Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 3, Nummer 1 (Juli 2023)
ISSN 2749-4411

VIELKO
Vietnamesisches Lernerkorpus

Thi Bao Van Ho
Universität Leipzig / University of Languages and International Studies – VNU

Abstract

Das *Vietnamesische Lernerkorpus* (*VieLko*) wird seit 2017 im Rahmen eines laufenden, gemeinsamen Projekts zwischen zwei Hochschulen in Vietnam entwickelt und gilt als die erste landesweit, systematisiert aufgebaute Datensammlung mit sprachlichen Produkten von vietnamesischen Studierenden des Deutschen als Fremdsprache (DaF). Das Korpus umfasst sowohl geschriebene als auch gesprochene Lernerdaten: Die geschriebenen Daten wurden bisher mit vier Annotationsarten versehen, die gesprochenen Daten bleiben stets in ihrer digitaler Rohform. Seit Juni 2022 steht *VieLko* externen Forschenden sowie Interessierten mit beschränkter Zugänglichkeit zur Verfügung: Mithilfe einer lokal gespeicherten Kopie des Korpus haben Nutzer*innen die Möglichkeit, Konkordanzen in Bezug auf folgende Aspekte zu erstellen und auszusuchen: Tokenisierung, Lemmatisierung, Wortart, erste Zielhypothese. In der nächsten Entwicklungsphase (2023–2024) wird angestrebt, den freien Zugang zu den vorhandenen Korpusdaten auf einem Repositorium zu ermöglichen, sowie den Umfang des Datenbestandes zu erweitern.

Keywords: Korpus; Lernerkorpus; Deutsch als Fremdsprache; Lernersprache

Abstract

The *Vietnamese Learner Corpus* (*VieLko*) came into development in 2017 as a result of an on-going, joint project between two universities in Vietnam and is the first nationwide, systematically established database with data from Vietnamese students learning German as a Foreign Language. The corpus comprises not only written, but also spoken data: The written data have undergone linguistic processing and now contain four types of annotation, while the spoken data are left intact in their digital raw form. Since June 2022, *VieLko* has been, with limited access, available to the public through the use of a local copy that enables users to create and search for concordances with regard to the following aspects: tokenisation, lemmatisation, parts of speech, and first level target hypothesis. In the next phase of development (2023–2024), the existing corpus data will be published on a repository with free access to external users, while new data will also be collected to widen the scope of the corpus.

Keywords: corpus; learner corpus; German as a Foreign Language; learner language

1. Über das Projekt

Das Lernerkorpus *VieLko* entstand im Rahmen des DAAD-geförderten bilateralen *Germanistischen Partnerschaftsprogramms* (*GIP*) zwischen zwei Hochschulen in Vietnam – nämlich der ULIS[1] und der HANU[2] – und der Universität Leipzig sowie der Justus-Liebig-Universität Gießen. Bei *VieLko* handelt es sich um die parallele Entwicklung von zwei Teilkorpora – nämlich dem ULIS-Teilkorpus und dem HANU-Teilkorpus, die dieselbe Infrastruktur teilen, aber einzelne Unterschiede bezüglich des Umfangs der Primärdaten sowie des Anteils der annotierten Daten aufweisen. Beide Teilkorpora verfügen über ein sogenanntes ‚Kernkorpus', das ausschließlich aus schriftlichen Daten besteht und u. a. auf der Ebene der ersten Zielhypothese annotiert wird (s. Abbildung 1).

Eine Übersicht des Datenbestands beider Teilkorpora befindet sich im Folgenden:

[1] University of Languages and International Studies – Vietnam National University, Hanoi, Vietnam.
[2] Hanoi University, Hanoi, Vietnam.

