SeiKo

Korpus Seiteneinsteiger:innen in Vorbereitungsklassen

Handbuch zu Korpusdesign und Datenaufbereitung

Julia Schlauch, Aylin Braunewell, Jana Gamper

https://www.uni-giessen.de/seiko

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Deutsch als Zweit- und Fremdsprache

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Vorwort

Das vorliegende Handbuch beschreibt das Seiteneinsteiger:innen-Korpus (SeiKo), sein Design, die Datenerhebungen sowie das Vorgehen der Datenaufbereitung und die Korpusnutzung.

SeiKo[1] wurde an der Justus-Liebig-Universität Gießen im Zeitraum von 2020 bis 2026 erstellt und von 2023 bis 2026 von der Deutschen Forschungsgemeinschaft (DFG) im Rahmen des Projekts "Erwerb und Ausbau von Nominalgruppen durch neu zugewanderte jugendliche Lerner:innen" (Projekt-Nr. 527339472, Leitung: Jana Gamper) unterstützt. Ziel des Korpus ist es, die Entwicklung der sprachlichen Fähigkeiten von neu zugewanderten Schüler:innen zu modellieren. Ein besonderer Schwerpunkt der Annotationen liegt dabei auf dem Ausbau von Nominalgruppen sowie der Verbstellung.

An der Korpuserstellung beteiligt waren:

- Prof. Dr. Jana Gamper (Projektleitung)
- Julia Schlauch (Wissenschaftliche Mitarbeit, Korpuskonzeption, Datenerhebung, Datenaufbereitung, Qualitätssicherung, Dokumentation)
- Aylin Braunewell (Wissenschaftliche Mitarbeit, Datenaufbereitung, Qualitätssicherung, Dokumentation)
- Martin Klotz (Beratung zur technischen Datenaufbereitung und Qualitätssicherung)
- Anna Gäde (Transkription)
- Dominik Kainz (Datenerhebung und Transkription)
- Felix Luckau (Datenerhebung)
- Laura Meige (Transkription und Segmentierung)
- Anna Vogel (Datenerhebung und Transkription)
- Pauline Wardenski (Transkription und Segmentierung)

Dieses Handbuch und SeiKo zitieren

Bitte zitieren Sie dieses Handbuch wie folgt:

Schlauch, Julia; Braunewell, Aylin; Gamper, Jana (2026): SeiKo: Korpus Seiteneinsteiger:innen in Vorbereitungsklassen. Handbuch zu Korpusdesign und Datenaufbereitung. (Version 1.0-preview). Zenodo. DOI: 10.5281/zenodo.21495943

Bitte zitieren Sie SeiKo wie folgt:

Schlauch, Julia; Braunewell, Aylin; Gamper, Jana; Klotz, Martin (2026). SeiKo: Ein Lernerkorpus neu zugewanderter Schüler:innen (Seiteneinsteiger:innen) in Vorbereitungsklassen (Version XX). Zenodo. DOI: 10.5281/zenodo.21476830

[1] Die Korpuswebseite finden Sie unter: https://www.uni-giessen.de/seiko.

1 Allgemeines

1.1 Hinweise zur Korpusnutzung

SeiKo ist nach Anmeldung für die Nutzung zu Lehr- und Forschungszwecken verfügbar. Eine Übersicht zum Zugang findet sich in Kapitel 1.5.

Wenn Sie mit SeiKo arbeiten, bitten wir Sie, uns auf Publikationen aufmerksam zu machen, die wir dann sehr gerne auf der Korpuswebseite referenzieren.

1.2 Wie ist dieses Handbuch zu nutzen?

Das vorliegende Handbuch dient dazu, die Entstehung des Korpus nachzuvollziehen und die Korpusnutzung zu ermöglichen. Dafür ist es in zehn Teile gegliedert.

Kapitel 1 enthält eine allgemeine Übersicht über das Korpus, Angaben zum Korpusbau und zur Korpusnutzung sowie Hinweise zu diesem Handbuch. Kapitel 2 beschreibt das Korpusdesign und die Datenerhebung. Kapitel 3 stellt die Transkriptionskonventionen für die mündlichen und schriftlichen Daten mittels des genutzten Transkriptionstools (EXMARaLDA, Schmidt & Wörner, 2014) vor. In Tabelle 1.1 finden sich die unterschiedlichen Annotationsschritte, die in entsprechenden Kapiteln ausgeführt werden: Zunächst wird die manuelle Segmentierung auf Grundlage von Äußerungseinheiten beschrieben (Kapitel 4). Die automatisch durchgeführten und manuell korrigierten Schritte der Lemmatisierung und des POS-Taggings werden in Kapitel 5 dargestellt. Die manuelle Annotation der Nominalgruppen und der Verbstellung wird in Kapitel 6 bzw. Kapitel 7 beschrieben.[1] Anschließend enthält Kapitel 8 eine Übersicht zu Maßnahmen, die zur Qualitätssicherung ergriffen wurden, und berichtet Ergebnisse der errechneten Inter-Annotator-Agreements.

In Kapitel 1.3 und Kapitel 1.4 finden sich Angaben zum Korpusumfang und eine Übersicht der Arbeitsprozesse. In Kapitel 1.5 wird der Zugang zum Korpus beschrieben. Das Kapitel enthält auch eine Übersicht der Daten und Metadaten des Korpus.

Der Anhang umfasst das genutzte Stimulusmaterial (Kapitel 11.1). Im Anhang ist außerdem der genutzte Interviewleitfaden (Kapitel 11.2) sowie eine Übersicht der in diesem Handbuch genutzten Schriftkonventionen zu finden (Kapitel 11.3). Die Nutzungsbedingungen des Korpus finden sich in Kapitel 10.

Für Nutzer:innen, die eigene Annotationsebenen ergänzen wollen, stellen wir die .exb-files sowie die von uns genutzte Pipeline zur Verfügung (vgl. Kapitel 1.5).

[1] Um der Überzeugung gerecht zu werden „that the learner's system is worthy of study in its own right, not just as a degenerate form of the target system" (Bley-Vroman, 1983: 4) und Lerner:innensprache nicht an einer „native norm" (Gilquin, 2022: 87) zu messen, wurde im Korpus bewusst auf die Formulierung von Zielhypothesen verzichtet. Zugleich bleibt jede Annotation durch die ihr vorausgehenden Deutungen der sprachlichen Elemente auch interpretativ. Um auch ohne Zielhypothesen eine möglichst hohe Transparenz zu gewährleisten, sind alle Schritte der Annotation – von der Segmentierung über POS- und Lemma-Annotationen bis hin zur auf diesen Annotationen basierenden Bestimmung von Verbstellung und Nominalgruppen – im Korpus einsehbar und in diesem Handbuch umfassend beschrieben und damit in ihrer Entstehung nachvollziehbar gemacht.

Tabelle 1.1: Übersicht aller Annotationsebenen

Kapitel	Annotationsebene	Beschreibung	Beispielanfrage	Beispielergebnis auf ctok_part
Transkription	INT[v]	Ursprüngliche Transkriptionsspur der Interviewer:in	-	-
Transkription	INT[tok_int]	Tokenisierte Spur von INT[v]	-	-
Transkription	B02[v]	Ursprüngliche Transkriptionsspur der Teilnehmenden	-	-
Transkription	B02[tok_part]	Tokenisierte Spur der Teilnehmenden, hier z.B. von B02[v]	tok_part="weil"	weil
Transkription	[com]	Kommentarspur für alternative Transkriptions- oder Annotationsmöglichkeiten und weitere Anmerkungen	-	-
Transkription	B02[lay]	Layout	-	-
Transkription	B02[excl]	Ausgeschlossenes Material (z.B. Korrekturen, Wiederholungen)	excl="hes"	-
Segmentierung	B02[unit]	segmentierte Äußerungseinheiten	unit="DEC"	und er hat ein punkt für sein team
Segmentierung	B02[unit2]	Einschübe in Form von Äußerungseinheiten, analog zu [unit]	unit2="IRNON"	keine ahnung
Segmentierung	B02[subclause]	abhängige Nebensätze	subclause="SCADV"	wenn das eine e mail oder ein brief ist
Segmentierung	B02[subclause2]	abhängige Nebensätze	subclause2="SCREL"	wo der tee ist
Lemmatisierung & POS-Tagging	PART_CLEAN[ctok_part]	um [excl] und Pausenzeichen bereinigte Version von [tok_part]	ctok_part="und"	und
Lemmatisierung & POS-Tagging	PART_CLEAN[lemma]	manuell korrigierte Version von [auto_lemma]	lemma="spielen"	gespielt
Lemmatisierung & POS-Tagging	PART_CLEAN[pos]	manuell korrigierte Version von [auto_pos]	pos="PDAT"	diese
Nominalgruppen	PART_CLEAN[ngr]	Nominalgruppen	ngr="NGrADJ"	mein kleiner bruder
Nominalgruppen	PART_CLEAN[kern]	nominale Kerne	kern="KERN"	zimmer
Nominalgruppen	PART_CLEAN[ngr2]	Nominalgruppen	ngr2="NGrPREP"	dem jünge
Nominalgruppen	PART_CLEAN[kern2]	nominale Kerne	kern2="DET"	dem
Verben	PART_CLEAN[unit_cons]	Satzkonstituenten von unit	unit_cons="S"	er
Verben	PART_CLEAN[unit_stage]	Verbstellungsstufe von unit	unit_stage="INV"	und dann kommt die mutter wahrscheinlich
Verben	PART_CLEAN[unit2_cons]	Satzkonstituenten von unit2	unit2_cons="V"	warte
Verben	PART_CLEAN[unit2_stage]	Verbstellungsstufe von unit2	unit2_stage="SV"	ich glaub

Verben	PART_CLEAN[subclause_cons]	Satzkonstituenten von subclause	subclause_cons="O"	den spiel
Verben	PART_CLEAN[subclause_stage]	Verbstellungsstufe von subclause	subclause_stage="VEND"	dass ee wir den spiel verloren haben
Verben	PART_CLEAN[subclause2_cons]	Satzkonstituenten von subclause2	subclause2_cons="V"	sind
Verben	PART_CLEAN[subclause2_stage]	Verbstellungsstufe von subclause2	subclause2_stage="!SEP NOVEND"	dass eh sie sind so laut gewesen

1.3 Korpusgröße

SeiKo umfasst insgesamt 824 Texte (SeiKo1 = 462; SeiKo2 = 362) von insgesamt 17 Lerner:innen (SeiKo1 = 8; SeiKo2 = 9). Die Größe der Teilkorpora und des Gesamtkorpus wird in Tabelle 1.2 dargestellt.

Tabelle 1.2: Übersicht der Tokenzahl je Annotationsebene der Teilkorpora SeiKo1 und SeiKo2

Ebene	SeiKo1	SeiKo2	gesamt
tok_part	68124	83719	151843
ctok_part	43016	45612	88628
unit (Satzgefüge)	9438	9259	18697
stage (verbhaltige Sätze)	7197	7447	14644
Nominalgruppen	6089	6796	12885

1.4 Übersicht des Workflows

Abbildung 1.1 stellt den Ablauf der einzelnen Arbeitsschritte bei der Erhebung und Aufbereitung der SeiKo-Daten dar.

Am Anfang des Prozesses stand die Planung des Korpusprojektes sowie die Erstellung eines Korpusdesigns. Nach dem Genehmigungsverfahren[2] konnte mit der Datenerhebung begonnen werden. Anschließend an jeden Datenerhebungstermin wurden die Audiodateien abgelegt und archiviert; die geschriebenen Texte wurden eingescannt und ebenfalls archiviert. Es folgte die Transkription der Daten. Die transkribierten Daten wurden von einer anderen Bearbeiter:in im nächsten Schritt korrigiert sowie segmentiert. Anschließend wurden die Daten automatisch lemmatisiert und mit *part-of-speech*(POS)-Tags annotiert; hierfür wurde der TreeTagger (vgl. Schmid, 1994, 1995) verwendet. Danach wechselte erneut die Bearbeiter:in, sodass im gleichen Schritt eine Kontrolle der manuellen Segmentierung sowie der automatischen Lemmatisierung und des POS-Taggings vorgenommen werden konnte. Zudem wurden in diesem Schritt manuell Nominalgruppen und Verbkonstituenten und -stufen annotiert. Ein Teil der Daten wurde von zwei Bearbeiter:innen annotiert, sodass darauf ein Inter-Annotator-Agreement durchgeführt werden konnte; generell wurden Zweifelsfälle vor allem in anfänglichen Phasen der Segmentierung und Annotation gemeinsam diskutiert und aufgelöst. Über eine Annatto-Pipeline wird aus den einzelnen .exb-Daten und den Metadaten ein Annis-Korpus gebaut, das über eine lokale ANNIS-Schnittstelle (Krause & Zeldes, 2016)[3] oder Annimate[4] durchsucht werden kann.[5]

[2] Für Hinweise zu Forschungsvorhaben an Schulen und den Voraussetzungen für nachnutzbare Forschungsdaten vgl. z.B. Schwendemann et al. (2024).

[3] https://corpus-tools.org/annis/

[4] https://github.com/matthias-stemmler/annimate

[5] Die Pipeline zur automatischen Datenaufbereitung und zum Korpusbau ist unter https://github.com/seiko-korpus/seiko-pipeline/ verfügbar.

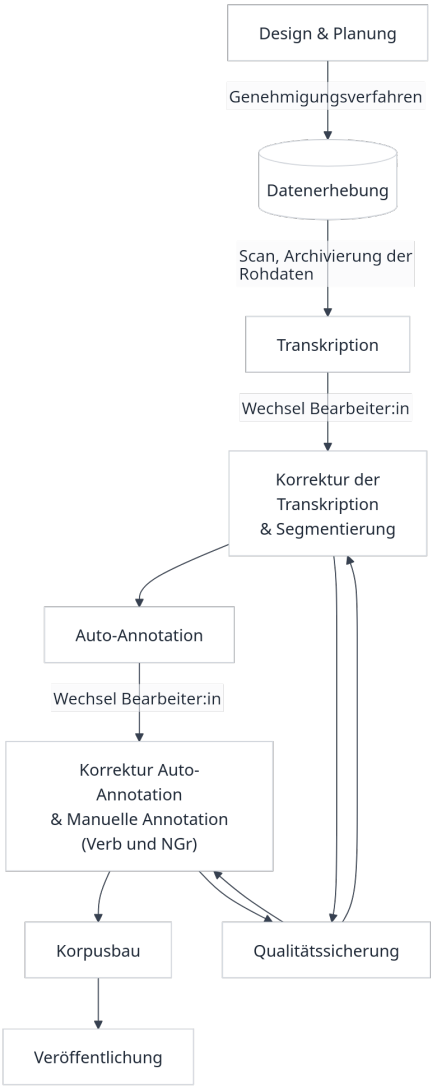

Abbildung 1.1: Ablauf der Arbeitsprozesse an SeiKo

1.5 Veröffentlichungsorte und Nutzungbedingungen

SeiKo ist unter bestimmten Nutzungsbedingungen (Kapitel 10) nutzbar. Der Zugang kann über ZENODO beantragt werden.[6]

Link zum Korpus: https://doi.org/10.5281/zenodo.21476830

Daneben ist eine Veröffentlichung in der DAKODA-Infrastruktur geplant.[7]

Die Datenverarbeitungspipeline ist hier veröffentlicht: https://github.com/seiko-korpus/seiko-pipeline/

Für eine Nutzung der Audiodaten beachten Sie bitte die Angaben in den Nutzungsbedingungen.

[6] Vgl. hierzu: https://help.zenodo.org/docs/share/access-requests/request-access/.
[7] https://dakoda.org/index.de.html.

2 Korpusdesign und Datenerhebung

Dieses Kapitel stellt die Grundlagen des Korpus- bzw. Erhebungsdesigns dar (vgl. Kapitel 2.1). Außerdem findet sich hier eine Übersicht zu den Datenerhebungen (vgl. Kapitel 2.2).

2.1 Korpusdesign und Ziele

Das Ziel bei der Erstellung von SeiKo war es, den Spracherwerb neu zugewanderter Schüler:innen zu begleiten, die in sogenannten *Intensivklassen* zur Deutschförderung an Sekundarschulen beschult werden. Diese Seiteneinsteiger:innen werden ungeachtet ihrer heterogenen Zusammensetzung auf der Grundlage als Gruppe konstruiert, dass sie zu geringe Deutschkompetenzen aufweisen, um am Regelunterricht teilnehmen zu können (vgl. Massumi & von Dewitz, 2015).

Das Korpus sollte die individuelle sprachliche Entwicklung einzelner Lerner:innen durch möglichst dichte Datenerhebungen sichtbar machen. Geplant wurde ein longitudinales Datenerhebungsschema mit Erhebungen im Abstand von etwa vier Wochen.

Herausforderungen bei der Konzeption der Datenerhebungen waren bestimmte praktische und organisatorische Gegebenheiten. Die Datenerhebungen durften insgesamt nicht mehr als 20 bis 30 Minuten pro Schüler:in in Anspruch nehmen, um die Beeinträchtigung der Lernzeit im Unterricht möglichst gering zu halten. Die Aufgaben selbst mussten so strukturiert sein, dass sie einerseits von Lerner:innen, die erst eine kurze Beschulungsdauer und eine kurze Kontaktzeit mit dem Deutschen aufweisen, trotz sehr geringer Deutschkenntnisse ohne Frust bearbeitet werden können und andererseits fortgeschrittene Lerner:innen nicht langweilen. Die Erhebungen mussten gleichzeitig gewährleisten, dass das erhobene Sprachmaterial sowohl über den Erwerbsverlauf einer Lerner:in als auch für die anderen Lerner:innen in der Gruppe vergleichbar ist. Daher wurde ein mehrfach gestuftes Erhebungsdesign konzipiert, das im Folgenden dargestellt wird.

Das Erhebungsdesign besteht aus einem individuellen Interviewsetting, das vom Vorgehen in Wiese (2020) inspiriert ist. Ziel ist die Erhebung mündlicher und schriftlicher Texte, die je zu einer bestimmten kommunikativen Aufgabe produziert werden sollen (vgl. Abbildung 2.1). Hintergrund dieser Variation in Modalität und Aufgabe sind die kontextuellen Bedingungen, in denen und für die die Schüler:innen lernen: Für das erfolgreiche Bestehen in der Regelklasse sind das Beherrschen zunehmend dekontextualisierter Sprache und der selbstständige schriftliche Ausdruck notwendig; Schüler:innen müssen eigenständig zwischen unterschiedlichen sprachlichen Registern variieren und sich dem sozialen Kontext anpassen können. Um die Entwicklung der entsprechenden Fähigkeiten modellieren zu können, wurden daher sowohl mündliche und alltagsnahe (Teil N und Q), mündliche und dekontextualisierte (Teil D) als auch schriftliche dekontextualisierte Daten (Teil T) erhoben, die die Registerdifferenzierung der Seiteneinsteiger:innen erfassen sollen.

