Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 2, Nummer 2 (Dezember 2022)
ISSN 2749-4411

FALKO
Eine Familie vielseitig annotierter Lernerkorpora des Deutschen als Fremdsprache

Hagen Hirschmann, Anke Lüdeling, Anna Shadrova, Dominique Bobeck, Martin Klotz, Roodabeh Akbari, Sarah Schneider, Shujun Wan
Humboldt-Universität zu Berlin

Abstract

Falko ist ein frei zugängliches Lernerkorpus des schriftsprachlichen Deutschen als Fremdsprache und umfasst nach jahrelanger Erschließung neuer Textressourcen und der Anreicherung mit diversen Annotationsebenen eine Reihe einzelner Korpora, die teilweise sehr komplex strukturiert sind. Im vorliegenden Beitrag stellen wir die komplexeste Datenressource aus der Reihe dieser Korpora vor – das *Falko-Essay*-Korpus, welches aktuell in einer neuen Version (3.0) erscheint und interessierten Forscherinnen und Forschern frei zur Verfügung steht.

Keywords: Lernerkorpora; Fremdspracherwerbsforschung; Mehrebenenkorpora; Korpusarchitektur

Abstract

Falko is a freely available learner corpus with written learner texts of German as a foreign language. After years of data acquisition and data processing on multiple layers of annotation, *Falko* consists of several different corpora with partly a very complex architecture. In this article, we want to introduce the most complex data resource among these corpora – the *Falko* essay corpus. It is currently released in a new version (3.0) and can be used openly and freely by researchers that are interested in this resource.

Keywords: learner corpus; second language acquisition; multi-layer corpora; corpus architecture

1. Einleitung

Das *Falko*-Korpus[1] mittlerweile seit fast zwei Jahrzehnten erstellt und weiterentwickelt. Heutzutage handelt es sich bei *Falko* um eine Reihe unterschiedlicher Korpora mit einem vergleichbaren Kernaufbau, die schriftliche Texte von Lernenden des Deutschen als Fremdsprache mithilfe von computerbasierten Analysemethoden qualitativ-quantitativ analysierbar machen. Den Kern der Familie bilden das Falko-Essay-Korpus mit 248 Lernertexten von DaF-Lernenden mit multiplen L1 und 95 muttersprachlichen Vergleichstexten und einigen architektonisch daran angelehnten Ressourcen. *Falko* besteht also aus einer Reihe von Korpora, die verschiedene Genres, Lernergruppen usw. abbilden. Die einzelnen Ressourcen werden auf der Webseite https://hu-berlin.de/falko-familie zusammenfasst. In Lüdeling et al. (2008) werden bereits wesentliche Designentscheidungen bei der Erstellung dieser Korpora vorgestellt und Anwendungsmöglichkeiten aufgezeigt. Im vorliegenden Beitrag präsentieren wir vor allem die Weiterentwicklung des Kernkorpus *Falko-Essay* seit dem Jahr 2008 und somit dessen aktuelle Version (3.0). Dabei möchten wir aufzeigen, wie nützlich versionierte Mehrebenenkorpora für eine sukzessive Erweiterung sind.

[1] https://hu-berlin.de/falko (28.10.2022).

2. Das Kerngerüst: Markierung von Abweichungen durch Zielhypothesen

Mit dem Aufkommen der großen internationalen Lernerkorpora wie dem *ICLE*-Korpus[2] kam das methodologische Ziel auf, in Textdaten von Lernenden sowohl die zielsprachlichen Strukturen als auch die nicht zielsprachlichen Strukturen analysierbar zu machen (vgl. Granger 2008). Erstere und Letztere müssen also, wenn sie in fortlaufenden Textdaten gemeinsam auftreten, jeweils identifizierbar gemacht werden. Ein grundlegendes Konzept von *Falko* ist gerade zu diesem Zweck die Erstellung von Zielhypothesen (vgl. Lüdeling 2008): Im Falle von grammatischen (und je nach Korpus ggf. auch stilistischen) Abweichungen wird eine explizite Zielform formuliert. Hierdurch wird jede weitergehende Kategorisierung der Abweichungen[3] transparent. Bei der Annotation sprachlicher Abweichungen fungiert die Angabe einer Zielform (die dem Konzept von Abweichungen bzw. Fehlern inhärent ist) als Vorverarbeitungsschritt für die Vergabe von Abweichungskategorien und gleichzeitig als Instrument des Nachvollziehbarmachens von Kategorisierungsentscheidungen bei späterer Nutzung der Korpusdaten. Ein weiterer wichtiger Beweggrund für die Annotation von Zielhypothesen entsteht bei dem Anspruch, die Wörter und Sätze im Korpus mit standardgrammatischen Analysen zu versehen (indem z.B. Wortarten und Satzkonstituentenfunktionen zugewiesen werden – s.u.). Solche Analysen sind in der Regel nur auf zielgrammatische Strukturen anwendbar, nicht aber auf grammatische Abweichungen. Vgl. zur weiteren Erläuterung Korpusbeleg (1) aus dem *Falko-Essay*-Korpus.

