

Dulko-Handbuch

Aufbau und Annotationsverfahren des
deutsch-ungarischen Lernerkorpus. Version 1.0

Christoph Beeh
Ewa Drewnowska-Vargáné
Péter Kappel
Bernadett Modrián-Horváth
Andreas Nolda
Orsolya Rauzs
György Scheibl

Universität Szeged / Szegedi Tudományegyetem
Institut für Germanistik / Germán Filológiai Intézet
Lehrstuhl für Germanistische Linguistik / Német Nyelvészeti Tanszék

2021

ISBN 978-963-306-767-3

DOI: <https://doi.org/10.14232/dulko-handbuch-v1.0>

© Die Verfasser, 2021

© Institut für Germanistik der Universität Szeged, Szeged 2021

1. Auflage, 2021

Alle Rechte vorbehalten. Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung der Copyright-Inhaber unzulässig und strafbar. Dies gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherungen und Verarbeitungen in elektronischen Systemen.

Verantwortlicher Herausgeber:

Attila Bombitz,

Leiter des Instituts für Germanistik

Dieses Werk ist lizenziert unter der Lizenz
Creative Commons Attribution 4.0 International (CC-BY 4.0).

Weitere Informationen finden Sie unter
<https://creativecommons.org/licenses/by/4.0/>.

Bitte zitieren Sie dieses Dokument als:

Christoph Beeh, Ewa Drewnowska-Vargáné, Péter Kappel, Bernadett Modrián-Horváth, Andreas Nolda, Orsolya Rauzs, György Scheibl (2021): Dulko-Handbuch. Aufbau und Annotationsverfahren des deutsch-ungarischen Lernerkorpus. Version 1.0. Szeged: Institut für Germanistik.
DOI: 10.14232/dulko-handbuch-v1.0

Inhalt

1 Daten und Struktur des Korpus	4
1.1 Teilkorpus DulkoEssay-v1.0	5
1.2 Teilkorpus DulkoTranslation-v1.0	9
2 Grundgedanken der Annotation	12
2.1 Fehler und Fehlerannotation	12
2.2 Annotationsverfahren	14
3 Annotationstools	17
4 Metadaten	20
4.1 Administrative Metadaten	21
4.2 Metadaten zum Korpusdesign	22
4.3 Metadaten zur Korpusannotation	24
4.4 Text-Metadaten	25
4.5 Lerner-Metadaten	26
5 Annotation	29
5.1 Tokenisierung (word-Ebene)	29
5.2 Satzspannen (S-Ebene und ZHS-Ebene)	30
5.3 Part-of-speech-Tagging (pos-Ebene und ZHpos-Ebene)	30
5.4 Lemmatisierung (lemma-Ebene und ZHlemma-Ebene)	31
5.5 Muttersprachliche Vorlage (trans-Ebene)	31
5.6 Zielhypothesen (ZH-Ebene)	31
5.7 Fehlerannotation	32
5.7.1 Orthographische Fehler (FehlerOrth-Ebene)	33
5.7.2 Morphologische Fehler (FehlerMorph-Ebene)	35
5.7.3 Syntaktische Fehlerkategorien (FehlerSyn-Ebene)	38
5.7.4 Lexikalische Fehler (FehlerLex-Ebene)	48
5.7.5 Semantische Fehler (FehlerSem-Ebene)	51
5.8 Abweichungen (ZHDiff-Ebene)	55
5.9 Inter-annotator agreement	57
5.10 Workflow und Qualitätskontrolle	58
6 Suchabfrage	60
6.1 Einfache Suchabfragen	61
6.2 Suche nach regulären Ausdrücken	62
6.3 Ebenenübergreifende Suchabfragen	62
6.4 Filterung nach Metadaten	63
6.5 Beispiele für Korpusabfragen	63
Literatur	66

1 Daten und Struktur des Korpus

Das am Lehrstuhl für Germanistische Linguistik der Universität Szeged erstellte deutsch-ungarische Lernerkorpus Dulko (Beeh et al. 2021) enthält schriftliche Daten fortgeschrittener ungarischer Deutschlerner, die unter anderem nach Wortarten, Lemmata, Zielhypothesen und Fehlerkategorien annotiert sind. Es handelt sich dabei um kontrolliert erhobene deutschsprachige Essays und Übersetzungen von Studierenden am Institut für Germanistik in Szeged mit Ungarisch als Muttersprache und Deutsch als Fremdsprache (Niveau: mindestens B2 gemäß GER/CEFR). Die Erhebung des Essay-Teilkorpus richtet sich grundsätzlich nach dem Korpusdesign von Falko (vgl. Abschnitt 2), um die Vergleichbarkeit der Daten mit anderen DaF-Korpora wie z. B. falkoEssayL2v2.4 sowie mit muttersprachlichen Vergleichskorpora (z. B. falkoEssayL1v2.3) zu gewährleisten. Das Übersetzungs-Teilkorpus stellt in dieser Hinsicht eine Neuheit dar.

C-Test	Essay-Teilkorpus (DulkoEssay-v1.0)	Übersetzungs-Teilkorpus (DulkoTranslation-v1.0)	Dulko-Korpus (Dulko-v1.0)
B2	28 Texte 9.450 Tokens	24 Texte 7.443 Tokens	52 Texte 16.893 Tokens
C1 (oder höher)	6 Texte 2.336 Tokens	6 Texte 1.781 Tokens	12 Texte 4.117 Tokens
Σ	34 Texte 11.786 Tokens	30 Texte 9.224 Tokens	64 Texte 21.010 Tokens

Tab. 1: Dulko-Gesamtkorpus

Alle Daten wurden im Zeitraum vom Wintersemester 2017/18 bis zum Wintersemester 2019/20 gesammelt. Das Korpus besteht aus insgesamt 64 Texten: 34 Essays und 30 Übersetzungen, die handschriftlich verfasst wurden. Die Rahmenbedingungen für die Erhebungen sind vergleichbar mit denen des Falko-Essay-Korpus: keine Vorbereitungszeit, keine Hilfsmittel, keine Textkenntnis. Die Bearbeitungszeit war aber kürzer (max. 60 Min.) und weder die Essay- noch die Übersetzungstexte des Dulko-Korpus wurden als Prüfungsleistung erhoben. (Eine ausführlichere Vorstellung des Workflows und der Qualitätskontrolle bei der Korpuserstellung findet sich weiter unten, vgl. Abschnitt 5.10). Der Umfang des ganzen Korpus beträgt 21.010 Tokens¹ (vgl. Tab. 1).

¹ Die Tokenzahlen in diesem Handbuch stammen aus den Metadaten der annotierten Lernertexte im EXB-Dateiformat. Diese wurden automatisch mit Hilfe der Annotationssoftware EXMARaLDA (Dulko) berechnet. Dabei zählen alle Ereignisse auf der word-Ebene in den EXB-Dateien (XPath: //tier[@category="word"]/event). Im Suchwerkzeug ANNIS können davon abweichende

Ausführliche Angaben über die Metadaten in Bezug auf die Verfasser (= Lerner) und die Texte der zwei Teilkorpora finden sich weiter unten (s. Tab. 2.1–3.4). Es handelt sich dabei um Folgendes: Datum der Erhebung, Anzahl und Geschlecht der Lerner, Studiengang der Lerner, C-Test-Ergebnis, Essay- bzw. Übersetzungsthema, Anzahl der Lernertexte sowie Anzahl der Tokens pro Text.

1.1 Teilkorpus DulkoEssay-v1.0

Größtenteils stimmen die Essay-Themen bei Dulko mit den Essay-Themen bei Falko überein. Als Grundlage für das Teilkorpus DulkoEssay-v1.0 wurden die folgenden 7 Datenerhebungen verwendet (vgl. Tab. 2.1):

Datum	Essaythema	Anzahl der Lernertexte		
		Lerner männlich	Lerner weiblich	Σ
2017-11-29	Feminismus	2	3	5
2017-12-04	Feminismus	1	4	5
2018-04-16	Entlohnung	1	4	5
2018-04-18	Entlohnung	1	6	7
2018-11-26	Kriminalität	1	3	4
2018-11-29	Kriminalität	1	6	7
2018-12-05	Kriminalität	0	1	1
Σ		7	27	34

Tab. 2.1: Allgemeine Übersicht über die Erhebungsdaten

Wie aus Tab. 2.2 ersichtlich ist, haben insgesamt 25 Lerner 34 Texte verfasst. Das aus 34 Texten bestehende Teilkorpus beinhaltet jeweils zwei verschiedene Texte von 7 Lernern, drei verschiedene Texte von einem Lerner und jeweils einen Text von 17 Lernern.

Tokenzahlen angezeigt werden, da hier auch ggf. leere Felder berücksichtigt werden, die auf der word-Ebene durch Umstellungen oder Einfügungen auf den ZH-Ebenen entstehen.

Lerner	25
Texte	34
Tokens	11786
Ø Tokens/Text	346,6

Tab. 2.2: Lerner-, Text- und Tokenzahlen

Im Folgenden (Tab. 2.3–2.5) werden ausführliche Angaben zu den Erhebungen pro Semester dargestellt. Die Abkürzung „BA“ bezieht sich auf die Lerner im Bachelorstudium; mit „BA-FT“ werden die Lerner im Bachelorstudium mit der Spezialisierung Dolmetscher-Übersetzer und mit „TO“ die Lerner im Lehramtsstudium bezeichnet.

Studiengang	Text	Erhebung	C-Test	Tokens
TO	Feminismus_1	2017-11-29 TO_4	147 (C1)	349
TO	Feminismus_3	2017-11-29 TO_4	103 (B2)	387
TO	Feminismus_4	2017-11-29 TO_4	141 (C1)	355
TO	Feminismus_7	2017-11-29 TO_4	116 (B2)	342
TO	Feminismus_9	2017-11-29 TO_4	121 (B2)	507
TO	Feminismus_10	2017-12-04 BA_TO-3	133 (C1)	342
TO	Feminismus_12	2017-12-04 BA_TO-3	112 (B2)	238
TO	Feminismus_13	2017-12-04 BA_TO-3	126 (B2)	391
BA-FT	Feminismus_19	2017-12-04 BA_TO-3	125 (B2)	290
TO	Feminismus_20	2017-12-04 BA_TO-3	104 (B2)	302

Tab. 2.3: Erhebungen im Wintersemester 2017/18

Studiengang	Text	Erhebung	C-Test	Tokens
TO	Entlohnung_3	2018-04-16 BA_3	98 (B2)	243
BA	Entlohnung_5	2018-04-16 BA_3	113 (B2)	371
TO	Entlohnung_6	2018-04-16 BA_3	100 (B2)	219
TO	Entlohnung_9	2018-04-16 BA_3	118 (B2)	435
TO	Entlohnung_10	2018-04-16 BA_3	115 (B2)	311
BA	Entlohnung_14	2018-04-18 TO_3	123 (B2)	408
TO	Entlohnung_16	2018-04-18 TO_3	120 (B2)	491
TO	Entlohnung_22	2018-04-18 TO_3	104 (B2)	256
BA	Entlohnung_26	2018-04-18	127 (B2)	421
BA	Entlohnung_27	2018-04-18	99 (B2)	312
BA	Entlohnung_30	2018-04-18	131 (C1)	397
BA	Entlohnung_31	2018-04-18	124 (B2)	368

Tab. 2.4: Erhebungen im Sommersemester 2018

Studiengang	Text	Erhebung	C-Test	Tokens
TO	Kriminalität_1	26-11-2018	118 (B2)	350
TO	Kriminalität_2	26-11-2018	121 (B2)	379
TO	Kriminalität_3	26-11-2018	100 (B2)	284
TO	Kriminalität_4	26-11-2018	131 (C1)	586
TO	Kriminalität_7	29-11-2018	114 (B2)	331
BA-FT	Kriminalität_8	29-11-2018	118 (B2)	421
TO	Kriminalität_9	29-11-2018	125 (C1)	307
TO	Kriminalität_10	29-11-2018	111 (B2)	261
TO	Kriminalität_11	29-11-2018	112 (B2)	241
TO	Kriminalität_12	26-11-2018	118 (B2)	328
TO	Kriminalität_13	29-11-2018	110 (B2)	256
TO	Kriminalität_15	05-12-2018	104 (B2)	307

Tab. 2.5: Erhebungen im Wintersemester 2018/19

1.2 Teilkorpus DulkoTranslation-v1.0

Als Grundlage für das Teilkorpus DulkoTranslation-v1.0 wurden die folgenden acht Datenerhebungen verwendet (vgl. Tab. 3.1):

Datum	Übersetzungsthema	Anzahl der Lernertexte		
		Lerner männlich	Lerner weiblich	Σ
2017-10-31	China	0	4	4
2017-10-31	Sprache	1	2	3
2017-11-02	China	1	1	2
2017-11-02	Sprache	1	2	3
2017-12-06	Weinen	1	7	8
2018-03-07	Wandern	1	3	4
2018-05-07	Weinen	0	5	5
2019-11-12	Wandern	0	1	1
Σ		5	25	30

Tab. 3.1: Allgemeine Übersicht über die Erhebungsdaten

Wie aus Tab. 3.2 ersichtlich ist, haben insgesamt 26 Lerner 30 Texte verfasst. Unter den 30 Texten des Teilkorpus sind jeweils zwei verschiedene Texte von 4 Lernern und jeweils ein Text von 22 Lernern enthalten.

Lerner	26
Texte	30
Tokens	9224
Ø Tokens/Text	307,5

Tab. 3.2: Lerner-, Text- und Tokenzahlen

Im Folgenden (Tab. 3.3–3.5) werden ausführliche Angaben zu den Erhebungen pro Semester dargestellt. Die Abkürzung „BA“ bezieht sich auf die Lerner im

Bachelorstudium, die Abkürzung „TO“ auf die Lerner im Lehramtsstudium und die Abkürzung „MA“ auf die Lerner im Masterstudium.

Studiengang	Text	Erhebung	C-Test	Tokens
BA	China_1	2017-10-31_BA-2	110 (B2)	461
BA	China_3	2017-10-31_BA-2	106 (B2)	408
BA	China_4	2017-10-31_BA-2	113 (B2)	382
BA	China_5	2017-10-31_BA-2	103 (B2)	265
BA	China_10	2017-11-02_BA-1	104 (B2)	298
BA	China_15	2017-11-02_BA-1	111 (B2)	447
BA	Sprache_1	2017-10-31_BA-2	109 (B2)	436
BA	Sprache_5	2017-10-31_BA-2	98 (B2)	180
BA	Sprache_6	2017-10-31_BA-2	123 (B2)	284
TO	Sprache_8	2017-11-02_BA-1	116 (B2)	355
BA	Sprache_10	2017-11-02_BA-1	117 (B2)	268
TO	Sprache_14	2017-11-02_BA-1	97 (B2)	222
MA	Weinen_2	2017-12-06_MA-1	132 (C1)	303
MA	Weinen_3	2017-12-06_MA-1	123 (B2)	309
MA	Weinen_4	2017-12-06_MA-1	103 (B2)	260
MA	Weinen_5	2017-12-06_MA-1	128 (B2)	270
MA	Weinen_6	2017-12-06_MA-1	126 (B2)	277
MA	Weinen_7	2017-12-06_MA-1	137 (C1)	319
MA	Weinen_9	2017-12-06_MA-1	125 (B2)	310
MA	Weinen_10	2017-12-06_MA-1	149 (C1)	314

Tab. 3.3: Erhebungen im Wintersemester 2017/18

Studiengang	Text	Erhebung	C-Test	Tokens
MA	Wandern_2	2018-03-07_MA-1	125 (B2)	300
MA	Wandern_4	2018-03-07_MA-1	132 (C1)	296
MA	Wandern_5	2018-03-07_MA-1	128 (B2)	292
MA	Wandern_6	2018-03-07_MA-1	137 (C1)	264
MA	Weinen_11	2018-05-07_BA-1	115 (B2)	297
BA	Weinen_12	2018-05-07_BA-1	124 (B2)	273
BA	Weinen_14	2018-05-07_BA-1	135 (C1)	285
BA	Weinen_15	2018-05-07_BA-1	114 (B2)	280
BA	Weinen_16	2018-05-07_BA-1	127 (B2)	285

Tab. 3.4: Erhebungen im Sommersemester 2018

Studiengang	Text	Erhebung	C-Test	Tokens
BA	Wandern_18	2019-11-12	102 (B2)	284

Tab. 3.5: Erhebung im Wintersemester 2019/20

2 Grundgedanken der Annotation

Die Annotation der Lernertexte bildet den zentralen Bereich des Dulko-Projektes und wird von geschulten Annotatorinnen (studentischen Hilfskräften) durchgeführt. Dies geschieht in Absprache mit den am Projekt beteiligten Wissenschaftler und Wissenschaftlerinnen, deren Muttersprachen entweder Ungarisch oder Deutsch sind.

Die Grundgedanken der Annotation werden in den folgenden beiden Abschnitten vorgestellt.

2.1 Fehler und Fehlerannotation

Sprachliche *Fehler* werden im Dulko-Projekt verstanden als Abweichungen zwischen einem Teil eines Texts im System der Lernaltersprache und einer (impliziten oder expliziten) *Zielhypothese* im System der Zielsprache (vgl. Lüdeling/Hirschmann 2015).

Die *Lernaltersprache* (*Interimsprache*) stellt ein spezifisches Sprachsystem dar, welches sich während des Lernprozesses herausbildet und sowohl über Kennzeichen der Zielsprache als auch über Kennzeichen der Muttersprache des Lerners verfügen kann. Darüber hinaus zeichnet sich die Lernaltersprache durch weitere Charakteristika aus, die weder in der Muttersprache des Lerners noch in der von ihm zu erlernenden Zielsprache vorhanden sind (vgl. Selinker 1972; Fekete 2016).

Die Fehlerannotation der Lernertexte in Dulko bezieht sich vorrangig auf Orthographie, Morphologie, Syntax, Lexik und Semantik. Demgegenüber werden umgangssprachliche, gesprochensprachliche sowie nationale Varianten der Standardvarietäten außerhalb Deutschlands nicht als Fehler betrachtet und demzufolge nicht annotiert. Somit wird darauf geachtet, bei der Fehlerkorrektur nur Abweichungen zu korrigieren, für deren Korrektur es kodifizierte Referenzwerke (vgl. Referenzwerke im Literaturverzeichnis) bzw. eine empirische Grundlage gibt. Im Zweifelsfall wurden vergleichende Korpusrecherchen vorgenommen und Muttersprachler konsultiert.