Jeder Erhebungszeitpunkt gliedert sich in vier entsprechende Aufgaben, die auf Grundlage einer Bildergeschichte als Stimulus produziert werden.

Die Bildstimuli stellen alltagsnahe Situationen in der Schule (*Hausaufgaben* und *Sportunterricht*), im häuslichen Umfeld (*Teekanne* und *Kuchen*) und im öffentlichen Raum (*Busticket* und *Fahrradtasche*) dar, in denen jeweils eine Problemsituation entsteht.[1] Zum Beispiel wird in der Geschichte

[1] Das gesamte Stimulusmaterial findet sich im Anhang unter Kapitel 11.1.

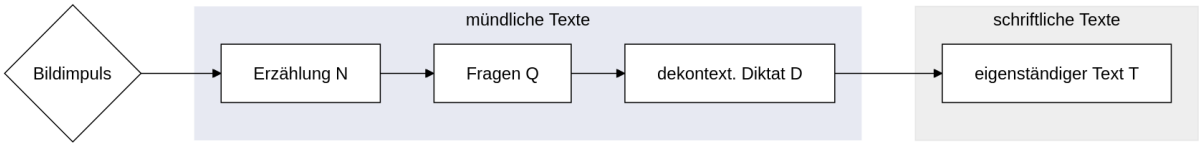

Abbildung 2.1: Übersicht des Interviewablaufs

Fahrradtasche der Diebstahl einer Fahrradtasche auf einem Markt beobachtet. Grundsätzlich werden im Erhebungsverlauf alle Bildergeschichten in einem Turnus als Stimuli angeboten (vgl. Abbildung 2.2), sodass jede Geschichte sich maximal viermal im gesamten Erhebungsverlauf wiederholt. (Die meisten Schüler:innen bemerkten nach ein paar Monaten, dass die Auswahl der Geschichte nicht wie zuvor suggeriert zufällig ist und dass sich die Geschichten wiederholen. In diesem Fall wurde den Schüler:innen das Vorgehen erklärt. In keinem Fall führte das zu einem Abfall an Motivation bei der Teilnahme.)

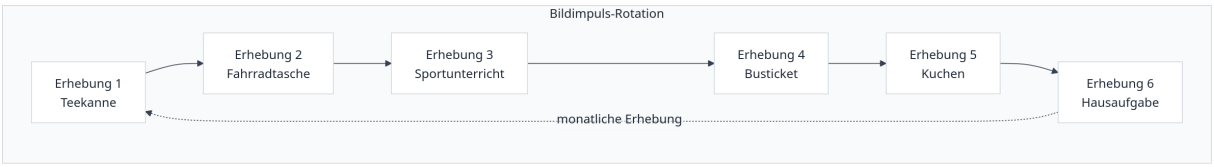

Abbildung 2.2: Rotation der Bildimpulse

Teil N - mündliche Erzählung

Zunächst wählen die Schüler:innen vermeintlich zufällig eine Geschichte aus, die in einem Umschlag verborgen ist, und werden dazu aufgefordert, die Bilder zunächst allein zu betrachten. Dann sollen sie für die Interviewer:in das dargestellte Geschehen nacherzählen, ohne die Bilder zu zeigen. Durch dieses Vorgehen sollen die Schüler:innen davon ausgehen, dass die Interviewer:in nicht weiß, welche Geschichte erzählt werden soll. Zusätzlich sollen die Bilder nicht offen gezeigt werden, damit kein geteilter Referenzraum entsteht und die Schüler:innen möglichst umfangreich erzählen. Die Interviewer:in soll hier lediglich mit ermunternden Rezeptionssignalen interagieren.

Teil Q - Fragen zur Erzählung

Wenn die Schüler:innen die Erzählung beendet haben, hat die Interviewer:in im zweiten Interviewteil die Gelegenheit, Rückfragen zu stellen und damit weitere Handlungselemente, Beschreibungen etc. zu erfragen.[2] Hier kann eine grundsätzliche Sicherung aller in der Bildergeschichte dargestellten Handlungsschritte erfolgen. Dabei wird die Geschichte auch auf den Tisch gelegt, sodass beide Gesprächsteilnehmer:innen die Bilder sehen können.

Teil D - Dekontextualisierung durch Diktieren eines Berichts

Im folgenden Interviewteil werden die Schüler:innen dazu aufgefordert, das Geschehen noch einmal für eine nicht anwesende Person zu berichten. Im Falle des obigen Beispiels der gestohlenen Fahrradtasche wird etwa erklärt, dass die bestohlene Frau den Vorfall selbst nicht beobachten konnte und nun einen Bericht für die Polizei benötigt. Das Vorgehen schließt dabei ein, dass die Interviewer:in diesen Bericht aufschreibt, während die Schüler:innen diesen diktieren sollen. Die Interviewer:in verschriftlicht den mündlichen Bericht nicht tatsächlich; es geht im Teil D um die mündliche Produktion.

[2]Der dafür genutzte Leitfaden für jede Geschichte findet sich in Kapitel 11.2.

Teil T - eigenständige schriftliche Textproduktion

Nach dem letzten Interviewteil sollen die Schüler:innen den Bericht schließlich noch einmal selbst verfassen. Die Idee der Trennung in Teil D und T besteht darin, dass durch den Teil D (ggf.) konzeptionell schriftliche Texte auch dann erhoben werden können, wenn die Schüler:innen selbst mit den Anforderungen des medialen Schreibens noch überfordert sind.

Zusätzlich wurde bei jeder Datenerhebung auch ein freier Interviewteil erhoben, der jedoch für das Korpus nicht aufbereitet wurde. In diesem wurden entweder Angaben zu den Metadaten erfragt oder über den Unterricht und Schulalltag sowie Freizeitaktivitäten gesprochen.

Metadaten

Die Erhebung von Metadaten des Korpus war durch das notwendige Genehmigungsverfahren leider stark beschränkt. Neben den erhebungsspezifischen Metadaten konnte für jede Schüler:in das Alter bei der ersten Erhebung, gesprochene Sprachen und deren Erwerbskontext, die Beschulungsdauer und der Beschulungskontext (exklusiver Besuch einer Intensivklasse, Teil- bzw. Vollintegration in die Regelklasse) für jede Datenerhebung erfasst werden.

2.2 Datenerhebung, Datenerhebungszeitpunkte

Die Datenerhebungen zu SeiKo fanden in zwei Erhebungswellen von Mai 2021 bis April 2023 (SeiKo1) und von Mai 2024 bis September 2025 (SeiKo2) statt. Ursprünglich konnten 25 Schüler:innen für die Teilnahme gewonnen werden, durch Fluktuation und beschränkte Zustimmung zur Publikation der Daten konnten schließlich 17 individuelle Datenreihen im veröffentlichten Korpus inkludiert werden.

Alle teilnehmenden Schulen, Lehrer:innen, Schüler:innen und deren Sorgeberechtigten wurden vorab über die Abläufe und Ziele der Datenerhebungen informiert und gaben ihr Einverständnis. (Alle Informationsschreiben und Formulare wurden dabei sowohl auf Deutsch als auch in Übersetzung in eine Erst- oder andere gewünschte Sprache vorgelegt.) Außerdem wurde eine offizielle Genehmigung der Schulaufsichtsbehörde für die Datenerhebung eingeholt.

Insgesamt wurden an 79 Tagen Datenerhebungen durchgeführt, aus denen eine Sammlung von 217 N-Texten, 196 Q-Texten, 215 D-Texten und 196 T-Texten resultiert. (Nicht bei jeder Erhebung haben alle Schüler:innen alle Textteile realisiert. Eine Übersicht der Gesamtkorpusgröße findet sich in Kapitel 1.3.)

Abbildung 2.3 zeigt die individuellen Datenerhebungszeitpunkte je Schüler:in (inkl. des jeweiligen Beschulungskontextes: IKL steht für *Intensivklasse*, TINT für *Teilintegration*, InteA für eine Intensivklasse an einer berufsbildenden Schule und RK für *Regelklasse*). Im Durchschnitt liegen 12,76 Erhebungszeitpunkte pro Schüler:in vor (min. 7, max. 20). Ungefüllte Punkte bei den Erhebungszeitpunkten markieren Erhebungen ohne T-Texte.

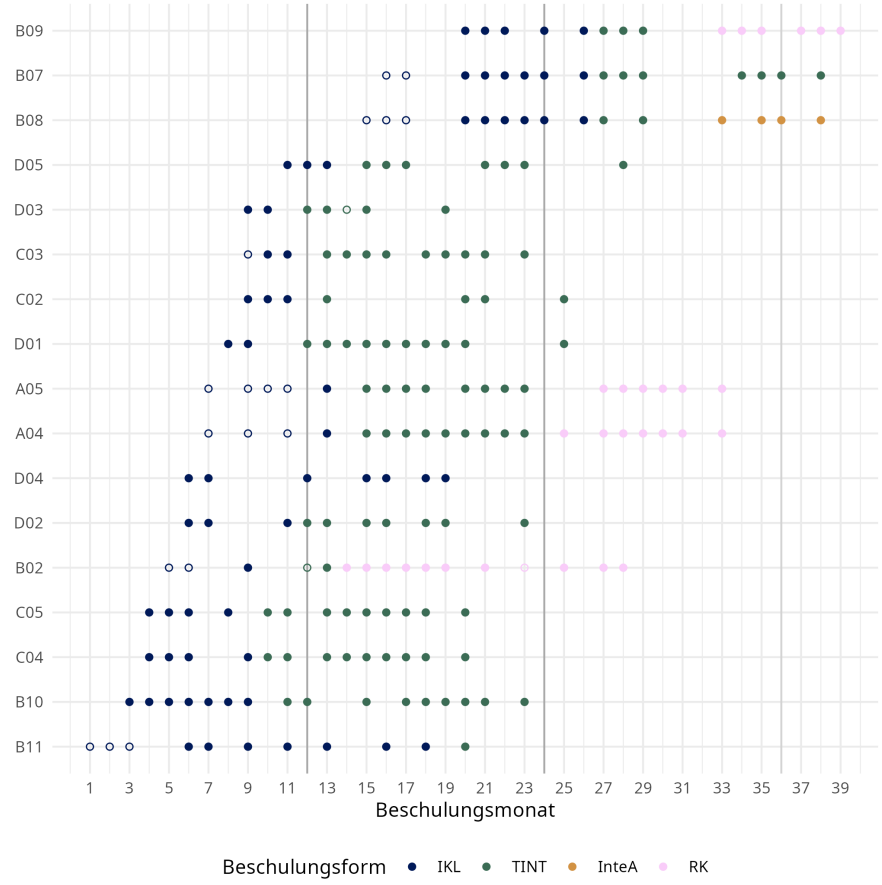

Abbildung 2.3: Übersicht der Datenerhebungszeitpunkte je Schüler:in inkl. Beschulungskontext

2.3 Metadaten

Eine individuelle Darstellung der Metadaten findet sich in den zugangsbeschränkten Korpusdaten (vgl. Kapitel 1.5) Eine Übersicht der Metadatenvariablen finden sich in Tabelle 2.1.

Tabelle 2.1: Übersicht der Metadatenvariablen

SeiKo-Variable	SeiKo-Werte	DAKODA-Variable	DAKODA-Werte
filename	Bsp. E01_A04_Teekanne_D		
refernce_file	Bsp. E01_A04		
SCHULE	A; B; C		
PSEUDONYM	Bsp. ORI		
PUBLIZIERBAR	VollständigPublizierbar; anonymPublizierbar_ohneAudio		
BESCHULUNG	IKL; TINT; RK; InteA		
PROBANDIN	Bsp. A04		
ALTER_E1	Bsp. 12		
L1	Bsp. Italienisch		
LF	Englisch; Französisch		
ERHEBUNG	Bsp. E01		
ERHEBUNGSDATUM	Bsp. 2020-01-01		

BESCHULUNGS-MONAT	Bsp. 8
STORY	Teekanne; Kuchen; Hausaufgabe; Sport-unterricht; Busticket; Fahrradtasche
PART	Narrative; Questions; Decontext; Text
partshort	N; Q; D; T
ZEITRAUM	SEIKO1; SEIKO2

Allgemein lag das Alter der Schüler:innen bei der ersten Erhebung lag im Durchschnitt bei 12,65 Jahren (min. 10, max. 16 Jahre). Die Beschulungszeit bei der ersten Erhebung lag im Durchschnitt bei 8,24 Monaten (min. 1, max. 20). Tabelle 2.2 zeigt die angegebenen Sprachkenntnisse (wobei vier Schüler:innen bei L1 mehrere Angaben machten).

Tabelle 2.2: Häufigkeit angegebener Erst- sowie Zweit-/Fremdsprachen

Sprache	L1	L2
Arabisch	5	-
Dari	2	-
Englisch	-	12
Französisch	-	1
Italienisch	2	-
Kurmanci	2	-
Polnisch	-	2
Rumänisch	1	-
Russisch	2	1
Somali	1	-
Türkisch	4	5
Ukrainisch	2	-

3 Transkription

Dieses Kapitel dokumentiert das Vorgehen bei der Transkription der Daten. Allgemeine Grundlagen werden in Kapitel 3.1 dargestellt, während das Vorgehen für mündliche Texte in Kapitel 3.2 sowie für schriftliche Texte in Kapitel 3.3 differenziert wird.

3.1 Grundlagen der Transkription

Zur Transkription der Audioaufnahmen sowie der Scans der handschriftlichen Texte der Schüler:innen wurde der EXMARaLDA-Partitur-Editor[1] (Schmidt & Wörner, 2014) genutzt.

Der Wortlaut der mündlichen wie schriftlichen Äußerungen wurde grundsätzlich in einfacher orthographischer Transkription wiedergegeben. Die Extension einer Äußerungseinheit entspricht dabei im Regelfall einem Event. (Bei Äußerungen, die durch viele Pausen unterbrochen werden oder bei Überlappungen mit Äußerungen der Interviewer:innen kann es an einzelnen Stellen nötig sein, Äußerungseinheiten in mehr als einem Event darzustellen. Für eine konsistente Segmentierung in Äußerungseinheiten vgl. Kapitel 4.)

Es gilt zu beachten, dass jede Transkription immer auch eine Interpretation des Gehörten sowie des schriftlichen Textes durch die Transkribierenden darstellt und notwendigerweise eine Reduktion des Gesamtumfangs der sprachlichen Äußerungen bedeutet (z.B. Verlust bestimmter intonatorischer Eigenschaften etc.). Transkriptionskonventionen dienen dazu, diese möglichst zu vereinheitlichen, zu vereinfachen und allgemein interpretierbar zu machen. Die Konventionen der SeiKo-Transkriptionen werden in Kapitel 3.2 für die gesprochenen und in Kapitel 3.3 für die geschriebenen Texte dargestellt.

Für jeden Text(teil) einer Datenerhebung (*Narrative*, *Questions*, *Decontext*, *Text*) einer Schüler:in werden verschiedene EXMARaLDA-Dateien angelegt. Die Dateinamen setzen sich aus folgenden Informationen zusammen:

ERHEBUNGSZEITPUNKT_SCHÜLER:IN-ID_STORY_INTERVIEWTEIL » z.B. E04_A06_Busticket_T

(In der Metadatenübersicht sind diese Informationen als ERHEBUNG, PROBANDIN, STORY, PART zu finden.)

Grundsätzlich wurden die Produktionen der Schüler:innen und der Interviewer:innen je in einer eigenen Transkriptionsspur dargestellt. (Die Spuren enthalten dabei folgende Pseudonyme als Sprecher:innen-ID zugewiesen: [INT] für die Interviewer:in, sowie eine individuelle ID - z.B. [B02] - für die Schüler:innen.) Nicht durch die Transkriptionskonventionen darstellbare Auffälligkeiten (z.B. Störgeräusche, Sprechqualität) oder alternative Transkriptionen wurden auf einer separaten Spur ([com]) vermerkt.

Sensible Textstellen (z.B. Nennung von Namen, Orten etc.) wurden in der Transkription pseudonymisiert wiedergegeben (z.B. NAME, ORT). In den Audioaufnahmen wurden solche Stellen durch Verrauschen unkenntlich gemacht. Eine Anonymisierung der Daten durch vollständige

[1] https://exmaralda.org/de/partitur-editor-de/

Verfremdung stimmlicher Merkmale (vgl. hierzu z.B. Meyermann & Porzelt, 2014) kann nicht vorgenommen werden, da die Daten sonst für viele linguistische Analysen nicht mehr nutzbar wären. (Zum Zugang zur vollständig pseudonymisierten und damit "stummen" Korpusversion sowie den Möglichkeiten zur Nutzung der Audiodaten vgl. Kapitel 10.)

Zur Sicherung der Qualität der Transkriptionen wurden alle Transkripte nach dem Vier-Augen-Prinzip von einem zweiten Teammitglied nachgehört oder überprüft. (Für eine Übersicht der qualitätssichernden Maßnahmen vgl. Kapitel 8.)

3.2 Transkription gesprochener Daten

Für die Transkription gesprochener Texte werden zunächst eigene Transkriptionsspuren für Schüler:in und Interviewer:in angelegt. Abbildung 3.1 zeigt eine typische Ansicht einer EXMARaLDA-Datei während der Transkription.