(1) *Mann muss sich mit diesen Theorien umgehen können* (...)

Der Korpusbeleg (1) ist eine authentische Lerneräußerung aus dem *Falko-Essay*-Korpus (Text cbs001_2006_09, URL zum Beleg: https://bit.ly/3Sks68h). Auch wenn sich viele Linguistinnen und Linguisten einig wären, dass das erste Wort in (1) eine orthographische Abweichung bei dem Indefinitpronomen *man* aufweist und das Verb *umgehen* mit dem nicht zielsprachlichen Reflexivpronomen *sich* realisiert wurde, lassen beide Abweichungen Spielraum für andere Interpretationen: Bei dem satzinitialen *Mann* könnte es sich auch um das Nomen handeln, welches ohne Artikel realisiert ist. Bei *sich umgehen* könnte es sich auch um einen lexikalischen Irrtum auf Seiten des Vollverbs handeln (*Man muss sich mit diesen Theorien abfinden können* ist grammatisch zielsprachlich). Je nach Annahme, was die ‚eigentlich intendierte' Form ist, ergibt sich also eine andere Beschreibung der Abweichung (und umgekehrt). Die Bearbeitung der in (1) gezeigten Lerneräußerung ist im *Falko-Essay*-Korpus folgendermaßen nach ihrer Zielsprachlichkeit bearbeitet.

word	Mann	muss	sich	mit	diesen	Theorien	umgehen	können
ZH1	Man	muss		mit	diesen	Theorien	umgehen	können
ZH1Diff	CHA		DEL					

Abbildung 1
Bearbeitung der in (1) aufgeführten Lerneräußerung auf der Ebene der ersten Zielhypothese (ZH1) und Kennzeichnung der Abweichung zwischen Lerneräußerung und Zielhypothese.

Während die Ebene ZH1 eine Interpretation der Lerneräußerung (dargestellt auf der Ebene *word*) festlegt, zeigt die Ebene ZH1Diff die systematischen Unterschiede (Editierdistanzen) zwischen der Lerneräußerung und der Ebene ZH1 (Änderungen von Wortformen werden durch das Kürzel CHA

[2] https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html (28.10.2022).
[3] Gemeint ist eine Kategorisierung der Abweichungen von der Zielvarietät, die z.B. nach Granger (2008: 268) „error analysis" genannt wird. Für eine Einordnung und Beschreibung verschiedener Klassifikationen von Abweichungskategorien bzw. „Fehler-Tagsets" vgl. Lüdeling / Hirschmann (2015).

für *change*, Löschungen von Wortformen durch DEL für *delete*, Einfügungen durch INS für *insert* und Verschiebungen durch MOV für *move* dargestellt). Durch diese Verarbeitung von Abweichungen (und ggf. unter Hinzunahme weiterer Beschreibungsebenen) können viele Typen von Abweichungen systematisch gefunden werden. Sehr spezifische bzw. grammatisch weitergehend interpretierte Abweichungstypen (z.B. abweichende Flexion bei Adjektiven o. Ä.) können mittels der aktuellen Annotationen vorgefiltert werden, müssen dann jedoch händisch weiterbearbeitet werden.