Als ein Beispiel für eine Abgrenzung zwischen Abweichungen, die im obigen Sinne als Fehler bzw. nicht als Fehler betrachtet werden, dient der folgende Auszug aus einem Lernertext:

- (1) *Natürlich gibt es solche Sportler, die ein Teil von seiner Geld für verschieden Vereine spendiert, aber auch so hat er/sie noch mehr Geld, als ein anderes Mensch.*

(Auszug aus dem Essay Entlohnung_5, Satzspanne 21, Sommersemester 2018, Universität Szeged)

Eine mögliche Zielhypothese zu (1) wäre:

- (2) *Natürlich gibt es solche Sportler, die einen Teil von ihrem Geld verschiedenen Vereinen spenden, aber auch so haben sie noch mehr Geld als andere Menschen.*

Neben der Zielhypothese in (2) gibt es noch zahlreiche andere mögliche Zielhypothesen zu (1). Im Allgemeinen wurde im Dulko-Projekt versucht, Zielhypothesen zu wählen, die so nah wie möglich am Lernertext bleiben. Insbesondere wurde nicht versucht, den Lernertext stilistisch zu verbessern.

Die folgenden Abweichungen zwischen (1) und (2) können als Fehler interpretiert werden:

- Das am Artikel *ein* sichtbare Genus der Nominalgruppe *ein Teil* ist falsch.
- Beim Pronomen *seiner* (in der Nominalgruppe *von seiner Geduld*), das sich auf das Antezedens *solche Sportler* bezieht, besteht ein Kongruenzfehler.
- Die beim Adjektiv *verschieden* (in der Nominalgruppe *für verschieden Vereine*) fehlende Flexion ist falsch.
- Beim Verb *spendiert* sind zwei Fehler zu verzeichnen: einerseits ein Wortbildungsfehler und andererseits ein Kongruenzfehler mit dem Antezedens *solche Sportler*.
- Bei der Phrase *für verschiedene Vereine spendieren* besteht ein Verbvalenzfehler.
- Bei der verbalen Phrase *hat er/sie* besteht ein Kongruenzfehler mit demselben Antezedens *solche Sportler*.
- Ein Zeichensetzungsfehler besteht bei dem Komma vor der Partikel *als* in der Vergleichskonstruktion *mehr Geld, als ein anderes Mensch*.
- Das am Zahlwort *anderes* sichtbare Genus des Substantivs *Mensch* in der Nominalgruppe *ein anderes Mensch* ist falsch.

Demgegenüber wird im obigen Beispiel die Struktur *solche* + Relativsatz (*solche Sportler, die ...*) *nicht* als relevanter Fehler betrachtet. Vielmehr handelt es sich hier im Wesentlichen um eine stilistische Abweichung.

Bei der Bestimmung von Abweichungen als Fehler haben wir uns auf gängige Referenzwerke gestützt (vgl. Literaturverzeichnis).

2.2 Annotationsverfahren

Das in Dulko verwendete Annotationsverfahren folgt im Prinzip den im Falko-Lernerkorpusprojekt an der Humboldt-Universität zu Berlin entwickelten Richtlinien (Reznicek et al. 2012). Anders jedoch als in Falko erfolgt die Fehler-Annotation in Dulko explizit auf mehreren Ebenen mit Hilfe von Fehlertags. Somit unterscheidet sich Dulko in folgenden relevanten Aspekten von Falko (vgl. Hirschmann und Nolda 2019; Nolda 2019):

1. Beim Dulko-Annotationsverfahren können beliebig viele Zielhypothesen angegeben werden.
2. Fehler und ihre Bereiche werden beim Dulko-Annotationsverfahren explizit mit Hilfe von Fehlerkategorien unterschiedlicher sprachlicher Ebenen annotiert.
3. Jeder Zielhypothese können beim Dulko-Annotationsverfahren Fehlerkategorien beliebiger sprachlicher Ebenen zugeordnet werden, z. B. FehlerOrth-, FehlerMorph-, FehlerSyn-, FehlerLex- und FehlerSem-Ebenen (vgl. Abschnitt 5 weiter unten).

Im Dulko-Annotationsverfahren werden zwei Hauptarten von Zielhypothesen unterschieden: *finale* Zielhypothese und *intermediäre* Zielhypothese. Was darunter zu verstehen ist, sei am folgenden Lernertext veranschaulicht:

- (3) *Wie in der ganzen Gesellschaft, auch in der Regierung sollte der Anzahl der Frauen 50 % sein [...].*
 (Auszug aus dem Essay Feminismus_4, Satzspanne 17, Wintersemester 2017/18, Universität Szeged)

In Abb. 1 sind zu diesem Lernertext zwei Zielhypothesen angegeben: eine intermediäre Zielhypothese 1 und eine finale Zielhypothese 2.

Lernertext:

*Wie in der ganzen Gesellschaft, auch in der Regierung **solte der Anzahl** der Frauen 50 % sein [...].*

Zielhypothese 1 (intermediär):

*Wie in der ganzen Gesellschaft **solte** auch in der Regierung **die Anzahl** der Frauen 50 % sein [...].*

Fehler: Zeichensetzung Wortstellung Genus

Zielhypothese 2 (final):

*Wie in der ganzen Gesellschaft **solte** auch in der Regierung **der Anteil** der Frauen 50 % sein [...].*

Fehler: Lexik

Abb. 1: Zielhypothesen und Fehleranalyse zu Beispiel (3)

Zwischen der intermediären Zielhypothese 1 und dem Lernertext können die folgenden Abweichungen als Fehler interpretiert werden:

- Das Komma ist ein Zeichensetzungsfehler.
- Die Stellung des Verbs *sollte* ist ein Wortstellungsfehler.
- Das am Artikel *der* sichtbare Genus der Nominalgruppe *der Anzahl* ist falsch.

Einen weiteren Fehler repräsentieren Abweichungen zwischen der Zielhypothese 2 und Zielhypothese 1:

- Die lexikalische Einheit *Anzahl* in der Nominalgruppe *der Anzahl* ist kontextuell falsch gewählt: 50 % ist nicht eine *Anzahl*, sondern ein *Anteil*.

Da *Anteil* ein anderes Genus hat als *Anzahl*, weicht auch der Artikel *der* in Zielhypothese 2 vom Artikel *die* in Zielhypothese 1 ab; zufälligerweise stimmt er mit dem Artikel *der* im Lernertext überein.

Zielhypothese 2 wird als final und somit als zielsprachlich betrachtet. Die intermediäre Zielhypothese 1 hingegen ist nicht in allen Teilen zielsprachlich; vielmehr enthält sie weniger Abweichungen als die finale Zielhypothese 2. Umgekehrt kann auf die intermediäre Zielhypothese 1 nicht verzichtet werden, da sonst der Genusfehler bei *der Anzahl* nicht repräsentiert wäre.

Wie das obige Beispiel illustriert, dienen intermediäre Zielhypothesen insbesondere dazu, Fehler zu repräsentieren, die bei der finalen Zielhypothese aufgrund überlappender Fehler ‚unsichtbar‘ bleiben.

Diese Zielhypothesen-Begriffe lassen sich folgendermaßen genauer bestimmen (vgl. Nolda 2019):

- Eine *finale Zielhypothese* zu einer lernersprachlichen Einheit *E* bei einer Lesart *L* ist eine mit *E* bei *L* synonyme muttersprachliche Einheit mit minimalen Abweichungen.
- Eine *intermediäre Zielhypothese* zu einer lernersprachlichen Einheit *E* bei einer Lesart *L* enthält weniger Abweichungen als eine finale Zielhypothese zu *E* bei *L*.

In praktischer Hinsicht ist eine finale Zielhypothese eine korrigierte Version des Lernertextes, die keine Fehler mehr enthält und zugleich relativ wenig vom Lernertext abweicht. Eine intermediäre Zielhypothese hingegen korrigiert den Lernertext nicht vollständig. Beide Arten von Zielhypothesen machen bestimmte Fehler im Lernertext in Form von Abweichungen zwischen Lernertext und Zielhypothese bzw. zwischen intermediärer Zielhypothese und finaler Zielhypothese explizit.

Die Unterscheidung zwischen finalen und intermediären Zielhypothesen erlaubt es darüber hinaus, eine technische Beschränkung des Partitur-Editors von

EXMARaLDA zu umgehen. Dort lassen sich auf derselben Fehlerebene keine sich überlappenden Fehlertags angeben. Als Workaround können die entsprechenden Fehlertags stattdessen unterschiedlichen Zielhypothesen zugeordnet werden (vgl. die Fehlertags Gen und Lex auf den FehlerLex-Ebenen in Abb. 2).

	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	
[word]	Wie	in	der	ganzen	Gesellschaft	,		auch	in	der	Regierung	sollte	der	Anzahl	der	Frauen	50	%	sein	,	
[S]	s17																				
[pos]	KOUS	APPR	ART	ADJA	NN	\$,		ADV	APPR	ART	NN	VMFIN	ART	NN	ART	NN	CARD	NN	VAINF	\$,	
[lemma]	wie	in	die	ganz	Gesellschaft	,		auch	in	die	Regierung	sollen	die	Anzahl	die	Frau	50	%	sein	,	
[ZH]	Wie	in	der	ganzen	Gesellschaft		sollte	auch	in	der	Regierung		die	Anzahl	der	Frauen	50	%	sein	,	
[ZHDiff]							DEL	MOVT					MOVS	CHA							
[ZHS]	s17																				
[ZHpos]	KOUS	APPR	ART	ADJA	NN			VMFIN	ADV	APPR	ART	NN		ART	NN	ART	NN	CARD	NN	VAINF	\$,
[ZHlemma]	wie	in	die	ganz	Gesellschaft			sollen	auch	in	die	Regierung		die	Anzahl	die	Frau	50	%	sein	,
[FehlerOrth]							ZS														
[FehlerMorph]																					
[FehlerSyn]								StV													
[FehlerLex]														Gen							
[FehlerSem]																					
[ZH]	Wie	in	der	ganzen	Gesellschaft		sollte	auch	in	der	Regierung		der	Anteil	der	Frauen	50	%	sein	,	
[ZHDiff]													CHA	CHA							
[ZHS]	s17																				
[ZHpos]	KOUS	APPR	ART	ADJA	NN			VMFIN	ADV	APPR	ART	NN		ART	NN	ART	NN	CARD	NN	VAINF	\$,
[ZHlemma]	wie	in	die	ganz	Gesellschaft			sollen	auch	in	die	Regierung		die	Anteil	die	Frau	50	%	sein	,
[FehlerOrth]																					
[FehlerMorph]																					
[FehlerSyn]																					
[FehlerLex]															Lex						
[FehlerSem]																					

Abb. 2: Annotation des Lernertext-Auszugs (3) im Partitur-Editor von EXMARaLDA

Bekanntlich gibt es oft verschiedene Möglichkeiten, Zielhypothesen zu einem Lernertext anzugeben. Wenn sich zwei finale Zielhypothesen zu einer lernersprachlichen Einheit E in der vorausgesetzten Lesart L unterscheiden, dann sprechen wir von *alternativen Zielhypothesen*. Analoges lässt sich bei intermediären Zielhypothesen unterscheiden. Das Dulko-Korpus enthält aktuell keine alternativen Zielhypothesen.

3 Annotationstools

Die erhobenen Lernertexte wurden im Dulko-Projekt mit Hilfe von [EXMARaLDA \(Dulko\)](#) annotiert. Diese unter einer Open-Source-Lizenz frei verfügbare Toolsammlung für den Partitur-Editor von [EXMARaLDA](#) (Schmidt 2004) stellt diverse Transformationsszenarios für die semiautomatische Annotation von Lernertexten zur Verfügung.

Jedes dieser Transformationsszenarios ruft ein XSLT-Stylesheet auf, das auf dem EXB-Dateiformat des Partitur-Editors operiert. Diese XSLT-Stylesheets können auch im Batch-Modus mit Hilfe eines der mitgelieferten Shell-Skripte ausgeführt werden.

Von den Transformationsszenarios sind hervorzuheben:

Transformationsszenario	Funktion	XSLT-Stylesheet(s)
Dulko: word-Spur (Lernertext)	Erzeugt oder aktualisiert eine Tokenisierung des Lernertexts.	exb2exb-word.xsl
Dulko: S-, pos- und lemma-Spuren (Lernertext)	Fügt S-, pos- und lemma-Spuren zur word-Spur hinzu oder aktualisiert diese.	exb2exb-tag.xsl exb2exb-s.xsl
Dulko: trans-Spur (Lernertext)	Fügt trans-Spur zur word-Spur hinzu oder aktualisiert diese.	ext2exb-trans.xsl
Dulko: ZH- und Fehler-Spuren (1. Zielhypothese)	Fügt ZH- und Fehler-Spuren für die 1. Zielhypothese hinzu oder aktualisiert diese und kopiert die word-Spur als Vorlage zur manuellen Bearbeitung. Über das Annotationswerkzeug können hier die Fehlerkategorien eingetragen und bei Bedarf angepasst oder erweitert werden.	exb2exb-zh.xsl exb2exb-fehler.xsl
Dulko: ZHS-, ZHpos- und ZHlemma-Spuren (1. Zielhypothese)	Fügt ZHS-, ZHpos- und ZHlemma-Spuren zur 1. ZH-Spur hinzu oder aktualisiert diese.	exb2exb-tag.xsl exb2exb-s.xsl

Tab. 4: Transformationsszenarios in EXMARaLDA (Dulko) (Teil 1/2)

Transformationsszenario	Funktion	XSLT-Stylesheet(s)
Dulko: ZHDiff-Spur (1. Zielhypothese)	Fügt ZHDiff-Spur zur 1. ZH-Spur hinzu oder aktualisiert diese; hier werden die vorgenommenen Abweichungen zwischen der ZH-Spur und word-Spur angezeigt.	exb2exb-diff.xsl
Dulko: ZH- und Fehler-Spuren (2. Zielhypothese)	Fügt ZH- und Fehler-Spuren für die 2. Zielhypothese und die darauf bezogenen Fehlerkategorien hinzu oder aktualisiert diese. Hierbei wird die 1. Zielhypothese als Vorlage zur manuellen Bearbeitung kopiert.	exb2exb-zh.xsl exb2exb-fehler.xsl
Dulko: ZHS-, ZHpos- und ZHlemma-Spuren (2. Zielhypothese)	Fügt ZHS-, ZHpos- und ZHlemma-Spuren zur 2. ZH-Spur hinzu oder aktualisiert diese.	exb2exb-tag.xsl exb2exb-s.xsl
Dulko: ZHDiff-Spur (2. Zielhypothese)	Fügt ZHDiff-Spur zur 2. ZH-Spur hinzu oder aktualisiert diese; hier werden die vorgenommenen Abweichungen zwischen der 1. ZH-Spur und der 2. ZH-Spur angezeigt.	exb2exb-diff.xsl
Dulko: ZH- und Fehler-Spuren (weitere Zielhypothese)	Fügt ZH- und Fehler-Spuren für eine weitere Zielhypothese hinzu.	exb2exb-zh.xsl exb2exb-fehler.xsl
Dulko: ZHS-, ZHpos- und ZHlemma-Spuren (weitere Zielhypothese)	Fügt ZHS-, ZHpos- und ZHlemma-Spuren zur letzten ZH-Spur hinzu oder aktualisiert diese.	exb2exb-tag.xsl exb2exb-s.xsl
Dulko: ZHDiff- Spur (weitere Zielhypothese)	Fügt ZHDiff-Spur zur letzten ZH-Spur hinzu oder aktualisiert diese.	exb2exb-diff.xsl

Tab. 4: Transformationsszenarios in EXMARaLDA (Dulko) (Teil 2/2)

Mit Hilfe dieser Transformationsszenarios kann ein Lernertext, der in die word-Spur des Dulko-Template [dulko.template.exb](#) eingegeben wurde, semiautomatisch annotiert werden. Mit dem Transformationsszenario „Dulko: word-Spur (Lernertext)“ wird die word-Spur tokenisiert. Das Transformationsszenario „Dulko: S-, pos- und lemma-Spuren (Lernertext)“ annotiert die word-Spur mit Wortarten-Tags auf der pos-Spur, mit Lemmata auf der lemma-Spur und mit Satzspannen auf der S-Spur. Mit dem Transformationsszenario „Dulko: ZH- und Fehler-Spuren (1. Zielhypothese)“ werden eine ZH-Spur mit einer Kopie der vorangegangenen der word-Spur sowie leere Fehler-Spuren (FehlerOrth, FehlerMorph, FehlerSyn, FehlerLex und FehlerSem) als Vorlage für die manuelle Bearbeitung erstellt. Abweichungen, die sich im Zuge der Bearbeitung ergeben und die als Fehler zu interpretieren sind, werden auf den einschlägigen Fehler-Spuren mit Hilfe von Fehlertags kategorisiert. Dabei kann das Annotationspanel von EXMARaLDA verwendet werden, für das die Datei [annotation-panel.xml](#) ein Fehlertagset zur Verfügung stellt. Das Transformationsszenario „Dulko: ZHS-, ZHpos- und ZHlemma-Spuren (1. Zielhypothese)“ annotiert die ZH-Spur mit Wortarten-Tags auf einer ZHpos-Spur, mit Lemmata auf einer ZHlemma-Spur und mit Satzspannen auf einer ZHS-Spur; Abweichungen zwischen der ZH-Spur und der word-Spur werden mit Hilfe des Transformationsszenarios „Dulko: ZHDiff-Spur (1. Zielhypothese)“ berechnet und auf einer ZHDiff-Spur ausgegeben. Die Annotation zusätzlicher Zielhypothesen erfolgt auf analoge Weise.

Mit weiteren Transformationsszenarios können die annotierten Lernertexte nach HTML und ANNIS exportiert werden:

- Dulko: HTML-Version: Annotierte Daten nach HTML exportieren.
- Dulko: ANNIS-kompatible Version: Redundante Annotationen löschen und Namensraum-Präfixe hinzufügen für den Export nach ANNIS.

Näheres zur Fehlerannotation von Lernertexten mit EXMARaLDA (Dulko) s. Abschnitt 5 sowie Nolda (2019).

4 Metadaten

Zu allen annotierten Lernertexten gehören Metadaten zur Lernerbiografie und zu den Textproduktionsbedingungen. Diese sind in Fragebögen erhoben und (pseudo)anonymisiert worden. Bei Erhebung und Speicherung zum Zwecke der wissenschaftlichen Auswertung sind wir gemäß der europäischen Datenschutz-Grundverordnung (DSGVO 2016) vorgegangen. Der Fragebogen ist deutschsprachig und wird von den Lernern auf Deutsch ausgefüllt. Die dazugehörige Einverständniserklärung gemäß DSGVO ist ungarischsprachig. Zwecks internationaler Vergleichbarkeit der Metadaten haben wir uns an den von Granger und Paquot (2017) vorgeschlagenen *Core metadata for learner corpora* (Draft 1.0) orientiert², die Metadaten-Variablen und -Werte in den folgenden fünf Kategorien enthalten:

1. administrative Metadaten
2. Metadaten zum Korpusdesign
3. Metadaten zur Korpusannotation
4. Text-Metadaten
5. Lerner-Metadaten

Für jede dieser Kategorien werden untenstehend Metadaten-Variablen mit Beispielwerten angegeben. In den Fällen, in denen die Werte von den Transformationsszenarios automatisch bestimmt werden oder vom Dulko-Template [dulko.template.exb](#) vorgegeben sind, werden nur diese angegeben. Bei Variablen, die bei einer Abfrage variiert werden können und für die es ein festes Vokabular gibt, werden mehrere Beispiele angegeben. In anderen Fällen wird ein ausgewählter Beispielwert angegeben.