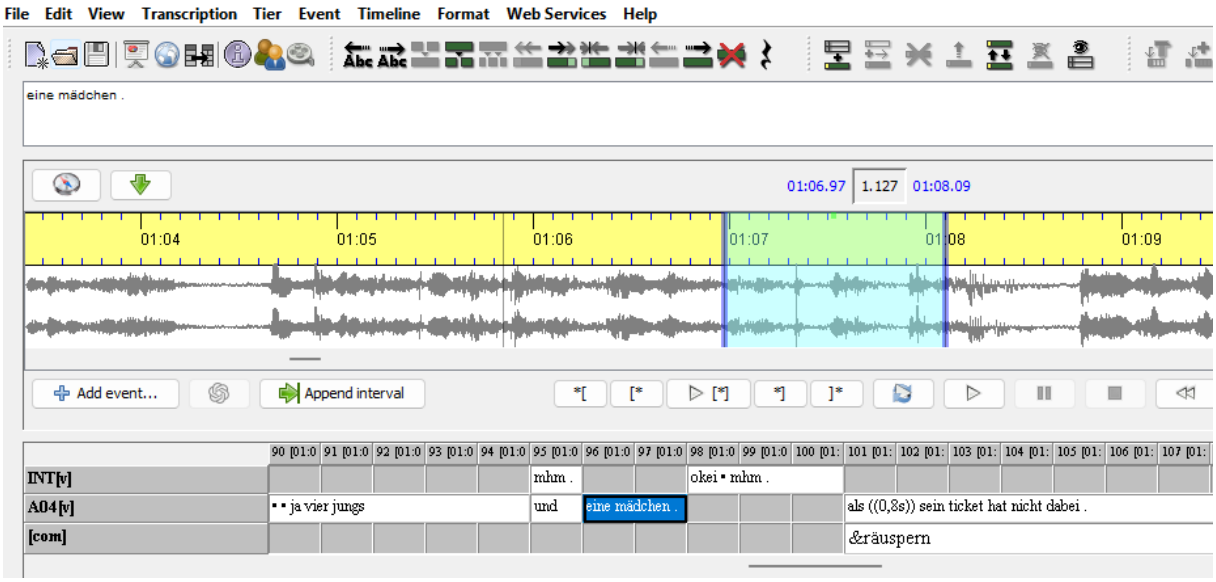

Abbildung 3.1: Ansicht des Partitur-Editors bei der Transkription mündlicher Texte

3.2.1 Äußerungsnahe Wiedergabe des Gesagten

Äußerungen wurden inklusive subordinierter Äußerungen in einfacher orthographischer Transkription wiedergegeben Rehbein et al. (2004). Großbuchstaben wurden dabei nicht verwendet. Subordination wird mit , kenntlich gemacht. Die Transkriptionszeichen (. , *?* und *!*) wurden für die Unterscheidung von deklarativen, interrogativen oder exklamativen Äußerungen genutzt. Entscheidend für die Zuordnung war neben der Intonation der Sprecher:innen auch die Interpretation durch die Transkribend:in. Wurde die finale Tonhöhenbewegung am Ende der Äußerung gehalten, wurde - transkribiert. Eventgrenzen wurden nach den gehörten Äußerungseinheiten gesetzt. (Längere Pausen, Störgeräusche u.a. sowie Überlappungen können auch in eigenen Events abgesetzt sein.)

Gesprochensprachliche Phänomene wurden möglichst originalgetreu wiedergegeben. Beispiele:
und er hat *ein* punkt für sein team (E02_B02) *ne* andere person (E02_A05)

Systematische (und obligatorische) Reduktionen der gesprochenen Sprache wurden nach orthographischen Standards transkribiert. Auch leichte sprachliche Verschleifungen wurden aufgelöst (z.B. *und dann*, nicht *undann*, *ich*, nicht *isch*).

Fremdsprachiges Material wurde möglichst originalgetreu wiedergegeben und pro Wort mit *[]* markiert. Beispiel: mit *[tea] [and]* buch (E01_B07)

Buchstabierte Abkürzungen wurden ohne Leerzeichen zwischen den Buchstaben ergänzt.

3.2.2 Unverständliches und Unklares

Für unklare Äußerungen, die mit # annotiert werden, ist potentiell eine alternative Transkription auf [com] angegeben. Diese ist mit einem vorangestellten *?* markiert.

Beispiel:

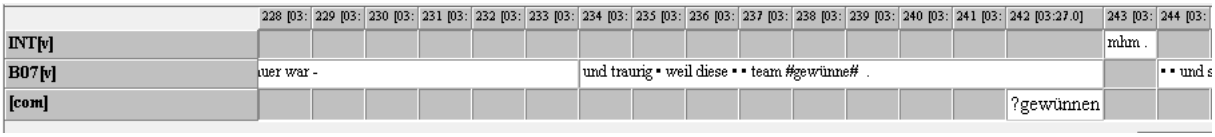

Abbildung 3.2: Unklare Transkription (E15_B07)

Gänzlich unverständliche Passagen wurden mit *XXX* markiert. Beispiel: die eine mann hat eis *XXX* mit hände (E02_B10)

3.2.3 Pausen

Zur vereinfachten Darstellung von Pausen wurden Pausenzeichen und exakte Angaben der Pausenlänge verwendet.

Ein kurzes Stocken im Redefluss kann mit einem Punktzeichen • transkribiert werden.

Eine Pause mit einer geschätzten Länge von bis zu einer halben Sekunde wird mit zwei Punktzeichen • • , eine geschätzte Pause von bis zu einer dreiviertel Sekunde mit drei Punktzeichen • • • transkribiert.

Längere Pausen wurden gemessen und numerisch in Doppelklammern angegeben. (z.B. *((2.5))* oder *((2.5s))*

Pausen innerhalb eines Wortes wurden auf [com] notiert.

Beispiel: kon((• •))trolleur (E10_B11)

3.2.4 Abbrüche und Selbstkorrekturen

Abbrüche auf Wortebene wurden mit / unmittelbar am Wortende transkribiert.

Beispiel: im fahrrad steckt *ei/* ein tüte (E15_B08)

Satzabbrüche wurden mit // am Ende des abgebrochenen Satzes und nach einem Leerzeichen markiert.

Beispiel: und dann // (E01_B09)

3.2.5 Überlappungen

Für überlappende Äußerungen wurde auf der jeweiligen Spur der Sprecher:innen der Audioabschnitt identifiziert und in entsprechenden Events transkribiert. Anschließend wurden die Eventgrenzen so gesetzt, dass die Extension der Überlappung auch in der Transkription sichtbar wird, wobei einzelne Worte nicht auf verschiedene Events aufgeteilt wurden (vgl. Event 52 in Abbildung 3.1).

Wenn eine Sprecher:in unterbrochen wird, wurde die Unterbrechung als Pause und Überlappung transkribiert.

3.2.6 Paralinguistische Phänomene

Paralinguistische Phänomene wie Lachen, Flüstern etc. wurden auf [com] notiert und mit [&] markiert (z.B. *&lachen* oder *&lachend*).

3.2.6.1 Fragen, Rückversicherungen, Zögern

Fragen und Rückversicherungen wurden standardisiert mit *mh ?*, *ne ?* und *hee ?* wiedergegeben.

Markierung von Zögern (*hesitation markers*) und hörbares Nachdenken (*turn holding*) wurden mit *em*, *mh*, *hm* und *ee* transkribiert.

3.2.6.2 Zustimmung und Negation

Zustimmung wurde mit *mhm*, *ehe*, *ah*, *aha*, *oke* oder *okei* wiedergegeben.

Negation wurde als *hmhm*, *eheh*, *nee* transkribiert.

3.3 Transkription geschriebener Texte

Die Transkription der geschriebenen Texte erfolgte auf Grundlage der gescannten handschriftlichen Produktionen der Schüler:innen. Bei schriftlichen Texten wurde bereits während der Transkription eine Tokenisierung der Texte vorgenommen. Abbildung 3.3 zeigt daher eine Transkriptionsspur für den originalen Text in seinem ursprünglichen Wortlaut [v], eine tokenisierte Spur [tok_part], in denen von den Lerner:innen vorgenommene Selbstkorrekturen umgesetzt sind (vgl. dazu Kapitel 3.3.4.2) sowie eine Spur zur Wiedergabe des Textlayouts [lay] und eine Spur zur Darstellung von Korrekturen [excl]. Optional konnte die [com]-Spur zum Vermerken von Kommentaren eingesetzt werden.

3.3.1 Textnahe Wiedergabe

Analog zu den mündlichen Texte wurden auch die geschriebenen Texte der Schüler:innen möglichst originalgetreu wiedergegeben. Die Schreibweisen werden dokumentiert, d.h. Orthographie und Interpunktion wurden nicht verändert. Dabei wurde auch die jeweilige Groß- und Kleinschreibung berücksichtigt. Die Eventgrenzen orientieren sich an der Interpunktion der Schüler:innen.

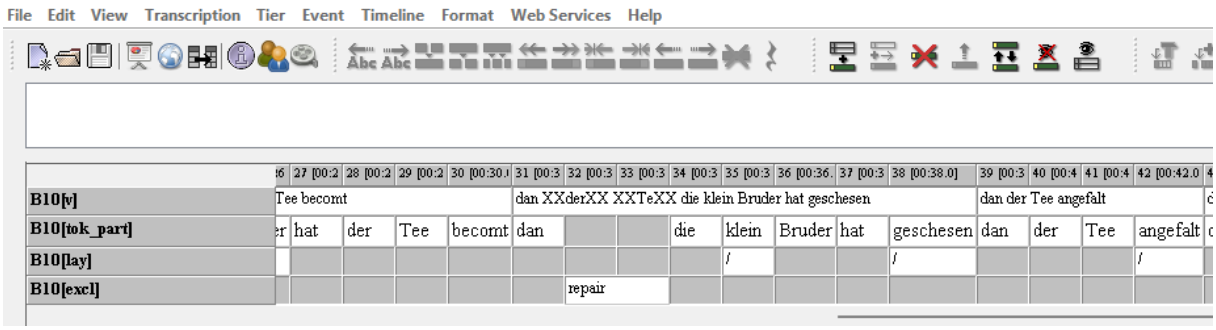

Abbildung 3.3: Ansicht des Partitur-Editors bei der Transkription schriftlicher Texte (E01_B10)

3.3.2 Unleserliches

Unleserliches wurde möglichst pro Buchstabe mit @ wiedergegeben: z.B. *ge@aufen*.

3.3.3 Korrekturen

Korrekturen werden grundsätzlich auf der [v]/Transkriptionsspur erfasst. In den tokenisierten Spuren steht nur die finale Fassung des Textes vgl. dazu auch Kapitel 3.3.4.2.

Durchgestrichene Wörter wurden mit XX XX transkribiert (vgl. auch Event 32 in Abbildung 3.3).

Beispiel: XXgelauXX

Durchgestrichene Buchstaben wurden mit X X wiedergegeben (vgl. auch Abbildung 3.4).

Beispiel: geXeXlaufen

Nachträglich ergänzte Buchstaben wurden mit <> markiert (vgl. auch Abbildung 3.4).

Beispiel: ge<lau>en

Ein nachträglich eingefügtes Wort wurde mit < markiert.

Beispiel: ich <*bin* gelaufen

Bei mehreren nachträglich eingefügten Wörtern wurde vor jedem Wort die Einfügung mit < < kenntlich gemacht.

Beispiel: ich < <bin < <dann gelaufen

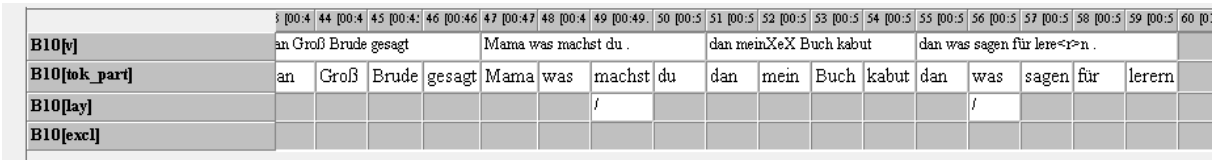

Abbildung 3.4: Durchgestrichene Buchstaben und nachträglich eingefügte Buchstaben (E01_B10)

3.3.4 Tokenisierung im Transkriptionsprozess und zusätzliche Spuren

Schriftliche Texte wurden bereits im Transkriptionsprozess tokenisiert. Die tokenisierte Spur wurde anhand der in Kapitel 3.3.4.2 dargestellten Vorgaben angepasst. Außerdem wurde eine zusätzliche Spur mit Hinweisen zum Layout ([lay]) und zu ausgeschlossenem Material ([excl]) ergänzt.

3.3.4.1 Layout

Die [lay]-Spur umfasst eine Markierung des Zeilenendes, von Absätzen, Leerzeilen und Einrückungen.

Das Zeilenende wurde mit / unter dem Event des letzten Worts dieser Zeile markiert. Bei Worttrennung wurde die Trennung (z.B. *gelau- / fen* und /) unter dem letzten Wortbestandteil dargestellt.

Absätze und Leerzeilen wurden mit // im Event des letzten Worts im Absatz transkribiert.

Leichte Einrückungen wurden mit *§*, starke Einrückungen mit *§§* im Event des eingerückten Worts bzw. des ersten eingerückten Worts wiedergegeben.

Große Lücken zwischen Wörtern wurden mit == im Event des Wortes nach dieser Lücke transkribiert.

3.3.4.2 Korrekturen als Vorbereitung für das automatische Tagging

Anders als bei der Transkription mündlicher Daten erfolgte für die schriftlichen Texte eine manuelle Korrektur der [tok_part]-Spur. Dies ist notwendig, um bessere Ergebnisse bei der automatischen Annotation (vgl. Kapitel 5) zu erzielen.

Korrekturtranskriptionen wie das Einfügen von Buchstaben oder Wörtern werden auf [tok_part] bereinigt. Vollständig korrigierte Wörter, die erkennbar von der Schreiber:in aus der “finalen Fassung” des Textes gestrichen wurden und nicht Teil der “letzten finalen Fassung des Textes” sein sollten, werden auf einer separaten Spur [excl] mit dem Tag *repair* markiert. Das heißt, [tok_part] enthält den Lerner:innentext in der “letzten Fassung”.

4 Segmentierung der Äußerungseinheiten

4.1 Spuren bei der Segmentierung

Für die Segmentierung wurden folgende zusätzliche Spuren eingerichtet:

Tabelle 4.1: Spuren bei der Segmentierung

[unit]	segmentierte Äußerungseinheiten
[subclause]	abhängige Nebensätze
[excl]	ausgeschlossenes Material (z.B. Korrekturen, Wiederholungen)

Bei Bedarf können zusätzlich folgende Spuren (ebenfalls als Annotationsspuren) erzeugt werden:

Tabelle 4.2: Optionale Spuren bei der Segmentierung

[unit2-x]	Einschübe in Form von Äußerungseinheiten, analog zu [unit]
[subclause2-x]	abhängige Nebensätze

4.2 Grundlage der Segmentierung

Um die Transkripte in Äußerungseinheiten zu segmentieren, die die Arbeit des automatisierten POS-Taggings und der Lemmatisierung erleichtern, wurde sich an der *Analysis of Speech Unit* von Foster et al. (2000) orientiert und die Operationalisierung der Segmentierungseinheiten fürs Deutsche spezifiziert und exemplifiziert. Das folgende Kapitel dokumentiert das Vorgehen und markiert allgemeine Modifikationen aus dem Vorgehen von Foster et al. (2000). Die Segmentierung beruhte v.a. auf syntaktischen Kriterien, wobei zur Disambiguierung teilweise auch semantische und/oder intonatorische Kriterien herangezogen wurden.

Jedes Token (außer Pausen) der [tok_part]-Spur ist von einer Annotation auf [unit] oder [excl] überspannt. Wo möglich, wurden Spannannotationen der Annotation einzelner Tokens vorgezogen. (So wurde etwa bei zwei Tokens, die als *false start* gewertet wurden, eine Spanne mit entsprechendem Tag über beide Tokenevents auf [excl] gezogen.) Grundsätzlich wurden Spannen auf [unit] so eng wie möglich angelegt, um einem möglichst minimalen Anspruch an Äußerungseinheiten gerecht zu werden.

4.3 Definition einer Äußerungseinheit (=unit)

Foster et al. (2000) orientieren sich bei der Definition ihrer AS-Unit an der t-unit. Eine Äußerungseinheit oder AS-unit (unit) ist eine Äußerung einer Sprecher:in, die aus einer eigenständigen/unabhängigen clause ("independent clause") oder einer "sub-clausal unit" zusammen mit allen abhängigen clauses/Nebensätzen besteht.

Eine eigenständige/unabhängige clause besteht aus mindestens einem Verb, das, anders als bei Foster et al. (2000), finit oder infinit sein kann, und maximal einem Verbalkomplex auf unit-Ebene. Weitere Verbalkomplexe können in abhängigen Nebensätzen auftreten.

Eine eigenständige Äußerungseinheit bilden Matrixsätze (vgl. Kapitel 4.5) auch dann mit den von ihnen abhängigen clauses, wenn diese keine Nebensatzwortstellung aufweisen (vgl. Kapitel 4.8).

Beispiel: sie hat gesagt ich hab keine fahrkarte (E10_B10)

Äußerungseinheiten wurden auf der Spur [unit] als Spannen annotiert. Dabei wurde differenziert zwischen eigenständigen verbhaltigen clauses und nicht verbhaltigen sub-clausal units. Die Tags ergänzen die Funktion der Äußerungseinheit oder den Typ der sub-clausal unit (vgl. Kapitel 4.7).

4.4 Unvollständige Äußerungseinheiten

Unvollständige units (und auch Nebensätze) können mit folgenden Tags differenziert werden (vgl. Kapitel 4.7):

Wenn eine unit nach Produktion eines Verbs ersichtlicherweise abgebrochen wurde, wird sie mit dem zusätzlichen Tag BREAK getagged.

Beispiel: die buch ist (E02_B02)

Handelt es sich um eine unvollständige, aber verbhaltige und nicht abgebrochene Äußerung, die durch den diskursiven Kontext sinnvoll erweiterbar ist, wird mit dem zusätzlichen Tag DISC gearbeitet.

Beispiel: was macht die mutter da? – *schimpfen* (E01_A05)

Beispiel: diesen schockmoment, das kann man richtig fühlen, wenn man in die tasche guckt und – *die hausaufgaben nicht hat* (E11_B02)

4.5 Matrixsätze

Als Matrixsätze werden alle Sätze bezeichnet, von denen eine subclause abhängig ist. Matrixsätze wurden entsprechend der auf [unit] eingerichteten Tags annotiert; die Eigenschaft eines Satzes als Matrixsatz wird durch die Integration einer subclause und die entsprechende Spurstruktur erfasst, die Tags selbst differenzieren zwischen unterschiedlichen Äußerungstypen und sind nicht explizit als Matrixsätze erfasst.

Nicht realisierte Matrixsätze, von denen eine subclause abhängig ist, wurden mit Ø auf [unit] dargestellt. Die Extension entspricht der Spanne auf [subclause].

Hauptsatzförmige Sätze, die von einem Matrixsatz abhängig sind, wurden als abhängige Sätze auf [subclause] erfasst (vgl. Kapitel 4.8 für das spezifische Vorgehen).

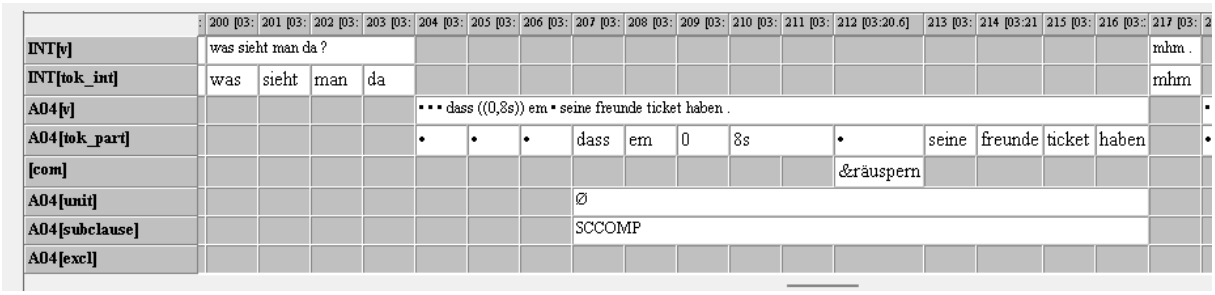

Abbildung 4.1: Nicht realisierter Matrixsatz (E16_A04)

4.6 Vor- und Nachlaufelemente

Topikalisierte NPs und adverbiale Prä-Vorfelder wurden als Teil der unit erfasst.