Die Zielhypothese könnte als Fehlerkorrekturebene verstanden werden, doch dieses Verständnis greift zu kurz und führt zu einer unpräzisen Vorstellung des Zwecks der Zielhypothese. Dieser ist, Abweichungen nach einheitlichen Kriterien zu behandeln und strukturell erfassen zu können. Ob durch diese Annotation eine Fehlerkorrektur erfolgt, wie sie im schulischen Kontext üblich ist, ist zweitrangig. Die Erstellung der Zielhypothesen ist genau genommen ein Normalisierungsschritt (zum Konzept der Normalisierung von Korpusdaten vgl. auch Hirschmann 2019: 29-30). Eines von vielen Beispielen für diesen Unterschied ist das im *Falko*-Korpus von einem Lernenden verwendete Wort *Arbeitslosigkeitszählen* (Text fk001_2006_08). Es würde im Lehr- und Lernkontext ggf. zu *Arbeitslosenzahlen* korrigiert werden, wird auf der Ebene der grammatischen Zielhypothese (ZH1) gemäß den *Falko*-Richtlinien[4] normalisiert zur morphologisch gesehen grammatischen Form *Arbeitslosigkeitszahlen*. Das *Falko-Essay*-Korpus verfügt über eine zweite Zielhypothesenebene, auf der stilistisch-lexikalische Abweichungen angezeigt werden und im konkreten Fall zu *Arbeitslosenzahlen* normalisiert wird.

3. Grammatische Standardanalysen

Um den Ansprüchen der Lernerkorpusmethodologie gerecht zu werden, benötigen Lernerkorpora neben der Kennzeichnung von Strukturen, die nicht zielsprachlich sind, grammatische Kategorisierungen, die auf die Zielsprache selbst zugeschnitten sind. Im Folgenden stellen wir die im *Falko-Essay*-Korpus verfügbaren standardgrammatischen Annotationen vor und möchten deren Relevanz für bestimmte Forschungsziele aufzeigen.

3.1 Wortarten und Lemmata

Die Einteilung aller denkbarer Wörter einer Sprache in eine definierte Menge an Kategorien ist eine der wichtigsten Grundlagen der beschreibenden Grammatik. Für das Deutsche existieren diverse Vorschläge für diese Kategorisierung, mit z.T. unterschiedlicher Zielstellung. Das Stuttgart-Tübingen-Tagset (STTS)[5] ist ein System von rund 50 hierarchisch organisierten Wortartenkürzeln, die von vielen automatischen Wortartanalyseprogrammen (Taggern) ausgegeben werden und mit denen die meisten Korpora des Deutschen analysiert sind. Durch diese Verarbeitung lassen sich polyfunktionale Wörter disambiguieren oder komplexe Syntagmen definieren und systematisch auffinden. Die Erfahrung zeigt, dass dieser Verarbeitungsschritt praktisch in jeder Korpusstudie hilfreich ist. Mit Blick auf den lernersprachlichen Beleg (1) ist zu erwähnen, dass die Form *Mann* für das Pronomen man zwangsläufig zu einem Konflikt führt; je nachdem, ob man die Form oder die Funktion bewertet, ist die Form *Mann* entweder als Nomen (STTS-Tag: NN) oder als Indefinitpronomen (STTS-Tag: PIS) zu kategorisieren. Dieses Problem gilt bei vielen nicht normgerechten Schreibungen. Erst auf der

[4] Reznicek et al. (2012), beziehbar unter https://hu-berlin.de/falko (28.10.2022).
[5] Detailliert beschrieben in Schiller et al. (1999), eine Zusammenfassung der Tags ist verfügbar unter https://bit.ly/3SDfAjL (28.10.2022).

Annotationsebene der Zielhypothese können gemäß der dort erfolgten Normalisierungen Werte vergeben werden, die nicht auf die beschriebene Weise zwischen verschiedenen Beschreibungsebenen konfligieren. Zur Beschreibung dieses konzeptionellen Problems vgl. auch Díaz-Negrillo et al. (2010). Die Annotationsrichtlinien in den *Falko*-Korpora sehen vor, dass in Fällen wie (1) Wortformen der Lerneräußerungen, die auf der Ebene der Zielhypothesen normalisiert werden, zunächst maßgeblich anhand ihrer Form kategorisiert werden.

Die Annotation von Lemmata (die Zuweisung von Grundformen zu allen Wortformen) dient vor allem der einheitlichen Abbildung von Lexemen, die im Deutschen aufgrund von Flexion nicht einheitlich repräsentiert werden. Für diesen Verarbeitungsschritt gilt dasselbe wie bei der Zuweisung von Wortarten: Bei orthographisch (bzw. morphologisch, lexikalisch usw.) nicht zielsprachlichen Formen werden Grundformen zunächst nicht hinsichtlich einer Zielform analysiert, weil dies gesondert auf der Ebene der Zielhypothesen erfolgt, sondern hinsichtlich ihrer Oberflächenform. Exemplarisch zeigen wir das Ergebnis der Wortarten- und Lemmaanalyse von (1) in Abb. 2.