		Bachelor										Master	
		Prüfungstext							VOR	HA	BA	HA	MA
		A1	A2	B1	B2	C1	DOL	ÜT					
ULIS-Teilkorpus	**M**	50	50				100		25				
	S		**50**	**50**	**25**			**60**	40		15	60	30
			KERNKORPUS										
HANU-Teilkorpus	**S**	130	80	**180**	130	130		355		35	33		

Abkürzungen - M: mündliche Daten, S: schriftliche Daten

Prüfungen - A1, A2, B1, B2, C1: Abschlussprüfungen zu den GER-entsprechenden DaF-Kursen; DOL: Abschlussprüfungen in Fachseminaren im Bereich Dolmetschen; ÜT: Abschlussprüfungen in Fachseminaren im Bereich Übersetzen

Weitere Prüfungsleistungen - VOR: Vortrag in Fachseminaren; HA: Hausarbeit in Fachseminaren; BA: Bachelorarbeit; MA: Masterarbeit

Abbildung 1
Übersicht über die Daten in beiden Teilkorpora von *VieLko*

2. Datenbestand

2.1 Metadaten

Bei der Erhebung der *VieLko*-Metadaten wurde besonders auf den Bildungshintergrund bzw. auf sprachbiographische Daten der Studierenden geachtet. Dabei wurden folgende Arten von Metadaten erhoben:

1. Allgemeine Daten zum Bildungshintergrund: Jahrgang, Studienrichtung, Institution, Art des Abschlusses, Studierendenmotivation
2. Allgemeine Daten zur Sprachbiographie: Muttersprache (L1) und Fremdsprachen (der Reihenfolge des Erwerbs nach jeweils gekennzeichnet als L2.1, L2.2, L2.3)
3. Spezifische Daten zum Fremdsprachenerwerb: Beginn des Spracherwerbs (an der Schule bzw. an der Universität), durchschnittliche Stundenzahl des Sprachunterrichts auf jeder Bildungsebene (an der Schule bzw. an der Universität), Stundenzahl des Sprachunterrichts bei Muttersprachler*innen (falls vorhanden);
4. Spezifische Daten zum DaF-Erwerb: Ort und Dauer von Aufenthalten im deutschsprachigen Raum (falls vorhanden).

Neben diesen Lerner-Metadaten wurden auch Daten zu den erhobenen Texten gesammelt, die auch in den annotierten Dateien mitkodiert wurden und daher aktuell in der frei zugänglichen Version des Korpus lesbar und durchsuchbar sind. Eine dritte Art von Metadaten ist im Projekt lediglich zur internen Nutzung archiviert, und beinhaltet Informationen zu den verantwortlichen Projekt-Mitwirkenden sowie Einzelheiten zum Format der Prüfungen.

