[anko] – Ansichtskartenkorpus
Entstehung, Aufbereitung und Anwendung

Kyoko Sugisaki, Nicolas Wiedmer, Selena Calleri

1 Einleitung

Dieser Artikel stellt das Ansichtskartenkorpus ([anko]) vor, ein annotiertes Korpus standarddeutscher und schweizerdeutscher Ansichtskarten. Das Ansichtskartenkorpus wurde manuell transkribiert, wobei Textbausteine (Gruß, Anrede usw.) annotiert und Metadaten (Ort des Schreibens, Datum usw.) erfasst wurden. Wir nutzen Technologien der natürlichen Sprachverarbeitung zur Lemmatisierung und Annotation mit Wortarteninformationen. In unserem Beitrag wollen wir auf die Herausforderungen der Digitalisierung eines handgeschriebenen Textes eingehen. Wir zeigen anhand einiger Fallbeispiele aus den Beiträgen dieses Sammelbandes die Eignung des Korpus für die Forschung in der Kultur- und Sprachwissenschaft auf.

Abb. 1a und 1b und 1c: Einsendung und Sammlung

Quelle: Privatbesitz (K. Sugisaki)

Der Beitrag beginnt mit dem Entwicklungsprozess des Korpus (Kapitel 2) von der Datensammlung bis hin zur Datenbank und den Zugriffsmöglichkeiten. In Kapitel 3 wird die Anwendung des Korpus in den Beiträgen dieses Sammelbandes zusammengefasst.

2 Von der Einsendung zur Datenbank

2.1 Einsendung und Sammlung

Die rund 12.000 Ansichtskarten, die als Datengrundlage für [anko] dienen, wurden seit 2009 nach einem Aufruf von Heiko Hausendorf von Privatpersonen für die linguistische Forschung gespendet. Die Karten wurden zum Teil durch sein Zürcher Forschungsteam in Privathaushalten abgeholt und zum Teil per Post zugestellt. Die Karten wurden nummeriert und in Archivschachteln am Deutschen Seminar an der Universität Zürich archiviert (Abbildung 1). Zu Beginn des durch den SNF (Schweizerischen Nationalfonds) und die DFG (Deutsche Forschungsgemeinschaft) geförderten Forschungsprojekts[1] wurden sie manuell eingescannt. Jeder digitalisierten Karte wurden zwei Identifikationsnummern, eine für die Schriftseite und eine für die Bildseite, zugeordnet. Die Nummerierung folgt dabei der Ablage in den Archivschachteln und hat darüber hinaus keine Bedeutung. Das Ansichtskartenkorpus wurde aus diesen Karten zusammengestellt.

Für die Kodierung wurde ein digitales Annotationstool (Abbildung 2) entwickelt. Dieses wurde für die Kartenauswahl (s. unten), die Transkription sowie für die manuelle Erfassung der Textbausteine und Metadaten eingesetzt. Im Annotationstool wurde jede gescannte Karte auf der Web-Oberfläche dargestellt und konnte so direkt abgeschrieben und erfasst werden. Diese Aufgabe wurde durch ein Schreibbüro in Deutschland anhand einer durch das Projektteam erstellten Gebrauchsanweisung (s. Anhang in diesem Sammelband) durchgeführt. Mit dem Annotationstool sollte eine hohe Effizienz und Effektivität des Schreibbüros sichergestellt und Qualitätskontrollen ermöglicht werden. Dabei wurde das Annotationstool als dynamische Web-Technologie implementiert, wodurch eine Kontrolle in Echtzeit, eine rasche Problembeseitigung und der Support des Schreibbüros gewährleistet werden konnten. Die Qualitätssicherung erfolgte durch studentische Mitarbeiter*innen in Zürich und Dresden. Es wurden einerseits Stichproben durchgeführt und andererseits maschinelle Validierungen der Annotationen vorgenommen.[2] Rund 12.000 in Standarddeutsch verfasste Ansichtskarten konnten so erfasst werden. Etwa 600 Ansichtskarten waren in Schweizerdeutsch verfasst. Aufgrund der Komplexität, die mit dem Erfassen von schriftsprachlichem Schweizerdeutsch einhergeht, wurden diese Karten gesondert an der Universität Zürich abgeschrieben,

1 »Textsortenentwicklung zwischen Standardisierung und Variation: Das Beispiel der Ansichtskarte. Text- und korpuslinguistische Untersuchungen zur Musterhaftigkeit privater Fern- und Alltagsschriftlichkeit« (SNF Nr. 160238) ist ein durch den SNF und die DFG gefördertes Kooperationsprojekt der Universität Zürich und der Technischen Universität Dresden und wurde von Prof. Dr. Heiko Hausendorf (Universität Zürich) und Prof. Dr. Joachim Scharloth (TU Dresden) geleitet.

2 Die Qualitätskontrolle wurde von Maaike Kellenberger (Universität Zürich), David Koch (Universität Zürich) und Jan Langenhorst (TU Dresden) durchgeführt und technisch von Selena Calleri (Universität Zürich) unterstützt.

wodurch ein zusätzliches Sub-Korpus entstanden ist [anko Schweizerdeutsch]. Die aus diesem Transkriptions- und Annotationsprozess entstandenen Daten wurden in einer MySQL-Datenbank gespeichert und dienten als Grundlage für die Erstellung einer XML-Datenbank sowie eines Online-Korpusabfragesystems (Kapitel 3).