² Variablen, die von Granger und Paquot (2017) vorgeschlagen, aber in Dulko nicht berücksichtigt werden, sind folgende: `edition`, `L1_comp_data`, `comp_corpus_included`, `comp_corpus_name`, `parsing_tool`, `written_author_type`, `written_multiple_author_ID`, `written_process_status`, `written_process_ID`, `spoken_task`, `multimodal_task`, `interaction_type`, `sound_file`, `sound_file_transcriber`, `timing`, `timing_duration`, `written_ref_tools_set`, `written_ref_tools_mono_dico`, `written_ref_tools_bil_dico`, `written_ref_tools_mono_conc`, `written_ref_tools_bil_conc`, `written_ref_tools_checker`, `written_ref_tools_grammar`, `written_ref_tools_other`, `official_language_testing_type`, `official_language_testing_result`, `text_proficiency_rating_scale`, `text_proficiency_rating`, `home_language`, `proficiency_level_CEFR_conversion`, `learner_proficiency_self_rated`, `occupation`, `socecStatus`, `language_instruction_primary`, `language_instruction_secondary`, `language_instruction_higher_edu`, `L2_exposure_written`, `L2_exposure_oral`, `learner_apptitude`, `apptitude_test_ID`, `apptitude_test_components`, `learner_intelligence`, `intelligence_test_ID`, `intelligence_test_components`, `learner_motivation`, `motivation_test_ID`, `motivation_test_components`.

Metadaten-Variablen, deren Werte Sprachen sind, werden von EXMARaLDA (Dulko) auf der Basis der Angaben in der Sprechertabelle des Partitur-Editors („Verwendete Sprache(n)“, „Erstsprache(n)“, „Zweitsprache(n)“) bestimmt. Dabei werden die vom Partitur-Editor verwendeten dreistelligen Sprachcodes nach ISO 639-2 (*deu*, *hun* etc.) übernommen.

Bei numerischen Werten wird, falls erforderlich, der Punkt als Dezimalzeichen verwendet.

4.1 Administrative Metadaten

Sämtliche administrativen Metadaten sind *vom Dulko-Template vorgegeben*:

Variablen	Werte und Bemerkungen
corpus_title	<i>Deutsch-ungarisches Lernerkorpus</i> Korpusname
corpus_acronym	<i>Dulko</i> Korpusakronym
distributor	<i>University of Szeged</i> verantwortliche Institution
availability	<i>free of charge</i> Verfügbarkeit des Korpus
licence	<i>CLARIN PUB+BY+SA+PRIV</i> Lizenz
character_markup	<i>UTF-8</i> Zeichenkodierung
markup_language	<i>XML</i> Markupsprache

Tab. 5.1: Variablen der administrativen Metadaten

4.2 Metadaten zum Korpusdesign

Variablen, die von den Transformationsszenarios automatisch bestimmt werden, sind folgende:

Variablen	Werte und Bemerkungen
corpus_size	1.234.567 tokens Korpusgröße in Tokens (Wörter, Zahlen, Satzzeichen)

Tab. 5.2: Variablen des Korpusdesigns, die von den Transformationsszenarios automatisch bestimmt werden

Variablen, die vom *Dulko-Template* vorgegeben sind:

Variablen	Werte und Bemerkungen
L2_target	<i>mono_L2</i> monolinguales oder multilinguales Lernerkorpus
L2_language	<i>deu</i> Sprache der Lernertexte
L1_language	<i>hun</i> gemeinsame Muttersprache der Lerner
corpus_mode	<i>written</i> sprachlicher Modus des Korpus
editorial_decisions	<i>Punctuation was normalised to ASCII characters.</i> vorgenommene Normalisierungen
transcription_guidelines	<i>Dulko-Handbuch</i> Richtlinien für die Transkription

Tab. 5.3: Variablen des Korpusdesigns, die vom *Dulko-Template* vorgegeben sind (Teil 1/2)

Variablen	Werte und Bemerkungen
written_process	<i>no</i> Enthält das Korpus verschiedene Versionen eines Textes desselben Lerner's?
longitudinal	<i>no</i> Ist das Korpus longitudinal?
proficiency_level	<i>intermediate</i> Sprachstand der annotierten Lernertexte
proficiency_level_type	<i>learner-based</i> Das Niveau bezieht sich auf den Sprachstand des Lerner's.
proficiency_level_descriptors	<i>TestDaF/onSET</i> Name der Sprachstandserhebung (im 1. Projektjahr <i>TestDaF</i> , danach <i>onSET</i>)
field	<i>general language</i> sprachliche Domäne (allgemeinsprachlich oder fachsprachlich)
official_language_testing	<i>no</i> Wurde das Korpus im Rahmen einer offiziellen Sprachprüfung erstellt?
comp_data	<i>no</i> Wurden zu Vergleichszwecken weitere Daten erhoben?

Tab. 5.3: Variablen des Korpusdesigns, die vom Dulko-Template vorgegeben sind (Teil 2/2)

4.3 Metadaten zur Korpusannotation

Variablen, die von den Transformationsszenarios *automatisch bestimmt* werden, sind folgende:

Variablen	Werte und Bemerkungen
annotation	<i>yes</i> Sind die erhobenen Texte annotiert?
pos_tagged	<i>yes</i> Sind die Texte nach Wortarten getaggt?
pos_tagset	<i>STTS</i> verwendetes Tagset
parsed	<i>no</i> Sind die Texte syntaktisch geparst?
error_annotated	<i>yes</i> Sind die Texte fehlerannotiert?
error_annotating_tool	<i>EXMARaLDA (Dulko)</i> Annotations-Werkzeug
annotation_other	<i>lemmata, sentence spans, target hypotheses, differences / lemmata, sentence spans, target hypotheses, differences, translated text</i> Annotationsebenen

Tab. 5.4: Variablen der Korpusannotation, die von den Transformationsszenarios automatisch bestimmt werden

Alle übrigen Variablen, die im Korpus berücksichtigt wurden, sind folgende:

Variablen	Werte und Bemerkungen
annotation_complete	<i>yes</i> Ist die Annotation abgeschlossen?

Tab. 5.5: Übrige Variablen der Korpusannotation

4.4 Text-Metadaten

Variablen, die vom *Dulko-Template* vorgegeben sind, sind folgende:

Variablen	Werte und Bemerkungen
country	<i>Hungary</i> Land der Erhebung
institution	<i>University of Szeged</i> Institution der Erhebung
written_ref_tools	<i>no</i> Durften bei der Erhebung Hilfsmitteln benutzt werden?

Tab. 5.6: Variablen der Text-Metadaten, die vom Dulko-Template vorgegeben sind

Alle übrigen Variablen, die im Korpus berücksichtigt wurden, sind folgende:

Variablen	Werte und Bemerkungen
text_ID	<i>Feminismus_3 / ...</i> Dateiname des Lernertextes (ohne Dateiendung)
text_title	<i>Feminismus in Ungarn / ...</i> Überschrift des Lernertextes (normalerweise leer)
date	<i>2017-10-31 / ...</i> Datum der Erhebung im ISO-Format
task_type	<i>in-class activity / ...</i> Kontext der Erhebung (Lehrveranstaltung, Prüfung ...)
task_instructions	<i>Übersetzen Sie ins Deutsche. / Diskutieren Sie die folgende Aussage: ...</i> Aufgabenstellung der Erhebung
written_task	<i>translation / argumentative essay</i> zu erstellende Textsorte

Tab. 5.7: Übrige Variablen der Text-Metadaten (Teil 1/2)

Variablen	Werte und Bemerkungen
task_document	<i>yes / no</i> Bezieht sich die Aufgabenstellung auf einen bestimmten zu bearbeitenden Text?
task_document_ID	<i>China / Sprache / ...</i> falls task_document = <i>no</i> : Dateiname des zu bearbeitenden Textes (ohne Dateiendung)

Tab. 5.7: Übrige Variablen der Text-Metadaten (Teil 2/2)

4.5 Lerner-Metadaten

Variablen, die vom *Dulko-Template* vorgegeben sind, sind in Tab. 5.8 aufgeführt. Alle übrigen Variablen, die im Korpus berücksichtigt wurden, können Tab. 5.9 entnommen werden.³ Falls dies für benutzerdefinierte Teilkorpora relevant erschien, wurden alle für das Korpus einschlägigen Werte angegeben.

Variablen	Werte und Bemerkungen
learner_proficiency_level_type	<i>internally rated</i> Art der Bewertung des Sprachstands des Lernalers (institutionsintern am Institut für Germanistik, Universität Szeged)
learner_proficiency_rating_scale	<i>TestDaF / onSET</i> Skala der Sprachstandserhebung (im 1. Projektjahr <i>TestDaF</i> , danach <i>onSET</i>)

Tab. 5.8: Variablen der Lerner-Metadaten, die vom Dulko-Template vorgegeben sind

Variablen	Werte und Bemerkungen
learner_ID	<i>7DDD003BC9450E4B84CCCA8B72F560CC / ...</i> (pseudo-)anonymisierte Lerner-ID
learner_status	<i>L2 learner / trainee translator</i> Status des Lernalers (Studierende mit der Spezialisierung Dolmetscher-Übersetzer: <i>trainee translator</i> ; sonst <i>L2 learner</i>)

Tab. 5.9: Übrige Variablen der Text-Metadaten (Teil 1/3)

³ Zwei Werte werden im Dulko-Korpus erhoben, die bei Granger/Paquot 2017 nicht aufgeführt sind: L2_region + L2_study_institutions.

Variablen	Werte und Bemerkungen
age	24 / ... Alter zum Zeitpunkt der Erhebung
gender	male / female Geschlecht des Lernalters laut Selbstauskunft
L1	hun Muttersprache
L2_other	yes / no Beherrscht der Lerner nach eigenen Angaben neben Deutsch (= L2_language) noch weitere Fremdsprachen?
L2_other_1	eng / rus / srp / ... erste Fremdsprache neben Deutsch (Gewichtung der Kompetenz gemäß Selbstauskunft des Lernalters)
L2_other_2	fra / ita / spa / ... zweite Fremdsprache neben Deutsch (Gewichtung der Kompetenz gemäß Selbstauskunft des Lernalters)
L2_other_3	dan / eus / ... dritte Fremdsprache neben Deutsch (Gewichtung der Kompetenz gemäß Selbstauskunft des Lernalters)
L2_other_4	lat / ... vierte Fremdsprache neben Deutsch (Gewichtung der Kompetenz gemäß Selbstauskunft des Lernalters)
L2_other_5	ita / ... fünfte Fremdsprache neben Deutsch (Gewichtung der Kompetenz gemäß Selbstauskunft des Lernalters)
learner_proficiency	103 / ... Punktestand im Sprachstandstest

Tab. 5.9: Übrige Variablen der Text-Metadaten (Teil 2/3)

Variablen	Werte und Bemerkungen
learner_level_CEFR_conversion	<i>B2 / C1 / ...</i> Sprachstand des Lerners nach dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER) (<i>B2</i> : ab 96 Punkten im Sprachstandstest; <i>C1</i> oder höher: ab 130 Punkten im Sprachstandstest)
study_level	<i>undergraduate / graduate</i> Studienniveau (<i>undergraduate</i> : BA bzw. 1. Teil des Lehramtsstudiums; <i>graduate</i> : MA bzw. 2. Teil des Lehramtsstudiums)
study_area	<i>German studies / German teacher training / translation studies</i> Studienfach (Germanistik, Lehramt Deutsch, Übersetzungswissenschaften)
L2_study_years	<i>11 / ...</i> Gesamtzahl der Jahre des besuchten Deutschunterrichts
L2_study_institutions	<i>primary school, secondary school, university / secondary school, university / ...</i> Institutionen, an denen Deutschunterricht besucht wurde (optionale Variable)
time_spent_L2_country	<i>3 / ...</i> Zeit in Monaten, während derer sich der Lerner im deutschsprachigen Ausland aufgehalten hat
L2_region	<i>Oberösterreich / Bayern, Baden-Württemberg / ...</i> Regionen im deutschsprachigen Ausland (i. d. R. Bundesländer oder Kantone), in denen sich der Lerner aufgehalten hat (optionale Variable)

Tab. 5.9: Übrige Variablen der Text-Metadaten (Teil 3/3)

5 Annotation

5.1 Tokenisierung (word-Ebene)

Der im Partitur-Editor geöffnete Lernertext wird in der ersten Spur maschinell tokenisiert. Satzzeichen gelten dabei als selbständige Tokens. Die durch die Tokenisierung entstandenen Spalten (in EXMARaLDA: Zeitpunkte) werden, mit Null angefangen, automatisch nummeriert. Die Nummerierung der Tokens eines Lernertexts ändert sich automatisch, wenn für die Zielhypothesenebene (ZH) neue Spalten eingefügt werden. Zum Beispiel (Lernertext: Weinen_6):

	0	1	2	3	4	5	6	7	8	9	10
[word]	Vor	Kurzem	,	eine	Forschung	wurde	mit	3000	Personen	gemacht	,
[S]	s1										
[pos]	APPR	NN	\$,	ART	NN	VAFIN	APPR	CARD	NN	VVPP	\$,
[lemma]	vor	Kurze	,	eine	Forschung	werden	mit	3000	Person	machen	,

	0	1	2	3	4	5	6	7	8	9	10	11
[word]	Vor	Kurzem	,		eine	Forschung	wurde	mit	3000	Personen	gemacht	,
[S]	s1											
[pos]	APPR	NN	\$,		ART	NN	VAFIN	APPR	CARD	NN	VVPP	\$,
[lemma]	vor	Kurze	,		eine	Forschung	werden	mit	3000	Person	machen	,
[ZH]	Vor	Kurzem		wurde	eine	Untersuchung		mit	3000	Personen	gemacht	,

Abb. 3: Automatische Nummerierung der Spalten

Bewusste Auslassungen in den Lernertexten (insbesondere in Form von Weißraum oder „...“) werden mit _ auf der word-Ebene repräsentiert. Wenn ein ganzes Wort ausgelassen wurde, ist _ ein eigenes Token. Ist ein Wortteil ausgelassen, ist _ Teil des entsprechenden Tokens. Für beide Typen lassen sich im Lernertext Sprache_8 Beispiele finden:

Es scheint so, als ob wir d_ einzige _ auf der Erde, d_ eine Sprache erlernen kann.

	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183
[word]	Es	scheint	so	,	als	ob	wir	d_	einzige	_	auf	der	Erde		,	d_	eine	Sprache	erlernen	kann	.
[S]	s12																				
[pos]	PPER	VVFIN	ADV	\$,	KOKOM	KOUS	PPER	NN	ADJA	XY	APPR	ART	NN		\$,	NN	ART	NN	VVINF	VMFIN	\$.
[lemma]	es	scheinen	so	,	als	ob	wir	d_	einzig	_	auf	die	Erde		,	d_	eine	Sprache	erlernen	können	.
[trans]	Úgy tűnik, mintha mi lennénk az egyetlen faj a Földön, amely el tud sajátítani egy nyelvet.																				
[ZH]	Es	scheint	so	,	als	ob	wir	die	einzige	Art	auf	der	Erde	wären	,	die	eine	Sprache	erlernen	kann	.
[ZHDiff]								CHA		CHA				INS		CHA					
[ZHS]	s12																				
[ZHpos]	PPER	VVFIN	ADV	\$,	KOKOM	KOUS	PPER	ART	ADJA	NN	APPR	ART	NN	VAFIN	\$,	PRELS	ART	NN	VVINF	VMFIN	\$.
[ZHlemma]	es	scheinen	so	,	als	ob	wir	die	einzig	Art	auf	die	Erde	sein	,	die	eine	Sprache	erlernen	können	.
[FehlerOrth]																					
[FehlerMorph]																					
[FehlerSyn]																					
[FehlerLex]																					

Abb. 4: Auslassungen im Lernertext

Bis auf die Auslassung von Flexionsendungen (F_{lex}) werden solche Auslassungen als lexikalische Fehler (L_{ex}) getaggt (s. Abschnitt 5.7.4).

5.2 Satzspannen (S-Ebene und ZHS-Ebene)

Der Lernertext und die Zielhypothesen werden bei Punkten, Doppelpunkten, Semikolons, Fragezeichen und Ausrufezeichen automatisch in Satzspannen gegliedert. Um eine möglichst korrekte Gliederung zu erreichen, soll zuerst die word-Spur überprüft und korrigiert werden, denn eine falsche Tokenisierung kann zu einer falschen Satzspannengliederung führen, etwa beim Abkürzungspunkt der Ordinalia wie im Lernertext Feminismus_7:

Satzspanne 8

	108	109	110	111	112	113	114	115	116	117	118	119
[word]	In	europäischen	Ländern	hat	die	Suffragette	Bewegung	im	19	-	20	.
[S]	s8											
[pos]	APPR	ADJA	NN	VAFIN	ART	NN	NN	APPRART	CARD	\$(CARD	\$.
[lemma]	in	europäisch	Land	haben	die	Suffragette	Bewegung	in	19	-	20	.

Satzspanne 9

	120	121	122	123	124
[word]	Jahrhundert	ganz	schnell	eskaliert	.
[S]	s9				
[pos]	NN	ADV	ADJD	VVPP	\$.
[lemma]	Jahrhundert	ganz	schnell	eskalieren	.

Abb. 5: Automatische Satzspannengliederung

Satzspanne 8

	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123
[word]	In	europäischen	Ländern	hat	die	Suffragette	Bewegung	im	19	-	20.	Jahrhundert	ganz	schnell	eskaliert	.
[S]	s8															
[pos]	APPR	ADJA	NN	VAFIN	ART	NN	NN	APPRART	CARD	\$(CARD	NN	ADV	ADJD	VVPP	\$.
[lemma]	in	europäisch	Land	haben	die	Suffragette	Bewegung	in	19	-	20.	Jahrhundert	ganz	schnell	eskalieren	.

Abb. 6: Manuelle Korrektur der falschen Satzspannengliederung

Titel werden nicht als Satzspannen identifiziert, sind aber am Anfang der Dokumente aufgeführt.