Beispiel: *dann* er hat lachen (E06_B11)

Nachlaufelemente wurden entweder als Teil der vorangegangenen unit erfasst oder bildeten eine eigene unit:

Sie wurden, sofern sie aus syntaktischer, intonatorischer und pragmatischer Perspektive als Nachfeldbesetzung gesehen werden können, als Teil der unit "eingespannt".

Beispiel: ich hab nicht geschwommen *leider* (E02_B02)

Auch tag questions wie *oder* wurden als Teil der vorigen unit gefasst, deren Annotation sich aus der Struktur der Äußerung ergibt; die Äußerung einer tag question führte nicht dazu, dass die gesamte Äußerung als Frage (WINT/INT) erfasst wird.

Eine lange Pause und/oder anaphorische Marker (z.B. *also*) führten hingegen dazu, die beiden Äußerungen als getrennt zu betrachten und das Nachlaufelement einzeln zu taggen. Das Nachlaufelement konnte hierbei oft als DISC erfasst werden (vgl. Kapitel 4.7.2).

Beispiel: die anderen sprechen mit den. *also mit diese mädchen* (E16_B10)

4.7 Tagübersicht für [unit]

4.7.1 verbhaltige units

In Tabelle 4.3 werden alle Tags für verbhaltige Units dargestellt.

Tabelle 4.3: Übersicht der annotierten verbhaltigen units

Tag	Beschreibung	Beispiel
DEC	Äußerungseinheit in Form eines Deklarativs	und dann hat der kleiner bruder n ball so geworfen (E19_A05)
INT	Äußerungseinheit in Form eines Interrogativs	hast du fahrkarte oder (E04_B10)
WINT	Äußerungseinheit in Form eines W-Interrogativs	was passiert (E05_B08)
IMP	Äußerungseinheit in Form eines Imperativs	Und dass wegen lass die leute in ruhe (E18_A04)
BREAK	(+) als ergänzendes Tag für abgebrochene, verbhaltige Äußerungen	aber ich hab ge/ (DECBREAK, E14_B07)

DISC	(+) als ergänzendes Tag für unvollständige verbhaltige Units, die durch diskursive Rekonstruktion sinnvoll erweitert werden können	diesen schockmoment, das kann man richtig fühlen, wenn man in die tasche guckt und – die hausaufgaben nicht hat (E11_B02)

4.7.2 nicht verbhaltige (*sub-clausal*) units

Nicht verbhaltige units (*sub-clausal units*) werden in Tabelle 4.4 dargestellt:

Tabelle 4.4: Übersicht der annotierten verblosen units

Tag	Beschreibung	Beispiel
DISC	eine oder mehrere Phrasen, die aus dem diskursiven/situationen Kontext zu einer vollen clause rekonstruierbar sind. Dieses Tag ist operationalisiert als eine alleinstehende clause, die aus dem diskursiven/situativen Kontext verständlich ist; z.B. als Antwort auf eine Frage.	was könnte das sein? – papier, oder? (E06_A05)
IRNON	reguläre, zielsprachenkonforme Äußerungen. Dazu zählen auch *ja, nein, oke* etc.). Mit IRNON markierte Strukturen sind dabei grundsätzlich nicht satzförmig rekonstruierbar.	ja, nein, oke
SCU	bruchstückhafte, nicht verbhaltige Äußerungen, bei denen es sich auch nicht um DISC, IRNON, *false starts* oder Abbrüche handelt. Dieses Tag erfasst klassische lerner:innensprachliche Äußerungen mit pragmatischem Gehalt. Diese Kategorie wird bei Foster et al. 2000 so nicht erfasst.	von feuer (E05_B08)
Ø	nicht realisierte Äußerungseinheit (Matrixsatz), von der eine subclause abhängt	

DISC kann also bei verbhaltigen units angehängt werden (vgl. Kapitel 4.3), markiert alleinstehend aber nicht verbhaltige und aus dem Kontext rekonstruierbare Äußerungen. IRNON können “minor utterances”, “irregular sentences”, “nonsentences” nach Quirk et al. (1985) sein. Das Sammeltag SCU wird in Abgrenzung zu den anderen verblosen Annotationen und Abbrüchen (vgl. Kapitel 4.11.1) vergeben und bei Foster et al. (2000) so nicht verwendet. Für nicht realisierte Matrixsätze vgl auch Kapitel 4.5.

4.8 Abhängige clauses / Nebensätze

Abhängige Sätze bestehen aus mindestens einem finiten oder infiniten Verb und einem weiteren Element (Subjekt, Objekt, Komplement oder Adverbial) (vgl. Foster et al. (2000)). Anders als bei Foster et al. (2000) zählen hier auch Subjunktionen und *zu*-Infinitive als weitere Elemente.

Nebensätze wurden auf der Spur [subclause] nach Typ in Form von Spannen unterhalb der [unit]-Spur annotiert. Die Differenzierung nach Typen ist nicht Foster et al. (2000) entnommen.

Auch Nebensätze mit abweichender Verbstellung wurden nach Funktion in Form von Spannen unterhalb der [unit]-Spur annotiert; Verbletztstellung ist also kein ausschlaggebendes Kriterium dafür, auf [subclause] getagged zu werden. Relevant war ihre Abhängigkeit von einem Matrixsatz.

Nebensätze, die von Matrixsätzen auf [unit2] abhängig sind, werden unabhängig davon auf [subclause] (bzw. [subclause2], wenn [subclause] schon "belegt" ist) annotiert.

Beispiel: dann mama hat hm ich hab vergessen *wie heißt das auf deutsch* tea zum diese junge geben (E01_C02)

Auch hauptsatzförmige Sätze, die von einem Matrixsatz abhängig sind, wurden auf [subclause] erfasst. Diese wurden mit dem Tag MCSUB markiert und einem Zusatztag, das, analog zu den Tags auf DEC, die Funktion der Äußerung annotiert.

Beispiel MCSUBDEC: ich glaube *ich habe namen falsch gesagt* (E12_B07)

Beispiel MCSUBWINT: und hier die fragt, *wie viel kostet ee so etwas* (E02_B09)

Die Entscheidung, ob der potenzielle MCSUB eingebettet ist oder nicht, ergab sich daraus, ob der potenzielle Matrixsatz alleinstehen kann. Wenn er nicht alleinstehen kann, liegt auf [subclause] ein MCSUB vor:

Beispiel: ich glaube *ich habe namen falsch gesagt* (E12_B07)

Hier liegt auf [subclause]-Ebene ein MCSUBDEC vor.

Können dagegen beide Sätze alleinstehen, wird DEC DEC erfasst.

Beispiel: *ich denke gerade. unterbrich mich nicht.* (konstruiertes Beispiel)

So wurde zum Beispiel auch wörtliche Rede als MCSUB und die sie einleitende Äußerung auf [unit] als Matrixsatz getagged.

Beispiel: der hat halt gesagt *was machst du* (E12_B07)

Zusätzlich differenzieren wir, analog zu den Tags auf [unit]:

1) verbhaltige, aber abgebrochene Nebensätze mit dem Tag-Zusatz BREAK (Intonation beachten!)

   Beispiel: weil die haben ee in bücher ge/ (E05_B09)

2) DISC: auf [subclause] nur bei verbhaltigen Äußerungen, die aus dem diskursiven Kontext heraus sinnvoll erweitert werden könnten.

   Beispiel: entschuldigung, dass ich diesen buk zerstört habe *oder ja zerstört* (E07_B02)

3) *subclausal units* auf [subclause], also nicht verbhaltige, aber eingeleitete und nach Funktion differenzierbare Nebensätze mit dem Tag-Zusatz SCU. Dieses Vorgehen weicht von Foster et al. (2000: 368) ab, die nicht verbhaltige subordinierte Einheiten exkludieren.

   Beispiel: weil ich kein ticket (E04_B07)

4.9 Tagübersicht für [subclause]

Tabelle 4.5 verzeichnet mögliche Tags der Annotationsebene [subclause]. Diese können sowohl Einbettungen in Form von Nebensätzen (Tags mit SC…) als auch in Form von Hauptsätzen umfassen (Tags mit MCSUB…).

Tabelle 4.5: Übersicht der annotierten Nebensätze

Tag	Beschreibung	Beispiel
SCREL	Relativsatz	sachen das sie verkaufen (E08_B09)
SCADV	eingebetteter Adverbialsatz	er wart bis das ball zu ihn kommt (E04_B09)
SCINDER	indirekter Fragesatz-	ein schülerin gucken was passiert (E06_B07)
SCCOMP	Komplementsatz	er guckt dass sie frau guckt oder nicht (E08_B10)
SCINFUM	eingebetteter Infinitivsatz	und der mann benutzt den perfekten moment um abzuhauen (E20_A04)
SCINF	nicht eingebetteter Infinitvsatz	sie sehen ein junge gucken für sie (E02_C05)
SCNON	nicht eingeleitete Nebensatz	und die beide will kuchen geht ofen (E05_D02)
BREAK	als ergänzendes Tag für abgebrochene, verbhaltige Äußerungen	Beispiele ab hier s. Tagübersicht für [unit]
DISC	als ergänzenders Tag für verbhaltige subclauses, die durch diskursive Rekonstruktion sinnvoll erweitert werden können	
SCU	als ergänzendes Tag für SCUs auf subclause-Ebene	
MCSUBDEC	Äußerungseinheit in Form eines Deklarativs	
MCSUBINT	Äußerungseinheit in Form eines Interrogativs	
MCSUBWINT	Äußerungseinheit in Form eines W-Interrogativs	
MCSUBIMP	Äußerungseinheit in Form eines Imperativs	
BREAK	als ergänzendes Tag für abgebrochene, verbhaltige Äußerungen	
DISC	als ergänzendes Tag für unvollständige verbhaltige Units, die durch diskursive Rekonstruktion sinnvoll erweitert werden können	
MCSUBDISC	eine oder mehrere Phrasen, die aus dem diskursiven/situativen Kontext zu einer vollen clause rekonstruierbar sind.	
MCSUBIRNON	irregular and non-sentences	
MCSUBSCU	Sammeltag: Andere sub-clausal units, die nicht über false starts, DISC oder IRNON erfassbar sind.	
MCSUBØ	nicht realisierte Äußerungseinheit, von der eine subclause abhängt	

4.10 Mehrfache Einbettung / Einschübe

Hauptsatzförmige Einschübe wurden auf [unit2] (im Folgenden Beispiel als DEC) annotiert:

Beispiel: das war *ich glaube* leichter als die andere drei (E07_B02)

Die Differenzierung zu MCSUB ergibt sich daraus, dass auf unit2 getaggte Äußerungen einen Einschub innerhalb eines Hauptsatzes darstellen, während Matrixsätze mit MCSUB Hauptsätze mit von ihnen abhängigen Äußerungen darstellen.

Mehrfach eingebettete Nebensätze wurden mit zusätzlichen Spuren [subclause2], [subclause3] annotiert. Wenn sich eine subclause auf [unit2] bezieht, wird diese unabhängig davon trotzdem auf [subclause] (oder [subclause2], wenn [subclause] schon "belegt" ist) annotiert.

4.11 Ausgeschlossenes Material

Abbrüche, Wiederholungen und Selbstreparaturen wurden von den weiteren Annotationsschritten ausgeschlossen. Sie wurden daher auf einer separaten Spur [excl] annotiert und klassifiziert. Ausnahmen bildeten verbhaltige Abbrüche, die auf [unit] und [subclause] entsprechend getagged und zusätzlich mit BREAK markiert wurden (vgl. Kapitel 4.4).

4.11.1 Abbrüche (*false starts*)

Abgebrochene, verblose Äußerungen (egal, ob es durch die Sprecher:in selbst oder eine Unterbrechung geschieht) wurden nicht als Äußerungseinheiten, sondern mit dem Tag falst als Abbrüche markiert.

Beispiel: und dann (E06_A05)

Verbhaltige Abbrüche wurden, anders als bei Foster et al. (2000) nicht exkludiert, sondern gemäß ihrer Funktion (also bspw. mit DEC) getagged und zusätzlich mit BREAK markiert (vgl. Kapitel 4.4).

4.11.2 Wiederholungen

Bei Wiederholungen wurde zwischen unterbrochenem Redefluss, Stottern sowie Halten des Rederechts auf der einen Seite und rhetorisch motivierten Wiederholungen auf der anderen Seite unterschieden; erstere wurden mit doub als Wiederholungen markiert.

Beispiel: *die* die kinder spielen (E01_A05)

Dabei wurde nur das jeweils letzte Wort bzw. die jeweils letzte Phrase nicht markiert. Rhetorisch motivierte Wiederholungen (Beispiel: *Das ist sehr sehr lecker*) werden als Teil der Äußerungseinheit angesehen und nicht auf [excl] markiert.

4.11.3 Selbst-Reparaturen / Korrekturen

Wird ein Teil einer Äußerung unmittelbar durch eine Reformulierung repariert / korrigiert, wurde der zu reparierende Teil nicht als Teil der Äußerungseinheit segmentiert, sondern mit repair auf [excl] getagged.

Beispiel: deine bruder *lernen* ee ein buch lesen (E01_B07)

Äußerungen, für die lediglich ein Korrekturbedarf angezeigt wird, wie z.B. *Das ist ein Hund – Nein!* oder auch *Das ist ein Hund – Nein! Kein Hund!* (konstruiertes Beispiel) wurden nicht ausgeschlossen.

4.11.4 Vorgelesene Passagen

Wird sprachliches Material von der Proband:in vorgelesen, aus dem Gedächtnis "zitiert", wie bei der Wiedergabe von Aufgabenstellungen, oder direkt von der Interviewer:in übernommen, wird dies mit dem Tag read ebenfalls exkludiert.

Beispiel: thema ist *wie hast du mutti kennengelernt* (E02_B02)

Beispiel: das ist ein korb, sagt man – *ein korb*? (E03_A04)

4.11.5 Hesitationsmarker

Isolierte, also außerhalb von units auftretende Hesitationsmarker (*em, mh, ee*) wurden mit dem Tag *hes* ebenfalls auf [excl] getagged und damit ausgeschlossen.

Äußerungsinterne Hesitationsmarker, Interjektionen, fremdsprachliches Material etc. wurde im folgenden Annotationsschritt auf der [pos]-Spur markiert bzw. korrigiert und differenziert.

4.11.6 Unterbrechungen / Scaffolding

Unterbrechungen durch die Gesprächspartner:in oder Dritte müssen bei der Segmentierung in Äußerungseinheiten nicht beachtet werden. Die Spanne wird ungeachtet der Unterbrechung über die von der Lerner:in geäußerte Einheit gezogen.

4.11.7 Tagübersicht für [excl]

Tabelle 4.6 zeigt eine Übersicht der Tags für die Annotationsebene [excl].

Tabelle 4.6: Übersicht der annotierten ausgeschlossenen Äußerungsanteile

Tag	Beschreibung
falst	False starts / Abbrüche
doub	Wiederholungen, die nicht rhetorisch motiviert sind. (Ausgeschlossen werden die vorangegangenen Wörter/Phrasen.)
repair	Wörter, Phrasen oder Äußerungseinheiten, die als Selbstreparatur markiert sind
read	Vorgelesene Äußerungen
hes	Hesitationssignal außerhalb von units

4.12 Ziel der Segmentierung und der Selektion ausgeschlossenen Materials

Am Ende dieses Verarbeitungsschrittes liegt ein Transkript vor, in dem alle Lerner:innen-Äußerungen mindestens über [unit] ODER [excl] getagged wurden. Außerdem wurden ggf. subclauses und verbale Koordination erfasst. Anhand der excl-Spur konnte anschließend eine "cleane" Spur ([ctok_part]) erzeugt werden, die nur Tokens enthält, die auch über [unit] erfasst wurden, und alles, was über [excl] erfasst wurde, ausschließt.

5 Lemmatisierung und POS-Tagging

5.1 Spuren bei Lemmatisierung und POS-Tagging

Diesen Annotationsschritten sind die folgenden Spuren mit einem neuen Sprecher *PART_CLEAN* zugeordnet:

Tabelle 5.1: Halbautomatische Annotationsspuren

[lemma]	Spur mit den durch den TreeTagger erzeugten und manuell korrigierten Lemmata
[pos]	Spur mit den durch den TreeTagger erzeugten und manuell korrigierten POS-Tags

5.2 Automatisierte Arbeitsschritte

Zunächst wurde über ein Skript geprüft, ob die Dateien verarbeitungsrelevante Fehler enthalten, die dann ggf. ausgebessert wurden.

Dann wurde die Spur [ctok_part] erstellt, die eine um Pausenzeichen und auf [excl] als auszuschließend markiertes Material bereinigte Kopie von [tok_part] umfasst.

Die Token auf [ctok_part] wurden anschließend mit Hilfe des TreeTagger Schmid (1995) (unter Einbezug der Satzspannen aus [unit]) anhand des STTS 2.0 (Westpfahl et al., 2017) mit einem POS-Tagging und einer Lemmatisierung versehen. Die automatisierte Annotation erfolgte über verschiedene Skripte mit Hilfe von *Annatto*[1] (vgl. Krause & Klotz, 2023).

5.3 Korrektur

Das POS-Tagging und die Lemmatisierung wurden auf den Spuren [lemma] und [pos], Kopien der automatisiert generierten Spuren, korrigiert. Die automatisiert generierten Spuren wurden anschließend gelöscht, können aber mithilfe der öffentlich zugänglichen Pipeline[2] reproduziert und somit für die Evaluation des Treetaggers für frühe Lerner:innendaten verwendet werden.

Tags wurden in Zweifelsfällen, wenn möglich, (auch) anhand funktionaler Kriterien korrigiert, bspw. in Bezug auf Verbformen (so kann z.B. “gib” als VVFIN anstatt als VVIMP getagged werden, wenn der Kontext diese Interpretation nahelegt).

Zweifelsfälle bei der Korrektur wurden im Vier-Augen-Prinzip besprochen; offensichtliche Versehen, die bei der Transkription oder Segmentierung entstanden sind, sowie Fehler durch die Pipeline wurden in diesem manuellen Korrekturschritt, sofern sie offensichtlich als Fehler zu erkennen sind

[1] https://github.com/korpling/annatto
[2] Die Pipeline ist über https://github.com/seiko-korpus/seiko-pipeline/ verfügbar.