word	Mann	muss	sich	mit	diesen	Theorien	umgehen	können
pos	NN	VMFIN	PRF	APPR	PDAT	NN	VVINF	VMINF
lemma	Mann	müssen	er|es|sie|Sie	mit	dies	Theorie	umgehen	können
ZH1	Man	muss		mit	diesen	Theorien	umgehen	können
ZH1pos	PIS	VMFIN		APPR	PDAT	NN	VVINF	VMINF
ZH1lemma	man	müssen		mit	dies	Theorie	umgehen	können

Abbildung 2
Bearbeitung der in (1) aufgeführten Lerneräußerung nach Wortarten und Lemmata.

Durch die in Abb. 1 und Abb. 2 ersichtlichen Analysen der Lernertexte in *Falko* können bereits einschlägige spracherwerbsbedingte Phänomene korpusbasiert untersucht werden, indem wesentliche Informationen zum grammatischen Status von Wörtern und Wortsequenzen definiert und beliebig miteinander verknüpft werden können.

3.2 Satzkonstituenten (Wortgruppen), topologische Bereiche (Felder und Klammern) und grammatische Bezüge (Dependenzen)

Im *Falko-Essay*-Korpus (sowie dem sehr ähnlichen Schwesterkorpus *Kobalt-DaF*) sind weitergehende syntaktische Kategorisierungen verfügbar. Sie gelten nur für die ZH1-Ebene und beziehen sich somit entweder auf zielsprachliche Äußerungen der Lernenden oder auf die Zielstrukturen, die auf der ZH1-Ebene explizit gemacht wurden. Bei Auswertungen der nachfolgend dargestellten Annotationen kann und muss also ggf. ein Bezug zwischen der originalen Lerneräußerung und der ZH1-Ebene hergestellt werden.

Die Annotation von Satzkonstituenten erfolgt über Spannen, die sich im Wesentlichen über nominale (Kategorie: NX), präpositionale (Kategorie: PX), oder adjektivische (Kategorie: ADJX)

Gruppen sowie Adverbgruppen (Kategorie: ADVX) erstrecken.[6] Die Anwendung des entsprechenden Annotationsschemas[7] führt bei dem Beispielbeleg (1) zu der in Abb. 3 gezeigten Analyse.

word	Mann	muss	sich	mit	diesen	Theorien	umgehen	können
ZH1	Man	muss		mit	diesen	Theorien	umgehen	können
ZH1Const	NX	VXFIN		PX			VXINF	VXINF
ZH1Const					NX			

Abbildung 3
Bearbeitung der in (1) aufgeführten Lerneräußerung nach Konstituenten.

Eine Verarbeitung wie diese gibt Aufschluss über syntaktische Verschachtelungen (je nach Verschachtelungstiefe bietet die Annotation bis zu sieben Konstituentenspannen) und somit die Komplexität von Strukturen. Phrasenbezogene Phänomene können mit dieser Verarbeitung außerdem gut definiert werden.

Als beispielhaften Anwendungsfall kann man nach präpositionalen Gruppen suchen, an deren linken Rand keine Präposition, sondern eine als *ADVX* ausgewiesene Einheit steht. Man erhält dadurch Präpositionalgruppen mit Fokuspartikeln wie (…) *erst nach den Fahrstunden* (…) (Text cbs003_2006_09). Die entsprechende Suchanfrage mit ihren Ergebnissen kann über diesen Link nachvollzogen werden: https://hu.berlin/falko-suche-const.

Die Aufteilung von Sätzen in die topologischen Bereiche Vorfeld (VF), Mittelfeld (MF) und Nachfeld (NF) sowie die Klammerbereiche der linken Satzklammer (LK) sowie der rechten Satzklammer (VC für Verbkomplex) ist annotiert, wie in Abb. 4 gezeigt.

word	Mann	muss	sich	mit	diesen	Theorien	umgehen	können
ZH1	Man	muss		mit	diesen	Theorien	umgehen	können
ZH1TopoFields	VF	LK		MF			VC	

Abbildung 4
Bearbeitung der in (1) aufgeführten Lerneräußerung nach topologischen Bereichen.

Ähnlich wie bei der vorangegangenen Wortgruppenanalyse der Fall, können durch diese Verarbeitung Phänomene an bestimmten Stellungsbereichen definiert werden. Will man bspw. Adverbien finden, die alleine das Vorfeld abdecken, kann man dies mit dieser Anfrage bewerkstelligen: https://hu.berlin/falko-suche-topo1.

