und beschreibt das im Entstehen begriffene Korpus des geschriebenen Deutschen in Südtirol. Dabei wird insbesondere auf den inhaltlichen Aufbau, die erhobenen Metadaten sowie die Organisation und Architektur des Gesamtsystems eingegangen. Zudem werden einige Anwendungsmöglichkeiten des Korpus Südtirol kurz vorgestellt, etwa für die Sprachdidaktik oder für varietätenlinguistische Untersuchungen, die auf der Grundlage eigens entwickelter Werkzeuge durchgeführt werden und deren Ergebnisse einen wichtigen Beitrag zur Beschreibung der deutschen Sprache in Südtirol leisten.

1 Einführung

Die deutsche Sprache ist in der Autonomen Provinz Bozen – Südtirol (Italien) der italienischen gleichgestellt (Autonomiestatut Art. 99, 100; Autonome Provinz Bozen-Südtirol 2006). Von den ca. 500.000 Einwohnern Südtirols sind etwa zwei Drittel deutscher Muttersprache, ein Viertel italienischer und der Rest ladinischer oder anderer Muttersprache. Deutsch wird in nahezu allen Bereichen des öffentlichen Lebens parallel zum Italienischen verwendet – in den Medien, in Schulen, Weiterbildungsinstituten oder in der öffentlichen Verwaltung (vgl. Lanthaler/Saxalber 1995; Abel 2009).

Zwischen den Varietäten der deutschen Standardsprache in Deutschland, Österreich, der Schweiz sowie in weiteren sogenannten nationalen Halbzentren wie Südtirol (vgl. Ammon 1995) bestehen Unterschiede auf verschiedenen Ebenen, von der Aussprache- über die Wort- bis zur Satzebene (vgl. Ammon et al. 2004; Abfalterer 2007). Die Initiative C4, die vier der deutschen Standardvarietäten in Europa gegenüberstellend dokumentieren soll, möchte einen Zugang zu Korpusdaten der folgenden Arbeitsstellen anbieten: DWDS (*Das Digitale Wörterbuch der deutschen Sprache des 20. Jhdts. der Berlin-Brandenburgischen Akademie der Wissenschaften*), AAC (*Austrian Academy Corpus der Österreichischen Akademie der Wissenschaften in Wien*), SCHWEIZER TEXT KORPUS des Deutschen Seminars der Universität Basel und Korpus Südtirol der Freien Universität Bozen und der Europäischen Akademie Bozen.

Das Korpus Südtirol dokumentiert die geschriebene deutsche Sprache in Südtirol erstmals in umfassender Weise mithilfe korpuslinguistischer Methoden und hat somit eine wichtige Archivfunktion. Des Weiteren wird es für varietätenlinguistische Untersuchungen verwendet sowie für die Förderung der Sprachaufmerksamkeit und in der Sprachdidaktik an Schule und Universität eingesetzt. Die Webseite <http://www.korpus-suedtirol.it> (Stand: Januar 2009) beinhaltet ausführliche Informationen über die Gesamtinitiative, Literaturangaben sowie den Link zur Suche in verschiedenen Teilkorpora.

Nach einer Beschreibung des inhaltlichen Aufbaus und der Gesamtarchitektur wird in diesem Beitrag ein kurzer Einblick in Anwendungsmöglichkeiten des Korpus Südtirol gegeben, deren

Ergebnisse einen wesentlichen Beitrag zur korpusbasierten Sprachbeobachtung, -analyse und -beratung leisten und den Weg zu einer neuen Sprachkultur in Südtirol öffnen helfen.

2 Korpusdaten und deren Aufbereitung

Um eine ausgeglichene und möglichst repräsentative Textsammlung zu erhalten, werden – auf der Grundlage einer umfassenden Bestandsaufnahme (s. Abschnitt 3) und in Übereinstimmung mit allen C4-Arbeitsstellen – vier Textsorten zu gleichen Teilen für das 20. Jahrhundert bis heute gesammelt: Belletristik, Gebrauchstexte, journalistische Prosa und Sachtexte. Teilweise liegen Texte bereits in elektronischer Form vor; vieles muss jedoch zunächst digitalisiert werden. Bisher wurden rund 70 Mio. laufende Textwörter vollständig aufbereitet und abfragbar gemacht. Ein Großteil davon (ca. 66 Mio. Wörter) besteht aus Texten der Tageszeitung *Dolomiten* der Jahre 1991–2006. Der Rest beinhaltet verschiedene belletristische Werke oder auch neuere Zeitschriften.