2.2 Korpus-Kodierung und manuelle Annotation

Abb. 2: Transkriptions- und Annotationstool

Quelle: https://postcards.linguistik.uzh.ch

2.2.1 Kartenauswahl

Die digitalisierten Ansichtskarten wurden nach den folgenden Kriterien unterschieden und weiter verarbeitet oder aussortiert:

- *Standarddeutsche oder schweizerdeutsche Karten:* Das Entscheidungskriterium für die Zuteilung zu einer dieser Kategorien war, in welcher Sprache der überwiegende Teil des Textmaterials verfasst wurde. In vielen Karten wurden einzelne schweizerdeutsche Anreden oder Grußelemente verwendet, wobei der Haupttext in Standarddeutsch verfasst wurde.
- *Urlaubskarten:* Da im Projekt die Kommunikation aus dem Urlaub mittels des Mediums Ansichtskarte im Fokus stand, wurden nur Karten mit einem Urlaubsbezug ins Korpus aufgenommen. Geburtstags- und Weihnachtskarten ohne einen solchen Bezug wurden daher aussortiert.

- *Karten mit Ansicht:* Der Untersuchungsgegenstand im Forschungsprojekt war die klassische Ansichtskarte mit einer Bild- und einer Schriftseite. Exemplare der Vorgängerin der Ansichtskarte, d.h. Postkarten ohne Bildseite, wurden aussortiert.
- *Gesendete Karten:* Da gerade der kommunikative Charakter der privaten Fernkommunikation mittels Ansichtskarte für die Forschungsfragen im Projekt zentral war, wurden Karten, die nicht versendet wurden, also solche mit leerem Adressfeld oder Karten, die aufgrund der Motive auf der Bildseite gesammelt wurden, aussortiert.

Die Karten, welche nicht als Datengrundlage des Forschungsprojekts geeignet waren, wurden zwar aussortiert, sind jedoch im Deutschen Seminar der Universität Zürich archiviert und können somit für andere Forschungsfragen in Nachfolgeprojekten genutzt werden.

2.2.2 Transkription

Die Mitteilungen wurden nach den folgenden Vorgaben für die Schrift- und Bildseite transkribiert:

- Keine orthographische Korrektur: Handgeschriebenes und Gedrucktes wurde orthographisch exakt so abgeschrieben, wie es auf der Karte ersichtlich war.
- Unleserliche Wörter oder Passagen wurden mittels der Markierung *[unclear]* gekennzeichnet.
- Zeichnungen oder komplexe Zeichen, die Teil eines Fließtextes sind, wurden so abgeschrieben, wie man sie vorlesen würde: z.B. wurde *8ung!* als »Achtung!« und ♥*Grüsse* als »Herzliche Grüsse« abgeschrieben. Abtippbare Emoticons, z.B. *:-)* wurden mittels Interpunktionszeichen und Klammern transkribiert.
- Die Anonymisierung wurde zur Sicherstellung des Datenschutzes direkt während der Transkription durchgeführt. Nachnamen wurden durch die Kodierung *[NN]* und vertrauliche Informationen (wie Mailadressen, Kontodaten, Telefonnummern) durch *[vertraulich]* ersetzt.
- Zeilenumbrüche wurden mittels einfachen Zeilenvorschubs markiert.
- Absatzumbrüche wurden mittels doppelten Zeilenvorschubs markiert. Dabei wurde alles, was sich visuell als Texteinheit abgrenzen ließ, als eigener Absatz markiert. So wurden eine Randnotiz oder eine Einrückung am Anfang oder in der Mitte einer Zeile als Absatz markiert.

2.2.3 Annotation der Textbausteine

Im Textkörper wurden während des Abschreibprozesses die folgenden immer wieder auftauchenden Textbausteinkategorien manuell annotiert, welche wertvolle Rückschlüsse auf den Wandel der Musterhaftigkeit der Textsorte Ansichtskarte zulassen:

- *Datum:* Auf vielen Ansichtskarten wird direkt am Anfang der Mitteilung oder nach der Ortsangabe (z.B. *Bonton den 10/10/99*) ein Datum angebracht.
- *Anrede:* Die meisten Ansichtskarten weisen eine Anrede auf, z.B. *Liebe Erika*.

- *Gruß:* Ebenfalls musterhaft für die Textsorte Ansichtskarte ist der Gruß bzw. eine Grußformel, welche(r) meistens vor der Unterschrift platziert ist, z.B. *viele Grüße* oder *bis bald*.
- *Unterschrift:* Fast jede Mitteilung weist eine Unterschrift auf. Dadurch wird die Karte »signiert« bzw. die Urheberschaft angegeben, was für die Pflege des sozialen Kontaktes zwischen Autor*in und Empfänger*in wesentlich ist.
- Die Annotationen wurden direkt im Text mit Hilfe von Markdowns gekennzeichnet. Anstatt eckige XML-Klammern haben wir spezielle Zeichenketten (Markdowns) für jede Textbausteinkategorie entwickelt, damit die Annotation dank besserer Lesbarkeit schnell mit geringerem Aufwand für das Schreibbüro durchgeführt werden konnte. Zum Beispiel wurde eine Anrede mit *|*Hoi Heidi*|* gekennzeichnet. Im Annotationstool wurde die Validität der Markdowns automatisch geprüft und farblich markiert dargestellt (erst nach einer validen Annotation erschien eine Anrede oder ein Gruß in der entsprechenden Farbe). So konnten die Markdowns leicht in korrekte XML-Tags transformiert werden und Transkribierende hatten ein direktes visuelles Feedback über die Validität der Annotation.