5.3 Part-of-speech-Tagging (pos-Ebene und ZHpos-Ebene)

Hier werden die Wortartenzugehörigkeit bzw. die wichtigsten grammatischen Kategorien der Tokens des Lernertexts und der Zielhypothesen mit dem Stuttgart-Tübingen-Tagset (STTS) maschinell bestimmt (vgl. Schiller et al. 1999).

5.4 Lemmatisierung (lemma-Ebene und ZHlemma-Ebene)

In den lemma- und ZHlemma-Spuren stehen unter jedem Token die entsprechenden unflektierten Lemmata sowie die wiederholten Satzzeichen.

5.5 Muttersprachliche Vorlage (trans-Ebene)

Bei den Übersetzungen wird in dieser Spur der jeweilige ungarische Ausgangstext angegeben. In der Regel ist die trans-Ebene genauso in Satzspannen unterteilt wie die entsprechende S-Ebene.

In Fällen, wo mehrere Sätze der Übersetzung (also des Lernertexts) einem einzigen Satz der muttersprachlichen Vorlage entsprechen, wird der ungarische Satz bei allen deutschen Sätzen in der trans-Ebene angegeben (vgl. Weinen_7, s1 und s2):

	0	1	2	3	4	5	6	7	
[word]	Vor	Kurzem	hat	man	eine	Forschung	gemacht	.	
[S]	s1								
[pos]	APPR	NN	VAFIN	PIS	ART	NN	VVPP	\$.	
[lemma]	vor	Kurze	haben	man	eine	Forschung	machen	.	
[trans]	Nemrég egy 3000 főt vizsgáló kutatást végeztek, amely során azt vizsgálták, hogy mennyit sír egy nő átlagosan élete különböző szakaszaiban.								

	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
[word]	Es	gab	3000	Teilnehmerinnen	und	es	wurde	untersucht	,	wie	viel	Zeit	Frauen	mit	Weinen	in	den	verschiedenen	Lebensphasen	verbringen	.
[S]	s2																				
[pos]	PPER	VVFIN	CARD	NN	KON	PPER	VAFIN	VVPP	\$,	KOUS	PIAT	NN	NN	APPR	NN	APPR	ART	ADJA	NN	VVIN	\$.
[lemma]	es	geben	3000	Teilnehmerin	und	es	werden	untersuchen	,	wie	viele	Zeit	Frau	mit	Wein/Weinen	in	die	verschieden	Lebensphase	verbringen	.
[trans]	Nemrég egy 3000 főt vizsgáló kutatást végeztek, amely során azt vizsgálták, hogy mennyit sír egy nő átlagosan élete különböző szakaszaiban.																				

Abb. 7: Ein Satz im Ungarischen – mehrere Sätze in der Übersetzung

Falls in der Übersetzung mehrere Ganzsätze des Ausgangstextes zu einem zusammengezogen werden, werden die betreffenden ungarischen Sätze in einer Satzspanne angeführt (z. B. China_1, s16):

	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293
[word]	Zum	Glück	gelang		mir	auch	,	ein	Fahrrad	zu	leihen	,	so	konnte	ich	die	Innenstadt	schneller	entdecken	.
[S]	s16																			
[pos]	APPRART	NN	VVFIN		PPER	ADV	\$,	ART	NN	PTKZU	VVIN	\$,	ADV	VMFIN	PPER	ART	NN	ADJD	VVIN	\$.
[lemma]	zu	Glück	gelingen		ich	auch	,	eine	Fahrrad	zu	leihen	,	so	können	ich	die	Innenstadt	schnell	entdecken	.
[trans]	Szerencsére nekem is sikerült biciklit bérelnem. Így sokkal gyorsabban lehetett felfedezni a belvárost.																			

Abb. 8: Mehrere Sätze im Ungarischen – ein Satz in der Übersetzung

5.6 Zielhypothesen (ZH-Ebene)

Auf den ZH-Ebenen wird der Lernertext korrigiert. Dabei wird die letzte, fehlerfreie Textvariante eventuell in mehreren Schritten (d.h. Zielhypothesen) erreicht, damit die einzelnen Fehler nachvollziehbar bleiben (vgl. Abschnitt 2.2). Als Beispiel dient der Lernertext Kriminalität_2, in dem die erste ZH-Ebene die intermediäre Zielhypothese

ist und die zweite ZH-Ebene die finale Zielhypothese enthält (weitere Spuren wurden hier der Einfachheit halber ausgelassen):

	153	154	155	156	157
[word]	Im		Hinsicht	der	Staat
[S]	s9				
[pos]	APPRART		NN	ART	NN
[lemma]	in		Hinsicht	die	Staat
[ZH]	In		Hinsicht	des	Staates
[ZH]	Aus	staatlicher	Sicht		

Abb. 9: Intermediäre und finale Zielhypothese

Es wird immer die kleinstmögliche Korrektur vorgenommen. Beim Korrigieren können auf der ZH-Ebene Spalten zusammengezogen, aufgespalten oder eingefügt werden, diese Änderungen werden dann auf der ZHDiff-Ebene ebenfalls getaggt (s. Abschnitt 5.8):

	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274
[word]	Sie	haben	viele	online	Seiten	und	andere	Möglichkeiten		über		Feminismus	zu		lernen	.
[S]	s13															
[pos]	PPER	VAFIN	PIAT	ADJA	NN	KON	PIS	NN			APPR		NN	PTKZU	VVIN	\$.
[lemma]	Sie	haben	viele	online	Seite	und	andere	Möglichkeit			über		Feminismus	zu	lernen	.
[ZH]	Man	hat	viele	Online-Seiten	und	andere	Möglichkeiten	,	etwas	über	den	Feminismus	zu	lernen	.	
[ZHDiff]	CHA	CHA		MERGE					INS	INS		INS				

Abb. 10: Zusammenziehung (MERGE) und Einfügung (INS) von Spalten auf ZH-Ebene (Feminismus_3)

	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213
[word]	Eine	aktuelle	Studie	zeigt	zum	Beispiel	,	dass	in	Ungarn	immernoch	die	klassischen	Genderstereotypen	und	Genderrollen	herrschen	.	
[S]	s12																		
[pos]	ART	ADJA	NN	VVFIN	APPRART	NN	\$.	KOUS	APPR	NE	VVFIN		ART	ADJA	NN	KON	NN	VVFIN	\$.
[lemma]	eine	aktuell	Studie	zeigen	zu	Beispiel	,	dass	in	Ungarn	immernoch	die	klassisch	Genderstereotypen	und	Genderrollen	herrschen	.	
[ZH]	Eine	aktuelle	Studie	zeigt	zum	Beispiel	,	dass	in	Ungarn	immer noch	die	klassischen	Genderstereotype	und	Genderrollen	herrschen	.	
[ZHDiff]											SPLIT			CHA					

Abb. 11: Aufspaltung (SPLIT) von Spalten auf ZH-Ebene (Feminismus_4)

5.7 Fehlerannotation

Abweichungen zwischen der word-Ebene und der ersten ZH-Ebene bzw. zwischen zwei aufeinander folgenden ZH-Ebenen, die als Fehler zu interpretieren sind, werden manuell mit Hilfe von Fehlertags kategorisiert. Bei jeder Zielhypothese werden eigene Fehlerebenen generiert (FehlerOrth, FehlerMorph, FehlerSyn, FehlerLex und FehlerSem). Über das Annotationswerkzeug von EXMARaLDA (Dulko) können hier die Fehlertags eingetragen werden.

Folgefehler, d. h. Abweichungen zwischen dem Lernertext und der ZH, die nur als Folge einer anderen Korrektur entstehen, werden nicht getaggt.

Mehrfachtagging: Wenn für einen bestimmten Fehler mehrere alternative Fehlerkategorien in Frage kommen, werden alle (bei der gleichen oder bei mehreren

ZH) getaggt. Im Fall des Lernertexts Kriminalität_8 kann z. B. *ein Unterschied* in Satzspanne 17 sowohl als Valenzfehler (es *gibt* + Nominativ) als auch als Genusfehler (Neutrum statt Maskulinum) interpretiert werden (s. Abb. 12):

	315	316	317	318	319	320	321
[word]	Ich	denke	,	es	gibt	ein	Unterschied
[S]	s17						
[pos]	PPER	VVFIN	\$,	PPER	VVFIN	ART	NN
[lemma]	ich	denken	,	es	geben	eine	Unterschied
[ZH]	Ich	denke	,	es	gibt	einen	Unterschied
[ZHDiff]						CHA	
[ZHS]	s17						
[ZHpos]	PPER	VVFIN	\$,	PPER	VVFIN	ART	NN
[ZHlemma]	ich	denken	,	es	geben	eine	Unterschied
[FehlerOrth]							
[FehlerMorph]							
[FehlerSyn]						ValV	
[FehlerLex]						Gen	

Abb. 12: Mehrfachtagging bei der gleichen Zielhypothese

Mehrere Tags stehen auch bei Fehlern, die durch die Änderung von mehr als einer (grammatischen) Kategorie korrigiert werden müssen, etwa im Fall von *die einzige Wesen auf den Erde* – ZH: das einzige Wesen auf der Erde (Sprache_5) wird sowohl Gen als auch SemRel getaggt.

Wenn es bei der Korrektur zum Austausch von Lexemen kommt (z. B. Kon-Fehler, s. 5.7.3 oder lexikalische Fehler, s. 5.7.5), ist für die Zuweisung der jeweiligen Tags grundsätzlich die erwartete Form (die ZH) ausschlaggebend. So wird z. B. die Verwendung von *was für ein* anstatt von *welcher* nicht mit Phr, sondern mit Lex getaggt, obwohl *was für ein* eine phraseologische Einheit darstellt:

- Mit *was für eine Sprache* – ZH: in welcher Sprache (Sprache_10)

Im Folgenden werden die einzelnen Fehlerebenen detailliert behandelt. Angegeben werden die Fehlertags, ihre Erklärungen und einige Beispiele aus dem annotierten Korpus (Beeh et al. 2021). Unterstreichungen in den Beispielen markieren den Skopus des jeweiligen Fehlertags.

5.7.1 Orthographische Fehler (FehlerOrth-Ebene)

Auf dieser Ebene wurden orthographische und Interpunktionsfehler getaggt, wofür der Rechtschreibduden (2017) und das Regelwerk des Rats für deutsche Rechtschreibung (2018) als Referenzwerke dienten.

GKS: Groß- und Kleinschreibung

Mit diesem Tag werden Wörter markiert, in denen große Buchstaben durch kleine ersetzt werden sollten oder umgekehrt, um den Rechtschreibregeln zu entsprechen:

- *In Schweden habe ich ganz feministisch gefühlt, Ungarische Menschen lachen an ...* – ZH: *ungarische* (Feminismus_3)
- *In Kleinasien ist es anders: da leben die Menschen ...* – ZH: *Da* (Feminismus_10)

GZS: Getrennt- und Zusammenschreibung

Wörter oder Wortgruppen, die hinsichtlich der Regeln der Getrennt- und Zusammenschreibung falsch geschrieben wurden, werden mit diesem Fehlertag versehen:

- *die Suffragette Bewegung ...* – ZH: *Suffragetten-Bewegung* (Feminismus_7)
- *immernoch* – ZH: *immer noch* (Feminismus_4)

WS: Wortschreibung

Wörter, in denen Buchstaben fehlen und/oder überflüssige Buchstaben vorkommen oder Buchstaben verwechselt werden, bekommen diesen Fehlertag:

- *Ich glaube, dass scon seit langer Zeit ...* – ZH: *schon* (Feminismus_19)
- *spührt* – ZH: *spürt* (Feminismus_9)
- *einige beliebe Reiseziele* – ZH: *beliebte* (China_3)

In manchen Fällen liegt sowohl eine falsche Wortschreibung als auch ein Wortbildungsfehler vor:

- *seehliche Verletzungen* – ZH: *seelische* (Weinen_6)

Hier wird das Wort wegen des überflüssigen *h* im Wortstamm als WS und wegen des falschen Suffixes als Wortb getaggt. Wenn ein und dieselbe Schreibung sowohl als Wortschreibungsfehler als auch als Wortbildungsfehler interpretiert werden kann, wird nur ein Wortbildungsfehler angegeben (s. Abschnitt 5.7.2).

ZS: Zeichensetzung

Der Tag ZS wird bei Fehlern verwendet, welche die Setzung von Satzzeichen auf Satz- und Wortebene betreffen. Bei Zeichensetzungsfehlern auf Satzebene wird das

falsche Satzzeichen auf der word-Ebene und/oder die auf der ZH-Ebene eingefügte Korrektur mit ZS getaggt (in den Beispielen mit Unterstrich markiert):

- *Das Thema_ Feminismus interessiert mich ...* – ZH: *Das Thema Feminismus* (Feminismus_12)
- *Ich glaube_im Ungarn haben die Frauen ...* – ZH: *Ich glaube, in Ungarn ...* (Feminismus_12)

Bei Zeichensetzungsfehlern auf Wortebene, z. B. bei Abkürzungen, wird das entsprechende Wort mit ZS getaggt:

- zB. – ZH: z. B. (Entlohnung_16)

Bei paarigen Satzzeichen (Kommas, Gedankenstrichen, Klammern) werden die falsch gesetzten oder nicht gesetzten Teile einzeln als Fehler annotiert.

Bei mit Komma verbundenen parataktischen Teilsätzen, die logisch gesehen besser mit Semikolon oder Punkt verbunden wären, wird das Komma in der Zielhypothese durch ein Semikolon ersetzt, weil dies weniger invasiv ist (vgl. Weinen_5, Spalte 32):

Overview

word

ZH

FehlerOrth

FehlerMorph

FehlerLex

Details

word

s2

txt::pos

txt::lemma

txt::trans

ZH1::ZH

ZH1::ZHDiff

ZH1::ZHS

ZH1::ZHpos

ZH1::ZHlemma

ZH1::FehlerOrth

ZH1::FehlerMorph

ZH1::FehlerLex

Die	Endergebnis	wirkt	erstmal	vielleicht	schockierend	.	es	kam	heraus	.	dass	die	Frauen	16	Monaten	ihrer	Lebens	mit	Weinen	verbringen	.
Das	Endergebnis	wirkt	erstmal	vielleicht	schockierend	.	es	kam	heraus	.	dass	die	Frauen	16	Monate	ihrer	Lebens	mit	Weinen	verbringen	.
						ZS															
																Flex					
Gen																					

Details

word

s2

txt::pos

txt::lemma

txt::trans

ZH1::ZH

ZH1::ZHDiff

ZH1::ZHS

ZH1::ZHpos

ZH1::ZHlemma

ZH1::FehlerOrth

ZH1::FehlerMorph

ZH1::FehlerLex

Die	Endergebnis	wirkt	erstmal	vielleicht	schockierend	.	es	kam	heraus	.	dass	die	Frauen	16	Monaten	ihrer	Lebens	mit	Weinen	verbringen	.
s2																					
ART	NN	VVFIN	ADV	ADV	ADJD	\$.	PPER	VVFIN	PTKVZ	\$.	KOUS	ART	NN	CARD	NN	PPOSAT	NN	APPR	NN	VVINF	\$.
die	Endergebnis	wirken	erstmal	vielleicht	schockierend	.	es	kommen	heraus	.	dass	die	Frau	16	Monat	ihr	Leben	mit	Wein Weinen	verbringen	
A végeredmény elsőre talán megdöbbentő lehet, kiderült ugyanis, hogy a nők egész életükben 16 hónapot töltenek el sírással.																					
Das	Endergebnis	wirkt	erstmal	vielleicht	schockierend	.	es	kam	heraus	.	dass	die	Frauen	16	Monate	ihrer	Lebens	mit	Weinen	verbringen	.
CHA							CHA								CHA						
s2							s3														
ART	NN	VVFIN	ADV	ADV	ADJD	\$.	PPER	VVFIN	PTKVZ	\$.	KOUS	ART	NN	CARD	NN	PPOSAT	NN	APPR	NN	VVINF	\$.
die	Endergebnis	wirken	erstmal	vielleicht	schockierend	.	es	kommen	heraus	.	dass	die	Frau	16	Monat	ihr	Leben	mit	Wein Weinen	verbringen	
						ZS															
																Flex					
Gen																					

Abb. 13: Verwendung des Semikolons in der Zielhypothese

5.7.2 Morphologische Fehler (FehlerMorph-Ebene)

Unter den morphologischen Fehlerkategorien werden die Abweichungen bezüglich der Flexion und Wortbildung zusammengefasst. Für die Komparation sowie für die Verwendung von Auxiliaren gibt es eigene Fehlerkategorien.

Flex: Flexion

Mit Flex werden Flexionsfehler verschiedener Art getaggt, die innerhalb einer Wortform auftreten. Der Hauptverwendungsbereich des Tags ist die Markierung von falschen Flexionsklassen; dies betrifft insbesondere verbale (z. B. *darfen/dürfen*),

adjektivische (z. B. *besser/bessere Jobs*) und nominale Flexionsfehler (z. B. *unsere Politikern/Politiker*).

Die Abgrenzung von ‚puren‘ Flexionsfehlern von anderen, syntaktischen und lexikalischen Fehlern ist allerdings oft schwierig, da manche Fehlertypen (v. a. Valenz- und Genusfehler) in einem implikativen Verhältnis mit Flexionsfehlern stehen, da sich Erstere in der Regel mit einer abweichenden Flexion der regierten Elemente oder der Determinative einhergehen. Wenn ein Flexionsfehler durch die Änderung einer einzigen Kategorie – z. B. Genus – behoben werden kann, wird dieser Fehler als Genusfehler betrachtet und nicht zusätzlich mit Flex getaggt. Wenn z. B. im Komplement einer Adposition eine falsche Artikel- oder Adjektivform entweder als Flex oder als ValAP getaggt werden könnte, wird der Fehler als ValAP getaggt, sofern das Komplement mit dieser falschen Form einen anderen als den regierten Kasus in korrekter Flexion realisiert (z. B. *während ihre Weltentdeckung*, ZH: während ihrer Entdeckung der Welt, Weinen_6).

Wenn die Wörter einer Nominalgruppe teilweise im richtigen Genus, Kasus bzw. Numerus stehen, wird lediglich das falsch flektierte Element mit Flex getaggt, z. B. *dass in die anderen Ländern* (ZH: dass in den anderen Ländern, Entlohnung_27) – kein SemRel für die ganze Phrase, sondern Flex für den Artikel.