Die Texte wurden tokenisiert und linguistisch mit Wortart- und Lemma-Informationen annotiert (TreeTagger, Schmid 1994). Der Chunker YAC (Kermes 2003) diente teilweise zur weiteren Annotation von Chunks und morphosyntaktischen Annotationen, z.B. Kasus und Numerus von Substantiven.¹ Verwendete Abfragesysteme sind die Korpuswerkzeuge des Instituts für Maschinelle Sprachverarbeitung der Universität Stuttgart: CorpusWorkBench (Christ 1994) und Corpus Query Processor (im Folgenden: CQP; Evert 2005), auf denen die eigens entwickelte Benutzeroberfläche (s. Abschnitt 5) aufbaut.

3 Bestandsaufnahme und Metadatenerfassung

Die Bestandsaufnahme-Datenbank, die alle in das Korpus aufgenommenen und aufzunehmenden Texte als Einzeleinträge (d.h. jeweils einen eigenen Eintrag für einen einzelnen Text, z. B. ein Buch, einen Aufsatz oder einen Zeitungsartikel, eines bestimmten Autors) erfasst und jeweils deren Metadaten (wie Autor, Textsorte, Jahr, Seitenzahl etc.) beinhaltet, besteht derzeit aus ca. 294.000 Einträgen, sprich dokumentierten Südtiroler Texten. Diese wurden und werden von Literaturexperten halbautomatisch mithilfe eigens angepasster Werkzeuge erstellt, was das Verarbeiten einer großen Datenmenge erlaubt und einen wichtigen Schritt im Rahmen der Zusammenstellung, Dokumentation und Digitalisierung von Südtiroler Texten darstellt. Abb. 1 zeigt eine Übersicht der bisherigen Artikelverteilung auf die einzelnen Textsorten und Dekaden, die als Orientierung für die weitere Anreicherung der Bestandsaufnahme dient.

¹ Für weitere Details zur Korpusaufbereitung vgl. z.B. Lemnitzer/Zinsmeister (2006).