(wenn etwa auf [excl] Markiertes nicht durch die Pipeline gelöscht wurde oder wenn offensichtlich zu exkludierendes Material nicht auf [excl] markiert oder offensichtlich ein falsches Annotationstag vergeben wurde), eigenständig und ohne weitere Besprechung korrigiert.

5.4 Lemmatisierung

Die Lemmatisierung orientiert sich grundlegend an den Richtlinien für das TIGER-Korpus (vgl. Proisl et al., 2019). Davon abweichendes Vorgehen ist in den folgenden Richtlinien jeweils als solches gekennzeichnet.

5.4.1 Normalisierung der Wortformen

Umgangssprachliche Wortformen wurden möglichst oberflächennah normalisiert, in schriftlichen Texten wurden orthographische Fehler korrigiert, Wortneuschöpfungen wurden normalisiert, aber nicht korrigiert (vgl. Proisl et al., 2019: 2).

Abkürzungen wurden, anders als bei den TIGER-Richtlinien (ebd.) nicht normalisiert. Kontraktionen wurden aufgelöst (*ums -> um es, son -> so ein, machste -> machst du*) und in einem Event einzeln und mit Leerzeichen zwischen den beiden Token lemmatisiert. Dieses Vorgehen unterscheidet sich ebenfalls von Proisl et al. (2019), die keine Auflösung von Kontraktionen vorgeben.

Wenn in der Transkription schriftlicher Daten Wörter zusammengeschrieben wurden, die zielsprachlich nicht zusammengeschrieben werden (bspw. *ausversehen*), wurden diese im gleichen Event durch Leerzeichen getrennt lemmatisiert (hierfür sehen Proisl et al., 2019 keine Regelung vor). Wenn Wörter auseinandergeschrieben wurden, die zielsprachlich zusammengeschrieben werden (*bus fahr Karte*, E10_B09_T), wurden diese auf der Lemma- und POS-Spur in jeweils einem Event zusammengeführt.

Wenn Fehler in der Transkription sichtbar wurden (bspw. *verliern* statt *verlieren*), wurden diese im Transkript und auf allen Spuren korrigiert. (Bei der Auswertung des automatischen Taggings/Lemmatisierens muss dementsprechend erwähnt werden, dass manche Token ggf. aufgrund von Transkriptionsfehlern nicht richtig getagged/lemmatisiert werden konnten.)

Wenn Wörter wegen einer wortinternen Pause in zwei Events aufgeteilt wurden, wurden die entsprechenden Events auf [pos] und [lemma] zusammengeführt und es wurde jeweils ein Tag pro Spur vergeben. Bei zwei möglichen Alternativen wurden beide, jeweils gefolgt von einem *or*, im gleichen Event lemmatisiert (z.B. *kontroll > KontrolleurorKontrolleor*). (Dieses Vorgehen orientiert sich an dem im STTS 2.0 für die Markierung möglicher Alternativen vorgesehenen Fragezeichen, das allerdings technische Schwierigkeiten bereitete und daher durch das *or* ersetzt wurde.) In geschriebenen Texten, in denen Wörter (nur) aufgrund graphematischer Interpretation und/oder unter Rückgriff auf den Bildstimulus und die mündlichen Daten zu rekonstruieren sind, werden die entsprechenden Rekonstruktionen lemmatisiert.

5.4.2 Substantive

NN: Substantive wurden in der 1. Person Singular Nominativ lemmatisiert. Ausnahmen bilden Pluraliatantum, die im Nominativ Plural lemmatisiert wurden. Substantivierte Formen wurden entsprechend ihrer Basiswortart lemmatisiert; substantivierte Infinitive dementsprechend als verbaler Infinitiv (*das Lesen > lesen*). Derivationen aus Formen des Partizip I (*die Überzeugende > überzeugen*) und Partizip II (*der Überzeugte > überzeugen*) wurden in ihrer verbalen Grundform

lemmatisiert, was sich vom Vorgehen bei Proisl et al. (2019) unterscheidet. Unserem Vorgehen nach wurden bei der Suche nach *überzeugen* auf [lemma] demnach alle Treffer angezeigt, die sich aus dem Lemma oder Derivationen davon ergeben; eine weitere Differenzierung nach Wortarten ist über [pos] möglich. Aus Adjektiven abgeleitete Substantive wurden in ihrer prädikativen Grundform lemmatisiert, bspw. *die Schöne* also als *schön* [und nicht, wie bei Proisl et al. (2019): 4, in der schwachen Form des Nominativs Singular Maskulin]. Grammatikalisierte Substantivierungen werden als Substantive lemmatisiert, was so in den TIGER-Richtlinien nicht ausdifferenziert wird.

NE: Eigennamen werden im Nominativ Singular lemmatisiert; Eigennamen ohne Singularform werden im Nominativ Plural lemmatisiert.

5.4.3 Adjektive

Adjektive werden in ihrer unflektierten Form im Positiv der prädikativen Form lemmatisiert. Falls es diese Form nicht gibt, wird die starke Form des Nominativs Singular Femininum (im Gegensatz zu Maskulinum bei Proisl et al., 2019: 5) lemmatisiert. Attributiv verwendete Partizipien wurden in ihrem verbalen Infinitiv lemmatisiert (die *überzeugte* Wählerin > *überzeugen*), was sich vom TIGER-Vorgehen unterscheidet (vgl. ebd.) und eine umfassende Lemmasuche, bei Bedarf mit Differenzierung der Wortart über [pos], erlaubt (s. auch Kapitel 5.4.2).

5.4.4 Zahlen

Zahlen werden in ihrer Grundform lemmatisiert.

5.4.5 Verben

Verben wurden im Infinitiv lemmatisiert. Kontrahierte Verben wie *schreibste* wurden im Infinitiv lemmatisiert. Der zweite Bestandteil der Kontraktion (Pronomen) wurde entsprechend der entsprechenden Regeln im gleichen Event mit Leerzeichen getrennt lemmatisiert, was sich vom TIGER-Vorgehen unterscheidet (vgl. Proisl et al., 2019: 2).

5.4.6 Artikel

Bestimmte Artikel werden als *die* lemmatisiert. Unbestimmte Artikel werden im Nominativ Singular Femininum (*eine*) lemmatisiert, was sich bzgl. des Genus von den TIGER-Richtlinien von Proisl et al. (2019) unterscheidet und u.a. im Output des Taggers begründet liegt.

5.4.7 Pronomen

Pronomen wurden, wenn möglich, im Nominativ Singular Femininum (statt Maskulinum wie bei Proisl et al., 2019: 7f.) lemmatisiert. Pronomen, die nur im Plural existieren (z.B. *beide*, *irgendwelche*), wurden in der indefiniten Form im Nominativ Plural Femininum lemmatisiert. Bei unflektierbaren Formen (*sich*, *wer*) wurde die Oberflächenform lemmatisiert.

Tabelle 5.2: Übersicht über die Lemmatisierung von Pronomen

Tag	Definition	Lemmatisierung	Beispiellemma
PDAT	Attribuierendes Demonstrativpronomen	Nominativ Singular Femininum	diese
PDS	Substituierendes Demonstrativpronomen	Nominativ Singular Femininum	die
PIAT	Attribuierendes Indefinitpronomen	Nominativ Singular / Plural Femininum / Oberflächenform	andere, irgendwelche, etwas
PIS	Substituierendes Indefinitpronomen ohne Determinierer	Nominativ Singular / Plural Femininum / Oberflächenform	eine, alle, etwas
PIDAT	Attribuierendes Indefinitpronomen mit Determinierer	Nominativ Singular Femininum	eine
PIDS	Substituierendes Indefinitpronomen mit Determinierer	Nominativ Singular Femininum / Plural / Oberflächenform	andere, beide, bisschen
PPER	Irreflexibles Personalpronomen	Nominativ Singular	Oberflächenform: er > er, ihnen > sie, ihm > er …
PPOSAT	Attribuierendes Possessivpronomen	Nominativ Singular Femininum	ihre, seine
PPOSS	Substituierendes Possessivpronomen	Nominativ Singular Femininum	meine, deine
PRELAT	Attribuierendes Relativpronomen	Nominativ Singular Femininum	deren
PRELS	Substituierendes Relativpronomen	Nominativ Singular Femininum	die
PRF	Reflexives Personalpronomen	sich	sich
PWAT	Attribuierendes Interrogativpronomen	Nominativ Singular Femininum	welche, wessen
PWS	Substituierendes Interrogativpronomen	Oberflächenform	wer, was
PWAV	Adverbiales Interrogativ- oder Reflexivpronomen	Oberflächenform	warum, wo

5.4.8 Adverbien

Adverbien wurden in ihrer Oberflächenform lemmatisiert. Bei Kontraktionen wie bei *sone* wurde das Adverb in seiner Oberflächenform lemmatisiert; der zweite Bestandteil der Kontraktion (Artikel) wurde, anders als bei Proisl et al. (2019, S. 10), ebenfalls lemmatisiert (gleiches Event, durch Leerzeichen getrennt).

5.4.9 Konjunktionen

Konjunktionen wurden in ihrer Oberflächenform lemmatisiert. Bei Kontraktionen wurde das Pronomen, anders als bei Proisl et al. (2019, S. 11), ebenfalls lemmatisiert (gleiches Event, durch Leerzeichen getrennt).

5.4.10 Adpositionen

Adpositionen wurden in ihrer Oberflächenform lemmatisiert.

APPRART: Bei Präpositionen mit Artikel wurden die Präposition und der Artikel separat lemmatisiert; der Artikel wurde also, anders als bei Proisl et al. (2019, S. 10), ebenfalls lemmatisiert (gleiches Event, durch Leerzeichen getrennt).

5.4.11 Partikeln

Partikeln wurden in ihrer Oberflächenform lemmatisiert.

5.4.12 Interjektionen, Diskursmarker (wie NGHES), ONO, Interpunktion

Das Lemma ist grundsätzlich die Oberflächenform. Normalisierbare Formen, die phonetisch transkribiert wurden (z.B. *oke*), wurden normalisiert (*okay*).

5.4.13 Fremdsprachliches Material

Lemma ist die Oberflächenform.

5.4.14 Unklares/Unverständliches

Bei unsicheren Transkriptionen wurde die Form, die als wahrscheinlichste transkribiert wurde, sofern sie im Kontext schlüssig ist, der Lemmatisierung zugrunde gelegt. *#fällt#* wurde also zu *fallen* lemmatisiert. Bei nicht auflösbaren unklaren Transkriptionen (z.B. *#anzo#*) wurde die Oberflächenform inklusive # transkribiert. Ganz Unverständliches, das als *XXX* transkribiert und auf [pos] als UI getagged wurde, wurde in dieser Form lemmatisiert. Dieses Vorgehen ist bei Proisl et al. (2019) nicht geregelt.

5.5 POS-Tagging

Das POS-Tagging verwendet das Stuttgart-Tübingen-Tagset 2.0 (STTS 2.0) (vgl. Westpfahl et al., 2017). Folgende Tags ergänzen das STTS 2.0:

5.5.1 Präfix- und Partikelverben

Für abgetrennte Präfixe, die zielsprachlich nicht abgetrennt werden können, ergänzen wir das Tag PTKVZU.

Für Präfix- und Partikelverben wird eine ergänzende Verbklasse *VP* eingeführt, für die sich analog zu den Vollverben (vgl. Westpfahl et al. (2017), 28) folgende Tags ergeben:

VPFIN, *VPIMP*, *VPINF*, *VPIZU*, *VPPP*

5.5.2 Kopula

Für Kopulae (*sein*, *werden*, *bleiben*) wird die Verbklasse VK ergänzt, hier ergeben sich die Tags:

VKFIN, *VKIMP*, *VKINF*, *VKIZU*, *VKPP*

5.5.3 *weil* als Diskursmarker vs. Subjunktion

Abweichend zum STTS 2.0 (26f.) wurde sich in der Differenzierung zwischen Diskursmarker *SEDM* und Subjunktion *KOUS* nicht an der Verbstellung orientiert, sondern an der Funktion des entsprechenden Wortes und an der Klassifikation der dazugehörigen Äußerung als eigenständig oder subordiniert, da die Verbstellung oft nicht zielsprachlich ist oder teilweise gar kein (finites) Verb vorhanden ist und die Äußerung dennoch als subordiniert und das entsprechende Token dementsprechend als *KOUS* klassifiziert werden kann.

Beispiele:

#weil#/KOUS ich ee kein ticket (E04_B07)

weil/KOUS ich ee vergessen mein ticket (E04_B07)

6 Annotation Nominalgruppen

6.1 Spuren bei der Annotation von Nominalgruppen

Nominalgruppen wurden auf [ngr] getagged. Auf [kern] wurde der Kern/Nukleus, das lexikalische Zentrum der NGr, erfasst sowie das Determinativum bzw. wurde dort ggf. die Information erfasst, dass ein obligatorisches Determinativum fehlt. Bei Bedarf (also Einbettung einer NGr in eine andere) konnten zweite usw. Spuren eingerichtet werden: [ngr2], [kern2] usw.

Tabelle 6.1: Annotationsspuren für Nominalgruppen

[ngr]	Nominalgruppen
[kern]	nominale Kerne

6.2 Grundlagen der Annotation von Nominalgruppen

Alle Nominalgruppen (NGr) wurden auf [ngr] basierend auf dem in Gamper (2022) genutzten Tagset erfasst. Um dem gerecht zu werden, dass Lerner:innen nominale Ausdrücke (und ihre Attributionen) nicht immer gemäß festgelegter Phrasenstruktur (vgl. Himmelmann, 1997) realisieren, wurde der Begriff der Nominalgruppe dem der -phrase vorgezogen (vgl. Eroms, 2016). Nominalgruppen müssen also nicht zielsprachlich realisiert sein und strikte Phrasenstrukturen aufweisen, um als NGr erfasst zu werden; das Gleiche gilt für potenzielle Attribute. So wurden etwa NGr mit nachgestellten Adjektiven, die klar als attributiv zum vorangegangenen Nomen erkannt werden können, als NGrADJ oder attributive Relativsätze mit abweichender Verbstellung als NGrREL getagged.

6.3 Abweichungen und Spezifizierungen zu Gamper (2022)

Anders als bei Gamper (2022) wurden **besitzanzeigende Präpositionalphrasen** mit *von* (bspw. *der hund von der frau*) nicht als postnominaler Genitiv (PGEN), sondern als Präpositionalphrase (PREP) annotiert.

Es wurden lediglich solche NGr als **mehrfach, also als NGrMULT** erfasst, bei denen sich zwei oder mehr Attribute auf den gleichen nominalen Kern beziehen (horizontale Attribution, vgl. Schmidt, 2011). Die Erweiterungen bei NGrMULT müssen nicht unterschiedlicher Art sein. Die dem NGrMULT folgenden einzelnen Tags geben die Attribute in der Reihenfolge, in der sie auftreten, wieder.

Beispiel: der perfekte momen um die Tute zu glauen (E20_A04; NGrMULTADJINF)

Vertikale Mehrfachattribution (vgl. Schmidt, 2011), also die Attribution eines nominalen Kerns, der selbst in eine attribuierte NGr eingebettet ist, ist durch die Annotation der Nominalgruppe auf der Spur [ngr2] erfasst.

Beispiel: die junge das das rote tshirt hat (E09_B11; NGrREL auf [ngr], NGrADJ auf [ngr2])

Anders als bei Gamper (2022) wurden für **Präpositionalphrasen** keine gesonderten Tags vergeben; auf oberster Ebene wird NGrPREP annotiert; in der Präpositionalphrase enthaltene NGrs werden als solche auf [ngr2] erfasst.

Als mit **Apposition** erweiterte NGr (NGrAPP) wurden auch Strukturen wie *bild vier* oder *nummer sechs* erfasst, die im Korpus häufig als kohärenzstiftende Mittel eingesetzt werden.

Außerdem wurde das Tag NGrSO eingeführt, das NGr mit dem Prädeterminativ *so* erfasst (vgl. Hirschmann, 2023). Auch wenn diese NGr nicht an sich als erweitert gelten, wurden sie, wenn sie gemeinsam mit Attributen auftreten, mit dem kombinierten NGrMULT-Tag erfasst, also etwa NGrMULTSOPREP. So können auch diese NGr differenziert abgefragt werden; die weitere Differenzierung kann dann in späteren Schritten vollzogen werden, indem etwa NGrMULTSO+ aus den mehrfach erweiterten NGr ausgeschlossen werden.

Zudem wurden NGrQ eingeführt, die "slightly extended nouns" (Klein & Dittmar, 1979: 134) erfassen. Diese NGr enthalten quantifizierende Determinatoren, die auch in Kombination mit anderen Determinatoren auftreten können, etwa *die beiden kinder*. Auf [kern] werden diese Quantifikatoren mit QUA erfasst.

Auf [kern] wurde der **lexikalische Kern der NGr**, also das Nomen, mit dem Tag KERN erfasst. Dem Kern zugehörige Determinative wurden auf der gleichen Spur mit DET getagged. Wenn ein Determinativum obligatorisch wäre und nicht realisiert wurde, wird das mit dem ergänzenden Tag 0DET ebenfalls erfasst. Wenn zwei Determinative realisiert wurden oder ein Determinativum realisiert wurde, wo ein Nullartikel zielsprachlich obligatorisch gewesen wäre, wurde das mit pDET erfasst. Das Tag QUA erfasst quantifizierende Determinatoren.

