Fehlerannotation Pipeline: SWIKO Korpus

Für alle fremdsprachlich und schulsprachlich deutschen Texte wurde eine Zielhypothese formuliert (siehe Dokument «ZielhypothesenAnleitung» für detaillierte Ausführungen zum Vorgehen). Auf Basis der vorhandenen Annotationen wurden die Texte anschliessend automatisch getaggt:

Orthographie

Kategorie	***Unterkategorie***	***Beispiel***	***Tag***
Gross-/Kleinschreibung	Satzanfang	*ich liebe Pferde.*	O_cap_bos
	Nomen	*computer, ferien*	O_cap_noun
	andere	*Acht, lieblingsTiere*	O_cap
Grapheme	Akzent / Umlaut	*Brôte, kostët, Schuler*	O_graph_acc
	1 Buchstabe zu viel	*Personnen*	O_graph_ad
	1 Buchstabe zu wenig	*nich, Tire*	O_graph_del
	1 Buchstabe anders	*bezucht, sotial*	O_graph_ch
	mehrere	*numere, helfbereit*	O_graph_mul
Wortgrenzen	getrennt	*Lieblings Musik, ge funden*	O_wdbd_mg
	verbunden	*Diewochenende, Inder*	O_wdbd_sp
	Bindestrich	*Liebst-du, gross-Vater*	

Hier ein Beispiel mit orthographischen Fehlern und entsprechenden Tags. Der ursprüngliche Originaltext der Lernenden (*tok*-Spur) wurde bereits in einem früheren Annotationsschritt als Vorbereitung auf das POS-Tagging orthographisch korrigiert (*ctok*-Spur; siehe Dokument «SWIKO_Transcription_Error-correction» für detaillierte Instruktionen) und entsprechend auf der *tag*-Spur vermerkt, ob es sich dabei um einen Rechtschreibfehler (*error*) oder ein Wort in einer anderen Sprache, spezifisch der Schulsprache (*schlg_item*) oder Englisch (*foreign_Eng*).

Für die orthographische Annotation wurden somit nur Token der *tok*- und *TH1*-Spur verglichen, die bereits in der *Tag*-Spur mit einem Tag versehen wurden.

	0	1	2	3	4	5	6	7	8	9
Ex000 [tok]	inder		schweitz	gipt	es	8,2	Milionen	Tiere	domestique	.
Ex000 [ctok]	In	der	Schweiz	gibt	es	8,2	Millionen	Haustiere	NA	.
Ex000 [lemma]	in	die	Schweiz	geben	es	@card@	Million	Haustier	NA	.
Ex000 [clemma]										
Ex000 [commonPOS]	PRP	DET	NP	VER:PRE	PRO:PER	NUM	NN	NN	NA	PUN
Ex000 [lg-specific POS]	PRP	ART	NE	VVFIN	PPER	CARD	NN	NN	NA	$.
Ex000 [cpos]										
Ex000 [tag]	error		error	error			error		schlg_item	
Ex000 [markup]										
Ex000 [comment]										
Ex000 [TH1]	In	der	Schweiz	gibt	es	8,2	Millionen	Haustiere		.
Ex000 [SEA]										
Ex000 [O_cap]	O_cap_bos		O_cap_noun							
Ex000 [O_graph]			O_graph_del	O_graph_ch			O_graph_add	O_graph_mul		
Ex000 [O_wdbd]	O_wdbd_sp							O_wdbd_mg		

Grammatik

Kategorie	***Unterkategorie***	***Beispiel***	***Tag****
Überflüssiges Wort		*Es gebe 8,2 Million [die] Haustiere*	G_ *pos*[1] _del (G_ART_del)
Fehlendes Wort		*In [der] Schweiz*	G_ *pos*[1] _add (G_ART_add)
Wort verändert	Flexion	*es [gebe], 8,2 [Million]*	G_ *pos*[1] _cha (G_VVFIN_cha) (G_NN_cha)
	Wortart	*Wir haben [nicht] électricité [warum] ich mag mangas*	G_POS_ *pos* _ *pos*[2] (G_POS_PIAT_PTKNEG) (G_POS_KOUS_PWAV)
	Wortwahl	*Ich [habe] 14 Jahre* *Eins [in] zwei Häusern*	G_ *pos*[1] _wch (G_VAFIN_wch) (G_APPR_wch)
Falsche Position	Auf Token-Ebene	In Schweiz es gebe 8,2 Million die Haustiere. → In der Schweiz gibt es 8,2 Millionen Haustiere.	G_ *pos* _movs[1, 3] + (G_VVFIN_movs) G_ *pos* _ movt[1, 3] (G_VVFIN_movt)
	Auf Satz-Ebene	In Schweiz es gebe 8,2 Million die Haustiere.	G_wordorder[4]