2.2.4 Annotation der Metadaten

Die folgenden Metadaten wurden während des Abschreibeprozesses anhand des Adressfelds, des Textkörpers, gedruckter Einheiten und des Poststempels manuell eingegeben:

- Autor*innen: Schreibende (Geschlecht und Anzahl), Ortschaft (Region und Land) und Datum (Poststempel oder handschriftliches Datum)
- Empfänger*innen: Angeschriebene (Anrede, Vornamen, Nachnahmen, Geschlecht und Anzahl), Ortschaft (Postleitzahl, Region und Land)
- Das Vorhandensein von Zeichnungen und Emoticons jeglicher Art wurde binär, das heißt als vorhanden oder nicht vorhanden, angegeben.

2.3 Automatische Annotation

Transkriptionen, Annotationen und Metadaten wurden in einer MySQL-Datenbank gespeichert und im Anschluss in eine XML-Datenbank umgewandelt. Die MySQL-Datenbankeinträge wurden als JSON exportiert und in eine XML-Datenbank transformiert. Dabei wurden folgende maschinelle Verarbeitungsschritte durchgeführt: Textsegmentierung, Part-of-Speech-Tagging, Lemmatisierung und Kompositadekomposition. Die einzelnen Schritte werden in den folgenden Kapiteln erklärt.

2.3.1 Preprocessing und Annotationsfehler

Um die mehrheitlich automatisierten Analyse- und Verarbeitungsprozesse durchführen zu können, wurde die Transkription einerseits manuell und andererseits automatisch überprüft und falls nötig korrigiert. Die häufigsten Fehler waren unabgeschlossene Markdown-Markierungen, welche durch ein Skript identifiziert und korrigiert wurden. Außerdem wurden die Metadaten in XML-Tags umgewandelt und/oder als XML-Attribute abgespeichert, damit sie durchsuchbar und abfragbar sind. So wurde z.B. [unclear]

in <unclear/> umgewandelt, um interpretative Annotationen vom eigentlichen Text unterscheiden zu können.

2.3.2 Textsegmentierung und Tokenisierung

Der Mitteilungstext wurde automatisch in die Segmente Absatz, Satz, und Wort zerlegt, damit sie durch eine XML-Repräsentation strukturiert werden konnten. Der Absatz wurde bereits während des Transkriptionsprozesses durch einfachen Zeilenvorschub markiert und konnte somit maschinell übernommen werden. Um automatisch einzelne Sätze und Wörter aus Absätzen segmentieren zu können, wurde ein eigenes CRF-Textsegmentierungssystem (CRF, für *Conditional Random Fields*) entwickelt, wobei die Eigentümlichkeiten der unredigierten handgeschriebenen Texte automatisch angepasst wurden (vgl. Sugisaki 2017). Für die Entwicklung wurde ein Zeitungskorpus TüBA D/Z (Tübinger Baumbank des Deutschen Zeitungskorpus, das Korpus von der *tageszeitung* (taz), vgl. Telljohann 2015) als Trainingsdaten eingesetzt und experimentell das beste Segmentations-CRF-Modell für die Ansichtskarten herausgearbeitet.

Während der Systementwicklung der automatischen Textsegmentierung wurde insbesondere beachtet, wie die Schreibweise auf Ansichtskarten sich von derjenigen in Zeitungstexten unterscheidet. Dies betrifft insbesondere die Verwendung von Interpunktionszeichen wie Punkt, Strichpunkt, Ausrufezeichen und Fragezeichen. Diese können auf Ansichtskarten nicht nur am Ende eines Satzes, sondern auch in der Mitte (wie z.B. in *Sat.1*) vorkommen. Daher sind sie für die automatische Textsegmentierung von hoher Relevanz. Konkret wurde für die Implementation berücksichtigt, dass ein paar Dutzend Abkürzungseigentümlichkeiten wie *herzl.* vorhanden sind. Ähnlich wie in anderen Privatkommunikationstexten ist die Kommunikation auf Ansichtskarten außerdem durch Smileys wie *;)* oder Wiederholungen wie *!!!!* geprägt. Solche Sprachphänomene sind in Zeitungstexten, welche der Standardorthographie folgen, praktisch nie vorhanden.

Abb. 3: Satzsegmentierungsproblem

Quelle: [anko] 90133

Außerdem haben wir festgestellt, dass eine automatische Satzsegmentierung von handschriftlichen Texten keine triviale Aufgabe ist – so kommt es auf Ansichtskarten immer wieder vor, dass ein Satz weder mit einem Interpunktionszeichen beendet ist noch eine genügend lange Einrückung zwischen Sätzen aufweist. Ein entsprechendes Beispiel ist in Abb. 3 dargestellt. Durch die Berücksichtigung der Textbausteinannotation konnte für die [anko]-Testdaten in der automatisierten Textsegmentation eine höhere Präzision (F1 Score: 0.96) erlangt werden als in einem bereits bestehenden System (PUNKT vgl. Kiss & Strunk 2006). Der F1-Wert wird durch die Kombination der Genauigkeit und der Trefferquote berechnet. Die detaillierte Evaluation ist in Sugisaki 2017 dargestellt.