6.3.1 ngr

Tabelle 6.2: Übersicht der annotierten Nominalgruppen

Tag	Beschreibung	Beispiel
NGrO	nicht erweiterte / einfache NGr (mit oder ohne Determinierer, nominal; keine Pronomen)	der lehrerin (E01_A04)
NGrCO	koordinierte, nicht erweiterte NGr (auch bei Aufzählungen)	diesem ball oder diesem papier (E12_D01)
NGrQ	nur mit ggf. zusätzlich zum Artikel vorhandenen Quantifikator erweiterte NGr	die zwei mädchen (E10_B10)
NGrADJ	NGr enthält attributives Adjektiv.	eine braune gürtel (E08_A05), die alte frau (E10_D02)
NGrPRT	NGr enthält Partizip.	eine gestorben blume (E08_C03), ein verstecke buch E12_B09)
NGrGEN	NGr enthält pränominales Genitivattribut.	marias buch (E11_B02)
NGrSO	NGr enthält pränominales "so".	so eine typ (E02_D04)
NGrPADV	NGr enthält postnominalen adverbiales Attribut.	diese junge hier (E09_B07)
NGrPREP	NGr enthält postnominale Präpositionalphrase.	die tasse mit dem tee (E01_B02)
NGrPGEN	NGr enthält postnominalen Genitiv.	das Buch seines Bruder (E01_C02)
NGrPEX	NGr enthält andere postnominale Attribute (wie-phrase, als-phrase).	sportlehrer wie so herr NAME (E09_C05)
NGrAPP	NGr enthält postnominale Apposition.	eine tasse tee (E01_B02)
NGrREL	NGr enthält Relativsatz.	ein frau welche wollen einkaufen (E02_C02)
NGrINF	NGr enthält Infinitivsatz.	keine zeit ein neuen zu backen (E11_A05)

NGrCOMP	NGr enthält Komplementsatz.	keine ahnung wo ist mein tasche (E08_D05)
NGrMULT	NGr hat mehr als ein Attribut auf einem Level.	eine andere name für salon (E13_C04)
ADJ	als ergänzendes tag bei NGrMULT	
PRT	als ergänzendes tag bei NGrMULT	
EXPRT	als ergänzendes tag bei NGrMULT	
GEN	als ergänzendes tag bei NGrMULT	
PADV	als ergänzendes tag bei NGrMULT	
PREP	als ergänzendes tag bei NGrMULT	
PGEN	als ergänzendes tag bei NGrMULT	
PEX	als ergänzendes tag bei NGrMULT	
APP	als ergänzendes tag bei NGrMULT	
REL	als ergänzendes tag bei NGrMULT	
INF	als ergänzendes tag bei NGrMULT	
COMP	als ergänzendes tag bei NGrMULT	

6.4 kern

Tabelle 6.3: Übersicht der annotierten Nominalgruppenkerne und -determinierer

Tag	Beschreibung	Beispiel
DET	Determinierer des Kerns	
KERN	Kern der NGr	
0DET	als ergänzendes Tag für zu determinierende, aber nicht determinierte Kerne, an KERN angehängt	frau kauft orange (E02_C05)
pDET	DET bei nicht zu determinierenden, aber determinierten Kernen	eine leut (E03_D01)
QUA	quantifier	ein bisschen angst (E02_B02)

7 Annotation Verben

7.1 Spuren bei der Annotation von Satzkonstituenten und Verbstellung

Tabelle 7.1: Annotationsspuren für Verbstellungsmuster

[unit_cons]	Satzkonstituenten von unit, jeweils entsprechend der Bezugsspur (z.B. unit2_cons; subclause_cons)
[unit_stage]	Verbstellungsstufe von unit, jeweils entsprechend der Bezugsspur (z.B. unit2_stage; subclause_stage)

7.2 Annotation Satzkonstituenten

Die Annotationen von Satzkonstituenten und Verbstellungsmustern (vgl. Kapitel 7) sind, soweit es für SeiKo sinnvoll erscheint, an den Annotationsspezifikationen des DAKODA-Projekts (vgl. Ruppenhofer et al., 2024) orientiert.

Eine Annotation der Satzkonstitutenten erfolgt nur für verbhaltige units. Das heißt DISC IRNON und SCU werden grundsätzlich nicht berücksichtigt.

Für jede vorhandene [unit]- und [subclause]-Spur muss eine korrespondierende [cons]-Spur angelegt werden. (Diese werden dann z.B: [unit_cons] oder [subclause_cons] genannt.)

Die Konstituenten werden als Spannen unter den jeweiligen [pos]-Events annotiert. Eine Übersicht der möglichen Tags wird in Tabelle 7.2 dargestellt.

Tabelle 7.2: Übersicht der annotierten Satzkonstituenten

Tag	Beschreibung
V	Verb
S	Subjekt (nominal, pronominal)
Sexpl	Subjekt expletiv
Scl	Subjekt satzförmig
Sext	zugehöriger Teil eines Subjekts
O	Objekt (nominal, pronominal)
Oexpl	Objekt expletiv
Ocl	Objekt satzförmig
Oext	zugehöriger Teil eines Objekts
A	Argument
Aext	zugehöriger Teil eines Arguments
Pred	Prädikativ
Predcl	Prädikativ satzförmig

cl	abhängiger Nebensatz
ext	DISCONT. Konstituenten, z.B. Oext
N	Negator
C	Junktor
X	Restkategorie: Modifizierer/Adjunkt oder Partikel
Xcl	Restkategorie satzförmig
@	unklare Elemente

Im Unterschied zu Ruppenhofer et al. (2024) unterscheiden wir verschiedene Verbtypen nicht auf [unit_cons], da diese auf [pos] differenziert getagged werden.

Falls eine Konstituente in mehreren Teilen realisiert wird, wird der Kopf mit dem regulären Tag annotiert. (+)ext markiert dann den zweiten Teil der Konstituente.

Negatoren werden nur als solche annotiert, wenn es sich um eine Satznegation handelt.

Beispiel: die Katze frisst nicht = [cons] SVN

Andernfalls wird der Negator als Teil der Konstituente annotiert.

Beispiel: Die Katze frisst nicht den Vogel, sondern die Maus = [cons] SVA

Unklare Elemente werden mit @ annotiert.

7.3 Annotation Verbstellung

Bei der Annotation der Verbstellungsmuster orientieren wir uns ebenfalls an den Spezifikationen von Ruppenhofer et al. (2024).

Für alle [unit]- und [subclause]-Spuren werden korrespondierende [stage]-Spuren angelegt. (z.B. [unit_stage], [subclause_stage], etc.).

Für jede verbhaltige unit wird dann auf der entsprechenden [stage]-Spur eine Spannannotation erstellt, die alle für die Zuordnung relevanten Konstituenten umfasst. Nicht konnektintegrierte Konnektoren (*und*, *oder*) werden dabei nicht mit eingeschlossen!

Weist eine unit Eigenschaften mehrere Verbstellungsmuster bzw. Abweichungen oder spezifischen Interpretationsformen dieser auf, werden alle relevanten Tags vergeben. Die einzelnen Tags und (einige) mögliche Kombinationen werden in Tabelle 7.3 dargestellt.

Tabelle 7.3: Übersicht der annotierten Verbstellungsstrukturen

Tag	Beschreibung	Beispiel
nostage	keine PT-Stufe zuzuordnen	sprechen mit der kinder (E01_B07)
nostagekor	keine PT-Stufe zuzuordnen, da aufgrund eines Koordinationskontextes eine syntaktische Reduktion vorliegt	(Dann ist die Mutter gekommen) und sagte (E19_A04)
SV	Subjekt und Verb ohne weiteres Element	wir kochen (E05_B10)
SVX	Subjekt und Verb mit folgendem Element, das kein Objekt ist	und sie versteht nicht (E06_C05)
SVO	kanonische Satzstellung in einem Deklarativsatz mit finitem Verb	und der junge em schmeißt n ball (E07_A05)
ADV	SVO oder SVX mit einem weiteren Element vor dem Subjekt im Vorfeld	und danach n die gibt eine junge ein tee (E07_A05)
SEP	Verbklammer mit gefülltem Mittelfeld	und ein junge hat tor gemacht (E09_B11)
NOSEP	nicht realisierte Separation in SEP-Kontext	Ein Junge hat verloren in Basketball (E09_C03)
SEPnoev	SEP mit leerem Mittelfeld (Nähestellung)	und der tasse is hingefallen (E01_A05)

SEPanti	obligatorische Objektsatz im Nachfeld, daher leeres Mittelfeld	Junge hat gesagte was sie müssen leise sein (E01_C02)
SEP ADV		und dann der kinder • ee macht kaffee runter (E01_A05)
NOSEP ADV		in diese bild wir • • • können sehen (E03_C02)
SEPnoev ADV		danach • • ee die junge willen umziehen (E09_C04)
SEPanti ADV		
INV	Subjektinversion	und da kommt die mutter (E01_A05)
INVpseudo	nach W-Adverb	wie sagt man das? (E08_A04)
INVin	eingebettete Inversionen	(und das wars) glaube ich (E07_C02)
INV SEP		vielleicht ee • wollte sie einfach mit er ee • • • lachen (E12_C04)
INVpseudo SEP		warum hat sie • nicht gesagt (E06_A04)
INVin SEP		
INV SEPnoev		dann gehen • • ee sitzen (E05_D04)
VEND	Verbendstellung im Nebensatzkontext	weil der verloren hat (E09_A04)
VENDnoev	wenn nur eine Konstituente vor dem finiten Verb steht (+ggf. Konnektor) oder in zielsprachlichen Kontexten (etwa V2-Nebensätze)	ob der telefoniert (E15_B10)
!VEND	VEND anstelle von MC-Strukturen	ich buch lesen (E01_D04)
!SVX NO-VEND(noev)	SVX Struktur in einem VEND-Kontext	
!SVO NO-VEND(noev)	SVO Struktur in einem VEND-Kontext	
!ADV NO-VEND(noev)	ADV Struktur in einem VEND-Kontext	
!SEP NO-VEND(noev)	SEP Struktur in einem VEND-Kontext	
!NOSEP NO-VEND(noev)		
!SEPnoev NO-VEND(noev)		
!INV NO-VEND(noev)	INV Struktur in einem VEND-Kontext	
!INVnoev NO-VEND(noev)		
!INVpseudo NO-VEND(noev)		
!SEP !INVpseudo NOVEND(noev)		
V1imp	für V1 Imperative	zeig mir • eure • buskart (E10_C04)
Ø	wenn auf unit eine nicht realisierte Äußerungseinheit annotiert ist, von der eine subclause abhängt	

Ruppenhofer et al. (2024) unterscheiden in ihren Annotationsspezifikationen grundsätzlich zwischen Strukturen, die als Kernindikatoren für die Verbstellungsstufen gelten können (core), und Strukturen, die in den meisten vorliegenden Arbeiten nicht als Indikatoren gewertet werden (peri). Wir ergänzen diese Zuordnungen nicht als eigene Analysekategorien, da die Merkmale aus anderen der vorliegenden Annotationen abgeleitet werden können.

Im Folgenden werden die Kriterien für die Zuordnung der Kernstufen und die Interpretation von Varianten dargestellt.

7.3.1 SVO

Als Indikatoren für die kanonische Verbstellung (SVO) werten wir alle Deklarativsätze, die ein Verb in der linken Satzklammer sowie ein lokales Subjekt und Objekt enthalten. (Abweichend von Ruppenhofer et al., 2024 wird SVO auch bei Kopulaverben annotiert.)

Enthält die Äußerung hingegen eine weitere Konstituente, die kein Objekt ist, wird SVX annotiert. Folgt nach dem Verb keine weitere Konstituente, ist das Tag SV.

SVO wird auch für Strukturen annotiert, die auf [subclause] mit MCSUB(+) getagged werden.

SVO/SVX wird auch bei Sätzen mit Interrogativpronomen, die als Subjekt des Satzes fungieren, annotiert.

Beispiel: Wer war das? Was ist los?

Treten SV/SVX/SVO-Strukturen in Kontexten auf, in denen sie ungrammatisch sind (VEND), werden sie als !SV, !SVX oder !SVO zusammen mit der nicht realisierten Struktur annotiert (NOVEND oder NOVENDnoev).

7.3.2 ADV

SVO-Strukturen bei denen ein weiteres Element neben dem Subjekt das Vorfeld besetzt, werden mit ADV annotiert. Wir unterscheiden dabei nicht zwischen verschiedenen Konstituenten oder der Position vor oder nach dem Subjekt im Vorfeld.

ADV wird jedoch nicht bei topikalisierten Nominalgruppen annotiert.

Beispiel: Die Katze sie frisst

Ebenso wird ADV nicht bei zielsprachlichen mehrfache Vorfeldbesetzungen gesetzt.(z.B. Konnektoren wie Die Katze aber schläft immer noch.)

Auch ADV kann ungrammatisch in Kontexten auftreten, in denen VEND obligatorisch wäre. Diese Strukturen werden dann mit !ADV NOVEND annotiert.

7.3.3 SEP

SEP-Strukturen zeichnen sich durch ein komplexes Prädikat in Distanzstellung bei besetztem Mittelfeld aus. Ein lokales Subjekt ist dafür nicht zwingend erforderlich.

Wird diese Separation nicht realisiert, wird die Struktur mit NOSEP getagged.

Bei leerem Mittelfeld und wenn kein weiteres Element auf den finiten Prädikatsteil folgt (Nähestellung), kann die Struktur nicht als Evidenz für Separation gewertet werden (SEPnoev).

Eine obligatorische Nähestellung liegt bei Verben vor, die einen Objektsatz fordern, der dann im Nachfeld realisiert wird. Solche Strukturen werden mit SEPanti annotiert.

Auch SEP-Strukturen können in MCSUB(+) [subclause] in Nebensätzen auftreten.

SEP kann zusammen mit ADV und INV und ebenso anstelle von VEND-Strukturen auftreten (!SEP NOVEND).

7.3.4 INV

Für den Nachweis einer Subjektinversion muss das Subjekt im Mittelfeld realisiert werden, während eine weitere Konstituente das Vorfeld besetzt.

Bei einem unbesetzten Vorfeld bei V1-Exklamativa wird ebenfalls INV annotiert.

Beispiel: Frisst die schon wieder einen Vogel!

Parenthetische Inversionsstrukturen (Beispiel: Die Katze, *glaube ich*, frisst den Vogel) werden in SeiKo mit INVin annotiert.

INVpseudo wird nach W-Adverbien annotiert; bei Sätzen mit Interrogativpronomen, die als Subjekt des Satzes fungieren, wird nicht INVpseudo, sondern SVO/SVX annotiert. (Bsp.: Wer war das? Was ist los?)

7.3.5 VEND

Die Verbendstellung ist grammatisch in Nebensatzkontexten. Ein Subjekt ist für die Zuordnung nicht notwendig.

Folgt auf einen Subjunktor kein weiteres Element (CSV statt CSXV auf [cons]) wird mit VENDnoev getagged.

Treten VEND-Strukturen in Kontexten auf, in denen z.B. SVO grammatisch wäre, werden diese mit !VEND annotiert.

Bei einigen Nebensätzen (z.B. nach Subjunktor *weil*) werden für das Deutsche auch V2-Strukturen im Nebensatz als zielsprachlich angesehen. Wir annotieren diese mit NOVENDnoev (meist im Zusammenhang mit der tatsächlichen Struktur (also z.B. SEP NOVENDnoev)

Beispiel: weil die Maus ist weggelaufen

7.3.6 V1

Neben den in Kapitel 7.3.4 beschriebenen V1-Strukturen wird V1imp für Imperative mit Verberststellung genutzt.

7.3.7 nostage

Kann keine Verbstellungsstruktur zugeordnet werden, wird mit nostage getagged. Ist die Zuordnung nicht möglich, weil durch syntaktische Reduktion (z.B. bei Koordinationsellipsen) obligatorische Konstituenten nicht realisiert werden, ist das Tag nostagekor (vgl. Kapitel 7.3.9).

7.3.8 Ø

Ø wird auf unit_stage annotiert, wenn auf unit ebenfalls Ø annotiert ist, was für eine nicht realisierte Äußerungseinheit, von der eine subclause abhängt, steht.

7.3.9 Koordination(sellipsen)

Koordinierte Äußerungseinheiten werden als getrennte Einheiten segmentiert.

Beispiel: *Die Katze schläft und die Maus frisst Käse*

Dies gilt auch bei doppelter Prädikation.

Beispiel: *Die Maus frisst einen Käse und schläft dann ein.*

sowie bei koordinierten Verbalphrasen

Beispiel: *Die Katze schläft und schnarcht.*

Das Kriterium für die Segmentierung ist die Verbhaltigkeit einer unit. Dieses Vorgehen unterscheidet sich von dem von Foster et al. (2000), die koordinierte Einheiten unter bestimmten Bedingungen (vgl. ebd.: 367) als Teil einer AS-Unit fassen.

Koordinierte Nominalgruppen werden nicht getrennt segmentiert.

8 Qualitätssicherung und Evaluation

Dieses Kapitel beschreibt das Vorgehen zur Sicherung der Konsistenz und Qualität der Datenaufbereitung.

8.1 Erarbeitung und Präzisierung der Richtlinien

Die Richtlinien für Segmentierung und Annotation wurden auf Grundlage theoretischer Vorarbeiten und bestehender Manuals erarbeitet und für die spezifischen Anforderungen des Korpus iterativ angepasst. Ein zentraler Bestandteil dieses Prozesses war der fortlaufende datennahe Austausch im Team: Annotationsentscheidungen wurden anhand konkreter Beispiele diskutiert, um Unklarheiten zu identifizieren, die Richtlinien zu präzisieren und notwendige Anpassungen vorzunehmen.

8.2 Vier-Augen-Prinzip

Im Prozess der Datenaufbereitung wurde, soweit möglich, nach dem Vier-Augen-Prinzip gearbeitet. Neben dem Austausch zu Problemfällen und der Diskussion eines Konsens wurde der Arbeitsprozess so strukturiert, dass zwischen zentralen Arbeitsschritten ein Wechsel der Bearbeiter:in erfolgte. So wurden die Transkriptionen bei der folgenden Segmentierung zusätzlich überprüft und Zweifelsfälle ggf. besprochen. Ebenso wurden die Segmentierungen zusammen mit den automatischen Annotationen während der manuellen Annotation geprüft und korrigiert.

8.3 Inter-Annotator-Agreements

Zusätzlich führten wir zu zwei Zeitpunkten eine Evaluation der Annotationen anhand der Berechnung eines Inter-Annotator-Agreements (IAA) durch.[1]

Zunächst erfolgte dies während der Datenaufbereitung von SeiKo1 mit einem Sample von 21 % der bis zu diesem Zeitpunkt vorliegenden Daten. Für die doppelt und unabhängig annotierten Stichprobendaten wurde ein IAA berechnet, das die hierarchischen Abhängigkeiten in mehrschichtigen Annotationsebenen als Baumstrukturen interpretiert und mit einem dafür geeigneten Maß vergleicht (vgl. Skjærholt, 2013).

Hierbei konnte für Krippendorff's alpha (Krippendorff, 1970) ein Wert von .73 für alle Verbstellungsstrukturen berechnet werden, für die Verbstellungsstrukturen eingebetteter AS-units ein Wert von .43. Für die Nominalgruppen (NGr) ergab sich ein ähnliches Bild: NGr auf erster Ebene, also solche, die nicht selbst in andere Nominalgruppen eingebettet sind, ergaben einen Wert von 0.86, während abhängige NGr lediglich bei .43 lagen.

[1]Für Beratung zu und Berechnung der Inter-Annotator-Agreements bedanken wir uns herzlich bei Martin Klotz!

Die Sichtung inkonsistenter Annotationen und ihre gemeinsame Diskussion dienten dazu, die Richtlinien zu präzisieren und die Übereinstimmung bei späteren Annotationen zu erhöhen. Nach Abschluss dieses iterativen Prozesses wurden die finalen Richtlinien auf alle Daten angewendet.

Nach abgeschlossener Annotation von SeiKo1 wurde ein über die Schüler:innen, Erhebungszeitpunkte und Textteile balanciertes Zufallssample aus ca. 21 % der Daten (20,9 % der [pos]-Token, 20,6 % [unit]-Token, 20,7 % [stage]-Token, 20,6 % [ngr]-Token) gezogen und unabhängig von zwei Annotator:innen annotiert. Das bedeutet, dass insgesamt 1488 verbhaltige Äußerungen und 1248 Nominalgruppen doppelt annotiert wurden und auf dieser Grundlage erneut ein IAA berechnet wurde.