Bei Mehrfachannotationen wird Flex im Allgemeinen außer Acht gelassen: Zum Beispiel kann die Abweichung in der Wortgruppe *es gibt ein Unterschied* (Kriminalität_8) sowohl als Valenz- als auch als Genusfehler beschrieben werden; die entsprechenden Stellen werden mit Gen bzw. ValV getaggt, nicht aber das Wort *ein* mit Flex. Auch weitere Fehler, z. B. die fehlerhafte Markierung des Possessors (s. Poss in 5.7.3) oder die semantisch unpassende Wahl des Numerus (s. Num in 5.7.5) sowie Kongruenzfehler verschiedener Art (s. Kongr-Tags in 5.7.3) sind als Abweichungen in der Flexion erkennbar; trotzdem wird auch in diesen Fällen nur der entsprechende syntaktische bzw. semantische Fehlertag verwendet.

Flex beim Substantiv:

- *unsere Politikern und Ministern denken, dass ...* – ZH: *Politiker und Minister* (Feminismus_3)
- *keine Schaden* – ZH: *Schäden* (Feminismus_4)

Flex beim Verb:

- *die Frauen da haben alles beginnen⁴* – ZH: *begonnen* (Feminismus_3)
- *sie darfen nicht studieren* – ZH: *dürfen* (Feminismus_11)

⁴ auch Infin, da hier zwei infinite Verbformen zusammenfallen

Flex beim Adjektiv:

- wer besser Jobs und equal Geld verdienen möchten – ZH: bessere (Feminismus_3)
- Die westliche Länder – ZH: westlichen (Feminismus_3)
- die nach einem Job suchenden Frau – ZH: suchende (Feminismus_20)
- zur östliche Region des Landes – ZH: östlichen (Feminismus_3)

Flex beim Determinativ:

- mit der neuen Generationen – ZH: den (Feminismus_12)

Wenn zwei oder mehr Adjektive in derselben Substantivgruppe falsch dekliniert werden, wird jedes Adjektiv separat mit Flex getaggt.

- In Zusammenhang mit meinem schwedischen Freiwilliges Soziales Jahr – ZH: mit meinem schwedischen Freiwilligen Sozialen Jahr (Feminismus_3)

Wortb: Wortbildung

Mit Wortb werden Abweichungen getaggt, bei denen nicht der Stamm, sondern Präfixe, Suffixe oder andere Wortbildungselemente (z. B. Fugenelemente) korrigiert werden müssen. Dies betrifft insbesondere eine Inkompatibilität zwischen Stamm und Affix, die doppelte Verwendung von Derivationsaffixen, Abweichungen bei den Fugenelementen in Komposita sowie Affixderivation statt Konversion oder umgekehrt.

Die Verwendung des Tags geschieht unabhängig davon, ob die abweichende Form auf der word-Ebene ein existierendes, ein mögliches oder kein mögliches Lexem des Deutschen darstellt (z. B. *Beschränkung* vs. *Einschränkung*, Feminismus_13; *prophezien* vs. *prophezeien*, Entlohnung_3). Die korrigierte Form auf der ZH-Ebene ist hingegen stets ein existierendes Lexem des Deutschen. Wenn gleichzeitig ein Wortbildungs- und ein Flexionsfehler vorliegt, werden diese in zwei ZH-n korrigiert (z. B. *Kriminalen werden* - ZH1: *Kriminellen werden* - ZH2: *Kriminelle werden*, Kriminalität_3).

- maximisiert – ZH: *maximalisiert* (Feminismus_4)
- ein patriarchalistische Staat – ZH: *patriarchalischer* (Feminismus_7)
- Beschränkung – ZH: *Einschränkung* (Feminismus_13)
- scheinen – ZH: *erscheinen* (Feminismus_13)
- während IT ... blütet – ZH: *blüht* (Entlohnung_16)
- Lebensstil – ZH: *Lebensstil* (Kriminalität_1)
- Geschichteunterricht – ZH: *Geschichtsunterricht* (Feminismus_4)

- *KrankepflegerInnen* – ZH: *KrankenpflegerInnen* (Entlohnung_3)

Nicht kontextgemäße Simplicia an der Stelle von Komposita (typischerweise im Fall fehlender Kompositionsglieder) werden jedoch nicht als Wortbildungsfehler, sondern als lexikalische Fehler gewertet (s. Lex in 5.7.5).

Wenn ein Wortbildungsfehler als Teil einer festen Wortverbindung auftritt, wird sowohl Wortb als auch Phr getaggt, z. B.:

- *benenne einige Beispiele* – ZH: *nenne* (Entlohnung_30)

Aux: Auxiliar

Mit Aux werden falsch gewählte Auxiliare (Perfekt- oder Passivauxiliare) getaggt.

- *Es ist gesagt* – ZH: *wird* (Feminismus_13)
- *hat die Suffragette Bewegung ... eskaliert* – ZH: *ist* (Feminismus_7)

Falls nur Personalendungen fehlerhaft realisiert werden, wird nicht Aux, sondern die entsprechenden Tags (Flex bzw. KongrSubj) gewählt.

Kompb: Komparativbildung

Kompb kennzeichnet fehlerhaft gebildete Komparation.

- *viel mehr entwickelt* – ZH: *viel entwickelter* (Feminismus_12)
- *weniger bessere Chancen* – ZH: *weniger gute Chancen* (Feminismus_9)
- *das größeste Problem* – ZH: *größte* (Entlohnung_22)

5.7.3 Syntaktische Fehlerkategorien (FehlerSyn-Ebene)

(I) Wortstellung

Wortstellungsfehler werden immer durch das Einfügen von neuen Spalten korrigiert. Gemäß Falko-Standards geschieht die Verschiebung der Elemente meist von rechts nach links (eine Ausnahme stellt der Fall dar, wenn kurze Einheiten – z. B. ein einzelnes Wort – um mehrere Wörter verschoben werden müssen). Der Skopus der Wortstellungstags umfasst den größtmöglichen kontinuierlichen Bereich vom ersten falsch gestellten Element (bzw. vom ersten eingefügten Element) bis zum letzten falsch gestellten Element (bzw. bis zum letzten eingefügten Element).

Overview	
word	ist nicht so einfach zu bestimmen , weil im Fernsehen oder in der Zeitung kann man sich nicht über alles informieren
ZH	ist nicht so einfach zu bestimmen , weil im Fernsehen oder in der Zeitung man sich nicht über alles informieren kann
FehlerSyn:	StV
ZH	ist nicht so einfach zu bestimmen , weil man sich im Fernsehen oder in der Zeitung nicht über alles informieren kann
FehlerSyn:	StMF
Details	
word	ist nicht so einfach zu bestimmen , weil im Fernsehen oder in der Zeitung kann man sich nicht
txt::S	s24
txt::pos	VAFIN PTKNEG ADV ADJD PTKZU VVINF \$, KOUS APPRART NN KON APPR ART NN VMFIN PIS PRF PTKNEG
txt::lemma	sein nicht so einfach zu bestimmen , weil in Fernsehen oder in die Zeitung können man sich nicht
ZH1::ZH	ist nicht so einfach zu bestimmen , weil im Fernsehen oder in der Zeitung man sich nicht
ZH1::ZHDiff	
ZH1::ZHS	s24
ZH1::ZHpos	VAFIN PTKNEG ADV ADJD PTKZU VVINF \$, KOUS APPRART NN KON APPR ART NN PIS PRF PTKNEG
ZH1::ZHlemma	sein nicht so einfach zu bestimmen , weil in Fernsehen oder in die Zeitung man sich nicht
ZH1::FehlerSyn	StV
ZH2::ZH	ist nicht so einfach zu bestimmen , weil man sich im Fernsehen oder in der Zeitung nicht
ZH2::ZHDiff	MOVT MOVT
ZH2::ZHpos	PIS PRF
ZH2::ZHlemma	man sich

Abb. 14: Korrektur von Wortstellungsfehlern – allgemeines Schema (Kriminalität_4)

Overview	
word	. Doch hängt die Belohnung mit der Schwierigkeit der Aufgaben so eng zusammen .
ZH	. Doch hängt so die Belohnung mit der Schwierigkeit der Aufgaben eng zusammen .
FehlerSyn:	StMF
Details	
word	. Doch hängt die Belohnung mit der Schwierigkeit der Aufgaben so eng zusammen .
txt::S	s4 s5
txt::pos	\$. ADV VVFIN ART NN APPR ART NN ART NN ADV ADJD PTKVZ \$.
txt::lemma	. doch hängen die Belohnung mit der Schwierigkeit die Aufgabe so eng zusammen .
ZH1::ZH	. Doch hängt so die Belohnung mit der Schwierigkeit der Aufgaben eng zusammen .
ZH1::ZHDiff	MOVT
ZH1::ZHS	s4 s5
ZH1::ZHpos	\$. KON VVFIN ADV ART NN APPR ART NN ART NN ADJD PTKVZ \$.
ZH1::ZHlemma	. doch hängen so die Belohnung mit der Schwierigkeit die Aufgabe eng zusammen .
ZH1::FehlerSyn	StMF

Abb. 15. Korrektur von Wortstellungsfehlern mit der Verschiebung eines relativ kurzen Elements nach rechts (Entlohnung_10)

StV: Stellung des Verbs

Mit StV wird die falsche Stellung des Finitums getaggt.

- *Ich denke, die Menschen sich entwickeln* – ZH: *entwickeln sich* (Feminismus_13)
- *In meiner Situation leider gibt es* – ZH: *gibt es leider* (Feminismus_3)

Ausklammerungen werden nicht korrigiert, da sie als umgangssprachliche Phänomene nicht unter den hier angewandten Fehlerbegriff fallen.

StPTKVZ: Stellung der Verbpartikel

- *weil ich den Wort nicht sprechen konnte aus* – ZH: *aussprechen konnte* (Beispiel aus Mebus, Gudula u. a. (1989): Sprachbrücke. Bd. 2. München, S. 147)

- *Denn man fängt die Frage zu stellen an – ZH: fängt an die Frage zu stellen*
(Beispiel aus falkoEssayL2v2.4)

StW: Stellung des Frageworts

- *ob eine Frau in ihre Lebensphasen durchschnittlich wie viel Zeit verbringt –*
ZH: *wie viel Zeit eine Frau in ihren Lebensphasen durchschnittlich mit Weinen verbringt* (Weinen_2)

StPTK: Stellung der Negations- oder Fokuspartikel

- *und deshalb nicht das Leben genießen kann – ZH: und deshalb das Leben nicht genießen kann* (Beispiel aus falkoEssayL2v2.4)
- *viele solche Traditionen auch – ZH: auch viele solche Traditionen* (Feminismus_3)
- *Es gibt Lehrer auch. – ZH: auch Lehrer* (Feminismus_11)

StMF: Stellung im Mittelfeld

Mit diesem Tag wird eine korrigierte Reihenfolge von Satzgliedern im Mittelfeld erfasst.

- *Frauen wollten die gleichen Rechte immer, ... – ZH: immer die gleichen Rechte* (Feminismus_3)
- *Die Frauen beschäftigen sich mit sich viel, ... – ZH: viel mit sich selbst* (Feminismus_9)

StN: Wortstellung in der Nominalgruppe

Mit diesem Tag werden falsche Stellungen in einer Nominalgruppe kennzeichnet.

- *deswegen kann ich nicht mit vielen Informationen, eine ausführliche Zusammenfassung schreiben – ZH: eine ausführliche Zusammenfassung mit vielen Informationen* (Feminismus_12)

Ist die Nominalgruppe Komplement einer Adposition, kann diese, falls erforderlich, ebenfalls umgestellt werden. Der Skopus des Tags umfasst dann zusätzlich die Adposition:

- *Meiner Lehrer Meinung nach – ZH: Nach der Meinung meines Lehrers* (China_5)

StND: Stellung der Teile eines Eigennamens oder einer Datumsangabe

- Hosszú Katinka – ZH: *Katinka Hosszú* (Feminismus_1)

HSt: Herausstellung

Mit HSt werden Herausstellungen wie Links- oder Rechtsversetzungen getaggt, die in inkorrekt Weise gebildet sind:

- Die westliche Länder, die Frauen da haben alles beginnen – ZH: *In den westlichen Ländern, da haben die Frauen...* (Feminismus_3)

(II) Valenz (Rektion)

Die Tags ValV, ValN, ValADJ und ValAP kennzeichnen fehlerhafte oder fehlende Realisierungen der Verb-, Substantiv-, Adjektiv- und Adpositionsvalenz. Valenzfehler sind meist erkennbar als Kasusfehler oder als Adpositionsfehler (falsche Präposition bzw. Postposition). In Fällen, wo ein falsch flektiertes Wort allein durch die valenzgerechte Änderung des Kasus korrigiert werden kann, wird ein Valenzfehler getaggt (dies betrifft sowohl reine Kasusreaktionen als auch Kasus in vom Verb festgelegten Präpositionalgruppen). Wenn ein Fehler sowohl als Kasus- als auch als Genusfehler interpretiert werden kann, werden beide Tags verwendet (z. B. *es gibt ein Unterschied* (Kriminalität_8). Bei falschen Präpositionen wird ValV getaggt, wenn die Präposition vom Verb festgelegt wird; bei anderen Ergänzungen (z. B. Direktiv- oder Lokalergänzungen) wird stattdessen SemRel getaggt (z. B. *nach einer kleineren Stadt ... zu reisen*, ZH *in eine kleinere Stadt*, China_4). Bei einer Adposition wird ein ungrammatischer Kasus des Komplements grundsätzlich mit ValAP getaggt (z. B. *nach ein Date*, Feminismus_3), während eine rein semantisch bedingte Abweichung mit SemRel getaggt wird (dies betrifft Wechselprepositionen wie in *an einen/einem speziellen Ort feiern*, Kriminalität_4).

Der Skopus der Valenztags erstreckt sich immer auf die ganze regierte Phrase. Der regierende Valenzträger selbst gehört nicht zum Skopus (er umfasst also bei ValV nicht das regierende Verb, bei ValN nicht das regierende Nomen, bei ValADJ nicht das regierende Adjektiv und bei ValAP nicht die regierende Adposition).

ValV: Verbvalenz (falscher Kasus, falsche Präposition etc.)

- *in der Regierungsprozess nehmen auch wenige Frauen teil* – ZH: *an dem* (Feminismus_4)
- *Ich habe oft darüber gehört, dass ...* – ZH: *davon* (Feminismus_13)

ValN: Substantivvalenz (falsche Präposition etc.)

- *dies hat meine Meinung daran erstärkt* – ZH: *darüber* (Feminismus_20)
- *auf diesen Fragen ... eine Antwort geben* – ZH: *auf diese Fragen* (Feminismus_9)
- *Im Gegensatz existieren* – ZH: *Im Gegensatz dazu* (Feminismus_10)

ValADJ: Adjektivvalenz (falscher Kasus, falsche Präposition etc.)

- *wegen Hormone produzierte Ungleichheit* – ZH: *wegen durch Hormone produzierter Unausgeglichenheit* (Weinen_6)

ValAP: Adpositionsvalenz (falscher Kasus)

- *nach ein Date* – ZH: *einem* (Feminismus_3)
- *für demselben Job* – ZH: *denselben Job* (Feminismus_15)

(III) (Ungrammatisches oder fehlendes) Korrelat

Korrelatfehler werden in den folgenden Fällen mit Kor-Tags getaggt:

- Es wurde ein Korrelat gesetzt, obwohl kein Korrelat möglich ist.
- Es wurde kein Korrelat gesetzt, obwohl ein Korrelat erforderlich ist.

KorKomp: zu einem Komplementsatz (Objektsatz etc.)

- *kann es behauptet werden, dass* – ZH: *kann behauptet werden* (Feminismus_1)
- *hören wie z. B.* – ZH: *wie z. B. davon* (Feminismus_1)
- *Feminismus hatte den Aufgabe, das zu zeigen, dass es kein großen Unterschied zwischen Männern und Frauen gibt* – ZH: *... , zu zeigen* (Feminismus_16)
- *In diesem Essay wird geschrieben, ob...* – ZH: *wird darüber geschrieben* (Feminismus_13)

KorSupp: zu einem Supplementsatz (Relativsatz etc.)

- *... kann ich kein Beispiel aus meinem Leben erwähnen, wenn...* – ZH: *dafür, dass* (Feminismus_17)
- *aber es sind nur kaum welche Fälle, wo* – ZH: *kaum Fälle* (Feminismus_12)
- *Im Gegensatz existieren auch solche Orte, wo die Gesellschaft matriarchal ist.* – ZH: *existieren auch Orte* (Feminismus_10)

(IV) (Ungrammatischer oder fehlender) Konnektor:

Bei falschen Konnektoren werden die Kon-Tags in den folgenden Fällen verwendet:

- Ein Konnektor (z. B. eine Subjunktion) fehlt.
- Ein Konnektor ist zu viel.
- Ein Konnektor hat die falsche Wortart.

Ansonsten (z. B. *weil* statt *obwohl*) werden falsche Konnektoren mit Lex getaggt.

Da es bei der Korrektur immer zu einem Lexemaustausch kommt, ist hier für das Tagging die erwartete Form (genauer gesagt, die auf der ZHpos-Ebene befindliche Kategorie) ausschlaggebend. Deshalb wird beispielsweise der Austausch einer Konjunktion auf der word-Ebene gegen eine Subjunktion in der ZH als KonKOU (und nicht als KonKON) getaggt:

- *Denn die Menschen wissen es nicht, wird in der Zukunft nichts verändern.* – ZH: *Da die Menschen es nicht wissen, wird sich...* (Entlohnung_26)

KonKOU: Subjunktion

- *hören wie z. B. Frauen kämpfen für ihre Rechte* – ZH: *davon, dass ... kämpfen* (Feminismus_1)

KonKON: Konjunktion

- *Sie können schon Schullehrerinnen, Ärztinnen, Bürgermeisterinnen sein.* – ZH: *und* (Feminismus_9)

(V) Kongruenz

Die Tags KongrSubj und KongrAnt kennzeichnen die fehlende Übereinstimmung grammatischer Kategorien, bei denen eine Kongruenz syntaktisch erforderlich ist. Bei der Kongruenz mit dem Subjekt (KongrSubj) handelt es sich um die verbalen Kategorien Person bzw. Numerus, bei der Kongruenz mit dem Antezedens (KongrAnt) in der Regel um den Numerus und/oder das Genus eines anaphorischen Pronomens.

KongrSubj: mit dem Subjekt (in Person oder Numerus)

- *dass die Differenzen zwischen Männer und Frauen gesellschaftlich aufgebaut ist* – ZH: *sind* (Feminismus_4)
- *Ich denke, dass es keine weibliche Politikern gaben* – ZH: *es ... gab* (Feminismus_14)

KongrAnt: mit dem Antezedens (in Numerus oder Genus)

- *in Ungarn bleibt die Situation so, wie es jetzt ist.* – ZH: *sie* (Feminismus_20; fehlende Genuskongruenz)
- *Rechte, mit der sie zufrieden sein könnten* – ZH: *denen* (Feminismus_12, fehlende Numeruskongruenz)
- *eine Studie, in dem analysiert wurde* – ZH: *der* (Weinen_12, fehlende Genuskongruenz)

Der Tag KongrAnt wird nur bei fehlender, aber syntaktisch erforderlicher Übereinstimmung von Antezedens und Anapher verwendet; für andere Fälle unpassender anaphorischer Pronomen steht der semantische Tag Ant zur Verfügung (s. 5.7.5).