2.3.3 Part-of-Speech-Tagging und Lemmatisierung

Für die Qualitätserhöhung des Part-of-Speech-Tagging für Ansichtskarten wurde ein eigener PoS-Tagger entwickelt (vgl. Sugisaki, Wiedmer & Hausendorf 2018). Analog zur automatischen Textsegmentierung haben wir ein Zeitungskorpus als Trainingsdaten verwendet. Prototypisch hat eine Textsorte eine eigene Verteilung der Part-of-Speech Tags. So wird z.B. *Liebe* auf Ansichtskarten häufig nicht als Nomen, wie in einem Zeitungstext, sondern als Adjektiv (*Liebe Erika*) in den Ansichtskarten benutzt. Um die domänenspezifische Verteilung zu umgehen, wurde ein manuell annotiertes Subkorpus von [anko] als Trainingsdaten für das CRF-Modell eingesetzt, als Zusatz zu den bestehenden handannotierten Zeitungskorpora wie TüBa D/Z oder Noah (für Schweizerdeutsch, vgl. Hollenstein & Aeppli 2014). Um die Präzision des Taggers zu erhöhen, wurden auch die Ausgaben der bestehenden Part-of-Speech-Tagger (TreeTagger vgl. Schmidt 1999 und Stanford Tagger, vgl. Toutanova et al. 2003) als Feature berücksichtigt. Unser Part-of-Speech-Tagger erlangt einen F1-Wert von 0.93, wobei der TreeTagger allein nur den F1-Wert von 0.86 erreicht hat. Für die Lemmatisierung haben wir jedoch den TreeTagger benutzt.

2.3.4 Komposita-Dekomposition

Für die Komposita-Dekomposition wurden die Nomen-Nomen-Komposita automatisch in ihre Morpheme zerlegt (vgl. Sugisaki & Tuggener 2018). Die Komposita sind sehr häufig in der deutschen Sprache und die Zerlegung bringt eine Vermehrung des Sprachmaterials und somit eine Erhöhung der Auffindbarkeit der einzelnen Kompositakomponenten mit sich. Um die Komposita automatisch zu zerlegen, wurde die Produktivität der Morpheme im Web-Korpus (SdeWaC, vgl. Faaß & Eckart 2013) gemessen und diese für die Gewichtung der Zerlegungsalternativen (wie z.B. *Hundesteuer* als *Hunde|steuer* oder *Hundes|teuer*) eingesetzt. Das entwickelte System erreichte den F1-Wert von 0.92.

2.4 Korpus als Forschungsressource[3]

Das Korpus wurde für die Nutzung des Korpus in zwei XML-Formate gebracht. Eines galt als Basis für die Indexierung in CQPweb (Hardie 2012) und eines als Basis für die

3 Die Implementierung, die in Kapitel 2.4 beschrieben wird, wurde durch Josephine Devi Obert (TU Dresden) durchgeführt.

Erstellung des Korpus im TEI-Format. Grundsätzlich unterscheiden sich die beiden Formate nur in der Struktur. Die beiden Formate werden auch parallel durch ein Skript aus den Daten der Datenbank erstellt.

2.4.1 Korpusabfragesystem

Als Korpusabfragesystem haben wir CQPweb (Hardie 2012) online für Forschende zur Verfügung gestellt.[4] CQPweb ist ein Konkordanz-Programm zur Korpusanalyse. Ein Suchergebnis-Beispiel mit der CQPweb Engine ist in Abbildung 4 dargestellt. Mit CQPweb kann ein Nutzer nicht nur nach der exakten Wortform, sondern auch nach spezifischen Textbausteinen (z.B. <s_disc=»gruss«> []+ </s_disc>) oder nach Lemmata (z.B. [lemma=»Hallo«]) und Part-of-Speech-Tags (z.B. [pos=»NN«]) suchen. Ebenfalls möglich ist eine Metadatensuche (z.B. <text_zielland=»Schweiz«>[]+</text_zielland>). Außerdem bietet das Tool nützliche Funktionen an wie Keywordanalysen und die Erstellung von Subkorpora für linguistische Vergleiche mit einer grafischen Benutzeroberfläche.

Abb. 4: CQPweb

Your query "Hallo" returned 499 matches in 490 different texts (in 614,685 words [12,289 texts]; frequency: 811.80 instances per million words) [0.158 seconds - retrieved from cache]

I< << >> >I Show Page: 1 Line View Show in random order New query Go!

