Dulko – auf dem Weg zu einem deutsch-ungarischen Lernerkorpus

Hagen Hirschmann
Berlin

Andreas Nolda
Szeged

Dulko ist ein im Aufbau befindliches fehlerannotiertes deutsch-ungarisches Lernerkorpus an der Universität Szeged. Es wird seit Sommer 2017 von der Alexander-von-Humboldt-Stiftung gefördert im Rahmen einer Institutspartnerschaft zwischen dem IDS und dem Institut für Germanistik an der Universität Szeged („Deutsch-ungarischer Sprachvergleich: korpustechnologisch, funktional-semantisch und sprachdidaktisch (DeutUng)“). Die in Dulko erhobenen Lernerdaten setzen sich zusammen aus kontrolliert erhobenen deutschsprachigen Essays und Übersetzungen aus dem Ungarischen ins Deutsche. Die Probanden sind Studierende am Institut für Germanistik der Universität Szeged mit Ungarisch als Muttersprache und Deutsch als erster oder zweiter Fremdsprache. Metadaten zur Lernerbiographie und zu den Textproduktionsbedingungen werden mit den Textdaten und den Annotationen im Format des jüngst von Granger und Paquot (2017) vorgeschlagenen Standard gespeichert.

Ziel der Erstellung des Lernerkorpus Dulko ist die Untersuchung spracherwerbstheoretischer Fragestellungen der folgenden Art: Welche spezifischen Erwerbsprobleme haben ungarische DaF-Lerner? Inwieweit spielen bei ihnen Transferprobleme eine Rolle, die sich in Form typischer Interferenzfehler manifestieren? Zur empirischen Untersuchung dieser Fragestellungen müssen sowohl grammatische als auch ungrammatische Strukturen in der Lernersprache auf verschiedenen sprachlichen Ebenen analysiert werden. Es stellt sich somit die korpuslinguistische Frage, wie sprachliche Abweichungen in einem Lernerkorpus transparent und nachvollziehbar durch die Erstellung von Zielhypothesen (Reznicek, Lüdeling und Hirschmann 2013) als Fehler interpretiert und mit Hilfe eines mehrdimensionalen Fehler-Taggings kategorisiert werden können.

Zur Illustration dieser korpuslinguistischen Frage soll der folgende Satz aus einem der in Dulko annotierten Lernertexte dienen:

(1) *Wie in der ganzen Gesellschaft, auch in der Regierung sollte der Anzahl der Frauen 50 % sein* […]. (Auszug aus einem Essay eines ungarischen DaF-Lerners, Wintersemester 2017/18, Universität Szeged)

Zu Beispiel (1) sind in Abbildung 1 zwei Zielhypothesen angegeben, die Abweichungen der Lernersprache vom muttersprachlichen Deutsch repräsentieren. Die mittels der intermediären Zielhypothese 1 repräsentierten Abweichungen können als Zeichensetzungs-Fehler, Wortstellungs-Fehler und Genus-Fehler kategorisiert werden. Die finale Zielhypothese 2 repräsentiert einen lexikalischen Fehler, dessen Skopus sich mit dem Skopus des Genus-Fehlers überlappt. Man beachte, dass es nicht möglich ist, sowohl den Genus-Fehler als auch den lexikalischen Fehler bei *der Anzahl* in ein und derselben Zielhypothese zu repräsentieren, da die ‚Korrektur‘ des lexikalischen Fehlers zu *der Anteil* den Genus-Fehler ‚unsichtbar‘ macht. Deshalb wird hier eine intermediäre Zielhypothese angesetzt, in der zwar der Genus-Fehler zu *die Anzahl* ‚korrigiert‘ ist, das im Kontext *50 %* unpassende Substantiv *Anzahl* jedoch vorläufig unverändert bleibt. Die Möglichkeit, einander ergänzende, kumulativ zu

Lernertext:

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

Zielhypothese 1 (intermediär):

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

Fehler: Zeichensetzung Wortstellung Genus

Zielhypothese 2 (final):

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

Fehler: Lexik

Abbildung 1: Zielhypothesen und Fehleranalyse zu Beispiel (1)

[tok]	**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]																			

Abbildung 2: Annotation von Beispiel (1) in EXMARaLDA (Dulko)

interpretierende Zielhypothesen anzugeben – neben einander ausschließenden, alternativ interpretierten Zielhypothesen – ist ein innovatives Charakteristikum des Annotationsverfahrens in Dulko.

