DUO_DE A1: An Annotated Corpus of Online Learning Material for Beginning Learners of German as a Foreign Language

Jammila Laâguidi[1], Vitaliia Ruban[1], Ronja Laarmann-Quante[1], Anastasia Drackert[2,3]
Ruhr University Bochum, Germany
[1]Faculty of Philology, Department of Linguistics
[2]Faculty of Philology, Institute for German Language and Literature
[3]g.a.s.t. (Society for Academic Study Preparation and Test Development)

Abstract

This paper describes the creation of DUO_DE A1, a corpus based on A1-level learning material from the Deutsch-Uni Online (DUO) language courses for German as a foreign language. We split the material into small segments and manually annotated each with fine-grained information such as the type of segment (e.g. task description, description of grammar), the medium (e.g. text, table, audio), the text units it contains (e.g. words, phrases, sentences) and other special features (e.g. marking cloze texts). Furthermore, we automatically tokenized, POS tagged and lemmatized the corpus and compared the performance of three models on these steps for different kinds of segments. We publish the created corpus in a manner that respects copyright, releasing all structural features, metadata and POS tags.

Keywords: German as a foreign language, corpus creation, learning material, beginning learners, Common European Framework of Reference

1. Introduction

When learning a foreign language, structured material like traditional textbooks or online learning material typically plays an important role. Especially in the beginning of the learning process, the learning material to a large degree determines what vocabulary and grammatical structures the learners are exposed to. Analyzing the learning material is thus of great interest to foreign language acquisition research, see e.g. the comprehensive overview of studies on the language in English as a foreign language (EFL) textbooks in Le Foll (2023). However, such studies can typically not be replicated because code and data are not made available, often for copyright reasons (Le Foll, 2023, p.66).

The goal of the present paper is to contribute to transparency in corpus linguistic research on language learning material. We present DUO_DE A1, a corpus compiled from an online course for learning German as a foreign language (GFL) on level A1 of the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001). While the texts themselves cannot be published for copyright reasons, we release all structural features, metadata and linguistic annotations (part of speech tags).

Preparing the learning material for corpus linguistic research requires a careful consideration of several aspects. For example, the material does not only consist of coherent standard German text. There are also e.g. word lists or tables or tasks that present incomplete sentences for didactic purposes (like cloze texts) or that require the learner to correct the word order. Such elements need to be identifiable in the corpus, as they might distort certain analyses e.g. regarding collocations or syntactic dependencies. Furthermore, the material does not only contain the primary learning content but also e.g. task instructions or metalinguistc information about grammar. These elements should be clearly distinguishable as they may play different roles depending on the research question.

In this paper, we describe how we addressed such aspects. We split up the learning material into small segments and manually annotated them with various information about the type of text (e.g. task instruction or grammar description), the medium (e.g. text, table or audio), the text units it contains (e.g. words, phrases or sentences) and other special features like marking cloze texts (Sec. 3). Furthermore, we automatically tokenized the texts and added part of speech (POS) tags

Proceedings of Leveraging Derived Text Formats to Unlock Copyrighted Collections for Open Science (DTF) @ LREC 2026, pages 51–62
12 May 2026.

and lemma information. We compare the performance of three models on these tasks for different types of segments (Sec. 4). We hypothesize that segments containing coherent sentences would be processed more accurately than those consisting of isolated words or phrases (see also Volodina et al., 2014). Finally, we present an analysis of the composition of the whole corpus (Sec. 5) and describe the JSON format in which it is stored (Sec. 6). For copyright reasons, the word forms and lemmas can only be made available for research purposes upon request. All other parts of the corpus including structural information, metadata and POS tags are made publicly available under a CC BY-NC-SA 4.0 license. The whole corpus titled *DUO_DE A1: An Annotated Corpus of A1-Level Learning Material from the Deutsch-Uni Online Courses for German as a Foreign Language* can be accessed via the following link: https://doi.org/10.5281/zenodo.19113347 (Laâguidi et al., 2026).

2. Related Work

Corpus-based studies on GFL learning material have typically worked with corpora compiled for the purpose of the particular study without sharing the data or annotations. The data is often pruned according to the specific research questions as the following examples show. Furthermore, a detailed description of the creation of the corpus is rarely the focus of the researchers.

Bautista Zambrana (2018) studied phraseological units in learning material. Their corpus consisted of one GFL textbook and workbook on A1 level, which they divided into three subcorpora (written, oral and exercises). They excluded single word forms, morphological units, task instructions, grammar reference sections and vocabulary lists. The remaining corpus comprised 20,806 tokens.

Behnke (2023) investigated how GFL textbooks deal with language change phenomena. They compiled a corpus from five GFL textbook series across CEFR levels A1 to C1. They excluded task instructions and tasks that explicitly targeted one of the phenomena under investigation. The paper does not provide information about the number of tokens.