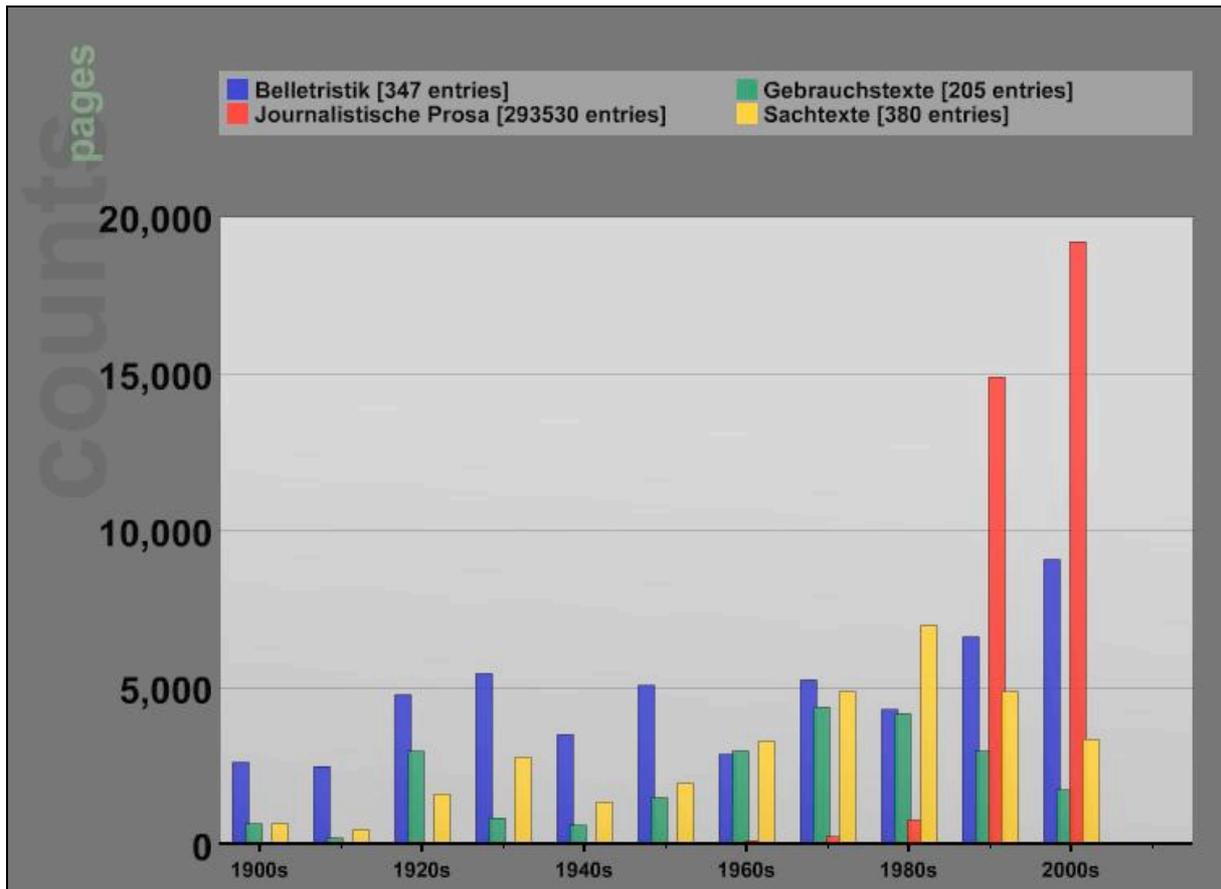


Abb. 1: Übersicht der Einträge in der Bestandsaufnahme des Korpus Südtirol, sortiert nach Textsorten und Dekaden, dargestellt auf der Grundlage von Seitenzahlen

Zu jedem Eintrag in der Bestandsaufnahme-Datenbank werden zwei Arten von Metadaten getrennt gesammelt, nämlich allgemeine zum Text einerseits (wie Titel, Untertitel, Publikationsort, Verlag, Publikationsjahr etc.) und spezielle zum Autor andererseits (wie Name, Geburtsjahr, Geburtsort etc.).

Eine Reihe von deskriptiven Metadaten zum Text wird dabei halbautomatisch erfasst. Für Zeitungsausgaben etwa kommen Werkzeuge zum Einsatz, die Artikel automatisch trennen und ihnen jeweils u. a. das Datum der Ausgabe zuweisen; weitere Angaben, wie z. B. Verlag o. Ä., müssen per Hand ergänzt werden. Bei Einzelwerken werden die Metadaten manuell in teilweise mühevoller, aber unvermeidlicher Detailarbeit in Bibliotheksarchiven gesammelt, was den Wert einer solchen Datensammlung deutlich macht. Ein Auszug der Metadaten zum Bucheintrag "Tschonnie Tschennett – Grobes Foul" ist in Abb. 2 zu sehen.

SourceTitle	Tschonnie Tschenett – Grobes Foul
SourceSubtitle	Kriminalroman
ConversionType	analog
SourceExtentPagecount	189
SourceLocation	Tessmann Bibliothek
SourcePublicationISBN	3-85218-148-8
SourcePublicationWhere	Innsbruck
SourcePublicationWho	Haymon
SourcePublicationYear	1993
TaxonomyPublicationType	Buch
TaxonomySourceFormat	analog (a)
TaxonomyOriginalTypeface	Antiqua
TaxonomyTextCategory	Belletristik

Abb. 2: Auszug aus der detaillierten Metadatenammlung für die Bestandsaufnahmeneinträge am Beispiel eines Buchs