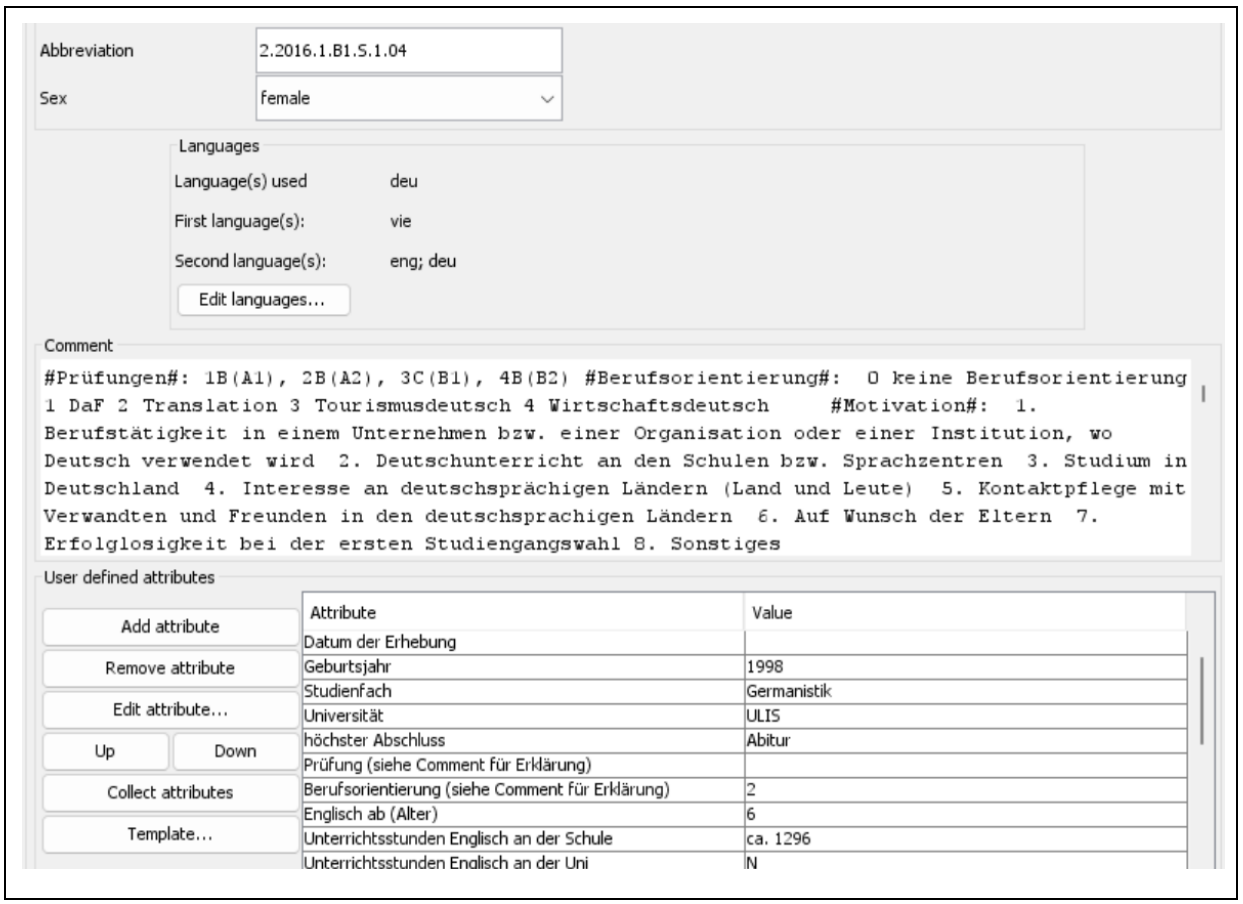

Abbildung 2
Beispiel für die Kodierung von Lerner-Metadaten in *EXMARaLDA* (*Dulko*)

2.2 Primärdaten

Die Primärdaten in *VieLko* sind vollständig digitalisiert und beinhalten Prüfungstexte sowie bewertete Prüfungsleistungen von vietnamesischen Studierenden, die ein Germanistikstudium an der ULIS bzw. an der HANU erfolgreich abgeschlossen haben. Zum Zweck der Datenerhebung haben sich die zwei Projekt-Teams auf eine Auswahl an Fachbereichen sowie Prüfungsarten geeinigt, die u. a. an beiden Hochschulen vorhanden sind und des Weiteren eine möglichst repräsentative Übersicht über den sprachlichen Erwerbsprozess der Proband*innen darstellen kann (s. Abbildung 1): Die Daten in den Bereichen A1-B2 umfassen Texte aus den Abschlussprüfungen in den Sprachkursen, die sich nach den entsprechenden GER-Sprachniveaus richten[3] und in den ersten vier Semestern des Bachelorprogramms an beiden Hochschulen zu besuchen sind. Ab dem fünften Semester beteiligen sich Studierende je nach Fachrichtung an verschiedenen Seminaren in den Bereichen Linguistik, Translation, Literatur- und Kulturstudien – und zusätzlich noch an einem Sprachkurs zum Zielniveau C1 an der HANU, wo sie weitere Prüfungsleistungen wie Vorträge, Hausarbeiten, Portfolioarbeiten oder Klausuren abschließen müssen. Am Ende des Bachelorprogramms haben Studierende die Wahl, eine Abschlussarbeit zu schreiben. Im Gegensatz dazu gibt es im Master-Bereich eine geringere Anzahl

[3] Dass sich diese Daten als A1-Texte, A2-Texte usw. bezeichnen lassen, bezieht sich direkt auf die Benennungen der Sprachprüfungen, bei denen die Texte erhoben wurden, und dient auf keinen Fall als Beweis, dass die Proband*innen zur Zeit der Erhebung die entsprechenden Niveaus erreicht haben.

von Leistungen: Als Korpusdaten wurden lediglich Hausarbeiten in den Fachseminaren und die obligatorischen Masterarbeiten erhoben. Diese Daten wurden wegen struktureller Unterschiede des Masterstudiengangs sowie der Vielzahl externer Studienbewerber*innen nur als Querschnittdaten erhoben, während im Bachelor-Bereich die Erhebung von sowohl Querschnitt- als auch Längsschnittdaten gelungen ist.