(VI) Weitere syntaktische Fehlerkategorien

Subj: (ungrammatisches oder fehlendes) Subjekt

Der Tag Subj wird in den folgenden Fällen verwendet.

1. Das Subjekt auf der word-Ebene wurde auf der ZH-Ebene getilgt (z. B. ein Subjekt in einer subjektlosen Struktur oder zwei Subjekte in einem Satz):
 - *Danach haben sie mir doch erlaubt, ich in eines solches fremdes Land allein zu fahren.* – ZH: *Danach haben sie mir doch erlaubt, in ein so fremdes Land allein zu fahren.* (China_5)
 - *dass man immer, wenn man hungrig ist, ist es die einzige Möglichkeit, zu stehlen* – ZH: *dass es immer, wenn man hungrig ist, die einzige Möglichkeit ist, zu stehlen* (Kriminalität_12)
 - *einige wer alle diese Traditione nicht mögen und sie wollen manche Sachen verändern* – ZH: *einige, die alle diese Traditionen nicht mögen und manche Sachen verändern wollen* (Feminismus_3)
2. Ein auf der word-Ebene fehlendes obligatorisches Subjekt wurde auf der ZH-Ebene eingefügt:
 - *Erst möchte erklären, was eigentlich Kriminalität ist.* – *Erst möchte ich erklären, was eigentlich Kriminalität ist.* (Kriminalität_3)
 - *ein Punkt ist, worüber sich großen Gedanken macht* – ZH: *... worüber man sich ...* (Feminismus_19)

Ein als Subjekt eingefügtes fixes es wird nicht mit Subj, sondern mit Exp1 getaggt.

Präd: Prädikat

Ein Prädikatsfehler (Präd) liegt in den folgenden Fällen vor.

1. In einem Teilsatz wird kein finites Verb verwendet:

- *ob es eher geschadet oder genutzt* – ZH: *geschadet oder genutzt hat* (Feminismus_9)

2. Ein Teilsatz enthält mehr als ein finites Verb:

- *es ist überall so war* – ZH: *es war überall so* (Entlohnung_9)

Alle übrigen Fehler bezüglich des Prädikats – wie etwa Flexions- oder Wortstellungsfehler – werden stattdessen mit den entsprechenden Tags versehen.

Infin: (In-)Finitheit

Abweichungen bezüglich der Finitheit/Infinitheit von Verbformen werden mit *Infin* getaggt. Dies beinhaltet die folgenden Fälle:

1. Statt einer finiten Verbform steht eine infinite Form:

- *denen die Eltern die erstrangigen Vorbilder seien sollten* – ZH: *sein* (Kriminalität_1)

2. Statt einer infiniten Verbform steht eine finite Form:

- *Ein sollte nie damit rechnet* – ZH: *Man sollte nie damit rechnen* (Kriminalität_4)

3. Statt einer bestimmten infiniten Verbform steht eine andere Art von infiniten Verbform (z. B. ein Partizip Perfekt statt eines Infinitivs, ein reiner Infinitiv statt eines *zu*-Infinitivs):

- *Zeit wieder arbeiten gehen* – ZH: *zu gehen* (Feminismus_13)
- *kann ich völlig einverstanden zu sein* – ZH: *sein* (Feminismus_7)
- *Emanzipation und Feminismus in Ungarn ist eher negativ bewertend* – ZH: *werden ... bewertet* (Feminismus_3)

Der Skopus des Fehlertags erstreckt sich jeweils auf den gesamten betroffenen Bereich der infiniten/finiten Formen.

Exp1: Expletivum/Vorfeld-es

Der Tag Exp1 kommt im Fall von falsch verwendeten Expletiva (Vorfeld-es oder fixes es) zum Einsatz. Dies betrifft insbesondere die folgenden Fälle:

1. Es wurde ein Expletivum (Vorfeld-es) gesetzt, wo es nicht möglich ist:
 - *dass es in anderen Ländern mehr Wert auf Feminismus gelegen wird* – ZH: *dass in anderen Ländern mehr Wert auf Feminismus gelegt wird* (Feminismus_1)
2. Es wurde ein fixes es gesetzt, wo keines möglich ist:
 - *früher galt es in Serbien* – ZH: *früher galt in Serbien, dass ...* (Feminismus_19)
3. Es wurde ein erforderliches fixes es nicht gesetzt:
 - *wie kalt im Wald sein kann* – ZH: *wie kalt es im Wald sein kann* (Wandern_2)

Ref1: (ungrammatische oder fehlende) Reflexivität

Mit Ref1 werden Abweichungen der folgenden Art getaggt:

1. Ein erforderliches Reflexivpronomen wurde nicht gesetzt:
 - *soll Ungarn auch noch viel entwickeln* – ZH: *sich ... entwickeln* (Feminismus_4)
2. Es wurde ein Reflexivpronomen verwendet, wo keines möglich ist:
 - *dass sich diese Angelegenheit in der Zukunft noch sichtbarer wird* – ZH: *dass diese Angelegenheit ...* (Feminismus_19)

Det: (ungrammatischer oder fehlender) Determinator

Der Tag Det (ungrammatischer oder fehlender Determinator) wird verwendet, wenn innerhalb einer Definitheitskategorie eine falsche Artikelverwendung vorliegt. Typische Fälle sind:

1. Bei einem indefinit zu interpretierenden Kontinuativum wurde ein Indefinitartikel gesetzt, oder bei einem indefinit zu interpretierenden Individuativum im Singular wurde kein Indefinitartikel gesetzt:

- *In den vergangenen Jahren hat dieses Effekt Rollenwechsel in den Familien hervorgehoben.* – ZH: einen Rollenwechsel (Feminismus_18)
- *Die Frauen haben Rechten und freies Willen ...* – ZH: einen freien Willen (Feminismus_9)

2. Es wurde mehr als ein Artikel gesetzt:

- *die den Chinesischer Großmauer* – ZH: *die Chinesische Mauer* (China_15)

Abweichungen, die die Definitheitskategorien selbst betreffen, werden bei den semantischen Fehlertags behandelt (s. Def in 5.7.5). Der Skopus des Tags Det ist immer nur ein einzelner Determinator.

Poss: Possessivität

Possessoren in Possessivkonstruktionen, die fehlerhaft realisiert wurden (etwa in einem falschen Kasus oder in Form einer *von*-Phrase), werden mit Poss getaggt. Der Skopus des Tags umfasst die fehlerhaft realisierte Phrase des Possessors:

- *in der ersten Periode unserem Leben* – ZH: *unseres Lebens* (Weinen_16)
- *um Hilfe und Unterstützung andere Menschen* – ZH: *anderer Menschen* (Entlohnung_16)
- *ein ganzes Jahr von ihrem Leben* – ZH: *ihres Lebens* (Weinen_12)

Komp: Komparation, Vergleichspartikel

Der Tag Komp kommt bei fehlenden oder überflüssigen Vergleichspartikeln zum Einsatz:

- *Viele östliche Länder, Arabien und China* – ZH: wie Arabien und China (Feminismus_3)

Eine Verwechslung der Vergleichspartikeln *wie* und *als* wird stattdessen mit Lex getaggt (s. Lex in 5.7.5).

Neg: Negation

Der Tag Neg kennzeichnet:

1. ein fehlendes Negationswort,
2. doppelte Negation oder
3. Abweichungen, die den Negationsartikel betreffen (Verwendung von *nicht ... ein/e/r/s* statt *kein/e/r/s*)

- nicht ... eine ausführliche Zusammenfassung – ZH: *keine* (Feminismus_12)
- *kann ich nicht so einen Fall nennen, wenn* – ZH: *keinen Fall* (Feminismus_13)

Ellip: Ellipse

Syntaktisch inkorrekte Ellipsen können explizit mit dem Tag Ellip getaggt werden. Dabei markiert Ellip jeweils eine kontinuierliche Folge von fälschlicherweise ausgelassenen Tokens. Im folgenden Beispiel ist die Auslassung von *existieren* im letzten Konjunkt mit Ellip markiert:

- *Kriminalität ist immer ein aktuelles Thema, das war in der Vergangenheit, existiert auch heute und das wird wahrscheinlich in der Zukunft.* – ZH: *das wird wahrscheinlich in der Zukunft existieren* (Kriminalität_2)

Wenn in einer Koordination von definit zu interpretierenden Substantiven unterschiedlichen Genus einer der Definitartikel weggelassen wurde, wird ebenfalls Ellip getaggt:

- *Oder viele können den Hass oder Eifersucht nicht ertragen* – ZH: *den Hass oder die Eifersucht* (Kriminalität_4)

5.7.4 Lexikalische Fehler (FehlerLex-Ebene)

Zu den lexikalischen Fehlern werden Abweichungen auf lexikalischer Ebene gezählt.

Gen: (Substantiv-)Genus

Ein Genusfehler liegt bei falscher Verwendung des Genus bei Substantiven vor.

Die Wahl der Artikelform (Maskulinum vs. Femininum vs. Neutrum) zeigt als zentraler Genus-Indikator, welches Genus der Lerner intendiert hat. Hier liegt demnach ein Genusfehler der Nominalgruppe und kein Flexionsfehler des Artikels vor. Da die Fehlerkategorie Gen (Substantivgenus) immer die gesamte Substantivgruppe betrifft, bildet diese immer den Skopus des Fehlertags, auch wenn die Abweichung nur an einem Determinator oder Adjektiv erkennbar ist.

- eine patriarchalistische Staat – ZH: *ein patriarchalischer Staat* (Feminismus_7)
- Verurteilung des Gesellschafts – ZH: *der Gesellschaft* (Feminismus_7)
- von der Beginn – ZH: *seit dem Beginn* (Feminismus_7)

In einigen Fällen kann der Fehler sowohl als Genus als auch als Kasusfehler kategorisiert werden. Diese Abgrenzungsproblematik ergibt sich z.T. aus den

Flexionseigenschaften des Determinators. Solche ambigen Fälle werden mit beiden Tags (Gen und ValV oder Poss) getaggt.

- Prozesse der Feminismus – ZH: *Ergebnisse des Feminismus* (Feminismus_13)

Lex: Wortwahl

Die Fehlerkategorie Wortwahl (Lex) bezieht sich auf die Verwendung im Deutschen nicht existierender oder kontextuell unpassender lexikalischer Einheiten oder auf Auslassungen derselben.

Bei Lex werden folgende Fälle differenziert:

1. falsche Verwendung eines (morphologisch einfachen) Lexems:

- kostete ich die berühmte Ente – ZH: *probierte ich die berühmte Ente* (China_1)
- Man soll aber auch berücksichtigen – ZH: *Man muss aber auch berücksichtigen* (Wandern_2)
- Es gibt immer mehrere Wanderer – ZH: *Es gibt immer mehr Wanderer* (Wandern_5)

2. phrasale Umschreibung statt Kompositum:

- einige chinesische Sprachunterricht – ZH: *einige Chinesischstunden* (China_1)
- den lokalen Menschen – ZH: *den Einheimischen* (China_1)
- mit Schreiben ins Tagebuch – ZH: *mit Tagebuchschreiben* (China_1)

3. Auslassung:

- auf den Gebieten – ZH: *in den Naturschutzgebieten* (Wandern_2)
- später sind die nervliche Aus- – ZH: *später sind die nervliche Erschöpfung* (Weinen_2)

Eine Fehlerannotation erfolgt bei Auslassungen nur, wenn durch die Auslassung nach unserer Fehlerauffassung ein Fehler entsteht (vgl. Abschnitt 2.1).

4. falsche Konjunktion/Subjunktion:

- Menschen lachen an mich als ich nach ein Date bezahlen möchte – ZH: *wenn* (Feminismus_3)

- Da im kalten Wetter verschlechtert – ZH: Aber im kalten Wetter verschlechtert sich (Wandern_5)
- während IT ... blüht, es gibt immer größere Anspruch für Mechanikern und Informatikern – ZH: solange (Entlohnung_16)

5. falsche Kompositionsglieder:

- die Sichtbarkeitsumstände sind schlechter – ZH: die Sichtverhältnisse sind schlechter (Wandern_6)
- in einem Lebenslang zu weinen – ZH: im Leben zu weinen (Weinen_2)
- in der verschiedene Phase ihres Lebenslauf – ZH: in den verschiedenen Phasen ihres Lebens (Weinen_4)

6. sog. falsche Freunde:

- das passende Anziehen – ZH: das passende Outfit (Wandern_4)
- ein ordinäres, weibliches Baby – ZH: ein normales weibliches Baby (Weinen_6)
- reagieren wir auf alles zu sensitiv – ZH: reagieren wir auf alles zu sensibel (Weinen_9)

7. fremdsprachliche Elemente:

- equal – ZH: gleich viel (Feminismus_3)

Zur Abgrenzung von Wortbildung und Wortwahl: Ein Wortbildungsfehler (Wortb) liegt vor, wenn Stamm und Wortbildungsmorphem inkompatibel sind, was zu einem nicht existierenden Wortbildungsprodukt führt wie bei *Uniformalität* (ZH: *Uniformität*) oder *ambitioses* (ZH: *ambitiöses*). Demgegenüber wird ein unpassender Stamm mit der Fehlerkategorie Wortwahl (Lex) getaggt wie bei *Unterschätzung* (ZH: *Geringschätzung*).

Speziell bei Übersetzungstexten gilt der Richtwert: Das übersetzte Produkt zählt, nicht die Genauigkeit der Übersetzung. Folglich werden reine Übersetzungsfehler nicht getaggt.

Grundf: falsche Grundform

Eine falsche Grundform (Grundf) liegt vor, wenn von einem Lexem bei der intendierten Bedeutung eine nicht existierende oder unpassende Grundform gewählt wird.

- Karrier – ZH: *Karriere* (Feminismus_9)
- Schmink – ZH: *Schminke* (Feminismus_20)

- Planete – ZH: *Planet* (Sprache_10)

Wenn zwei ähnlich lautende Lexeme mit Bedeutungsunterschied miteinander verwechselt werden, wird jedoch Lex (und nicht Grundf) getaggt, z. B.:

- *wenn man mehrere Geld braucht* – ZH: *mehr Geld* (Kriminalität_13)
- *Wie zum Beispiel Modellen* ZH: *wie zum Beispiel Models* (Entlohnung_22)

Phr: Wendung

Zu dieser Fehlerkategorie gehören Fehler in Phraseologismen, Kollokationen oder idiomatischen Wendungen.

- stehen außer der Gesetze – ZH: *verstoßen gegen die Gesetze* (Kriminalität_2)
- Verbrechen machen – ZH: *Verbrechen begehen* (Kriminalität_2)
- in Vergleich zu – ZH: *im Vergleich zu* (Feminismus_10)

Bei Korrekturen, die Einzellexeme vs. phraseologische Einheiten betreffen, ist immer die Zielhypothese ausschlaggebend:

1. Wenn in der Zielhypothese an der Stelle eines Einzellexems des Lernertextes eine phraseologische Einheit steht, wird Phr getaggt:
 - *viele können den Hass oder Eifersucht nicht ertragen und behandeln* – ZH: *viele können den Hass oder die Eifersucht nicht ertragen und damit umgehen* (Kriminalität_4)
2. Wenn in der ZH an der Stelle einer phraseologischen Einheit ein Einzellexem steht, wird Lex getaggt:
 - *mit was für eine Sprache* – ZH: *in welcher Sprache* (Sprache_10)
 - *dass es Beispiele darauf sind* – ZH: *dass es Beispiele dafür gibt* (Entlohnung_30)

5.7.5 Semantische Fehler (FehlerSem-Ebene)

Zu den semantischen Fehlern werden Bedeutungsfehler bei kontextspezifischen grammatischen Bedeutungen und semantischen Relationen gezählt.

Def: (semantisch unpassende) Definitheit

Die (In-)Definitheit der Substantivgruppe wird in der Regel am Artikelwort (Artikel, Demonstrativpronomen, Possessivpronomen etc.) ausgedrückt. Definitheit drücken auch Verschmelzungen aus Präposition + Artikel aus, die im Unterschied zur Fügung aus Präposition + Artikel typischerweise eine unspezifische oder generische Lesart haben (vgl. *ich bin im Kino* vs. *ich bin in dem Kino neben der Uni*). Sowohl das

Fehlen als auch die falsche Setzung dieser Definitheitsindikatoren (bzw. Spezifizitätsindikatoren) werden als semantischer Fehler aufgefasst.

Mit Def werden folgende Fälle getaggt:

1. die Verwendung eines definiten Artikelworts statt eines indefiniten Artikelworts bzw. einer artikellosen Form (oder umgekehrt)
 - *in westlichen Ländern* – ZH: *in den westlichen Ländern* (Feminismus_1)
 - *ich interpretiere Feminismus als* – ZH: *den Feminismus* (Feminismus_3)
2. die Verwendung einer mit dem bestimmten Artikel verschmolzenen Präposition statt einer unverschmolzenen Präposition (oder umgekehrt), falls diese nicht rein grammatisch bedingt ist wie bei einem Genusfehler⁵
 - *Ich glaube im Ungarn haben die Frauen* – ZH: *in Ungarn* (Feminismus_12)
3. die Verwendung einer mit dem bestimmten Artikel verschmolzenen Präposition statt der Fügung Präposition + Artikel (oder umgekehrt).
 - *in dem Pubertätsalter* – ZH: *im Pubertätsalter* (Weinen_3)
 - *in dem Fall der Kleinkinder* – ZH: *im Fall der Kleinkinder* (Weinen_11)

Im Unterschied zur semantischen Fehlerkategorie Def ist die syntaktische Fehlerkategorie Det auf Fälle beschränkt, wo aus syntaktischen Gründen ein (anderer) Determinator bzw. kein Determinator erscheinen muss (vgl. 5.7.3) wie in:

- *viele Kriminalen bekommen Geldstrafe* – ZH: *bekommen eine Geldstrafe* (Kriminalität_3)

Da die Fehlerkategorie Def die Definitheit der gesamten Substantivgruppe betrifft, bildet diese immer den Skopus des Fehlertags, auch wenn die Abweichung nur an einem Determinator oder Adjektiv erkennbar ist.

Wenn gleichzeitig eine feste Verbindung vorliegt, wird der Fehler zusätzlich als Phr (Wendung) getaggt.