Zuletzt wollen wir auf eine für viele Studien sehr nützliche Annotation der grammatischen Beziehungen zwischen Wörtern – Dependenzannotationen bzw. -parses – eingehen, die ebenfalls in *Falko*-Daten verfügbar sind. Dieser Annotationstyp ist der einzige, der nicht in Form von Zellen oder Spannen in tabellarischen Strukturen dargestellt werden kann, weil durch Dependenzparses beliebige gerichtete Beziehungen zwischen fortlaufenden Wörtern im Satz möglich gemacht werden müssen (was in diesem Format nicht darstellbar ist). Bei den zu vergebenden Beziehungswerten handelt sich

[6] Etwas redundanterweise werden Verben auf dieser Ebene auch markiert, flexionsmorphologisch spezifiziert und ggf. mit dazugehörigen Partikeln wie zu zusammengezogen.
[7] Es handelt sich bei den hier vorgestellten Konstituenten- und Stellungsfelderanalysen um eine Anwendung der im Annotationsschema der *Tübingen Treebank of Written German* (TüBa-D/Z) (vgl. Telljohann et al. 2017, https://bit.ly/2TQpGDX, 28.10.2022) durch eine Programmanwendung, die als *Topoparser* bezeichnet wird (vgl. Cheung / Penn 2009, https://bit.ly/2Fu8dsp, 28.10.2022).

um standardgrammatische Kategorien wie Subjekt (SUBJ), Präpositionalobjekt (OBJP) usw. Die dependenzgrammatische Analyse des Beispielsatzes (mit elidiertem ungrammatischen Reflexivpronomen auf der ZH1-Ebene) ist in Abb. 5 abgebildet.

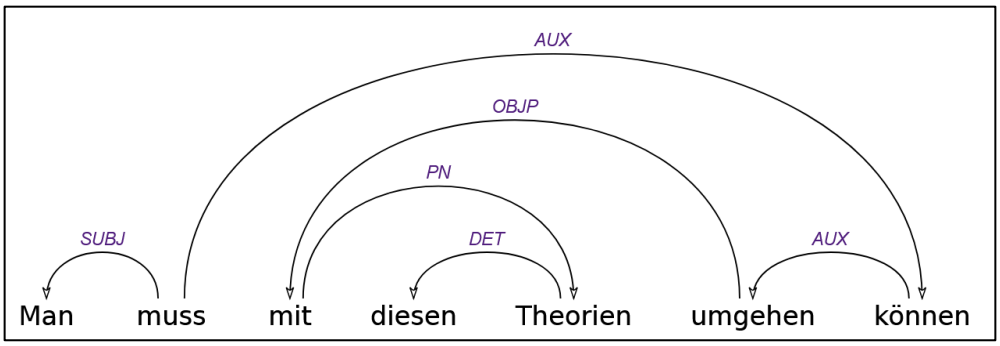

Abbildung 5
Bearbeitung der in (1) aufgeführten Lerneräußerung
(normalisierte Variante der ZH1-Ebene) mit dependenzgrammatischer Annotation.

Eine Anwendung, die diese Annotationen nutzt, stellen wir im nachfolgenden Abschn. 4 vor. In die Aufbereitung dieser Dependenzanalysen ist viel Arbeit geflossen, da die Dependenzannotationen, die ursprünglich mit dem automatischen Werkzeug MaltParser[8] (vgl. Nivre / Hall / Nilsson 2006) erstellt wurden, entsprechend fehleranfällig sind. Von Version 2 des Essaykorpus zur aktuellen Version 3 wurden die Annotationen manuell korrigiert.[9]

4. Forschungsspezifische Analysen und Anwendungsbeispiele

Die bislang erwähnten Korpusannotationen, durch die die Lernerdaten des *Falko*-Korpus angereichert wurden, sind Standardannotationen, die mit den genannten technischen Ressourcen allgemein für die Annotation von Korpusdaten zur Verfügung stehen (Voraussetzung ist allerdings häufig eine orthographisch und grammatisch normalisierte Textebene, die die ZH1-Ebene des *Falko*-Korpus bietet). Kurz wollen wir sehr spezifische, stark auf konkrete Forschungsziele abgestimmte Annotationen vorstellen, die ebenso den *Falko*-Daten hinzugefügt wurden und in der aktuell erscheinenden Version 3 allgemein nutzbar sind. Aus Platzgründen verzichten wir im Folgenden auf detaillierte Abbildungen von Annotationen, sondern verweisen auf die Korpusdokumentation, die auf der Webseite https://huberlin.de/falko herunterladbar ist und die Beschreibung der Annotationen enthält.

