Aufbau des chinesischen Deutschlernerkorpus (CDLK) und Erforschung der Schriftkompetenzentwicklung der Lernenden

das CDLK-Team (Forschungsgruppe von Prof. Dr. Li Yuan)

Zhejiang Universität, China

Motivation

Die Motivation für den Aufbau dieses Korpus liegt in der zunehmenden Verbreitung der deutschen Sprache in China seit der Jahrtausendwende. Trotz der großen Zahl chinesischer Deutschlernender fehlt bislang ein Korpus, das schriftliche Texte dieser Lernergruppe auf verschiedenen Niveaustufen systematisch erfasst. Dies hat zur Folge, dass es an empirischen Untersuchungen zu den sprachlichen Merkmalen chinesischer Deutschlernender mangelt.

Korpus-Design und Umfang

Im CDLK werden Texte mit unterschiedlichen Themen und verschiedenen Genres gesammelt, die von chinesischen Deutschlernern im Unterricht handschriftlich innerhalb von 30 Minuten ohne Hilfsmittel produziert werden.

	Thema 1	Thema 2		Thema 3
Schüler/-innen	Meine Familie	Meine Hobbys		Bildbeschreibung
Student/-innen			Arbeiten oder Weiterstudieren nach dem Abschluss	
	deskriptiv	deskriptiv	argumentativ	bildbeschreibend
	Thema 4	Thema 5		Thema 6
Schüler/-innen	Handy im Unterricht oder nicht	Ein besonderes Erlebnis		Meine Neujahrswünsche
Student/-innen				
	argumentativ	erzählend		deskriptiv

Tab. 1: Themen und Genres im CDLK

Die Texte stammen aus Lernenden von 27 Schulen und Universitäten in verschiedenen Regionen Chinas. Das Korpus deckt alle Niveaus bzw. Lernstufen ab und enthält sowohl querschnittliche als auch längsschnittliche Daten.
Metadaten im CDLK werden multidimensional gesammelt, sowohl unter Deutschlernenden, als auch unter Deutschlehrenden, von persönlichen Informationen, z. B. Alter, Geschlecht, über Sprachlerninformation wie Deutschlerndauer, Erlernen anderer Sprachen und Sprachniveau bis Unterrichtsinformationen wie Lernmaterialien.
Bisher beträgt der Umfang des Korpus CDLK insgesamt 10.103 Texte, wovon 5.814 Texte von Mittelschulen und 4.289 Texte von Universitäten stammen.

Annotation

Die Texte werden mit Hilfe verschiedener Programme in Bezug auf Lexik, Syntax und Fehler annotiert. Die lexikalische Annotation umfasst Wortart und Lemma eines Wortes, während sich die syntaktische Annotation primär an der Dependenzgrammatik orientiert.