No	Filename	Solution 1 to 50 Page 1 / 10		
1	101411	Overland - Gruppe unterwegs . Wir zelten bei unter 10 Grad lunclearl	Hallo	Rita . Viele Grüsse von unserer Flusskreuzfahrt durch Holland - noch !
2	160973	Kartenschreiben ! Liebe Grüsse , Ingrid + Dieti Samstag 17. 06. 2006	Hallo	Konrad Liebe Grüsse von unserer Rekognoszierung La Foldaz - Chamouix wir müssen
3	100729	und gemütliche Weihnachtstage . liebe Grüsse aus Savognin Priska lNNl 18. 7.	Hallo	zäme , wir waren am 15. 7. in einem ganz lässigen Wasserpark
4	302081	die sie zur Hochzeit bei uns zubrachte . Die besten Grüsse lunclearl	Hallo	Til Ich hoffe , Du hast die Ferien noch gut überstanden .
5	61225	Fr. 5 Ich war nochmals auf dem Weisshorn ! Alice + Frieder	Hallo	Ella es ist eine schöne Töffreise aber anstrengend . Da an der
6	301045	Strapazen keimkehren darfst . Mit lb. Gruss , küsst Dich Dein Clärli	Hallo	Ihr drei Zuhausegebliebenen ! Nun haben wir fast die Hälfte der Reise
7	220181	Dank und aus den Ferien die besten Grüsse I. lNNl u. Frau	Hallo	lieber Götli , Liebes Gotli , Heute waren meine ganze Familie ,
8	151037	Chur und es schaukelt heftig ! liebe Grüsse aus den Alpen Carla	Hallo	, hallo Gotti es Schneit es hat schon fast ein halben meter
9	151037	es schaukelt heftig ! liebe Grüsse aus den Alpen Carla Hallo ,	hallo	Gotti es Schneit es hat schon fast ein halben meter geschneit .
10	60981	bei angenehmen 29° ! Viele liebe Grüsse Eure Michelle Barbara lunclearl lunclearl	Hallo	Frau lNNl Ich habe es schön in Gaorle von NATASCHA Peter Resula
11	170513	liebe Grüsse von uns allen . Jasmine & Familie 24. 02. '15	Hallo	Beatrice ! Viele Grüße aus dem Erzgebirge ! Seit gestern ist wieder
12	302043	lunclearl lunclearl Liebe Ella Herzliche Grüsse aus Ischia Insel der Seele lunclearl	Hallo	Ihr Lieben ! Heute sind wir bei scheußlich schwülem Wetter in Berlin
13	81569	25 - 27°C . Herrlich ! Herzliche Grüße von hier von Nan	Hallo	Da id dä Lenzerheide hämmer immer schön's Wätter (7°C.) ,
14	300559	. auch an der Seine ist's schön	Hallo	, liebe Monika , Wiesbaden grüßt dich und dankt für die Urlaubsgrüße
15	101303	konnte . Herzliche Grüsse , Johannes lNNl Killarney , 14. 9. 04	Hallo	Momami ! Üs zwöi gods super do in Irland , zwösche grüener
16	81129	9. und 18. Juni bin ich in der Schweiz die lunclearl einer	Hallo	Kevin Wie immer haben wir natürlich kaum Zeit zum Schreiben . Also
17	300375	Herz - iche Grüsse 15. 7. 1979	Hallo	, Ihr Lieben , wir senden frohe Grüße aus der Rhön ,
18	81533	in Indonesien . Herrlicher Sandstrand und schöne Hotels . Joe + Suri	Hallo	Sarah , Hier in den Skiferien ist sehr schönes Wetter . Auch
19	210637	weiter , nach Luino . Auf ein fröhliches Wiedersehen hoffend . Thues	Hallo	zämä Ich schicke eu mega heissi Feriegrüess us Italie ! Venedig isch
20	20353	phantastisch schön . Viele Grüße an alle dort Werner 30. 08. 2000	Hallo	liebe Frau lNNl Aus einer kleinen Wochenreise von L'Auvergne senden wir Ihnen
21	100575	Von meinen Skiferien in Saas Fee sende ich Dir sonnige Grüsse Thesi	Hallo	Ihr zwei ! Sitze zur Zeit mit Tantchen in der Orangerie gleich
22	301047	Saas - Fee senden Malon , Bibi und Remo Herzl. Grüße Gitti	Hallo	Ihr Lieben Waldemar + ich sind heute in das wunderschöne Queldinburg gefahren
23	100717	einmalige Gegend . Herzliche Grüsse Rosmarie + lunclearl Claudia + Kerim Martin	Hallo	Rita u. Yvo Nochmals ein grössens Dankeschön für die SBB Gutscheine .

Quelle: https://pub.cl.uzh.ch/service/cqpweb/

4 Die Zugangsinformationen finden sich unter: https://www.ds.uzh.ch/de/projekte/ansichtskarten-projekt/

2.4.2 Korpus im TEI-XML-Format

Zusätzlich zum Online-Zugriff liegt [anko] zum Zeitpunkt der Veröffentlichung in einem XML-Format vor[5]. Das XML-Format folgt dem Standard der Text Encoding Initiative (TEI). Das TEI-XML-Format ist ein Standardformat für den Austausch der Textdatenbank und erleichtert anderen Forschenden in der Sprachwissenschaft, aber auch anderen Geisteswissenschaften wie Literaturwissenschaft und Geschichte, mit den Daten zu arbeiten.

In der XML-Datenbank ist die einzelne Ansichtskarte so strukturiert, dass für sie zwei verschiedene Datentypen auffindbar sind:

- **Metadaten**: Dies sind Informationen zum Text, die den beschriebenen Metadatenkategorien (Sektion 2.2.4) entsprechen.
- **Textdateien**: Dies ist der Mitteilungstext, der abgeschrieben und mit der Annotation (Sektion 2.2.3) ausgestattet wurde.

