Hagen Hirschmann (Berlin)/Andreas Nolda (Szeged)

Dulko – auf dem Weg zu einem deutsch-ungarischen Lernerkorpus

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-semantic 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 Lernerbiografie und zu den Textproduktionsbedingungen werden mit den Textdaten und den Annotationen im Format eines jüngst von Granger/Paquot (2017) vorgeschlagenen Standards gespeichert.

Ziel der Erstellung des Lernerkorpus Dulko ist die Untersuchung sprach-erwerbstheoretischer 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/Hirschmann 2013) als Fehler interpretiert und mithilfe 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.

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 orthografischen, morphologischen, syntaktischen, lexikalischen und semantischen Fehler-Tagsets auf den dafür zur Verfügung stehenden Fehler-Ebenen manuell kategorisiert. Dieser Verarbeitungsschritt wird so lange wiederholt, bis alle Zielhypothesen annotiert sind.

[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	S,		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 von Beispiel (1) in EXMARaLDA (Dulko)

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/Paquot, Magali (2017): Core metadata for learner corpora. Draft 1.0. Ms. Louvain-la-Neuve.
- Reznicek, Marc/Lüdeling, Anke/Hirschmann, Hagen (2013): Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In: Díaz-Negrillo, Ana/Ballier, Nicolas/Thompson, Paul (Hg.): Automatic treatment and analysis of learner corpus data. Amsterdam, S. 101–123.
- Reznicek, Marc et al. (2012): Das Falko-Handbuch: Korpusaufbau und Annotationen. Version 2.01. Ms. Berlin. Internet: www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2 (Stand: 26.1.2017).
- Schmid, Helmut (1997): Probabilistic part-of-speech tagging using decision trees. In: Jones, Daniel B./Somers, Harold L. (Hg.): New methods in language processing. London, S. 154–164.
- Schmidt, Thomas et al. (2016): EXMARaLDA Partitur-Editor: Manual. Version 1.6. Ms. Internet: www.exmaralda.org/pdf/Partitur-Editor_Manual.pdf (Stand: 27.3.2018).