Die *VieLko*-Primärdaten unterscheiden sich in einer weiteren fundamentalen Eigenschaft: So sind manche Daten gesprochensprachlicher Kommunikation entnommen, bei anderen handelt es sich um schriftliche Textdaten. Mündliche Daten sind nur im ULIS-Teilkorpus vorhanden und wurden in Prüfungen im DaF-Bereich (als Paarprüfung mit Aufgaben im Format des Goethe-Zertifikats), in Dolmetsch-Prüfungen (als Einzelprüfung mit Aufgaben zum simultanen Dolmetschen), sowie in linguistischen Seminaren (als Einzelleistung beim Gruppenvortrag) aufgenommen. Wegen technischer Beschränkungen wurden diese in den vorgegangenen Entwicklungsphasen des Projekts außer Betracht gelassen. Die Daten bleiben daher aktuell in ihrer digitalen Rohform als *mp3*-Dateien, d. h. sie wurden weder transkribiert noch annotiert, und enthalten keine Metadaten.

Bei dem schriftlichen Datenbestand wurden alle Prüfungstexte vom Original abgescannt und zu *txt*-Dateien in *UTF-8* digitalisiert. Hausarbeiten und Abschlussarbeiten wurden als *PDF*- oder *Word*-Dateien erhoben. Persönliche Daten und Eigennamen von Personen wurden vollständig anonymisiert. Zur weiteren Aufbereitung und Annotation wurde ein Ausschnitt der gesamten Daten für das Kernkorpus gewählt. Seitens der ULIS wurde eine gemischte Sammlung von Quer- und Längsschnittdaten aus dem DaF-Bereich (A2, B1, B2) und aus dem Übersetzen-Bereich ausgewählt, während sich das HANU-Team für den ganzen Bestand seiner B1-Texte entschieden hat (s. Abbildung 1). Da die Mehrheit dieser Texte aus verschiedenen Prüfungen stammte, lässt sich dabei eine Vielzahl an Textsorten finden: Briefe bzw. E-Mails, Blogkommentare, Forumsbeiträge (bei den DaF-Texten); Zeitungsartikel, Gebrauchsanweisungen usw. (bei den Übersetzungstexten).

	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
2.20... [word]	**weil**	**ich**	**eine**	**Fiebe**	**habe**	**und**	**ich**	**kann**	**nichts**	**machen**		**,**	**deshalb**	**kann**	**ich**	**nicht**	**meine**	**Arbeit**	
2.20... [S]																			
2.20... [pos]	KOUS	PPER	ART	NN	VAFIN	KON	PPER	VMFIN	PIS	VVINF		$,	PAV	VMFIN	PPER	PTKNEG	PPOSAT	NN	
2.20... [lemma]	weil	ich	eine	Fiebe	haben	und	ich	können	nichts	machen		,	deshalb	können	ich	nicht	mein	Arbeit	
2.20... [ZH]	**weil**	**ich**		**Fieber**	**hatte**	**und**	**ich**		**nichts**	**machen**	**konnte**	**,**	**deshalb**	**konnte**	**ich**		**meine**	**Arbeit**	**nicht**
2.20... [ZHDiff]			DEL	CHA	CHA			DEL			INS			CHA		MOVS			MOVT
2.20... [ZHS]																			
2.20... [ZHpos]	KOUS	PPER		NN	VAFIN	KON	PPER		PIS	VVINF	VMFIN	$,	PAV	VMFIN	PPER		PPOSAT	NN	PTKNEG
2.20... [ZHlemma]	weil	ich		Fieber	haben	und	ich		nichts	machen	können	,	deshalb	können	ich		mein	Arbeit	nicht