Die Korpuserstellung in Dulko ist methodisch angelehnt an das Lernerkorpus Falko (https://hu-berlin.de/falko/), das seit 2005 an der Humboldt-Universität zu Berlin entstanden ist. Dies betrifft sowohl automatische Verarbeitungsschritte wie die Annotation von Wortarten, Lemmata und Satzspannen als auch die manuelle Erstellung von Zielhypothesen. Im Unterschied zu Falko gibt es in Dulko ein explizites, mehrdimensionales Fehler-Tagging. Sämtliche Annotationen werden in Dulko mit Hilfe des EXMARaLDA-Partitureditors vorgenommen (http://exmaralda.org/; Schmidt *et al.* 2016). Hierzu wurde eine Variante namens „EXMARaLDA (Dulko)“ erstellt, die in XSLT implementierte Transformationsszenarien für die Annotation von Lernerdaten in Dulko zur Verfügung stellt (https://bitbucket.org/nolda/exmaralda-dulko/).

Abbildung 2 zeigt, wie mittels dieser Transformationsszenarien die Zielhypothesen und die Fehleranalyse in Abbildung 1 annotiert werden. Zunächst wird der Lernertext im Partitureditor maschinell tokenisiert (`tok`-Ebene). Die sich ergebenden Tokens werden durch internen Aufruf des TreeTaggers (Schmid 1997) nach Wortarten und Lemmata getaggt (`pos`-Ebene und `lemma`-Ebene); au-

ßerdem werden Satzspannen als solche identifiziert (S-Ebene). In einem zweiten Verarbeitungsschritt wird eine erste Zielhypothese unter Berücksichtigung der Falko-Guidelines (Reznicek *et al.* 2012) angegeben (ZH-Ebene), die gleichfalls automatisch mit Wortarten (ZHpos-Ebene), Lemmata (ZHlemma-Ebene) und Satzspannen (ZHS-Ebene) annotiert wird. Ein weiteres Transformationsszenarium erkennt Abweichungen zwischen Tokens der Zielhypothese und des Lernertexts (bzw. einer vorausgesetzten intermediären Zielhypothese) und markiert sie mit den in Falko dafür vorgesehenen Tags (ZHDiff-Ebene). Abweichungen, die als Fehler zu interpretieren sind, werden unter Verwendung eines in EXMARaLDA (Dulko) vordefinierten orthographischen, morphologischen, syntaktischen, lexikalischen und semantischen Fehler-Tagsets auf den dafür zur Verfügung stehenden Fehler-Ebenen manuell kategorisiert. Dieser Verarbeitungsschritt wird solange wiederholt, bis alle Zielhypothesen annotiert sind.

Die auf diese Weise im EXMARaLDA-Partitureditor erzeugten XML-Dateien lassen sich nach ihrer Fertigstellung in das Format des Suchwerkzeugs ANNIS (http://corpus-tools.org/annis/) konvertieren. Perspektivisch sollen sie darüber hinaus auch über das am IDS entwickelte Suchwerkzeug KorAP (https://korap.ids-mannheim.de) verfügbar gemacht werden.

Literatur

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

Reznicek, Marc, Anke Lüdeling und Hagen Hirschmann (2013). Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In *Automatic Treatment and Analysis of Learner Corpus Data*, hg. v. Ana Díaz-Negrillo, Nicolas Ballier und Paul Thompson, Amsterdam: Benjamins, 101–123.

Reznicek, Marc *et al.* (2012). Das Falko-Handbuch: Korpusaufbau und Annotationen. Version 2.01. Manuskript, Humboldt-Universität zu Berlin. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2 [26. Jan. 2017].

Schmid, Helmut (1997). Probabilistic part-of-speech tagging using decision trees. In *New Methods in Language Processing*, hg. v. Daniel B. Jones und Harold L. Somers, London: Routledge, 154–164.

Schmidt, Thomas *et al.* (2016). EXMARaLDA Partitur-Editor: Manual. Version 1.6. Manuskript. http://www.exmaralda.org/pdf/Partitur-Editor_Manual.pdf [27. März 2018].