Jeder Datenbankeintrag, der sich auf einen einzelnen Text (d.h. Artikel, Buch etc.) bezieht, wird mit dem separat erstellten Eintrag zu seinem Südtiroler Autor verlinkt. Diese Lösung wurde gewählt, um eine Vervielfachung der Autorenbeschreibungen zu vermeiden, da viele Autoren mit jeweils mehreren Werken vertreten sind. Abb. 3 zeigt den Autoren-Metadaten-Eintrag zum eben vorgestellten Bucheintrag. In die Autoredatenbank werden ausschließlich Südtiroler Autoren aufgenommen; entscheidungsleitend für die Eingliederung in diese Kategorie ist der Hauptaufenthaltsort.

FullName	Kurt Lanthaler
Surname	Lanthaler
ForeName	Kurt
YearOfBirth	1960
YearOfDeath	–
BirthPlace	Bozen

Abb. 3: Auszug der zum Datenbankeintrag in Abb. 2 gehörigen Autorenmetadaten

Die beschriebenen zwei Arten von Metadaten-Informationen in der Bestandsaufnahme-Datenbank, nämlich die Text- und die Autorenmetadaten, werden schließlich mit dem Text selbst verlinkt (s. Abschnitt 4). Für die Weiterverarbeitung zu einem abfragbaren Korpus wird alles in ein nach den Richtlinien der *Text Encoding Initiative*² (im Folgenden: TEI) gültiges XML-Dokument umgewandelt, wobei die Metadaten jeweils im XML-Header des Dokuments festgehalten sind. Ein Ausschnitt wiederum zum selben Beispiel ist in Abb. 4 zu sehen.

² <http://www.tei-c.org/release/doc/tei-p4-doc/html> (Stand: November 2008).

```

- <teiHeader id="id1191458784723">
  - <fileDesc>
    - <titleStmt>
      - <title>
        Tschonnie Tschennett - Grobes Foul [Machine readable version]
      </title>
    - <author>
      - <persName>
        <name type="full">Kurt Lanthaler</name>
        <foreName>Kurt</foreName>
        <surname>Lanthaler</surname>
      </persName>
    </author>
  - <respStmt>

```

Abb. 4: Auszug des zum Datenbankeintrag in Abb. 2 und 3 gehörigen XML-Dokuments

4 Organisation und Architektur des Gesamtsystems

Das System der Initiative Korpus Südtirol besteht aus zwei Hauptteilen, der Bestandsaufnahme-Datenbank und dem sogenannten Repositoryum.

Die Texte, die bereits digital vorliegen, werden im Repositoryum gespeichert, von dem anhand der Bestandsaufnahme/Metadatensammlung Subkorpora der vorhandenen Daten erstellt werden können. So wurde z.B. das Dolomiten-Korpus aus dem Repositoryum aufbereitet, indem zunächst automatisch alle Einträge der Bestandsaufnahme identifiziert wurden, die in dieses Korpus aufgenommen werden sollen. Die eigentlichen Texte, die mit den Datenbankeinträgen verlinkt und in XML-Dokumente umgewandelt wurden, wurden dann zu einem Ganzen zusammengefügt (s. Abschnitt 3).

Der angewandte Digitalisierungsprozess für analog vorliegende Texte orientiert sich an den *Open Archive Initiative Protocols for Reuse and Exchange of Content*³. Für die Neuaufnahme von Texten in das Repositoryum werden folgende Schritte durchgeführt: Der betreffende Eintrag in der Bestandsaufnahme wird identifiziert. Die zugehörigen Metadaten werden halbautomatisch auf Vollständigkeit und Korrektheit überprüft, wobei sie bis auf die Pflichtfelder (Textsorte, Jahr, Autor etc.) unvollständig sein können, und die Daten werden normalisiert; so werden z.B. Datumsangaben in ein einheitliches Format gebracht. Jeder Eintrag, der zu digitalisieren ist, erhält eine eindeutige Identifikationsnummer (im Folgenden: ID), damit später die entsprechenden Metadaten aus der Bestandsaufnahme hinzugefügt werden können. Der für die Digitalisierung Zuständige erhält die ID des Eintrags und ist – auf der Grundlage eigens dazu bereitgestellter Anleitungen – für die korrekte Lieferung eines digitalen Dokuments in XML-Format nach TEI-Richtlinien verantwortlich, das die ID enthält und kompatibel zu den vereinbarten C4-Richtlinien ist. Sobald die XML-Dateien auf Korrektheit und Vollständigkeit überprüft worden sind, werden sie dem Korpus Südtirol-Repositoryum hinzugefügt.