Eine solche TEI-XML-Datei ist in Abbildung 5 dargestellt. Die gesamte XML-Datei ist für jede Ansichtskarte separat (XML-Element *TEI*) gespeichert. Für jede Karte sind die Metadaten (XML-Element *teiHeader*) und der transkribierte Mitteilungstext (XML-Element *text*) vorhanden.

Als Metadaten sind neben Sprache (XML-Element *language*) die handerfassten Informationen über Sender und Empfänger (XML-Element *correspDesc*) nach den Gliederungs- und Benennungsstandards der XML-Elemente im TEI-Standard gespeichert.

Der Mitteilungstext ist mit der Angabe der Bild- oder Schriftseite (XML-Element *div* mit den Attributen *type*, dessen Wert als *recto* und *verso* angegeben wird) versehen, um die Schreibseite zu repräsentieren; für Absätze wurde das XML-Element *p* (Paragraph) und für Zeilen das XML-Element *ab* mit Kombination von *lb* (line break) verwendet. Die Textbausteine sind als Attribute (*type*) mit den Werten *greeting*, *signed*, *salutation* oder *dateline* in die Zeilen eingebettet, welche ursprünglich die Markdowns der Transkription waren.

Ziel aller Annotationen war es, sowohl strukturelle als auch formatspezifische Merkmale zu codieren und abfragbar zu machen. Die verschiedenen Zugänge zu den Daten (CQPweb und TEI) haben es ermöglicht, dass Forschende mit unterschiedlichen disziplinären Hintergründen und Kenntnissen mit den Daten arbeiten können und diese ohne größere Hürden zugänglich sind, da in beiden Fällen disziplinäre Standards befolgt wurden.

5 Auch diese Zugangsinformationen finden sich unter: https://www.ds.uzh.ch/de/projekte/ansichtskartenprojekt/

Abbildung 5a und 5b: Beispiel TEI-XML

```
<teiHeader> [21 lines]
<TEI>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>10001</title>
      </titleStmt>
      <publicationStmt>
        <publisher>
          <orgName role="hostingInstitution">Deutsches Seminar, Universität
            Zürich,Schweiz</orgName>
          <orgName role="hostingInstitution">Institut für Germanistik, Technische Universität
            Dresden, Deutschland</orgName>
        </publisher>
        <date>2017</date>
      </publicationStmt>
      <sourceDesc>
        <p>Korpus des Forschungsprojekts "Textsortenentwicklung zwischen Standardisierung und
          Variation: Das Beispiel der Ansichtskarte. Text- und korpuslinguistische Untersuchungen
          zur Musterhaftigkeit privater Fern- und Alltagsschriftlichkeit." Die hier aufgeführten
          Ansichtskarten wurden nach einem Aufruf als Spende an Prof. Hausendorf, Uni Zürich
          geschickt.</p>
      </sourceDesc>
    </fileDesc>
    <profileDesc>
      <langUsage>
        <language ident="ger">Standard German</language>
      </langUsage>
      <handNotes>
        <handNote>
          <figure style="standard" type="recto"/>
          <figure style="standard" type="verso"/>
        </handNote>
      </handNotes>
      <creation>
        <date>1971-03-02</date>
      </creation>
      <settingDesc>
        <setting>
          <p/>
        </setting>
      </settingDesc>
      <correspDesc>
        <correspAction type="sent">
          <state type="sex">
            <desc>M</desc>
          </state>
          <state type="number">
            <desc>1</desc>
          </state>
          <location>
            <settlement type="city">Savognin</settlement>
            <country key="CHE"/>
          </location>
        </correspAction>
        <correspAction type="received">
          <state type="sex">
            <desc>Unknown</desc>
          </state>
          <state type="number">
            <desc>unknown</desc>
          </state>
          <roleName type="honorific">FAM</roleName>
          <name>          </name>
          <address>
            <postCode>8006</postCode>
            <location>
              <settlement type="city">Zürich</settlement>
              <country key="CHE"/>
            </location>
          </address>
        </correspAction>
      </correspDesc>
    </profileDesc>
  </teiHeader>

  <text>
    <body>
      <div1 type="recto">
        <div2 type="p">
          <ab>VON unserem</ab>
          <lb/>
          <ab>Skiausflug</ab>
          <ab type="greeting">viele<lb/>GRüsse sendet<lb/>Euch</ab>
        </div2>
        <div2 type="p">
          <ab type="signed">Matthias</ab>
        </div2>
      </div1>
    </body>
  </text>
</TEI>
```

Quelle: [anko] 10001

3 Anwendung des Korpus

In diesem Kapitel fassen wir zusammen, wie das Korpus [anko] für die Beiträge in diesem Sammelband genutzt wurde. Die folgende Tabelle ist eine Zusammenfassung im Hinblick auf folgende Fragen: (1) Welches Format von [anko] wurde als Zugang zum Korpus benutzt? (2) Was war das Ziel der Untersuchung, für die [anko] als Grundlage diente? (3) Wie wurden die Ansichtskarten im Korpus ausgewertet und/oder analysiert?