- *in Vergleich* – ZH: *im Vergleich* (Feminismus_10)
- *in Ausland* – ZH: *im Ausland* (Feminismus_20)

⁵ *Graffiti am Stadtwand* - ZH: *Graffiti an der Stadtwand* (Kriminalität_4)

Num: (semantisch unpassender) Numerus

Mit Num wird ein falscher Numerus von Substantiven oder Substantivgruppen getaggt, d. h. Singular statt Plural oder umgekehrt. Da die Fehlerkategorie Num jeweils die gesamte Substantivgruppe betrifft, bildet diese immer den Skopus des Fehlertags, auch wenn die Abweichung nur an einem Determinator oder Adjektiv erkennbar ist.

- *aus der USA* – ZH: *aus den USA* (Feminismus_4)

Sexus: falscher Sexus (unpassendes konzeptuelles Genus)

Mit Sexus werden Fehler von Substantiven in Sexusopposition in Bezug auf männliche/weibliche Referenz getaggt. Es geht dabei praktisch um Fehler, bei denen ein maskulines Substantiv für weibliche Referenzobjekte gebraucht wird. So dürfte im heutigen Deutsch eine Referenz auf eine weibliche Person mit einer maskulinen Personenbezeichnung im Singular unüblich sein, wenn eine movierte Form existiert:

- *sofort ist die Frau ein Feminist* – ZH: *eine Feministin* (Feminismus_3)

Ant: falsche Bezugnahme auf ein Antezedens

Mit Ant werden semantische Fehler bei der Bezugnahme auf ein Antezedens durch Pronomina getaggt. Die Anwendung des Tags umfasst insbesondere folgende Fälle:

1. die Verwendung einer pronominalen statt einer nominalen Ausdrucksform, wenn nicht eindeutig bestimmbar ist, auf welches Antezedens sich das Pronomen bezieht:
 - *Die Kulturen sind so verschieden, dass die Emanzipation nicht in allen Ländern möglich ist. Also, es ist eine interessante Frage, ob es eher geschadet oder genutzt.* – ZH: *ob der Feminismus eher geschadet oder genutzt* (Feminismus_9)
2. die Wahl eines unpassenden Pronomens, z. B. bei fehlerhafter Setzung eines anadeiktischen Pronomens statt eines Korrelats oder bei anderen Verwechslungen von *es* und *das*.⁶
 - *Das ist umstritten, dass diese Verträge für die Kunden vorteilhaft sind.* – ZH: *Es ist umstritten, ob ...* (Entlohnung_14)
 - *Im Kalte wird unsere Koordination deutlich schlechter, und es führt zu Unfällen.* – ZH: *In der Kälte wird unsere Koordination deutlich schlechter, und das führt zu Unfällen.* (Wandern_6)

⁶ Die ungrammatische Setzung bzw. das Fehlen des Korrelats wird demgegenüber mit KorKomp getaggt (s. 5.7.3).

- ... sie sollten schön singen und sich mit den Kindern beschäftigen. Es war die Grundkonzeption vor allem in England – ZH: Das (Feminismus_9)

Die fehlende Übereinstimmung zwischen Pronomen und Antezedens wird demgegenüber mit dem Tag KongrAnt gekennzeichnet, wenn sie syntaktisch erforderlich ist (s. 5.7.3).

Mod: (semantisch unpassender) Modus

Mit der Fehlerkategorie Mod werden Fehler im Modusgebrauch getaggt. Dies betrifft die Verwechslung der Verbalmodi Indikativ, Konjunktiv I, Konjunktiv II und Imperativ. Einen Sonderfall stellen indirekte Aufforderungen dar, bei denen in der ZH neben dem Moduswechsel auch ein Modalverb erforderlich ist (s. letztes Bsp.). Im Falle eines Verbalkomplexes wird Mod immer am Finitum getaggt.

- *Es scheint so, als ob die chinesische Ratte ein sehr neugieriges Tier ist* – ZH: *ein sehr neugieriges Tier wäre* (Indikativ statt Konjunktiv, China_1)
- *Wenn es aber ihm gesagt worden sei* – ZH: *Wenn es aber ihm gesagt worden wäre* (Konjunktiv I statt Konjunktiv II, Sprache_8)
- *dass es unsere Planete ist und gehen sie weg* – ZH: *dass es unser Planet ist und sie weggehen sollen* (Imperativ statt Indikativ, Sprache_10)

Temp: (semantisch unpassendes) Tempus

Mit Temp werden Zeitstufenfehler im Gebrauch der Tempora getaggt (also falscher Zeitbezug, nicht notwendigerweise grammatisch falsches Tempus). Abweichungen im Tempusgebrauch in derselben Zeitstufe (d. h. Varianz in den grammatischen Tempora) werden nicht als Fehler betrachtet (z. B. Perfekt statt Präteritum). Im Konjunktiv II wird die Verwechslung von Konjunktiv Präteritum mit Konjunktiv Plusquamperfekt ebenfalls mit Temp getaggt. Dementsprechend wird die Differenz Indikativ Präteritum vs. Konjunktiv Präteritum mit Temp (und Mod) getaggt. Im Falle eines Verbalkomplexes wird Temp immer am Finitum bzw. am Tempushilfsverb getaggt.

- *Demgegenüber kommt die Wirkung dieser Aktivisten ziemlich spät ins Ungarn.* – ZH: *kam* (Feminismus_7)
- *Der Grund dafür war, ...* – ZH: *ist* (Feminismus_20)
- *Die Forscher würden ausgelacht* – ZH: *die Forscher wären ausgelacht worden* (Sprache_10)

SemRe1: (lexikalische Realisierung einer) semantische(n) Relation

Mit SemRe1 werden semantisch bedingte Abweichungen getaggt. Hierzu gehören:

1. falsche Präpositionen in Ergänzungen, sofern die Präposition nicht vom Verb festgelegt wird (z. B. bei Lokaler Ergänzungen), sowie in Angaben (z. B. Temporalangaben):
 - *die Wanderung zwischen den Steinen* – ZH: *die Wanderung auf den Steinen* (Wandern_2)
 - *auf den Gebieten* – ZH: *in den Naturschutzgebieten* (Wandern_2)
 - *in kaltem Wetter* – ZH: *bei kaltem Wetter* (Wandern_4)
 - *Kanze wollte sich selbst englisch lernen* – ZH: *von sich selbst* (Sprache_1)
 - *vom Kühlschrank die Orange rausnehmen* – ZH: *aus dem Kühlschrank* (Sprache_1)
2. semantisch bedingte Kasusfehler bei Wechselpräpositionen:
 - *in einem anderen, weiten Land allein zu fahren* – ZH: *in ein anderes ...* (China_15)
 - *dass wir die einzige ... auf die Erde sind* – ZH: *auf der Erde* (Sprache_14)
 - *an einen speziellen Ort feiern werde* – ZH: *an einem speziellen Ort* (China_4)

GenV: Genus Verbi

Mit GenV werden falsche Verbmodi getaggt. Dies betrifft Fälle, in denen Aktiv statt Passiv verwendet wird (oder umgekehrt).

- *Unter der Geschichte der Menschheit sind sich 6000 Sprachen entwickelt werden* – ZH: *haben sich entwickelt* (Sprache_8)

Demgegenüber wird das falsche Auxiliar im Passiv (d. h. die Verwendung von Zustandspassiv statt Vorgangspassiv oder umgekehrt) mit Aux getaggt (vgl. 5.7.2).

- *Es ist gesagt* – ZH: *es wird gesagt* (Feminismus_13)

5.8 Abweichungen (ZHDiff-Ebene)

Abweichungen zwischen der ZH-Spur und der word-Spur werden mit den Abweichungstags INS, DEL, CHA, SPLIT, MERGE, MOVS bzw. MOV T markiert (vgl. Reznicek et al. 2012: 60 f., Nolda 2019):

- Verändertes Token (CHA)
- Fehlendes Token (INS)
- Überflüssiges Token (DEL)
- Geteiltes Token (SPLIT)
- Token mit anderer Position (MOVS/MOVT)
- Token zusammenführen (MERGE)

Bewegungen nach links sind präferiert (wie in Falko, s. Reznicek et al. 2012: 40), max. zwei Einzelwörter können aber auch nach rechts bewegt werden, wenn viele Tokens von der Bewegung betroffen sind.

Overview																
word	Die	Situation	ist			in	Ungarn	hat	in	den	letzten	Jahrzehnten	nicht	genug	geändert	.
ZH	Die	Situation		hat		in	Ungarn		in	den	letzten	Jahrzehnten	nicht	genug	geändert	.
FehlerSyn			Präd	StV												
ZH	Die	Situation		hat	sich	in	Ungarn		in	den	letzten	Jahrzehnten	nicht	genug	geändert	.
FehlerSyn				Refl												
Details																
word	Die	Situation	ist			in	Ungarn	hat	in	den	letzten	Jahrzehnten	nicht	genug	geändert	.
txt::S	s2															
txt::pos	ART	NN	VAFIN			APPR	NE	VAFIN	APPR	ART	ADJA	NN	PTKNEG	ADV	VVPP	\$.
txt::lemma	die	Situation	sein			in	Ungarn	haben	in	die	letzt	Jahrzehnt	nicht	genug	ändern	.
ZH1::ZH	Die	Situation		hat		in	Ungarn		in	den	letzten	Jahrzehnten	nicht	genug	geändert	.
ZH1::ZHDiff			DEL	MOVT				MOVS								
ZH1::ZHS	s2															
ZH1::ZHpos	ART	NN		VAFIN		APPR	NE		APPR	ART	ADJA	NN	PTKNEG	ADV	VVPP	\$.
ZH1::ZHlemma	die	Situation		haben		in	Ungarn		in	die	letzt	Jahrzehnt	nicht	genug	ändern	.
ZH1::FehlerSyn			Präd	StV												
ZH2::ZH	Die	Situation		hat	sich	in	Ungarn		in	den	letzten	Jahrzehnten	nicht	genug	geändert	.
ZH2::ZHDiff					INS											
ZH2::ZHpos					PRF											
ZH2::ZHlemma					sich											
ZH2::FehlerSyn					Refl											

Abb. 16: Bewegung nach links (Entlohnung_3)

Overview

word

Ihr

Leben

ist

nicht

immer

hoffnungslos

,

aber

fehlt

die

Ausbruchspunkt

aus

dieser

Situation

und

ZH

Ihr

Leben

ist

nicht

immer

hoffnungslos

,

aber

der

Ausbruchspunkt

aus

dieser

Situation

fehlt

und

FehlerSyn

STV

FehlerLex

Gen

ZH

Ihr

Leben

ist

nicht

immer

hoffnungslos

,

aber

die

Ausstiegsmöglichkeit

aus

dieser

Situation

fehlt

und

FehlerLex

Lex

Details

word

Ihr

Leben

ist

nicht

immer

hoffnungslos

,

aber

fehlt

die

Ausbruchspunkt

aus

dieser

Situation

und

txt::S

s10

txt::pos

PPOSAT

NN

VAFIN

PTKNEG

ADV

ADJD

\$,

KON

VVFIN

ART

NN

APPR

PDAT

NN

KON

txt::lemma

Ihr

Leben

sein

nicht

immer

hoffnungslos

,

aber

fehlen

die

Ausbruchspunkt

aus

dies

Situation

und

ZH1::ZH

Ihr

Leben

ist

nicht

immer

hoffnungslos

,

aber

der

Ausbruchspunkt

aus

dieser

Situation

fehlt

und

ZH1::ZHDiff

MOVS

CHA

MOVT

ZH1::ZHS

s10

ZH1::ZHpos

PPOSAT

NN

VAFIN

PTKNEG

ADV

ADJD

\$,

KON

ART

NN

APPR

PDAT

NN

VVFIN

KON

ZH1::ZHlemma

Ihr

Leben

sein

nicht

immer

hoffnungslos

,

aber

die

Ausbruchspunkt

aus

dies

Situation

fehlen

und

ZH1::FehlerSyn

STV

ZH1::FehlerLex

Gen

ZH2::ZH

Ihr

Leben

ist

nicht

immer

hoffnungslos

,

aber

die

Ausstiegsmöglichkeit

aus

dieser

Situation

fehlt

und

ZH2::ZHDiff

CHA

CHA

ZH2::ZHlemma

Ausstiegsmöglichkeit

ZH2::FehlerLex

Lex

Abb. 17: Bewegung nach rechts (Kriminalität_9)

Abweichungen, die vom Stylesheet [exb2exb-diff.xsl](#) nicht innerhalb einer Satzspanne automatisch auflösbar sind, werden mit den provisorischen Tags MOVS/DEL oder MOV/INS versehen und müssen manuell disambiguiert werden.

5.9 Inter-annotator agreement

Die Praktikabilität der Fehlerannotation wurde mit Hilfe einer Studie zum *inter-annotator agreement* unter Verwendung einer früheren Version dieses Handbuchs überprüft. Zwei geschulten Annotatorinnen (studentischen Hilfskräften) wurde ein Übersetzungstext vorgelegt, der bereits die zugeordneten Zielhypothesen enthält. Die Aufgabe der Annotatorinnen bestand darin, aufgrund des Vergleichs der word- und der ZH-Spuren unabhängig voneinander die Fehlerannotation zu erstellen. Die Stärke der Übereinstimmung ihrer Annotate wurde nach den folgenden Kriterien bestimmt:

- Wie stark unterscheidet sich die Anzahl der von den Annotatorinnen verwendeten Tags bei den zu berücksichtigenden Abweichungen zwischen dem Lernertext und der ZH?
- Wie stark unterscheiden sich die von den Annotatorinnen eingesetzten Fehlertags qualitativ?
- Wie stark unterscheidet sich der Skopus der von den Annotatorinnen eingesetzten Fehlertags?

Der so festgestellte prozentuale Übereinstimmungswert (Anzahl Übereinstimmungen / Gesamtzahl der Fälle) beträgt 87,1%.⁷

5.10 Workflow und Qualitätskontrolle

An der gesamten Erstellung des Korpus waren sieben Linguisten und Linguistinnen am Institut für Germanistik der Universität Szeged als Projektmitglieder beteiligt. Für die Digitalisierung und die Textannotation sowie teilweise für die Archivierungsarbeit wurden in jedem Projektjahr jeweils zwei bis drei Germanistikstudentinnen als Hilfskräfte eingestellt. Die Auswahl der Hilfskräfte erfolgte immer in einem Bewerbungsverfahren.

Die Erstellung des Dulko-Korpus hat sich aus folgenden teilweise nacheinander, teilweise parallel miteinander verlaufenden Arbeitsphasen zusammengesetzt:

1. Feststellung der Themen in den zu erhebenden Essays und Übersetzungen
2. Texterhebung, Metadatenerhebung und Archivierungsarbeit (inkl. Digitalisierung der erhobenen Materialien)
3. Auswahl der Lernertexte aus der erhobenen Menge für die Annotation
4. Annotation der ausgewählten Lernertexte
5. automatisierte Konvertierung der annotierten Texte in eine mit der Suchplattform ANNIS kompatible Version (s. Abschnitt 6)

Alle Lernertexte (inkl. Metadaten und Einverständniserklärung) wurden unter kontrollierten Umständen (vgl. Abschnitt 1) in Seminaren, teilweise aber auch in Prüfungen in einer handschriftlichen Form erhoben und dann digitalisiert. Aus der jeweils erhobenen Textmenge wurden im Zufallsverfahren Essays und Übersetzungen für deren Annotation ausgewählt.

Die Hilfskräfte haben die ausgewählten Lernertexte immer unter Anleitung und systematischer Kontrolle der sie jeweils betreuenden Projektmitglieder digitalisiert und annotiert. Bei Fragen bzw. Unstimmigkeiten bezüglich der Zuweisung der Fehlertags wurde das Projektteam zurate gezogen. Die Zielhypothesen wurden von zwei am Projekt beteiligten Linguisten lektoriert, die zugleich deutsche Muttersprachler sind.

Parallel mit der Annotation der Lernertexte erfolgte das Optimalisieren des Tagsets, da einerseits einige Fehlertypen erst nach der Bearbeitung einer größeren Textmenge aufgetaucht oder überhaupt gar nicht vorgekommen sind. Auf einige Fehlertags konnte somit verzichtet werden bzw. wurden bestimmte Fehlertags durch andere ersetzt. So z. B. wurde der Tag KongrDet, d.h. „Kongruenz mit dem

⁷ Die IAA-Analyse soll in einer späteren Projektpublikation ausführlicher beschrieben werden. Die Dokumentation kann auf Anfrage zugeschickt werden.

Determinator“ gelöscht, da sich diese Fehlerkategorie im Korpus als schwer abgrenzbar von Flexionsfehlern herausgestellt hat. Demzufolge wurde der Tag KongrDet durch den Tag Flex ersetzt. Andererseits wurden aber auch neue Fehlertags eingeführt, wie z. B. SemRel für semantisch bedingte Abweichungen (vgl. Beispiele dafür in Abschnitt 5.7.5).

Die letzte Phase der Bearbeitung der annotierten Lernertexte bestand in deren Konvertieren in eine Version, die eine Abfrage der Texte mit dem Suchwerkzeug ANNIS ermöglicht (s. Abschnitt 6).

6 Suchabfrage

Das Dulko-Korpus (Beeh et al. 2021) kann mit dem Suchwerkzeug [ANNIS](https://korpling.org/annis/) (Krause/Zeldes 2016) unter der folgenden URL kostenlos abgefragt werden: <https://korpling.org/annis/>. Es wird erwogen, es künftig auch über das am IDS entwickelte Suchwerkzeug KorAP (<https://korap.ids-mannheim.de>) verfügbar zu machen.

ANNIS ist ein browserbasiertes Suchwerkzeug für komplexe linguistische Mehrebenenkorpora, mit dem man sprachliche Phänomene unterschiedlicher Ebenen gemeinsam durchsuchen kann. ANNIS dient mittlerweile für eine Reihe von Korpora als Suchplattform, u. a. auch für die Lernerkorpora der Falko-Familie: FalkoEssayL2, FalkoGeorgetownL2, FalkoSummaryL2, FalkoWHIGL2, KanDeL, kobaltL2 usw. (<https://korpling.german.hu-berlin.de/falko-suche/>).