Zur Bestimmung des Inter-Annotator-Agreements wurde für die hierarchisch strukturierten Ebenen ([ngr], [unit_cons], [subclause_cons]) Krippendorffs alpha auf Basis von Tree-Edit-Distance bestimmt Pawlik & Augsten (2012). Die Ergebnisse finden sich in Tabelle 8.1.

Tabelle 8.1: Übersicht über das IAA hierarchisch strukturierter Ebenen in SeiKo1.

Annotationsebene	(alle Token)	(nur annotierte Token)
Nominalgruppen	0.87	0.89
Konstituenten unit	0.90	0.89
Konstituenten subclause	0.90	0.94

Die Berechnung des IAA der [stage]-Annotationen war nicht darauf angewiesen, ihre hierarchische Struktur zu berücksichtigen, da Spannen und Struktur dieser Annotationen bereits durch den vorigen Annotationsschritt, die Segmentierung, für beide Annotator:innen gleichermaßen vorgegeben waren. Das IAA für diese Strukturen betrug 0.9 für Annotationen auf einer [unit]-Ebene bzw. 0.93 auf einer [subclause]-Ebene.

Die aus dem IAA entstandene Übersicht über doppelt annotierte Strukturen wurde außerdem genutzt, um systematische Unterschiede zu identifizieren, Unschärfen erneut zu besprechen und zu glätten, und gemeinsam beschlossene Korrekturen in den Daten vorzunehmen.

Auf Basis dieser Ergebnisse und unter Beibehaltung des Vier-Augen-Prinzips wurden anschließend die Daten aus SeiKo2 annotiert.

9 Literaturverzeichnis

Bley-Vroman, Robert. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, *33*(1), 1–17. https://doi.org/10.1111/j.1467-1770.1983.tb00983.x

Eroms, Hans-Werner. (2016). Zur Geschichte und Typologie komplexer Nominalphrasen im Deutschen. In Mathilde Hennig (Hrsg.), *Komplexe Attribution: Ein Nominalstilphänomen aus sprachhistorischer, grammatischer, typologischer und funktionalstilistischer Perspektive* (S. 21–55). De Gruyter. https://doi.org/10.1515/9783110421170

Foster, Pauline, Tonkyn, Alan, & Wigglesworth, Gillian. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Gamper, Jana. (2022). Ausbau nominaler Strukturen in der Sekundarstufe I. Eine textkorpusanalytische Studie. *Korpora Deutsch als Fremdsprache*, *2*(2), 13–42. https://doi.org/10.48694/kordaf.3551

Gilquin, Gaëtanelle. (2022). One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching*, *55*(1), 87–99. https://doi.org/10.1017/S0261444821000094

Himmelmann, Nikolaus P. (1997). *Deiktikon, Artikel, Nominalphrase: Zur Emergenz syntaktischer Struktur*. De Gruyter. https://doi.org/10.1515/9783110929621

Hirschmann, Hagen. (2023). Das pränominale so und seine lexikalischen Varianten im Deutschen – eine syntaktische, konstruktions- und variationslinguistische Beschreibung. In Marc Felfe, Dagobert Höllein, & Klaus Welke (Hrsg.), *Regelbasierte Konstruktionsgrammatik* (S. 183–230). De Gruyter. https://doi.org/10.1515/9783111334042-006

Klein, Wolfgang, & Dittmar, Norbert. (1979). *Developing Grammars* (Willem J. M. Levelt, Hrsg.; Bd. 1). Springer. https://doi.org/10.1007/978-3-642-67385-6

Krause, Thomas, & Klotz, Martin. (2023). *Annatto*. https://github.com/korpling/annatto

Krause, Thomas, & Zeldes, Amir. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, *31*(1), 118–139. https://doi.org/10.1093/llc/fqu057

Krippendorff, Klaus. (1970). Bivariate Agreement Coefficients for Reliability of Data. *Sociological Methodology*, *2*, 139. https://doi.org/10.2307/270787

Massumi, Mona, & von Dewitz, Nora. (2015). *Neu zugewanderte Kinder und Jugendliche im deutschen Schulsystem: Bestandsaufnahme und Empfehlungen*. https://international.uni-koeln.de/sites/international/aaa/92/92pdf/92pdf_REFUGEES_PUBL_MI_ZfL_Studie_Zugewanderte_im_deutschen_Schulsystem_final_screen.pdf

Meyermann, Alexia, & Porzelt, Maike. (2014). *Hinweise zur Anonymisierung von qualitativen Daten. Version 1.0*. Leibniz-Institut für Bildungsforschung und Bildungsinformation.

Pawlik, Mateusz, & Augsten, Nikolaus. (2012). *RTED: A Robust Algorithm for the Tree Edit Distance*. https://doi.org/10.48550/ARXIV.1201.0230

Proisl, Thomas, Dykes, Natalie, Heinrich, Philipp, Kabashi, Besim, & Evert, Stefan. (2019). *Lemmatisierungsrichtlinien. Version 2019-04-26*.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, & Svartvik, Jan. (1985). *A Comprehensive grammar of the English language*. Longman.

Rehbein, Jochen, Schmidt, Thomas, Meyer, Bernd, Watzke, Franziska, & Herkenrath, Annette. (2004). *Handbuch für das computergestützte Transkribieren nach HIAT*. Institut für Deutsche

Sprache.

Ruppenhofer, Josef, Schwendemann, Matthias, Portmann, Annette, & Wisniewski, Katrin. (2024). *Specification of Processability Theory's Developmental Stages.*

Schellhardt, Christin, & Schroeder, Christoph. (2015). *MULTILIT: Manual, criteria of transcription and analysis for German, Turkish and English.*

Schmid, Helmut. (1994). *Probabilistic Part-of-Speech Tagging Using Decision Trees.* Proceedings of International Conference on New Methods in Language Processing.

Schmid, Helmut. (1995). *Improvements in Part-of-Speech Tagging with an Application to German.* Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Schmidt, Jürgen Erich. (2011). *Die deutsche Substantivgruppe und die Attribuierungskomplikation.* Niemeyer. https://doi.org/10.1515/9783110958515

Schmidt, Thomas, & Wörner, Kai. (2014). EXMARaLDA. In *Handbook on Corpus Phonology* (S. 402–419). Oxford University Press.

Schwendemann, Matthias, Portmann, Annette, Schlauch, Julia, Gamper, Jana, & Wisniewski, Katrin. (2024). *Forschungsdaten für alle. Rechtliche Möglichkeiten und Herausforderungen bei der Veröffentlichung von Lernerkorpora. 4*(2), 149–176. https://doi.org/10.48694/kordaf.4138

Skjærholt, Arne. (2013). Influence of preprocessing on dependency syntax annotation: speed and agreement. *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, 28–32.

Skjærholt, Arne. (2014). A chance-corrected measure of inter-annotator agreement for syntax. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 934–944. https://doi.org/10.3115/v1/P14-1088

Westpfahl, Swantje, Schmidt, Thomas, Jonietz, Jasmin, & Borlinghaus, Anton. (2017). *STTS 2.0: Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS).*

Wiese, Heike. (2020). Language Situations: A Method for Capturing Variation within Speakers' Repertoires. In Yoshiyuki Asahi (Hrsg.), *Methods in Dialectology XVI* (S. 105–117). Peter Lang.

10 SeiKo-Nutzungsbedingungen

10.1 Gegenstand der Nutzungsbedingungen

Diese Nutzungsbedingungen regeln den Zugang zu SeiKo und die Nutzung der darin bereitgestellten Daten. Dazu gehören insbesondere Transkriptionen, Annotationen und Metadaten sowie – soweit gesondert freigegeben – die Audiodateien.

Der Zugang wird ausschließlich nach Prüfung und Genehmigung einer individuellen Zugriffsanfrage gewährt. Mit der Beantragung des Zugangs und der anschließenden Nutzung der Daten erkennt die nutzungsberechtigte Person diese Nutzungsbedingungen an.

10.2 Zugangsberechtigung

Der Zugang wird ausschließlich natürlichen Personen gewährt, die sich authentifiziert und Angaben zu ihrer institutionellen Zugehörigkeit sowie zum vorgesehenen Nutzungszweck gemacht haben.

Die Zugangsberechtigung ist persönlich und nicht übertragbar. Zugangsdaten, Freigabelinks oder heruntergeladene Korpusdateien dürfen nicht an andere Personen weitergegeben werden.

Weitere Personen, die im Rahmen eines Forschungs- oder Lehrvorhabens mit den Daten arbeiten wollen, müssen jeweils eine eigene Zugangsberechtigung beantragen.

10.3 Zulässige Nutzungszwecke

Das SeiKo-Korpus darf ausschließlich zu wissenschaftlichen, nicht kommerziellen Zwecken genutzt werden. Zulässig ist insbesondere die Nutzung für:

- wissenschaftliche Forschung,
- wissenschaftliche Lehre,
- Studium und wissenschaftliche Qualifikationsarbeiten.

Jede darüber hinausgehende Nutzung, insbesondere eine kommerzielle Nutzung, Auftragsforschung mit kommerziellem Verwertungsinteresse oder die Verwendung zur Entwicklung kommerzieller Produkte und Dienstleistungen, bedarf der vorherigen schriftlichen Zustimmung der Lizenzgeber:in.

Die Daten dürfen nur für den in der Zugriffsanfrage angegebenen Zweck verwendet werden. Wesentliche Änderungen oder Erweiterungen des Vorhabens sind dem SeiKo-Projekt mitzuteilen und können eine erneute Genehmigung erfordern.

10.4 Zulässige Bearbeitung der Daten

Im Rahmen des genehmigten Vorhabens dürfen die bereitgestellten Daten:

- heruntergeladen und auf geeigneten Systemen gespeichert,
- vervielfältigt, konvertiert und technisch aufbereitet,
- ausgewertet und mit anderen Forschungsdaten verknüpft,
- korrigiert, ergänzt und mit zusätzlichen Annotationen versehen werden.

Diese Erlaubnis umfasst ausschließlich die interne Bearbeitung für das genehmigte Forschungs-, Lehr- oder Studienvorhaben. Sie berechtigt nicht zur Weitergabe oder Veröffentlichung der bearbeiteten Daten.

10.5 Verbot der Weitergabe und Veröffentlichung

Das SeiKo-Korpus sowie einzelne Korpusdateien oder wesentliche Teile daraus dürfen weder in ihrer ursprünglichen noch in einer veränderten oder ergänzten Form:

- an nicht zugangsberechtigte Dritte weitergegeben,
- veröffentlicht oder verbreitet,
- in öffentlich zugängliche Repositorien hochgeladen,
- als Anhang oder Supplement zu einer Publikation bereitgestellt,
- über Webseiten, öffentliche Code-Repositorien oder andere Plattformen zugänglich gemacht werden.

Dies gilt auch für bearbeitete Fassungen, konvertierte Dateien, korrigierte Transkriptionen, zusätzliche Annotationen und andere abgeleitete Korpusversionen, sofern diese eine Rekonstruktion oder eigenständige Nutzung der ursprünglichen Korpusdaten ermöglichen.

Die Weitergabe aggregierter, anonymisierter Analyseergebnisse, die keine Rekonstruktion einzelner Korpusdaten oder Rückschlüsse auf beteiligte Personen ermöglichen, bleibt zulässig.

10.6 Veröffentlichung zusätzlicher Annotationen und Korpusversionen

Bearbeitete Korpusfassungen, zusätzliche Annotationen, Korrekturen oder Erweiterungen können dem SeiKo-Projekt zur Prüfung und dauerhaften Aufnahme in das Korpus angeboten werden.

Eine Veröffentlichung solcher Daten kann nach vorheriger schriftlicher Vereinbarung über das SeiKo-Projekt erfolgen. Dabei werden insbesondere die Qualitätssicherung, die Versionierung, die Dokumentation sowie die angemessene Nennung der beteiligten Personen gesondert geregelt.

Eine eigenständige Veröffentlichung oder Weitergabe solcher Fassungen außerhalb des SeiKo-Projekts ist nur mit vorheriger schriftlicher Zustimmung der Lizenzgeber:in zulässig.

10.7 Wissenschaftliche Publikationen und Korpusausschnitte

Die Veröffentlichung wissenschaftlicher Ergebnisse, die mithilfe von SeiKo gewonnen wurden, ist zulässig.

In wissenschaftlichen Publikationen, Vorträgen und Qualifikationsarbeiten dürfen einzelne, für die wissenschaftliche Argumentation erforderliche Korpusausschnitte wiedergegeben werden, sofern:

1. der Umfang der Wiedergabe durch den wissenschaftlichen Zweck gerechtfertigt ist,
2. keine zusammenhängenden oder umfangreichen Teile des Korpus reproduziert werden,
3. die betroffenen Personen nicht identifiziert oder reidentifiziert werden können,
4. die Quelle und die verwendete Korpusversion vollständig angegeben werden und
5. keine Rechte Dritter oder zusätzliche Schutzbestimmungen entgegenstehen.

Für die Veröffentlichung von Abbildungen der Primärtexte, längeren Transkriptpassagen oder Ausschnitten aus Audiodateien ist grundsätzlich die vorherige schriftliche Zustimmung des SeiKo-Projekts erforderlich.

Gesetzlich zulässige Nutzungen, insbesondere das Zitatrecht, bleiben unberührt.

10.8 Quellenangabe

Bei jeder Publikation, Präsentation oder Qualifikationsarbeit, die auf der Nutzung von SeiKo beruht, ist das Korpus unter Angabe der verwendeten Version und des zugehörigen DOI zu zitieren.

Die empfohlene Zitierweise lautet:

Schlauch, Julia; Braunewell, Aylin; Gamper, Jana (2026): SeiKo - Korpus Seiteneinsteiger:innen in Vorbereitungsklassen. Handbuch zu Korpusdesign und Datenaufbereitung. Version XX. Zenodo. DOI: 10.5281/zenodo.21495943

Nutzende werden gebeten, das SeiKo-Projekt über aus der Korpusnutzung entstandene Publikationen zu informieren. Entsprechende Hinweise können an seiko@uni-giessen.de gesendet werden. Die Publikationen können anschließend in die Bibliographie auf der SeiKo-Webseite aufgenommen werden.

10.9 Schutz der beteiligten Personen

Jeglicher Versuch, die im Korpus erfassten Personen zu identifizieren oder pseudonymisierte Daten mit anderen Informationen zu verknüpfen, um die Identität einer Person festzustellen, ist untersagt.

Zufällig bekannt gewordene identifizierende oder besonders schutzbedürftige Informationen dürfen nicht veröffentlicht oder weitergegeben werden. Ein entsprechender Vorfall ist dem SeiKo-Projekt unverzüglich mitzuteilen.

10.10 Datensicherheit

Die nutzungsberechtigte Person hat die Korpusdaten durch angemessene technische und organisatorische Maßnahmen vor unbefugtem Zugriff, Verlust und unbeabsichtigter Offenlegung zu schützen.

Die Daten dürfen nur auf Geräten, Servern und Speichersystemen verarbeitet werden, bei denen der Zugriff auf die am genehmigten Vorhaben beteiligten und zugangsberechtigten Personen beschränkt ist.

Eine Übermittlung an externe Dienstleister:innen oder die Eingabe der Daten in externe Cloud-, KI- oder Analysedienste ist nur zulässig, wenn die Vertraulichkeit der Daten und die Einhaltung dieser Nutzungsbedingungen nachweislich gewährleistet sind und erforderlichenfalls eine vorherige Zustimmung des SeiKo-Projekts eingeholt wurde.

Sicherheitsvorfälle, Datenverluste oder unbefugte Zugriffe sind unverzüglich an seiko@uni-giessen.de zu melden.

10.11 Beendigung der Nutzung

Nach Abschluss des genehmigten Vorhabens, nach Ablauf oder Widerruf der Zugangsberechtigung sowie auf Aufforderung der Lizenzgeber:in sind die heruntergeladenen Korpusdaten und alle davon angefertigten Kopien zu löschen.

Davon ausgenommen sind veröffentlichte Forschungsergebnisse sowie vollständig anonymisierte und aggregierte Auswertungsergebnisse, aus denen die ursprünglichen Korpusdaten nicht rekonstruiert werden können.

10.12 Widerruf des Zugangs

Bei einem Verstoß gegen diese Nutzungsbedingungen kann der Zugang zum Korpus mit sofortiger Wirkung gesperrt oder widerrufen werden. In diesem Fall sind sämtliche Korpusdaten und Kopien unverzüglich zu löschen.

Weitergehende Rechte der Lizenzgeber:in und Rechte betroffener Personen bleiben unberührt.

10.13 Beantragung des Zugangs

Der Zugang zu den regulär bereitgestellten SeiKo-Daten ist über die Zugriffsanfrage des entsprechenden Zenodo-Eintrags zu beantragen.[1]

In der Anfrage sind mindestens folgende Angaben zu machen:

- Name und institutionelle Zugehörigkeit,
- Funktion oder Status der anfragenden Person,
- Beschreibung des Forschungs-, Lehr- oder Studienvorhabens,
- vorgesehene Art und Dauer der Datennutzung,
- gegebenenfalls Namen weiterer beteiligter Personen.

[1] Vgl. hierzu: https://help.zenodo.org/docs/share/access-requests/request-access/

Erfordert das Vorhaben die Arbeit mit Audiodateien, ist zusätzlich eine Anfrage an seiko@uni-giessen.de zu richten. Für diese Daten können weitergehende Zugangs- und Nutzungsbedingungen gelten.

10.14 Rechte an den Daten

Durch die Gewährung des Zugangs werden keine Eigentums-, Urheber- oder sonstigen ausschließlichen Rechte an den Korpusdaten übertragen. Sämtliche Rechte verbleiben bei den jeweiligen Rechteinhaber:innen.

Die nutzungsberechtigte Person erhält ausschließlich die in diesen Nutzungsbedingungen ausdrücklich genannten, nicht ausschließlichen und nicht übertragbaren Nutzungsrechte.