Tab. 1: Anwendung des Korpus im Sammelband

Ziel der Untersuchung	**Auswertung/Analyse**
Diekmannshenke: [anko]-Beispielkorpus, private Ansichtskarten	
Ansichtskarte als Text-Bild-Botschaften	Ansichtskarte von der Entstehung bis zu ihren modernen elektronischen Verwandten wie E-Card und E-Mail
Gansel: [anko]-CQP	
Zeit (Rosa 2005; Luhmann (2016[1995]); Nasehi (2008[1993])); Elias (1988))	1) »Zeit« und ihre Kollokationen; 2) »Zeit zum X« und »Zeit um X zu Y«) und ihre exemplarische Analyse; 3) Zeitsemantische Ausdrücke (»vorgestern«, »gestern«, »heute«, »gegenwärtig«, »im Augenblick«, »jetzt«, »morgen«, »übermorgen«, »länger bleiben/noch länger«) exemplarisch analysiert
Hausendorf: [anko]-CQP, private Ansichtskarten	
Lesbarkeit der Ansichtskarte	Lesbarkeit aus der Hand, Lesbarkeit des Feriengrußes, Lesbarkeit der Welt als Sehenswürdigkeit im textlinguistischen Modell (Hausendorf et al. 2017) exemplarisch analysiert
Kellenberger: Korpus [anko], [anko]-CQP	
Formen und Funktionen von Bildverweisen und deren Bedeutung für die Textsorte Ansichtskarte	Manuelle exemplarische Analysen und CQP-Analysen der Verbindung zwischen Bild- und Textraum, Bildverweise explizit, implizit oder fehlend, Musterhaftigkeit der Bildverweise durch Platzierung im Text
Koch: [anko]-CQP	
Konstruktion von »Afrika« als Urlaubsraum in Ansichtskarten (Urlaubsraum, Fremdes/Differenz, Zugehörigkeit)	Exotischer und gewöhnungsbedürftiger Urlaubsraum, Darstellung von fremden Menschen und Armut auf Ansichtskarten aus Afrika exemplarisch analysiert
Langenhorst: [anko]-XML	
Textlänge und Abkürzungen	Verkürzungspraktik korpuslinguistisch analysiert und ihre Ergebnisse visuell dargestellt

Merten: [anko]-CQP	
Stancetaking-Praktik	»Grüße Präposition«, »Nach/nachdem X« und ihre exemplarische Analyse in der kognitiven Grammatik (Langacker 2008)
Müller & Bender: [anko]-XML (Vertikalisiert)	
Kontragenerische Textpraktik (KGTP): KGTP sind »schriftliche Praktiken, mit denen man gegen die Textmustererwartungen anschreibt, die mit einer kommunikativen Gattung verbunden sind«	Ein Subkorpus nach mehreren Suchmustern wie »allerdings«, »jedoch«, »leider«, evaluative Schlüsselwörter, »Arbeit« sowie »Regen« mit der binären Kategorie (KGTP oder nicht) annotiert und qualitativ exemplarisch nach Raible (1992) analysiert
Naef, Wiedmer, Sugisaki: [anko]-XML	
Urlaubsframe/Themen (Wetter, Aktivitäten, Ort, Wissen, Unterkunft, Essen und Trinken, Gefühle, Wahrnehmung, Vorkommnisse, Hin- und Rückreise, Kennenlernen neuer Leute, Urlaubsgrund, Urlaubsart, Extra-Diegetisches)	Ein Subkorpus mit 14 thematischen Kategorien annotiert und Principal Component Analysis (PCA) durchgeführt
Scharloth: Korpus [anko], komplexe n-Gramme	
Identifikation von Textmustern durch Paraphrasenanalyse	Ansichtskartenexte als rituelle Texte, Intertextualität durch Verknüpfung von Paraphrasen, Erstellung komplexer n-Gramme pro Ansichtskarte, Analyse ähnlicher Cluster (Textmuster), Wandel von Textmustern als Indikator soziokulturellen Wandels
Sugisaki: [anko]-XML und Spracherkennung	
Codeswitching (Standarddeutsch vs. Schweizerdeutsch/Urlaubsortssprache)	Codeswitching im textlinguistischen Modell (Hausendorf & Kesselheim (2008) und Hausendorf et al. (2017)) exemplarisch analysiert
Wiedmer: Subkorpus [anko Thema]	
Formulierungsmuster zwischen Selbstdarstellung und Kontaktnützlichkeit – Beschreibung sportlicher Leistung oder »Mitführung« der Leser*in	Manuelle thematische Annotation und exemplarische Analysen von Beschreibungen sportlicher Aktivitäten auf Ansichtskarten; Rekonstruktion des »Urlaubsframes« im Zusammenhang mit Erwartungen an den Urlaubsort
Wolff: [anko]-Beispielkorpus, private Ansichtskarten	
Ansichtskarte aus einer ethnomethodologischen Perspektive	Soziale Lesbarkeit, Sense-Making Maschine, Rezipienten-Orientierung, oder Rezipient-spezifische Gestaltung sowie die Kombinationen des Bilds, Texts und der Briefmarke der Ansichtskarten exemplarisch analysiert

4 Zusammenfassung

In diesem Artikel haben wir den Aufbau und die Nutzungsmöglichkeiten des Korpus [anko] beschrieben. Das Ansichtskartenkorpus beinhaltet ca. 12.000 Ansichtskarten, die digitalisiert, abgeschrieben und mit Textkörper- und Metadaten annotiert wurden. Die automatische linguistische Segmentation (Tokenisierung) und Annotation (Part-of-Speech und Morphologie) wurde speziell für die Textsorte Ansichtskarte angepasst. Das Korpus steht als TEI-XML und für die CQPweb-Suche zur Verfügung.