ANNIS hat eine eigene Suchabfragesprache AQL (ANNIS Query Language, s. <https://corpus-tools.org/annis/aql.html>).⁸ Sie ist vielen anderen Suchabfragesprachen ähnlich, indem sie auf Variable-Wert-Paaren und Relationen zwischen ihnen basiert. Die Recherche in Dulko funktioniert im Prinzip wie bei den Falko-Korpora. Dies soll anhand der einfachen Suche nach Wortformen mit Filtern nach Metadaten (Englisch als weitere Fremdsprache neben Deutsch) illustriert werden:

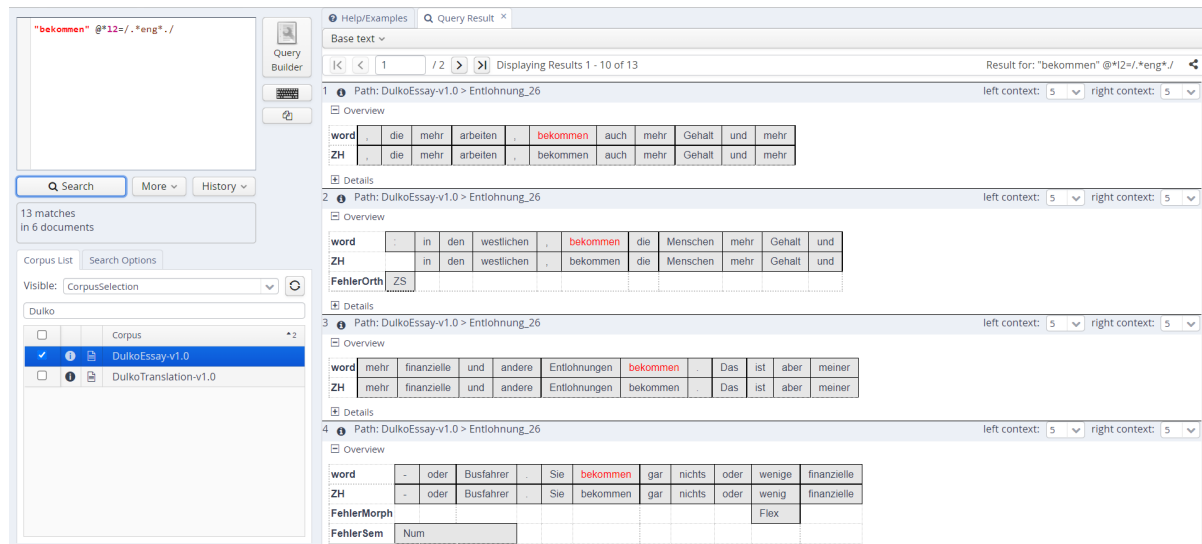


Abb. 18: ANNIS-Suchmaske von Dulko

Nach der Auswahl des Korpus (hier: DulkoEssay-v1.0, s. blaue Unterlegung links unten) soll im Suchbereich (oben links) die Suchabfrage angegeben werden. Nach

⁸ Die Suchabfragesprache AQL kann in mehreren Versionen verwendet werden. Der Suchabfrage für ANNIS4 "bekommen" @*12=/. *eng*./. in Abb. 18 entspricht die in der Tab. 6 angegebene Suchabfrage für ANNIS3: word="bekommen" & meta::12=/. *eng*./. Es ist empfehlenswert, unter "Search Options" den Wert "AQL (compatibility mode)" einzustellen, um beide Typen von Suchanfragen verwenden zu können.

einem Klick auf „Search“ bekommen wir im Ergebnisfenster (rechts) die als KWIC angezeigten Treffer, die exakt die in Anführungszeichen eingetragene Zeichenkette enthalten.

Mit ANNIS können Korpora mit unterschiedlichen Annotationskonventionen verarbeitet werden, je nach Korpus können auch korpuspezifische Ebenen durchsucht werden. Das Übersetzungsteilkorpus von Dulko enthält z. B. eine trans-Spur für die Angabe des ungarischen Ausgangstextes bei Übersetzungen (vgl. dazu Abschnitt 5). Im Folgenden sollen die wichtigsten Suchmöglichkeiten zusammengefasst werden.

In den einzelnen Ebenen kann nach verschiedenen Werten gesucht werden, die der jeweiligen Annotationsebene zugewiesen wurden (s. 6.1). Dabei kann nicht nur nach konkreten Werten, sondern auch nach regulären Ausdrücken gesucht werden (s. 6.2). Zudem ist auch die Suche nach der Beziehung zwischen einzelnen Variable-Wert-Paaren möglich (s. 6.3). Die Ergebnisse können nach Metadaten gefiltert werden (s. 6.4). Abschließend werden einige Beispiele für die Korpusabfrage aufgeführt, die den Einstieg in die Arbeit mit Dulko erleichtern sollen (s. 6.5).

6.1 Einfache Suchabfragen

Die allgemeine Form für die einfache Suche ist: `Variable="Wert"`. Bei der Suchabfrage können alle Annotationsebenen, die das Korpus enthält, als Variable angegeben werden. Ein Überblick findet sich im Korpus-Informationsfenster („Available Annotations“), s. dazu auch die Beschreibung der einzelnen Annotationsebenen in Abschnitt 5. In den einzelnen Ebenen kann nach verschiedenen Werten gesucht werden, die der jeweiligen Annotationsebene zugewiesen wurden, vgl. folgende Beispiele:

- `word="Anzahl"` (Suche nach Wortformen in den Lerneräußerungen, z. B. *Anzahl*)
- `lemma="beginnen"` (Suche nach einem Lemma, z. B. *beginnen*)
- `pos="ADJA"` (Suche nach Wortarten (POS), z. B. nach attributiven Adjektiven wie *schönen*)
- `ZH="Anteil"` (Suche nach Wortformen auf der Zielhypothesenebene, z. B. *Anteil*)
- `ZHDiff="INS"` (Suche nach *edit tags*, z. B. nach einer Einfügung auf der Zielhypothesenebene)

6.2 Suche nach regulären Ausdrücken

Die allgemeine Form für die Suche nach regulären Ausdrücken ist: Variable=/Wert/. Beispiele für die Suche nach regulären Ausdrücken sind:

- word=/de./ (Suche nach *der/den/dem/des* etc.),
- lemma=/beg.*/ (Suche nach allen Lemmata, die mit *beg-* beginnen)
- word=/ein.+/ (Suche nach *eine/einen/einem/einer/eines* etc.)
- word=/dass?/ (Suche nach *das/dass*)
- word=/(H|h)eute/ oder word=/[Hh]eute/ (Suche nach *heute/Heute*)

6.3 Ebenenübergreifende Suchabfragen

Bei komplexen Suchabfragen können mehrere Variable-Wert-Paare aus unterschiedlichen Annotationsebenen miteinander in Beziehung gesetzt werden. Zwischen den Paaren muss immer eine Beziehung hergestellt werden (z. B. mit = für Abdeckung, wenn die beiden Variable-Wert-Paare auf dieselbe Spalte/Tokens referieren). Auf die einzelnen Variable-Wert-Paare bezieht man sich mit # der Reihe nach, die Variable-Wert-Paare werden mit & verbunden. Sucht man zum Beispiel nach dem Lemma *zu* auf der Zielhypothesenebene (in unserem Beispiel #1), das auf der Wortartenebene als APPR (Präposition) getaggt wurde (in unserem Beispiel #2), ist die Suchabfrage ZHlemma="zu" & ZHpos="APPR" & #1=_#2 zu empfehlen. Im Folgenden finden sich die wichtigsten Beziehungen im Überblick sowie Beispiele für ebenenübergreifende Suchabfragen:

- #1.#2 (#1 direkt gefolgt von #2)
- #1.*#2 (#1 indirekt gefolgt von #2)
- #1=_#2 (#1 und #2 referieren auf dieselbe Spalte dasselbe Token)
- #1_i_#2 (#1 enthält #2)

Beispiele:

- ZHpos="APPR" & ZHlemma="d" & #1.#2 (Präposition, direkt gefolgt von einem bestimmten Artikel)
- ZHpos="VAFIN" & ZHpos="VVINF" & #1.*#2 (finites Auxiliärverb vor einem infiniten Vollverb)
- ZHlemma="zu" & ZHpos="APPR" & #1=_#2 (*zu* als Präposition)
- FehlerLex="Gen" & ZHpos="ADJA" & #1_i_#2 (attributive Adjektive, die im Skopus eines Genusfehlers enthalten sind)

6.4 Filterung nach Metadaten

Die Ergebnisse können nach Metadaten gefiltert werden. Es können alle erhobenen text- bzw. lernerbezogenen Metadaten durchsucht werden. Die allgemeine Form für die Suche nach Metadaten ist: `meta::Variable="Wert"`. Möchte man etwa die Tokenanzahl in einem Subkorpus bestimmen, kann man die folgende Suchabfrage formulieren:

- `tok & meta::learner_level_CEFR_conversion="B2"` (alle Tokens im Subkorpus zum Sprachstand B2 nach dem GER).

Man kann Suchausdrücke aus verschiedenen Annotationsebenen mit der Suche nach Metadaten kombinieren, vgl. die folgenden Beispiele:

- `FehlerLex="Gen" & meta::learner_level_CEFR_conversion="B2"` (alle Genusfehler im Subkorpus zum Sprachstand „B2“ nach dem GER)
- `FehlerSyn=/St.*/ & meta::L2_other_1="eng"` (alle Wortstellungsfehler in allen Korpus-texten, die von Versuchspersonen mit Englisch als weiterer Fremdsprache verfasst wurden)
- `FehlerSyn=/St.*/ & meta::learner_ID="5B38D8B519C31FBE372E6F758E99FECE"` (alle Wortstellungsfehler in allen Korpus-texten, die von der Versuchsperson mit der Learner_ID 5B38D8B519C31FBE372E6F758E99FECE verfasst wurden)

6.5 Beispiele für Korpusabfragen

Sucheingabe	Bemerkung	Beispieltreffer
<code>word="hat"</code>	exakte Wortform in den Lerneräußerungen	<i>hat</i>
<code>lemma="haben"</code>	lemmabasierte Suche in den Lerneräußerungen	<i>hat, Hat, haben, Haben, hätte, Hätte ...</i>
<code>word=/[Zz]um/</code>	Wortform mit Klein- oder Großschreibung	<i>Zum, zum</i>
<code>word=/[Zz]um/ & word="Beispiel" & #1.#2</code>	Wortform mit Klein- oder Großschreibung steht direkt vor einer anderen Wortform	<i>Zum Beispiel, zum Beispiel</i>

Tab. 6: Beispiele für Korpusabfragen (Teil 1/3)

Sucheingabe	Bemerkung	Beispieltreffer
word!=[Zz]um/ & word="Beispiel" & #1.#2	negierte Suche nach einer Wortform mit Klein- oder Großschreibung, direkt vor einer anderen Wortform	<i>als Beispiel, ein gutes Beispiel, die beste Beispiel ...</i>
word="ob" & trans=/.*hogy.* / & #2_i_#1	Wortform in Übersetzungstexten, die im einschlägigen Satz des ungarischen Ausgangstextes das Wort <i>hogy</i> enthalten	<i>ob</i>
lemma=/beg.* /	alle Lemmata, die mit <i>beg-</i> beginnen	<i>beginnen, beginnt, Beginn, begreifen, begeht ...</i>
word="bekommen" & meta::l2=/.*eng.* /	einfache Suche mit Filtern nach Metadaten (L2 Englisch)	<i>bekommen</i>
lemma="die" & meta::learner_ID="8DA38A136C95FF543F2C96E7582BE8FC"	lemmabasierte Suche in allen Texten von einer Testperson	<i>Der, der, Die, die, Das, das, Den, den, Dem, dem, Des, des</i>
ZH="hat"	exakte Wortform auf der Zielhypothesenebene	<i>hat</i>
ZHlemma="eine"	lemmabasierte Suche auf der Zielhypothesenebene	<i>ein, Ein, eine, Eine, einen, Einen ...</i>
pos="KOUS"	Abfrage nach Wortart	<i>als, dass, weil ...</i>
pos="ART" . pos="ADJA" . pos="NN"	Suche nach direkter Abfolge bestimmter Wortarten	<i>ein alter Mann, den hohen Anzahl, ein gutes Beispiel ...</i>
pos="KOUS" & word!="dass" & #1=_#2	Abfrage nach Wortart und negierte Suche nach einer Wortform bei demselben Token	<i>als, weil, wenn ...</i>

Tab. 6: Beispiele für Korpusabfragen (Teil 2/3)

Sucheingabe	Bemerkung	Beispieltreffer
ZHDiff="CHA"	Suche nach einem <i>edit tag</i> (Änderung auf der ZHDiff-Ebene)	<i>ein → eine</i>
FehlerLex="Gen"	Fehlertyp (lexikalische Ebene, Genusfehler)	<i>ein neues Weg, die beste Beispiel ...</i>
FehlerLex="Gen" & lemma="eine" & #1_i_#2	Lemma im Skopus eines Fehlertags	<i>ein neues Weg, eine Job ...</i>
FehlerSyn="StV" & pos="KOUS" & S=/s.*/ & #3_i_#1 & #3_i_#2 & #2.*#1	Verbstellungsfehler und unterordnende Konjunktion in demselben Satz	<i>Wenn man wird arbeitslos ...</i>
FehlerSyn=/St.*/	alle Fehlertypen aus dem Bereich Wortstellung (syntaktische Ebene)	<i>viele solche Traditionen auch</i>
FehlerSyn=/Kor.*/ & ZHDiff="DEL" & #1_=_#2	alle Korrelatfehler, bei denen das Korrelat auf der ZH-Ebene gelöscht wurde	<i>Viele denken es noch, dass → Viele denken noch, dass</i>
FehlerMorph=/.*/	alle Fehlertypen auf einer Fehlerebene (morphologische Ebene)	<i>besser Jobs, diese Traditione, erhöht ...</i>
FehlerMorph=/.*/ & pos="NN" & #1_i_#2	alle Fehler auf einer Fehlerebene bei einer Wortart (morphologische Ebene, normales Nomen)	<i>Traditione, Ungefriedenheit ...</i>
FehlerSyn="ValV" & pos="APPR" & ZHDiff="CHA" & #1_i_#2 & #2_=_#3	alle Verbvalenzfehler, bei denen auf der ZH-Ebene eine Präposition geändert wird	<i>in einem juristischen Prozess teilnehmen ...</i>

Tab. 6: Beispiele für Korpusabfragen (Teil 3/3)

Literatur

Korpus

Beeh, Christoph/Drewnowska-Vargáné, Ewa/Kappel, Péter/Modrián-Horváth, Bernadett/Nolda, Andreas/Rauzs, Orsolya/Scheibl, György (2021): Dulko: Deutsch-ungarisches Lernerkorpus. Version 1.0. Szeged: Universität Szeged, Institut für Germanistik.
<https://korpling.org/annis/> [17. 11. 2021].

Referenzwerke

Dudenredaktion (2017): Duden – Die deutsche Rechtschreibung. 27., völlig neu bearbeitete und erweiterte Auflage. Berlin: Dudenverlag (Der Duden in 12 Bänden, Bd. 1).

Hennig, Mathilde (Hrsg.) (2016): Das Wörterbuch der sprachlichen Zweifelsfälle. Richtiges und gutes Deutsch. 8., vollständig überarbeitete Auflage. Berlin: Dudenverlag (Duden. Bd. 9).

Institut für Deutsche Sprache: "Wörterbuch zur Verbvalenz". Grammatisches Informationssystem grammis. DOI: 10.14618/evalbu.
<https://grammis.ids-mannheim.de/verbvalenz> [28. 2. 2021].

Rat für deutsche Rechtschreibung: Deutsche Rechtschreibung. Regeln und Wörterverzeichnis. Mannheim, 2018.
<https://www.rechtschreibrat.com/regeln-und-woerterverzeichnis/> [28. 2. 2021].

Schumacher, Helmut/Kubczak, Jacqueline/Schmidt, Renate/de Ruiter, Vera (2004): VALBU – Valenzwörterbuch deutscher Verben. Tübingen: Narr (Studien zur Deutschen Sprache, 31).

Wöllstein, Angelika/Dudenredaktion (Hrsg.) (2016): Duden. Die Grammatik. Unentbehrlich für richtiges Deutsch. 9., vollständig überarbeitete und aktualisierte Auflage. Berlin: Dudenverlag (Der Duden in 12 Bänden, Bd. 4).

Sekundärliteratur

Datenschutz-Grundverordnung (DSGVO) (2016): Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG.
<http://data.europa.eu/eli/reg/2016/679/oj> [28. 2. 2021].

Fekete, Olga (2016): Komplexität und Grammatikalität in der Lernaltersprache: eine Längsschnittstudie zur Entwicklung von Deutschkenntnissen ungarischer Muttersprachler. Münster, New York: Waxmann (Mehrsprachigkeit, 42).

Granger, Sylviane/Paquot, Magali (2017): Core metadata for learner corpora. Draft 1.0. Manuscript. Louvain-la-Neuve: Université catholique de Louvain.

- Hirschmann, Hagen/Nolda, Andreas (2019): Dulko – auf dem Weg zu einem deutsch-ungarischen Lernerkorpus. In: Eichinger, Ludwig/Plewnia, Albrecht (Hrsg.): Neues vom heutigen Deutsch: Empirisch – methodisch – theoretisch (Institut für Deutsche Sprache, Jahrbuch 2018). Berlin: de Gruyter, 339–342.
- Krause, Thomas/Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: Digital Scholarship in the Humanities 31, 118–139.
<http://dsh.oxfordjournals.org/content/31/1/118> [28. 2. 2021].
- Lüdeling, Anke/Hirschmann, Hagen (2015): Error annotation systems. In: Granger, Sylviane/Gilquin, Gaëtanelle/Meunier, Fanny (Hrsg.): The Cambridge Handbook of Learner Corpus Research, Cambridge: Cambridge University Press, 135–157.
- Nolda, Andreas (2019): Annotation von Lernerdaten mit EXMARaLDA (Dulko). Manuskript, Berlin: Berlin-Brandenburgische Akademie der Wissenschaften.
https://andreas.nolda.org/publications/nolda_2019_annotation_lernerdaten.pdf [28. 2. 2021].
- Reznicek, Marc/Lüdeling, Anke/Krummes, Cedric/Schwantuschke, Franziska/Walter, Maik/Schmidt, Karin/Hirschmann, Hagen/Andreas, Torsten (2012): Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. Berlin: Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik – Korpuslinguistik.
<https://hu.berlin/falko-handbuch> [28. 2. 2021].
- Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Manuskript. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung und Universität Tübingen, Seminar für Sprachwissenschaft.
<http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf> [28. 2. 2021].
- Schmidt, Thomas (2004): EXMARaLDA – ein Modellierungs- und Visualisierungsverfahren für die computergestützte Transkription gesprochener Sprache. In: Buchberger, Ernst (Hrsg.): Proceedings of Konvens 2004, Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence, 5. Wien: Österreichische Gesellschaft für Artificial Intelligence.
https://www.exmaralda.org/files/Konvens_Paper.pdf [28. 2. 2021].
- Selinker, Larry (1972): Interlanguage. In: International Review of Applied Linguistics in Language Teaching 10:3, 209–231.