11 Anhang

11.1 Stimulusmaterial

11.1.1 Teekanne

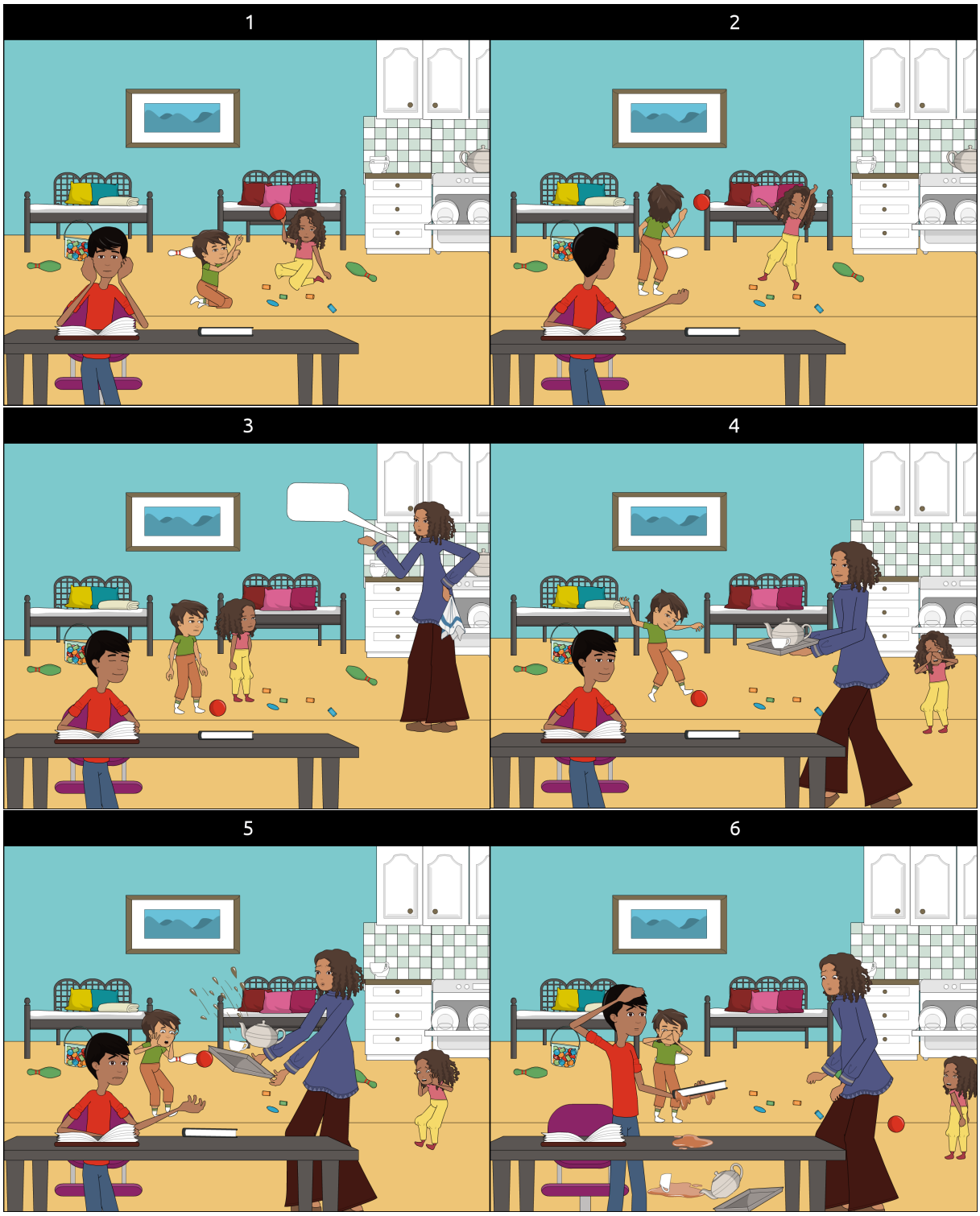

Abbildung 11.1: Stimulusmaterial *Teekanne*

11.1.2 Fahrradtasche

Abbildung 11.2: Stimulusmaterial *Fahrradtasche*

11.1.3 Sportunterricht

Abbildung 11.3: Stimulusmaterial *Sportunterricht*

11.1.4 Busticket

Abbildung 11.4: Stimulusmaterial *Busticket*

11.1.5 Kuchen

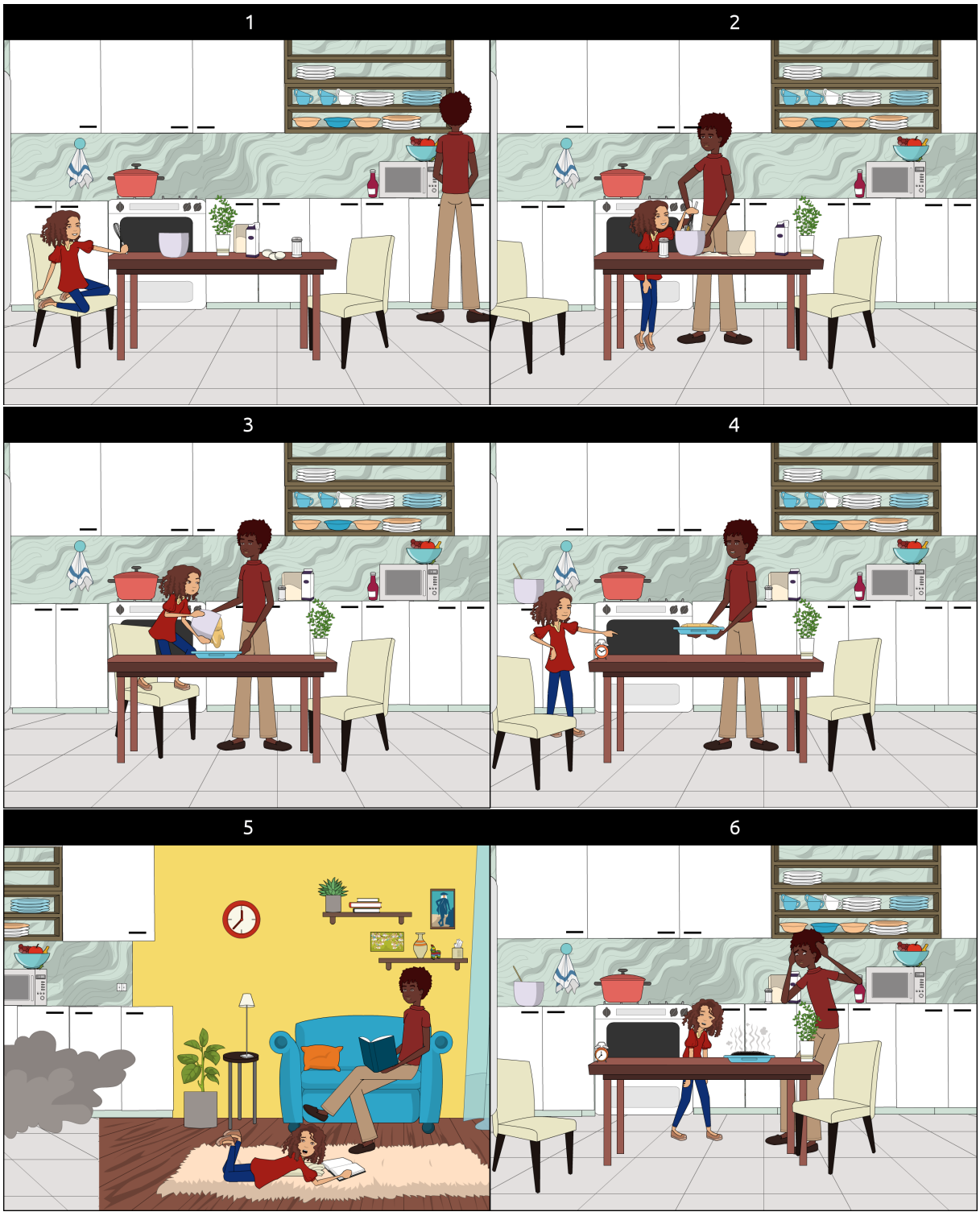

Abbildung 11.5: Stimulusmaterial *Kuchen*

11.1.6 Hausaufgabe

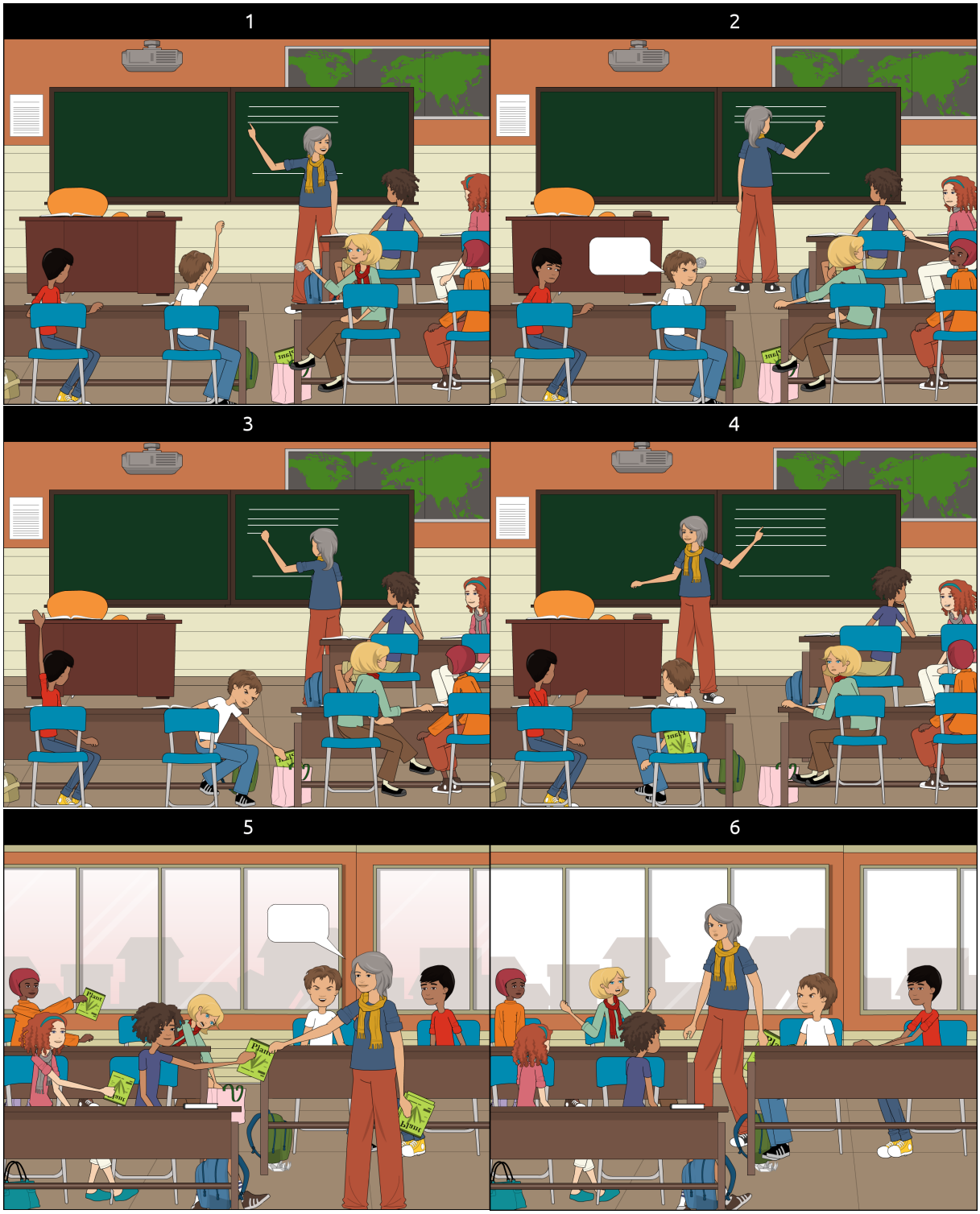

Abbildung 11.6: Stimulusmaterial *Hausaufgabe*

Alle Bildergeschichten wurden mit einer nicht mehr verfügbaren Version von Pixton[1]: erstellt.

[1]Vgl. für die aktuelle Version: https://www.pixton.com/de/.

11.2 Stimulusleitfaden (Q-Part)

11.2.1 Teekanne

Tabelle 11.1: Interviewfragen (Q) zum Stimulus „Teekanne“

Anlass	Interviewfrage
Bild 6	Wie schaut die Frau/ der Junge? Warum schaut die Frau/ der Junge so? Und was denkt der Junge? Und wie ist das passiert?
	Hier sieht man das ja.
Bild 5	Was sieht man hier? Und warum / wie passiert das genau? Was denkt die Frau wohl?/ was die Kinder?
Bild 4	Und was machen sie hier? Ach, die spielen im Zimmer Ball.
Bild 3	Und was passiert hier? Was sagt die Frau? Und warum sagt sie das?
Bild 2	Was macht der Junge? Warum?
Bild 1	Wo sind sie denn eigentlich? Was macht der Junge? Was machen die Kinder? Was glaubst du, was denkt der Junge?
Teil D	Du hast jetzt ja ganz genau gesehen, was passiert ist. Das Buch ist jetzt kaputt. Stell dir vor, der Junge geht in deine Klasse. Kannst du deiner Lehrerin sagen, was passiert ist?

11.2.2 Kuchen

Tabelle 11.2: Interviewfragen (Q) zum Stimulus „Kuchen“

Anlass	Interviewfrage
Bild 6	Wie schauen der Junge / das Mädchen? Warum schauen die so? Was denken sie? Was ist (genau) passiert?
	Hier sieht man das ja.
Bild 5	Und was machen sie hier? Warum machen die das? Was passiert? Und wie / warum ist das passiert? Was denken der Junge/ das Mädchen?
Bild 4	Und was machen sie hier? Was sagt das Mädchen?
Bild 3	Und was machen sie hier? Was glaubst du, warum backen sie den Kuchen?
Bild 2	Und was machen sie hier? Warum macht der Junge das?
Bild 1	Wo sind die denn eigentlich? Was machen die da?
Teil D	Der Nachbar riecht das Feuer und ruft die Feuerwehr. Die Feuerwehr weiß nicht was passiert ist, aber du hast es ja jetzt genau gesehen. Kannst du dem Feuerwehrmann erklären, was passiert ist?

11.2.3 Sportunterricht

Tabelle 11.3: Interviewfragen (Q) zum Stimulus „Sportunterricht“

Anlass	Interviewfrage
Bild 6	Wer ist der Mann? Wie schaut der Mann/ der (Sport-)Lehrer? Und warum schaut der so? Was denkt er? Und wie ist das passiert?
	Hier sieht man das ja.
Bild 5	Was passiert hier genau? / Was siehst du hier? Was denkt der Junge wohl? / was denken die Jungen? Und warum macht der Junge das?
Bild 4	Und was machen sie hier? Ach, die hatten Sportunterricht und ziehen sich um.
Bild 3	Und was passiert hier? Was denken die Jungen?
Bild 2	Und was machen die Jungen hier? Was sagt der Junge?
Bild 1	Wo sind sie denn eigentlich?
Teil D	Der Sportlehrer weiß nicht was passiert ist, aber du hast es ja jetzt ganz genau gesehen. Kannst du dem Lehrer erklären, was passiert ist?

11.2.4 Hausaufgaben

Tabelle 11.4: Interviewfragen (Q) zum Stimulus „Hausaufgaben“

Anlass	Interviewfrage
Bild 6	Was sagt das Mädchen? Wie schaut die Lehrerin hier? Was denkt sie? Was denkt der Junge? Und was ist davor (genau) passiert?
	Hier sieht man das ja.
Bild 5	Was passiert hier? Warum macht die Lehrerin das? Wie schaut das Mädchen? Warum schaut sie so?
Bild 4	Und was machen sie hier?
Bild 3	Und was ist hier los? Was macht der Junge? Wie macht er das? Was glaubst du, was denkt der Junge?
Bild 2	Und was passiert hier? Warum macht der Junge das? Was denkt er/ sie?
Bild 1	Wo sind die denn eigentlich? Was machen die Kinder?
Teil D	Die Lehrerin hat nicht alles gesehen. Aber du hast alles ganz genau beobachtet. Kannst du der Lehrerin sagen, was passiert ist?

11.2.5 Busticket

Tabelle 11.5: Interviewfragen (Q) zum Stimulus „Busticket“

Anlass	Interviewfrage
Bild 6	Wer ist das? [Mann] Wer ist das? [Mädchen] Was denkt er/ sie? Und warum stehen die da? / Was ist passiert?
	Hier sieht man sie ja im Bus.
Bild 5	Was passiert hier genau? / Was siehst du hier? Was denkt das Mädchen? Was sagt der Mann? Und warum sagt der Mann das?
Bild 4	Und was sieht man hier? Ach, die beiden haben aber ein Ticket.
Bild 3	Und was passiert hier? Was sagt der Mann?
Bild 2	Und was machen sie hier?
Bild 1	Und wo sind sie hier? Und was machen sie? Was siehst du noch?

Teil D	Das Mädchen kommt jetzt zu spät zur Schule. Stell dir vor, sie geht in deine Klasse. Eure Lehrerin fragt sich, wo das Mädchen ist. Du hast im Bus ja alles gesehen, kannst du deiner Lehrerin sagen, was passiert ist?

11.2.6 Fahrradtasche

Tabelle 11.6: Interviewfragen (Q) zum Stimulus „Fahrradtasche“

Anlass	Interviewfrage
Bild 6	Wer ist das? [Frau] Wie schaut die Frau denn? Da war doch auch ein Mann, oder? Wo ist der? Und was denkt die Frau? Was ist da (genau) passiert? / Wie ist das genau passiert?
	Hier sieht man das ja.
Bild 5	Und was macht der Mann hier? Und warum macht er das?
Bild 4	Und was macht der Mann hier? Und wie macht er das? Was glaubst du, was denkt der Mann?
Bild 3	Was macht die Frau hier? Und warum macht sie das? Was siehst du noch?
Bild 2	Was macht die Frau? Warum?
Bild 1	Wo ist das denn eigentlich? Was macht die Frau? Was macht die anderen Leute?
Teil D	Die Frau ruft die Polizei, aber eigentlich weiß sie ja nicht was passiert ist. Du hast das aber ganz genau gesehen. Kannst du dem Polizisten noch einmal alles ganz genau erzählen?

11.3 Schriftkonventionen

Um die Lesbarkeit zu erleichtern, wurden für dieses Manual einheitliche Schriftkonventionen erarbeitet, die im Folgenden aufgeführt sind:

- Anweisungen mit Befehlen aus Exmaralda *kursiv* (*Transcription -> Add Token Tiers*)
- Spur in eckigen Klammern ([unit])
- Tags in Großbuchstaben (DEC); Tags der [excl]-Spur (auszuschließendes Material) in Kleinbuchstaben
- Fundorte für Belege in GROßBUCHSTABEN (z.B. E03_B07 für Erhebungszeitpunkt 03 bei Proband:in B07)
- *kursiv* bei Beispielen mit zusätzlichem sprachlichen Kontext; hier wird der relevante Teil des Ausschnitts kursiv gesetzt
- farbliche Markierung von Beispielen: Beispieltext
- Sprachbeispiele werden entsprechend der Transkriptionskonventionen abgebildet; Pausen und exkludiertes Material können ausgelassen werden; wenn nötig, können Satzzeichen hinzugefügt werden
- Sprecher:innenwechsel in Beispielen werden mit einem Halbgeviertstrich (–) abgebildet