Wie oben schon am Beispiel des Dolomiten-Korpus beschrieben, können Subkorpora nun erstellt werden, indem anhand der Informationen aus der Bestandsaufnahme digitale Dokumente in TEI-XML-Format aus dem Repositoryum sowie ihre Metadaten aus der Bestandsaufnahme-Datenbank exportiert und verknüpft werden. Diese Textauswahl wird nach dem jeweiligen Zweck des zu erstellenden Korpus getroffen (z.B. Texte ausschließlich einer

³ <http://www.openarchives.org/ore> (Stand: Januar 2009).

Textgattung, eines Autors, eines Jahres). Sobald die Daten ausgewählt sind, werden linguistische Annotationen halbautomatisch mithilfe verschiedener Werkzeuge hinzugefügt (s. Abschnitt 2). Diese linguistische Aufbereitung wird am ausgewählten Subkorpus durchgeführt, nicht am Originaltext im Repositorium selber, da je nach Korpus-Zweck unterschiedliche Eigenschaften annotiert werden. Um für manuelle Korrekturen Mehrfacharbeit zu vermeiden, können bestimmte Besonderheiten, z.B. systematische Taggerfehler, dokumentiert und bei einer erneuten Annotation desselben Textes automatisch berücksichtigt werden. Die so angereicherten Daten werden schließlich für das verwendete Suchwerkzeug CQP indiziert und können dadurch abgefragt werden.

5 Abfrageoberfläche und Benutzerzugang

Die Registrierung zur Benutzung der Abfrageoberfläche kann unter <http://www.korpus-suedtirol.it>, Menüpunkt *Suche in den Korpora*, durchgeführt werden.

In dieser Benutzeroberfläche können drei verschiedene Arten von Suchoberflächen verwendet werden. Die *Einfache Suche* ist für Einsteiger ohne Korpuserfahrung konzipiert und für Einzelwortsuchen ausreichend. Ihr Suchfenster ist ähnlich einer Internetsuchmaschine zu benutzen und erlaubt es, die Korpora gewissermaßen zu 'googeln'. Die *Erweiterte Suche* spricht Nutzer an, die bereits Erfahrung mit Korpusarbeit und linguistisches Wissen haben. Hier können komplexere Anfragen gestellt werden, die auch die Annotationen in die Suche mit einbeziehen, z.B. "alle Adjektive, die vor dem Wort *Stimmzettel* vorkommen". Die *Expertensuche* erlaubt Anfragen mit regulären Ausdrücken nach der CQP-Syntax, wenn z.B. alle Nomina mit dem Präfix *Vor-* gefunden werden wollen.

6 Anwendungsmöglichkeiten

Ein Korpus des geschriebenen Deutschen in Südtirol trägt wesentlich zur Förderung des Sprachbewusstseins und Festigung der Sprachkultur bei, und zwar durch Ergebnisse linguistischer Untersuchungen sowie durch angepasste Didaktikkonzepte. Neben Linguisten, die Varietätenvergleiche durchführen und z.B. Variantenlexika erweitern, verwenden Sprachdidaktiker das Korpus Südtirol, das neben Südtiroler Texten auch Modelltexte enthält, für angewandte Korpuslinguistik im L2-Unterricht und in der Lehrerbildung, etwa an der Freien Universität Bozen, und machen systematisch u. a. auf den Gebrauch rekurrenter Sprachmuster und konnotative Dimensionen von Sprache aufmerksam. Ausgangspunkt für Rechercheaufgaben bilden dabei insbesondere Formulierungen in Lernertexten, die anhand von Korpusdaten überprüft werden (vgl. z. B. Zanin 2007).

