eurac research

Das Schülertexte-Korpus KoKo

Andrea Abel, Aivars Glaznieks, Egon W. Stemle, Lionel Nicolas
Institut für Angewandte Sprachforschung
Eurac Research, Bozen, Italien
linguistics@eurac.edu

Hintergrund

- Sprachkompetenzen als Schlüsselkompetenzen in Ausbildung und Beruf
- Geschriebene Texte als Möglichkeit der Bewertung von Sprach- und Schreibkompetenz(en)
- Plurizentrische Perspektive auf die deutsche Sprache
- Fehlende frei zugängliche Korpora mit Schülertexten als Basis für linguistische Analysen

Das KoKo-Korpus ist das Ergebnis eines Projekts mit dem Titel: „Bildungssprache im Vergleich: korpusunterstützte Analyse der Sprachkompetenzen bei Lernenden im deutschen Sprachraum". Dieses Projekt hat zwischen 2010 und 2015 bildungssprachliche Kompetenzen von OberschülerInnen im deutschen Sprachraum vergleichend untersucht.

Korpusaufbau und Annotationen

Allgemeine Informationen:
- Alle Texte aus dem Jahr 2011
- Alter der SchreiberInnen zur Zeit der Datenerhebung zwischen 17 und 19 Jahren (vorletzte Klasse abitur- bzw. maturaführender Schulen)
- Repräsentatives Sample von Schulen aus Nordtirol, Südtirol und Thüringen

Datenerhebung:
- Während des regulären Unterrichts, dieselbe Erörterungsaufgabe für alle Klassen, Transkription der handschriftlichen Texte
- Fragebogenerhebung sozio-demographischer und sprachbiographischer Daten

Annotationen:
- Manuelle Annotationen:
  - Textstrukturelle Eigenheiten (Absätze, Überschriften, Eigenkorrekturen)
  - Orthographiefehler (Rechtschreibung und Interpunktion)
  - grammatikalische Fehler
  - lexikalische Auffälligkeiten
- Automatische Annotationen: Token, Satz, Wortart, Lemma
- Metadaten:
  - Personenbezogen: sozio-demographische und sprachbiographische Daten
  - Texbezogen: Textevaluationsschema mit 62 „Fragen an den Text"

Korpuszugang

Das KoKo-Korpus steht in unterschiedlichen Versionen zur Verfügung und kann sowohl über eine ANNIS-Oberfläche durchsucht als auch für wissenschaftliche Zwecke als XML-Datei heruntergeladen werden.
Weitere Informationen zum Korpuszugang unter: https://clarin.eurac.edu/

Korpusbeschreibungen

- Glaznieks, A., L. Nicolas, E. W. Stemle, A. Abel & V. Lyding (2014): Establishing a Standardised Procedure for Building Learner Corpora. In: *Apples - Journal of Applied Language Studies* 8 (3), 5-20, Special Issue on *Learner Language, Learner Corpora: From corpus compilation to data analysis*, Jarmo Harri Jantunen, Sisko Brunni & Marianne Spoelmann (eds).
- Abel, A., A. Glaznieks, L. Nicolas & E. W. Stemle (2014): KoKo: an L1 Learner Corpus for German. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, 26–31 May, 2014, 2414-2421.
- Abel, A., A. Glaznieks, L. Nicolas & E. W. Stemle (2016): An extended version of the KoKo German L1 Learner Corpus. In *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016. 5-6 December 2016, Napoli,* Anna Corazza, Simone Montemagni, & Giovanni Semeraro, (eds). Torino: Academia University Press, 13-18.

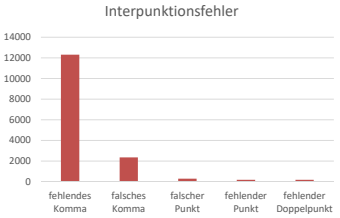
Zahlen zum Korpus

Das Korpus besteht aus circa 811.000 Tokens. Die Verteilung über die beteiligten Erhebungsgebiete gibt folgende Tabelle wieder:

Erhebungsgebiet	alle SchülerInnen		SchülerInnen mit Deutsch als L1	
	Tokens	Texte	Tokens	Texte
Nordtirol	~233.000	457	~206.000	404
Südtirol	~222.000	520	~193.000	451
Thüringen	~354.000	521	~317.000	464
k. A.	2.000	5	-	-
TOTAL	~811.000	1.503	~716.000	1.319

Tabelle 1: Korpusübersicht

Im gesamten Korpus sind insgesamt etwa 29.000 orthographische Fehler annotiert worden. Die häufigsten Fehler geben folgende Übersichten wieder:



Graphik 1: Interpunktionsfehler

Rechtschreibfehler

Fehlerkategorie	Anzahl	pro Text
Großschreibung	4.297	2,85
Getrenntschreibung	2.843	1,89
Weglassen	2.515	1,67
Verwechslung	1.374	0,91
Hinzufügen	1.034	0,69
andere	891	0,59
Eigennamen	525	0,35
TOTAL	13.479	8,97

Tabelle 2: Rechtschreibfehler

In einem Subkorpus von 597 Texten wurden über 2700 grammatische Fehler annotiert. Für 980 Texte stehen außerdem über 60.000 lexikalische Annotationen zur Verfügung, die sich entweder auf Einzelwörter oder auf phraseologische Einheiten beziehen können.

Grammatikfehler

Kategorie	Anzahl	pro Text
Korrespondenz	1697	2.84
Flexion	246	0.41
Vollständigkeit	381	0.64
Redundanzen	69	0.11
Anakoluth	127	0.21
Wortstellung	110	0.18
nicht kategorisierbar	111	0.18
TOTAL	2741	4.59

Tabelle 3: Grammatikfehler

lexikalische Annotationen

Bezugselement		
Einzelwort	Neologismen/Okkasionalismen	4.670
	arg. Adverbien und Konnektoren	14.345
Phrasem	referentiell	18.708
	kommunikativ	4.824
	strukturell	2.704
Auffälligkeiten	semantisch	8.397
	stilistisch	236
	formal	1.923
	metalinguistisch	1.412

Tabelle 4: lexikalische Annotationen

Neben den linguistischen Annotation stehen für alle Texte personenbezogene Metadaten zur Verfügung:
- Erstsprache
- Geschlecht
- Schultyp
- Herkunftsregion
- Deutschnote

Darüber hinaus stehen für ein Subkorpus von 596 Texten textbezogene Metadaten zur qualitativen Textevaluation zur Verfügung, die folgende Aspekte bewerten:
- Formale Vollständigkeit der Texte
- Inhalt der Texte
- Formale und linguistische Mittel der Textgestaltung
- Allgemeiner Eindruck