4.1 Komplexe Verben

Um den Gebrauch von Präfix- und Partikelverben bei fortgeschrittenen DaF-Lernenden im Kontrast zu deutschen MuttersprachlerInnen herauszuarbeiten (vgl. Lüdeling / Hirschmann / Shadrova 2017), wurden auf mehreren Annotationsebenen diese Verbklassen explizit gemacht (Partikelverben sind anhand der bislang vorgestellten Annotationen nur im Fall abgetrennter Verbpartikeln eindeutig iden-

[8] www.maltparser.org (28.10.2022).
[9] Dies wurde durch das Programm *WebAnno* (vgl. Eckart de Castilho et al. 2016, https://webanno.github.io/webanno/, 28.10.2022) ermöglicht.

tifizierbar). Auf diese Weise sind alle komplexen Verben, auch solche mit nicht abtrennbarem Erstbestandteil (Präfixverben) sowie Partikelverben mit abtrennbarer, aber nicht abgetrennter Verbpartikel unmittelbar am Verbstamm identifizierbar und auch nicht zielsprachliche Bildungen konkret auswertbar. Als schlaglichtartiges Beispiel für eine komplexere Korpussuchanfrage, die diese Annotationen zusammen mit den in Abschn. 3 vorgestellten Dependenzannotationen nutzt, ist mit dem Link https://hu.berlin/falko-suche-verb-abw einsehbar. Auf diese Weise findet man gezielt Fälle wie *Wissen auferarbeiten* (für *Wissen aufarbeiten*, Text cbs004_2006_09), *einen Weg anbahnen* (für *einen Weg bahnen*, Text cbs010_2007_10) oder *Theorien ausüben* (für *Theorien anwenden*, Text cbs010_2006_09), also Verb-Objekt-Gefüge mit einem Präfix- oder Partikelverb als Kopf, bei deren Kombination in den Korpusdaten ein semantisches, lexikalisches, morphologisches oder einfach phraseologisches Problem attestiert wurde.

4.2 Komplexität in der Nominaldomäne

Das Teilprojekt C04[10] des Sonderforscherbereichs 1412 Register[11] untersucht Registerkompetenz im Fremdspracherwerb anhand des grammatischen Phänomens Modifikation. Ein Teilziel des Forschungsplans ist, komplexe Nomina in den Lernerkorpora *Falko* und *Kobalt-DaF* (mitsamt den entsprechenden L1-Vergleichsdaten) auszuweisen, um so interindividuelle und gruppenspezifische Variation zu bemessen (methodische Probleme, die mit diesen Typen der Variation verbunden sind, werden in Shadrova et al. (2021) bezogen auf die Verwendung bestimmter Wortbildungstypen bei *Falko*-L2- und -L1-Daten im Vergleich besprochen). In diesem Rahmen wurden sämtliche Nomina im Korpus wortbildungsmorphologisch annotiert, indem komplexe Nomina als solche kenntlich gemacht wurden (Simplizia ebenso) und die relevanten Prozessschritte, die Wortbildungsprodukte hervorgebracht haben, explizit gemacht wurden.

4.3 Reflexivität

Im Forschungsprojekt *Crosslingual Language Varieties* (CLV)[12] werden solche sprachlichen Phänomene untersucht, bei denen in der Mehrsprachigkeit grammatische Konflikte auftreten können. Ein Phänomenbereich ist Reflexivität, welche in verschiedenen Zielsprachen unterschiedlich markiert wird. Ein Analyseziel in den deutschsprachigen Lernerdaten, die im Projekt untersucht werden, ist, Reflexivität auf verschiedene semantisch-lexikalische Ursachen zurückzuführen, so dass anschließend der Sprachgebrauch von Fremdsprachlernenden auf Spezifika bei der Verwendung reflexiver Strukturen analysiert werden kann. Interessante Fälle wie Korpusbeleg (1), in dem ein kontextuell ungrammatisches Reflexivpronomen verwendet wird, können anhand der erstellten Annotationen systematisch gefunden und in entsprechenden Analysen berücksichtigt werden.