Das Korpus steht Sprachlernenden in Schule und Universität sowie allgemein Sprachinteressierten frei im Internet zur Verfügung. Dort können häufige Wortverbindungen oder Satzkonstruktionen sowie Südtiroler Besonderheiten und deren Kontextsätze gefunden werden. Darüber hinaus ist es möglich, einzelne Vorkommen in verschiedenen Korpora einer Sprache einander gegenüberzustellen. So kann man derzeit beispielsweise die Verwendung von Einzelwörtern, wie etwa "Kondominium", sowohl im Südtiroler Dolomiten-Korpus als auch im deutschen allgemeinen Korpus DeWaC (großes deutsches Webkorpus, s. Baroni/Kilgarriff 2006) vergleichend untersuchen; die Ergebniskontexte unseres Beispiels machen deutlich, dass *Kondominium* in Südtirol in der Bedeutung von 'Wohnblock/Haus mit Eigentumswohnungen' verwendet wird, während im bundesdeutschen Sprachraum die völkerrechtliche Bedeutung im Sinne von 'Gebiet unter der gemeinsamen Herrschaft mehrerer Staaten' vorzufinden ist.

Erste varietätenlinguistische Untersuchungen anhand des Korpus Südtirol sind in Abel/Anstein (2008) beschrieben. Solche Studien bilden die empirische Grundlage für eine umfassende Beschreibung des Südtiroler Deutschen, die etwa in Wörterbüchern, Namenslis-

ten oder Rechtschreibprüfungsprogrammen Anwendung findet, wobei nicht nur die lexikalische (so besonders in der Vergangenheit), sondern auch die syntagmatische Ebene der Sprache im Mittelpunkt des Interesses steht. Ein Beispiel für letztere Ebene stellt die Wendung *in einem zweiten Moment* dar, die im DeWaC nicht vorkommt, in den Südtiroler Texten aber verwendet wird, um etwa 'in einem nächsten Schritt / zu einem späteren Zeitpunkt' auszudrücken, was sehr wahrscheinlich vom italienischen *in un secondo momento* herrührt. Die Architektur eines im Entstehen begriffenen Systems namens *Vis-À-Vis*, das Korpora verschiedener Varietäten miteinander vergleicht und halbautomatisch Unterschiede auf verschiedenen linguistischen Beschreibungsebenen ermittelt, wird in Anstein (im Druck) gezeigt.

Die Ergebnisse solcher vergleichender Studien ermöglichen es, Entwicklungstendenzen des Deutschen in Südtirol aufzuzeigen und mögliche Defizite frühzeitig zu erkennen. Sie stellen eine Hilfe für sprach- und bildungspolitische Entscheidungsträger dar und können in die Lehrerbildung und Sprachdidaktik einfließen. So kann das Bewusstsein für Unterschiede, für Standard und Norm, bei Lehrenden und Lernenden geschärft bzw. die Analysekompetenz bei Lehrenden und die Normenkompetenz bei Lernenden gefördert werden (vgl. z.B. Egger/Lanthaler 2001; Riehl 1994).

7 Ausblick

In der ersten Arbeitsphase der Initiative Korpus Südtirol konnten die Grundlagen für die Korpusentwicklung geschaffen und erste Anwendungen erprobt werden. Im weiteren Fortgang ist eine umfassende Korpusanreicherung zur gesteigerten Ausgeglichenheit der Daten auf der Grundlage der erstellten Bestandsaufnahme geplant. Mit einem ausgewogenen Subkorpus ist das Korpus Südtirol an der C4-Initiative beteiligt. Zusätzlich sollen weitere Spezialkorpora hinzugefügt werden, z.B. Lernertexte und parallele Korpora. Die Annotationen werden erweitert, auch durch die Anpassung der Annotationswerkzeuge an die Südtiroler Varietät. Des Weiteren werden Instrumente zum systematischen Vergleich von Varietätenkorpora weiterentwickelt und umfassende Analysen durchgeführt. Parallel dazu sollen die Anwendungen in der Sprachdidaktik ausgebaut und die Ausbildung von MultiplikatorInnen vorangetrieben werden.