Abbildung 3
Beispielannotation für die Daten im Kernkorpus

Richtlinien zur Datenaufbereitung und -annotation des Kernkorpus wurden in Anlehnung an die fehlerannotierten Korpora *MERLIN* (vgl. Boyd et al. 2014) und *FALKO* (vgl. Reznicek et al. 2012; Hirschmann et al. 2022: 142-144) entworfen. Mithilfe des Korpuswerkzeugs *EXMARaLDA* (*Dulko*) (vgl. Hirschmann / Nolda 2019) wurden die Daten zweifach annotiert:

1. Die vier ersten Annotationsspuren erfolgten automatisch anhand von *Dulko*-Transformationen. Das sind: Tokenisierung, Lemmatisierung, Satzspanne und POS-Tagging (*parts of speech tagging*). Diese Annotationsebenen (insbesondere die POS-Annotation) wurden nachgeprüft, um Fehler des automatischen Taggers vor dem nächsten Schritt der Annotation zu beheben.
2. Die Annotation der ersten Zielhypothese (ZH1) – in Anlehnung an das Konzept der minimalen Zielhypothese (vgl. Lüdeling et al. 2008: 70; Reznicek 2012: 38-39) – wurde manuell anhand von mehreren Annotationsspuren in *EXMARaLDA* (*Dulko*) durchgeführt. Als Triangulationsmaßnahme wurde jeder

originale Text von zumindest drei Personen behandelt: Die ZH1-Annotation wurde zuerst von zwei vietnamesischen Mitwirkenden gemeinsam mit einem/einer muttersprachlichen Interrater*in erstellt, und schließlich mit dem ganzen Team besprochen, bevor die Endversion in *EXMARaLDA* (*Dulko*) eingetragen werden durfte. Dann wurde zusätzlich dazu die ZH-Diff-Spur hinzugefügt, die die Änderungen der ersten Zielhypothese im Vergleich zum Originaltext zeigt (s. Abbildung 3).

3. Korpusnutzung und Weiterentwicklung

Die bisherigen Entwicklungsphasen von *VieLko* wurden als Vorbereitung angedacht, daher ließ sich der Fokus des Korpusaufbaus auf die Einrichtung grundlegender Infrastruktur sowie auf die Pilotierung des Annotationsschemas legen. Weil die bereits annotierten Daten nur einen kleinen Anteil der gesamten Primärdaten ausmachen, und zumal das Annotationsschema selbst noch weiterer Überprüfung bzw. Verfeinerung bedarf, schien es für beide Projektteams nicht das vornehmliche Ziel zu sein, das Korpus in seinem aktuellen Zustand online bereitzustellen. Stattdessen wurde eine lokale Kopie mithilfe des *EXMARaLDA*-Begleitprogramms *COMA* erstellt, die auf Anfrage an Interessierte als Ganzes oder je nach individuellen Bedürfnissen distribuiert werden kann.

RegEx (Annota... | Annotation: pos | Regex: (PTK|VVI)ZU

#	S	Speaker	Left Context	Match	Right Context	pos
1	✓	2.2016.5.ÜF.S.1.28	ve Veränderung , um Bedarf von der neue Phase der Entwicklung	abzudecken	. (5) Hanoi ist mit Bürogebäuden und Handelszentrum als ein	VVIZU
2	✓	2.2016.1.B1.SBr.1.19	eit zum Thema " Globalisierung und Multikulturalität " später	abzugeben	. Letzte Woche war ich 6 Tage im Krankenhaus , weil ich eine	VVIZU
3	✓	2.2016.1.B1.SBr.1.19	eit zum Thema " Globalisierung und Multikulturalität " später	abzugeben	. Letzte Woche war ich 6 Tage im Krankenhaus , weil ich eine	VVIZU
4	✓	2.2016.1.B1.SBr.1.33	rank . Ich muss sofort nach Vietnam fliegen , um meine Mutter	aufzupassen	. Danke für Ihre Verständnis . Ich wünsche Ihnen einen schöne	VVIZU
5	✓	2.2016.5.Ü.S.1.04		stattzufinden	. Es gab im	VVIZU
6	✓	2.2016.5.Ü.S.1.04		stattzufinden	. Es gab im	VVIZU
7	✓	2.2016.1.B2.S.1.32	bitte sagen , ob das benötigt ist , um eine ihre Sprachreise	teilzunehmen	? Mit freundlichen Grüßen Vorname	VVIZU
8	✓	2.2016.1.B2.S.1.28	und , meinen Kindern nicht erlauben zu würden , an eine Reise	teilzunehmen	. Abschließend möchte ich darauf hinweisen , dass die Sprache	VVIZU
9	✓	2.2016.1.B1.SFb.1.30	um Beruf	zu	finden . Heutzutage ist	PTKZU
10	✓	2.2016.1.B2.S.1.42	/-innen . Es ist eine Chance , um junge Erwachsene im Ausland	zu	gehen können . Ich habe gelesen , dass es eine Sprachreisen i	PTKZU
11	✓	2.2016.1.B2.S.1.24	kurs in meiner Heimatstadt teil , anstatt im Ausland Sprachen	zu	lernen . Es ist mit Sicherheit so , dass Lernen im Ausland vi	PTKZU
12	✓	2.2016.1.B2.S.1.24	während ich jetzt keine Bedingungen habe , Lernen im Ausland	zu	erfüllen . Trotzdem möchte ich noch einmal wiederholen , dass	PTKZU
13	✓	2.2016.1.B2.S.1.24	nen , um gut z.B. mit den Mitarbeitern , Kunden , Vorgesetzen	zu	kommunizieren . Außerdem wäre es mit Sprachen einfacher , die	PTKZU