4.4 Argumentationsstruktur

Shujun Wan untersucht in ihrem Dissertationsprojekt die Spezifika von Argumentationsstrategien vergleichend bei chinesischen DaF-Lernenden und deutschen MuttersprachlerInnen. Zu diesem

[10] https://sfb1412.hu-berlin.de/de/projekte/c04/ (28.10.2022), gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1412, 416591334.
[11] https://sfb1412.hu-berlin.de/de/ (28.10.2022).
[12] https://hu-berlin.de/clv (28.10.2022).

Zweck wurden dem *Kobalt-DaF*-Korpus, das dem *Falko-Essay*-Korpus sehr vergleichbar ist, aber homogenere Lernergruppen, u.a. chinesische L1-Sprechende, beinhaltet, Annotationen nach der Rhetorischen Strukturtheorie (RST)[13] hinzugefügt, die in Lüdeling et al. (2021) genauer vorgestellt und bildlich dargestellt werden.

5. Nutzungsperspektiven und Schlussbemerkungen

Die vorgestellten Annotationen stellen nur einen Ausschnitt aus den in der aktuellen *Falko*-Version verfügbaren Analysen dar. Wir haben versucht zu zeigen, dass diese Analysen unterschiedlich spezifisch sind. Die zuerst (in Abschn. 2 und 3) vorgestellten Annotationen kommen wahrscheinlich in vielen Korpusstudien zur Anwendung. Die in Abschn. 4 erwähnten Annotationen wurden aufgrund sehr spezifischer Forschungsziele erstellt, stehen aber genauso allen Nutzerinnen und Nutzern zur Verfügung. Die erwähnten Annotationen werden in aktuellen Korpusversionen zusammengeführt und auf zweierlei Weise verfügbar gemacht:[14] Erstens sind die Daten, die in Tabellenkalkulationsprogrammen abbildbar sind, als xlsx- sowie exb-Dateien herunterladbar (letzteres Format ist das Ein- und Ausgabeformat des *EXMARaLDA-Partitureditors*).[15] Vor allem sind die Korpusdaten jedoch in einem online zugänglichen Sucherinterface namens *ANNIS*[16] verfügbar und können dort direkt betrachtet, durchsucht und ausgewertet werden.[17] Zur Einführung in die Anfragesprache, mit der man die Korpusdaten systematisch durchsuchen kann, wurden Foliensätze erstellt, die vor allem für Einführungsveranstaltungen zur Suche in den *Falko*-Korpora genutzt werden.[18]

In unserer Korpusvorstellung haben wir versucht, den aktuellen Aufbereitungsstand der *Falko*-Kernkorpora darzustellen. Die Möglichkeit, sukzessive und über viele Jahre hinweg Analysen zu bereits bestehenden und veröffentlichten Korpusdaten hinzuzufügen, wurde durch drei wesentliche Maßnahmen erzielt: die Nutzung transparenter XML-basierter Dateiformate (wie dem *EXMARaLDA*-Speicherformat), die im Grunde beliebig viele Annotationsebenen mit flexiblen Bezügen untereinander zulassen, die Versionierung der Korpusdaten und die Erstellung eines Konversionsframeworks ‚Pepper', welches eine flexible Weiterverarbeitung und Zusammenführung verschiedener Korpusspeicherformate erlaubt.[19]

Literatur und Ressourcen

Cheung, Jackie C. K. / Penn, Gerald (2009): Topological Field Parsing of German. *Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics (ACL)*. Singapur, 64-72.

Díaz-Negrillo, Ana / Meurers, Detmar / Valera, Salvador / Wunsch, Holger (2010): Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. In: *Language Forum* 36: 1-2 (Corpus Linguistics for Teaching and Learning), 139-154.

[13] www.sfu.ca/rst/01intro (28.10.2022).
[14] Vgl. auch https://hu.berlin/falko-zugang (28.10.2022).
[15] Siehe https://exmaralda.org/de/ (28.10.2022) bzw. Schmidt / Wörner (2014).
[16] Vgl. Krause / Zeldes (2016), https://corpus-tools.org/annis/ (28.10.2022).
[17] Die Webadresse der Online-Instanz mit Lernerkorpora ist https://hu.berlin/annis-falko (28.10.2022).
[18] Vgl. https://hu.berlin/falko-folien (28.10.2022).
[19] Vgl. https://corpus-tools.org/pepper/ (28.10.2022) bzw. Zipser / Romary (2010).