Zudem haben wir zusammengefasst, wie das [anko]-Korpus in diesem Sammelband für die Untersuchungen genutzt wurde. Es wurde die Vielfalt der korpuslinguistischen, textlinguistischen und kognitivlinguistischen Analysen sowie eine große Breite an Untersuchungsthemen und -ansätzen im Sammelband aufgezeigt. Das Korpus [anko] wird im Jahr 2022 und 2023 veröffentlicht,[6] und es wird erwartet, dass es nicht nur sprachwissenschaftlichen und kulturwissenschaftlichen, sondern auch soziologischen, medienwissenschaftlichen oder kommunikationswissenschaftlichen Forschungen als Datengrundlage dienen kann.

Literatur

Elias, Norbert (1988): Über die Zeit. Arbeiten zur Wissenssoziologie II. Frankfurt a.M.: Suhrkamp.

Faaß, Gertrud /Eckart, Kerstin (2013). SdeWaC – a corpus of parsable sentences from the web. In Iryna Gurevych/Chris Biemann/Torsten Zesch (Hg.): Language Processing and Knowledge in the Web. Springer Berlin Heidelberg: Berlin, Heidelberg, 61–68.

Hardie, Andrew (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17 (3), 380–409.

Hausendorf, Heiko/Kesselheim, Wolfgang (2008): Textlinguistik fürs Examen. Göttingen: Vandenhoeck & Ruprecht.

Hausendorf, Heiko/Kesselheim, Wolfgang/Kato, Hiloko/Breitholz, Martina (2017): Textkommunikation. Ein textlinguistischer Neuansatz zur Theorie und Empirie der Kommunikation mit und durch Schrift. Berlin/Boston: De Gruyter.

Hollenstein, Nora/Aepli, Noemi (2014): Compilation of a Swiss German dialect corpus and its application to PoS tagging. In Marcos Zampieri/Liling Tan/Nikola Ljubešić/Jörg Tiedemann (Hg.): Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects. 85–94

Kiss, Tibor/Strunk, Jan (2006). Unsupervised multilingual sentence boundary detection. In: Computational Linguistics, 32 (4), 485–525.

Langacker, Ronald W. (2008): Cognitive Grammar. A basic introduction. New York: Oxford University Press.

6 Die Zugangsinformationen finden sich unter: https://www.ds.uzh.ch/de/projekte/ansichtskartenprojekt/

Luhmann, Niklas (2016 [1995]): Protestbewegungen. In: Niklas Luhmann: Protest. Systemtheorie und soziale Bewegungen. Hg. und eingeleitet von Kai-Uwe Hellmann. Frankfurt a.M.: Suhrkamp, 201–215.

Nassehi, Armin (2008[1993]): Die Zeit der Gesellschaft. Auf dem Weg zu einer soziologischen Theorie der Zeit. Neuauflage mit einem Beitrag »Gegenwarten«. Wiesbaden: Verlag für Sozialwissenschaften.

Raible, Wolfgang (1992): Junktion. Eine Dimension der Sprache und ihre Realisierungsformen zwischen Aggregation und Integration, Heidelberg: Winter (Sitzungsberichte der Heidelberger Akademie der Wissenschaften. phil.-hist. Klasse, Jg. 1992, Bericht 2).

Rosa, Hartmut (2005): Beschleunigung. Die Veränderung der Zeitstrukturen in der Moderne. Frankfurt a.M.: Suhrkamp Taschenbuch.

Schmid, Hermut (1999). Improvements in part-of-speech tagging with an application to German. In Susan Armstrong/Kenneth Church/Pierre Isabelle/Sandra Manzi/Tzoukermann Evelyne/David Yarowsky (Hg.): Natural Language Processing Using Very Large Corpora, Kluwer Academic Publishers: Dordrecht, 13–26.

Sugisaki, Kyoko (2017): Word and sentence segmentation in German: Overcoming idiosyncrasies in the use of punctuation in private communication. In: Rehm Georg/Declerck Thierry (Hg.): Language Technologies for the Challenges of the Digital Age, Cham:Springer International Publishing, 62–71.

Sugisaki, Kyoko/Don, Tuggener (2018): German compound splitting using the compound productivity of morphemes. In: Adrien Barbaresi/Hanno Biber/Friedrich Neubarth/Rainer Osswald (Hg.): Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018), 141–147.

Sugisaki, Kyoko/Wiedmer, Nicolas/Hausendorf, Heiko (2018): ANKO – a picture postcard corpus: Transcription, annotation and part-of-speech tagging. In: Nicoletta Calzolari et al. (Hg.): Proceeding of the 11th International Conference on Language Resources and Evaluation (LREC'18), 255–259.

Telljohann, Heike; Hinrichs, Erhard W./Kübler, Sandra/Zinsmeister, Heike/Beck, Kathrin (2015): Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Technical Report, Universität Tübingen.

Toutanova, Kristina/Klein, Dan/Manning, Christopher D./Singer, Yoram (2003): Feature-rich part-of-speech tagging with a cyclic dependency network. In: Marti Hearst/Mari Ostendorf (Hg.): Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL ›03). 173–180.