Hier ein Beispiel mit grammatischen Fehlertags. Für die automatische Annotation grammatischer Phänomene wurden die *tok*- und *TH1*-Spuren unter Berücksichtigung der *clemma*- und *cpos*-Spuren verglichen. Tokens, welchen bereits ein Orthographie-Tag zugeordnet wurde, erhalten automatisch keine weiteren Grammatik-Tags.

	0	1	2	3	4	5	6	7	8	9	10	11
RIxxx [tok]	In		Schweiz		es	gebe	8,2	Million	die	Haustiere	.	
RIxxx [ctok]	In		Schweiz		es	gebe	8,2	Million	die	Haustiere	.	
RIxxx [lemma]	in		Schweiz		es	geben	@card@	Million	die	Haustier	.	
RIxxx [clemma]				geben								
RIxxx [commonPOS]	PRP		NP		PRO:PER	VER:PRE	NUM	NN	DET	NN	PUN	
RIxxx [lg-specific POS]	PRP		NE		PPER	VVFIN	CARD	NN	ART	NN	$.	
RIxxx [cpos]		ART										
RIxxx [tag]												
RIxxx [markup]							NUM					
[comment]												
RIxxx [TH1]	In	der	Schweiz	gibt	es		8,2	Millionen		Haustiere	.	
RIxxx [SEA]												
RIxxx [G_add]		G_ART_add										
RIxxx [G_cha]						G_VVFIN_cha		G_NN_cha				
RIxxx [G_del]									G_ART_del			
RIxxx [G_pos]				G_VVFIN_movs		G_VVFIN_movt						
RIxxx [G_wordorder]	G_wordorder											

[1] Anstelle des « *pos* » wird die Wortart aus der lg-specific POS-Zeile eingefügt
[2] Anstelle der « *pos* » werden die Wortarten des ursprünglichen sowie zielsprachengerechten Tokens eingefügt.
[3] Die ursprüngliche- (_movs für start) und Zielposition (_movt für target) werden mit entsprechenden Tags vermerkt.
[4] Dieses Tag spannt über einen ganzen Satz, der mindestens ein Positions-Fehler-Tag auf Tokenebene enthält.

Zweifelsfälle

In einem letzten Schritt wurden sämtliche Wortpaare, welche im Rahmen der *ctok*-Spur korrigiert wurden, manuell kontrolliert, da es sich dabei oft um Zweifelsfälle handelt. Zum Beispiel kann es sich bei *habst* statt *hast* durchaus um einen Rechtschreibfehler handeln; man könnte aber auch argumentieren, dass es sich um einen Deklinationsfehler und somit ein grammatisches Phänomen handelt. Ähnlich u.a. bei *geschreiben-geschrieben* oder *geglauben-geglaubt*. Auch Fehlbildungen wie *helfbereit* statt *hilfsbereit* fallen in diese Kategorie. Oder eine falsche Wortart wie *das* statt *dass* könnte ebenfalls unterschiedlich interpretiert werden.

Um solche Fälle entsprechend zu taggen, wurde zuerst eine Liste mit entsprechenden Wortpaaren automatisch extrahiert. Anschliessend wurden manuell sämtliche Paare markiert, welche auch einer grammatischen Fehlerkategorie zugeordnet werden könnten (z.B. Flexion (G_*pos*_cha) für *habst-hast* oder Wortart (G_POS_*pos_pos*) für *das-dass*). Da diese Einschätzung auch subjektiv ist, wurde sie von zwei Forscherinnen separat durchgeführt und anschliessend abgeglichen. Die endgültige Fassung wurde dann wieder eingelesen und die Tags automatisch ergänzt.

Solche Zweifelsfälle sind somit die einzigen Token, welche sowohl grammatische als auch orthographische Tags aufweisen; je nach Forschungsfrage können sie entsprechend in die Auswertung miteinbezogen oder ausgeschlossen werden.