Mit diesen Schritten möchte die Initiative Korpus Südtirol einen Beitrag zur Spracharchivierung, -dokumentation, -beobachtung, -analyse und -didaktik in einem mehrsprachigen Grenzgebiet leisten und zugleich bei allen Nutzern die Neugier für und die Freude an sprachlichen Phänomenen wecken sowie den spielerischen Umgang mit Sprache anregen.

Literatur

- Abel, Andrea (2009): "Mehrsprachigkeit in Südtirol: Alles bleibt anders?". *Geographische Rundschau* 3: 20–27.
- Abel, Andrea/Anstein, Stefanie (2008): "Approaches to Computational Lexicography for German Varieties". In: Bernal, Elisenda/DeCesaris, Janet (eds.): *Proceedings of the XIIIth Euralex International Congress*. Barcelona, Universitat Pompeu Fabra: 251–260.
- Abfalterer, Heidemaria (2005): *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht. Lexikalisch-semantische Besonderheiten im Standarddeutsch Südtirols*. Dissertationsschrift, Universität Innsbruck.
- Ammon, Ulrich (1995): *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten*. Berlin/New York: de Gruyter.
- Ammon, Ulrich et al. (2004): *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin: de Gruyter.

- Anstein, Stefanie (im Druck): "VIS-À-VIS – a System for the Comparison of Linguistic Varieties on the Basis of Corpora". In: *Proceedings of LULCL II*. Bozen.
- Autonome Provinz Bozen-Südtirol (ed.) (2006): *Das neue Autonomiestatut*. Bozen (Landespresseamt).
- Baroni, Marco/Kilgarriff, Adam (2006): "Large linguistically-processed Web corpora for multiple languages". *11th Conference Companion of EACL 2006, East Stroudsburg PA: ACL*: 87–90. http://clic.cimec.unitn.it/marco/publications/eacl2006/dewac_eacl.pdf.
- Christ, Oliver (1994): "A Modular and Flexible Architecture for an Integrated Corpus Query System". In: *Proceedings of COMPLEX 1994, 3rd Conference on Computational Lexicography and Text Research*. Budapest: 23–32.
- Egger, Kurt/Lanthaler, Franz (2001) (eds.): *Die deutsche Sprache in Südtirol. Einheitssprache oder regionale Vielfalt*. Wien/Bozen: Folio.
- Evert, Stefan (2005): "The CQP query language tutorial". Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench>.
- Kermes, Hannah (2003): "Off-line (and On-line) Text Analysis for Computational Lexicography". Dissertationsschrift, Universität Stuttgart.
- Lanthaler, Franz/Saxalber, Annemarie (1995): "Die deutsche Standardsprache in Südtirol". In: Muhr, Rudolf et al. (eds.): *Österreichisches Deutsch. Linguistische, sozialpsychologische und sprachliche Aspekte einer nationalen Variante des Deutschen*. Wien: Hölder-Pichler-Tempsky: 289–305.
- Lemnitzer, Lothar/Zinsmeister, Heike (2006): *Korpuslinguistik*. Tübingen: Narr.
- Riehl, Claudia Maria (1994): "Das Problem von 'Standard' und 'Norm' am Beispiel der deutschsprachigen Minderheit in Südtirol". In: Helfrich, Uta/Riehl, Claudia Maria (eds.): *Mehrsprachigkeit in Europa – Hindernis oder Chance?* Wilhelmsfeld, Egert: 149–164.
- Schmid, Helmut (1994): "Probabilistic Part-of-Speech Tagging Using Decision Trees". In: *Proceedings of International Conference on New Methods in Language Processing*. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger>.
- Zanin, Renata (2007): "Korpusinstrumente für Deutsch als Zweitsprache". In: Krumm, Hans-Jürgen/Portmann-Tselikas, Paul R. (eds.): *Theorie und Praxis. Österreichische Beiträge zu Deutsch als Fremdsprache 10, 2006. Schwerpunkt: Aufgaben*. Innsbruck: Studienverlag.