Abbildung 4
Korpussuche in *EXAKT* (*EXMARaLDA*) nach Infinitivkonstruktionen im gesamten Kernkorpus

Um diese Kopie des Korpus zu durchsuchen, könnten Nutzer*innen wieder ein Programm des *EXMARaLDA*-Toolsets verwenden (s. Abbildung 4): Bei der Suche nach allen Infinitivkonstruktionen anhand einer *RegEx*-Anfrage auf der POS-Ebene ergeben sich z. B. 140 Treffer, unter denen nur eine kleine Minderheit Kollokationen mit Präfixverben umfasst. Trotz begrenzter Suchmöglichkeit bietet *EXAKT* grundlegende Funktionalitäten eines Korpusrecherchetools und wird daher für Nutzer*innen mit geringen Erfahrungen in der Korpuslinguistik empfohlen (Die lokale *VieLko*-Kopie wird zusammen mit Anleitungen für dieses Tool verteilt).

Für komplexere Suchverfahren müssten die Korpusdaten manuell exportiert und anhand von weiteren Softwares verarbeitet werden. Empfehlenswert sind u. a. der Texteditor *Notepad++* (wegen dessen vollständiger Unterstützung von *RegEx*-Suchen), oder das Korpusrecherchetool *AntConc*, das neben der Erstellung von Konkordanzen und *RegEx*-Kompatibilität weitere analytische Funktionen wie Konkordanzenplot, *n-gram*-Suche oder Wortliste ermöglicht. In Abbildung 5 lassen sich zwei Beispiele für solche Suchanfragen finden. Beide Suchanfragen beziehen sich auf die ZH1-Diff-Spur (vgl. Abbildung 3): Im Beispiel 5a wird nach allen Artikeln gesucht, die bei der Erstellung der Zielhypothese weggelassen wurden. Im Beispiel 5b wird anhand von einem Konkordanzenplot die Frequenz von syntaktischen Änderungen in der Zielhypothese gezeigt.