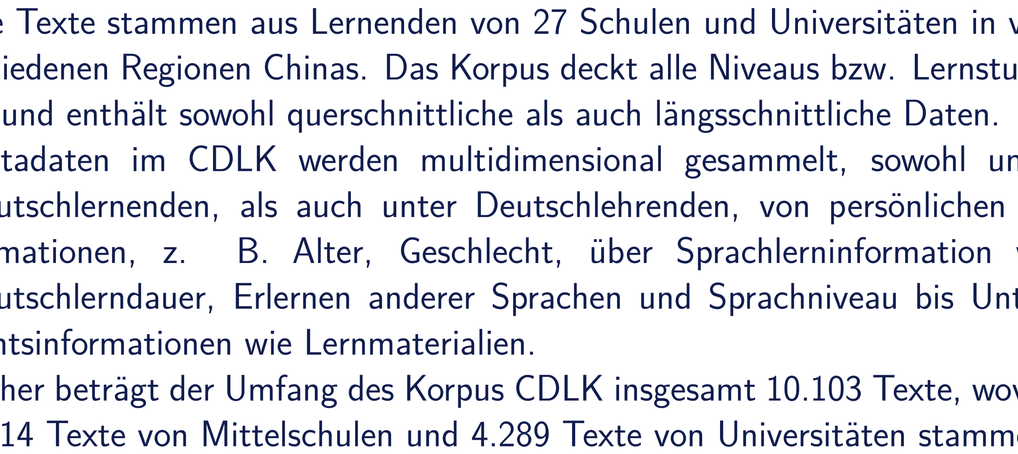

Abb. 1: Annotationsverfahren

Die Fehlerannotation erfolgt ausschließlich manuell: Zunächst wird der Originaltext korrigiert, um die Zielhypothese (als korrekt angenommene Version) aufzustellen. Anschließend wird jeder Fehler lokalisiert und einer Fehlerklasse zugeordnet.
Das Fehlerklassifizierungssystem des CDLK umfasst fünf Hauptfehlerkategorien: Orthographie, Morphosyntax, Syntax, Lexik und Semantik.

Jede Kategorie wird weiter in mehrere Unterkategorien unterteilt, was zu insgesamt 63 Unterkategorien führt.

	47	48	49	50	51	52	53	54	55	56
AUT [word]	**.**	**Mein**	**Mutter**			**machmal**	**tiffst**		**Freundin**	**.**
AUT [S]		s7								
AUT [pos]	$.	PPOSAT	NN			ADJD	VVFIN		NN	$.
AUT [lemma]	.	mein	Mutter			machmal	tiffst		Freundin	.
AUT [ZH]	**.**	**Meine**	**Mutter**	**trifft**	**sich**	**manchmal**		**mit**	**Freundinnen**	**.**
AUT [ZHDiff]		CHA		INS	INS	CHA	DEL	INS	CHA	
AUT [ZHS]		s7								
AUT [ZHpos]	$.	PPOSAT	NN	VVFIN	PRF	ADV		APPR	NN	$.
AUT [ZHlemma]	.	mein	Mutter	treffen	sich	manchmal		mit	Freundin	.
AUT [FehlerOrth]						WS	WS			
AUT [FehlerMorph]		Flex					Flex			
AUT [FehlerSyn]				StV–	ValV		StV	ValV0		
AUT [FehlerLex]										
AUT [FehlerSem]									Num	

Abb. 2: Beispiel einer Fehlerannotation

CDLK-basiertes Schreibfeedbacksystem: Dr. Write

Dr. Write ist im Wesentlichen ein domänenspezifisches Sprachmodell, das für Feedback zum Schreiben auf Deutsch entwickelt wurde. Drei zentrale Merkmale unterscheiden es von einem allgemeinen Sprachmodell. Erstens wurde es auf einem Korpus von nahezu 3.800 fehlerannotierten Aufsätzen chinesischer Deutschlernender trainiert, was die Spezifität und Relevanz des Feedbacks erheblich steigert. Zweitens verfolgt es einen indirekten Feedback-Ansatz, der auf sokratischer Fragetechnik basiert und die Lernenden dazu anregt, über ihre Schreibfehler zu reflektieren und diese eigenständig zu korrigieren. Drittens bietet es drei Feedbackstufen (schwach, mittel, stark), die sich am jeweiligen Kompetenzniveau der Lernenden orientieren und somit individualisierte Lernbedürfnisse berücksichtigen.

Abb. 3: Beispielinteraktion mit Dr.Write

Verwendung von CDLK bei der L2-Erwerbsforschung

Beispielstudie 1: Zur syntaktischen Komplexität des Schriftdeutschen chinesischer Deutschlerner(innen) - Eine korpusbasierte Profilanalyse. (Wu/Li 2022)
Beispielstudie 2: Negativer syntaktischer Transfer im Schriftdeutschen als Tertiärsprache von chinesischen Lernenden - Eine korpusbasierte Querschnittstudie. (Li/Zhang 2024)
Beispielstudie 3: Zum Nachfelderwerb im Schriftdeutschen bei chinesischen Deutschlernenden - Eine korpusgestützte longitudinale Untersuchung. (Huang/Li 2024)
Beispielstudie 4: Gebrauch deutscher Präpositionen bei chinesischen Deutschlerner/-innen im Rahmen der konzeptuellen Metapherntheorie – am Beispiel von um und in im Vergleich zu deutschen Muttersprachler/-innen und Deutschlerner/-innen mit anderen Erstsprachen. (Li/Zhao 2024)
*Für weitere Details scannen Sie bitte den QR-Code.