Eckart de Castilho, Richard / Mújdricza-Maydt, Eva / Yimam, Seid M. / Hartmann, Silvana / Gurevych, Iryna / Frank, Anette / Biemann, Chris (2016): A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In: *Proceedings of the LT4DH workshop at COLING 2016*, 76-84.

Granger, Sylviane (2008): *Learner corpora*. In: Lüdeling, Anke / Kytö, Merja (Hg.): *Corpus Linguistics. An International Handbook*. Bd. 1. Berlin: de Gruyter, 259-275.

Hirschmann, Hagen (2019): *Korpuslinguistik. Eine Einführung*. Stuttgart: Metzler.

Krause, Thomas / Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: *Digital Scholarship in the Humanities 2016* 31.

Lüdeling, Anke (2008): *Mehrdeutigkeiten und Kategorisierung. Probleme bei der Annotation von Lernerkorpora*. In: Walter, Maik / Grommes, Patrick (Hg.): Fortgeschrittene Lernervarietäten. Tübingen: Niemeyer, 119-140.

Lüdeling, Anke / Doolittle, Seanna / Hirschmann, Hagen / Schmidt, Karin / Walter, Maik (2008): Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache 2(2008)*, 67-73.

Lüdeling, Anke / Hirschmann, Hagen (2015): *Error annotation systems*. In: Granger, Sylviane / Gilquin, Gaëtanelle / Meunier, Fanny (Hg.): *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 135-158.

Lüdeling, Anke / Hirschmann, Hagen / Shadrova, Anna (2017): Linguistic Models, Acquisition Theories, and Learner Corpora: Morphological Productivity in SLA Research Exemplified by Complex Verbs in German. In: *Language Learning 67*, 96-129.

Lüdeling, Anke / Hirschmann, Hagen / Shadrova, Anna / Wan, Shujun (2021): Tiefe Analyse von Lernerkorpora. In: Lobin, Henning / Witt, Andreas / Wöllstein, Angelika (Hg.): *Deutsch in Europa. Sprachpolitisch, grammatisch, politisch*. [Jahrbuch des Leibniz-Instituts für Deutsche Sprache 2020]. Berlin u.a.: de Gruyter, 253-283.

Nivre, Joakim / Hall, Johan / Nilsson, Jens (2006): MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (*LREC 2006*), 2216-2219.

Reznicek, Marc / Lüdeling, Anke / Krummes, Cedric / Schwantuschke, Franziska / Walter, Maik / Schmidt, Karin / Hirschmann, Hagen / Andreas, Torsten (2012): *Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 2.0*. Technischer Bericht. Humboldt-Universität zu Berlin, Berlin. https://hu.berlin/falko-handbuch_v2 (28.10.2022).

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technischer Bericht. Institut für maschinelle Sprachverarbeitung, Stuttgart. www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (28.10.2022).

Schmidt, Thomas / Wörner, Kai (2014): EXMARaLDA. In: Durand, Jacques / Gut, Ulrike / Kristoffersen, Gjert (Hg.): *Handbook on Corpus Phonology*. Oxford: Oxford University Press, 402-419.

Shadrova, Anna / Linscheid, Pia / Lukassek, Julia / Lüdeling, Anke / Schneider, Sarah (2021): A Challenge for Contrastive L1/L2 Corpus Studies: Large Inter- and Intra-Individual Variation Across Morphological, but

Not Global Syntactic Categories in Task-Based Corpus Data of a Homogeneous L1 German Group. In: *Frontiers in Psychology, Section Language Sciences*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.716485/full (28.10.2022).

Telljohann, Heike / Hinrichs, Erhard / Kübler, Sandra / Zinsmeister, Heike (2017): *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Technischer Bericht. Universität Tübingen. https://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-1201.pdf (28.10.2022).

Zipser, Florian / Romary, Laurent (2010): A model oriented approach to the mapping of annotation formats using standards. In: *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*. Malta. http://hal.archives-ouvertes.fr/inria-00527799/en/ (28.10.2022).

Biographische Notiz: Die genannten Autorinnen und Autoren sind Mitarbeiterinnen und Mitarbeiter am Institut für deutsche Sprache und Linguistik der Humboldt-Universität zu Berlin, Fachbereich für Korpuslinguistik und Morphologie.

Kontaktanschrift:
Dr. Hagen Hirschmann
Humboldt-Universität zu Berlin
Institut für deutsche Sprache und Linguistik
Unter den Linden 6
10099 Berlin
hagen.hirschmann@hu-berlin.de