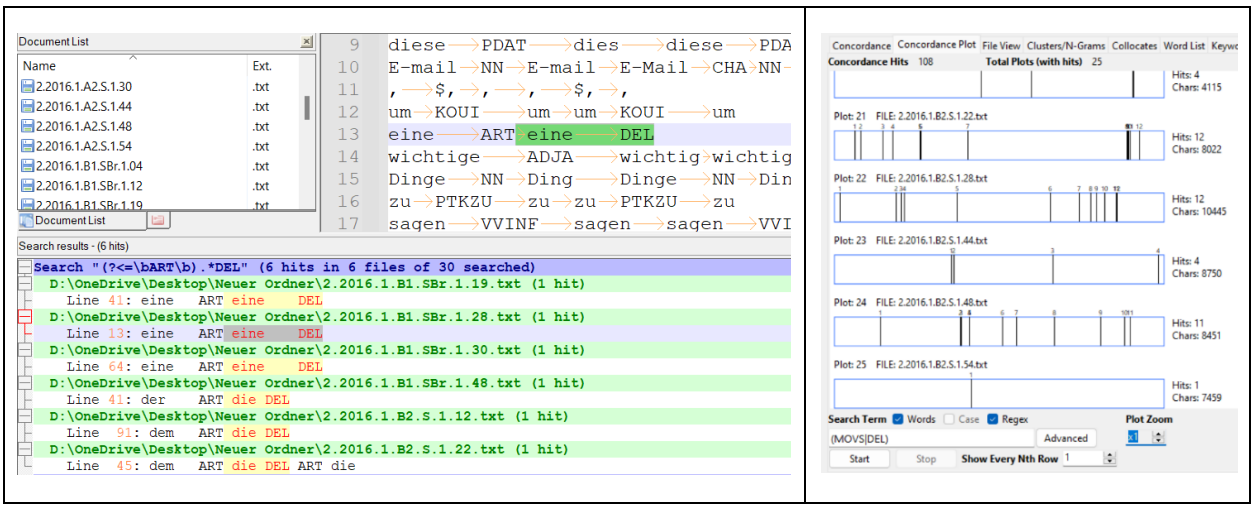

Abbildung 5
Komplexe Suche der Längsschnittdaten anhand von *Notepad++* (5a, links) und *AntConc* (5b, rechts)

Für die nächste Phase des Projekts (2023–2024) ist geplant, den Umfang des Kernkorpus zu erweitern und dessen annotierte Daten, samt Metadaten frei zugänglich auf einer Online-Plattform zu veröffentlichen, um die Nutzerfreundlichkeit beim Umgang mit dem Korpus zu erhöhen. Erwünscht sind u. a. auch der Einsatz eines Versionisierungssystems und die Erhebung neuer Primärdaten.

Literatur und Ressourcen

Boyd, Adriane et al. (2014): The MERLIN corpus: Learner Language and the CEFR. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), European Language Resources Association (ELRA). Reykjavik: European Language Resources Association (ELRA), 1281-1288.

Hirschmann, Hagen / Nolda, Andreas (2019): Dulko – auf dem Weg zu einem deutsch-ungarischen Lernerkorpus. In: Eichinger, Ludwig / Plewnia, Albrecht (Hrsg.): *Neues vom heutigen Deutsch: Empirisch – methodisch – theoretisch*. Institut für Deutsche Sprache: Jahrbuch 2018. Berlin u.a.: de Gruyter, 339-342.

Hirschmann, Hagen / Lüdeling, Anke / Shadrova, Anna / Bobeck, Dominique / Klotz, Martin / Akbari, Roodabeh / Schneider, Sarah / Wan, Shujun (2022): FALKO. Eine Familie vielseitig annotierter Lernerkorpora des Deutschen als Fremdsprache. In: *Korpora Deutsch als Fremdsprache* 2: 2, 139-148. https://doi.org/10.48694/kordaf.3552.

Lüdeling, Anke / Doolittle, Seanna / Hirschmann, Hagen / Schmidt, Karin / Walter, Maik (2008): Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache* 2 (2008), 67-73.

Reznicek, Marc / Lüdeling, Anke / Krummes, Cedric / Schwantuschke, Franziska / Walter, Maik / Schmidt, Karin / Hirschmann, Hagen / Andreas, Torsten (2012): *Das Falko-Handbuch. Korpusaufbau und Annotationen*. Version 2.01. HU Berlin. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/ (16.07.2023).

Biographische Notiz: Ho, Thi Bao Van ist seit 2017 im ULIS-Korpusteam als Projektassistentin tätig. Seit 2020 promoviert sie am Herder-Institut der Universität Leipzig zum Thema „Kohäsion und Kohärenz in Abschlussarbeiten vietnamesischer Studierenden“, wo sie auch mit Daten aus *VieLko* arbeitet.

Kontaktanschrift:
Thi Bao Van Ho
Universität Leipzig
Arno-Nitzsche-Str. 40
542, 04277 Leipzig
Deutschland
vanhtb@ulis.vnu.edu.vn / van.ho@uni-leipzig.de