A different example is the DAFLex project (François et al., 2021). They used a corpus of GFL textbook reading activities and simplified readers to build a CEFR-graded lexicon of receptive vocabulary. The texts were lemmatized and POS tagged, resulting in a lexicon of 41,646 lemma-POS pairs. While the corpus is not public, there is an online tool that allows users to check the frequency of a given word according to the CEFR levels. Furthermore, it can analyse a text and assign a CEFR level to each word (François et al., 2021).

For other languages, there are a few well-prepared and documented textbook corpora. The corpus of Textbook Material (TeMa, Meunier and Gouverneur, 2009) consists of 724,174 words from 32 volumes of English for general purposes coursebooks. It is stored in XML format and comes with a detailed annotation scheme focusing on vocabulary exercises. The corpus itself is not published but sections of it can be accessed for research purposes upon request.[1] The Textbook English Corpus (TEC, Le Foll, 2023) was compiled from 43 EFL coursebooks and comprises 3,023,958 words. It is stored in XML format and all material is annotated for register. While the texts are not available for copyright reasons, all metadata, annotations and code for processing the corpus are published.

A textbook corpus with extensive annotations on various levels is the Corpus of CEFR-based Textbooks as Input for Learner Levels' modelling (COCTAILL, Volodina et al., 2014). It was compiled from twelve Swedish as a foreign language coursebooks and comprises 708,589 tokens. Text passages are annotated for topics and genres and all other material has annotations about target skills, target competences, activity types, activity formats and linguistic units. Furthermore, the data was automatically POS tagged, lemmatized and annotated for syntactic dependencies. While for copyright reasons the corpus is not freely available, for research purposes it can be browsed and parts of the corpus can be downloaded as a bag of sentences upon request.

[1] https://www.uclouvain.be/en/research-institutes/ilc/cecl/tema

3. Data

The DUO_DE A1 corpus consists of data from *Deutsch-Uni Online (DUO)* ('German-University online')[2], a language learning platform for learning German on CEFR levels A1 through C1. Its content focuses on German within a university context. For the DUO_DE A1 corpus, we used all learning materials from the A1 level.[3]

3.1. Structure of the Material

In the following, we describe the internal structure of the learning material. The A1 level consists of two **courses** A1.1 and A1.2. Each course comprises six chapters that each represent a **scenario** such as *Ein Kochabend mit Freunden* 'An evening of cooking with friends' (A1.1. ch. 3). Each chapter is divided into six **phases**. Each of these phases consists of multiple **learning activities** that deal with an overarching topic, e.g. how to use the verb *(to) like* while talking about the seasons (A1.1, ch. 5, phase 2, learning activity 2). Within each learning activity, students are given multiple **tasks** that correspond to the respective topic and have a different focus (grammar, vocabulary, etc.). The kinds of tasks vary greatly and include, for example, cloze texts, listening exercises, or free writing tasks. Each task can be further subdivided into different **segments**, containing different kinds of texts. For example, there can be a task description, further input, e.g. an audio with a transcription, and an editing field for the learner's answer.

3.2. Data Extraction

The data was extracted manually by the first two authors of this paper. We annotated the data based on **segments** of tasks as described above, i.e. the segments serve as our *unit of observation* (Le Foll, 2023, p.76). For the most part, this meant copy-pasting any data that can be read by the learners or transcriptions of texts that were presented as audio. For images containing text that could not be copy-pasted, we typed the text as accurately as possible. For tables, we structured the content into paragraphs for titles and table content.

3.3. Segment-Level Metadata

Each segment is annotated with structural information and some additional information pertaining to the segment as a whole. In this paper and in the published JSON format (see Sec. 6), we refer to this as metadata, which must not be confused with corpus-level metadata such as the language and creators of the corpus, which can be obtained directly from the data repository (Laâguidi et al., 2026). The annotated information comprise the following:

- **course** A1.1 or A1.2
- **scenario** 1-6
- **phase** 1-6
- **learning activity** 1-6 (varies)
- **task** 1-14 (varies),
- **segment**, type of segment, one of
  - title
  - task type
  - situation and instructions
  - task description
  - media
  - editing field
  - further information
  - model solution
  - additional information
  - description of grammar
  - description of learning content
  - list of new expressions
- **medium**, one of
  - text
  - table
  - audio
  - video
- **text unit**, one of
  - words
  - phrases
  - sentences
  - dialogue
  - other (e.g. suffixes)
- **comments**, optional, one or multiple of
  - multiple choice cloze text
  - incomplete cloze text

[2] https://www.deutsch-uni.com/de/

[3] We extracted the data from files from which the content had been entered into the online platform. There may be minor differences between these files and the final online version.

  - intentionally incorrect grammar
  - wrong word order
  - unsegmented
  - model solution
  - partly crossed out
  - duplicate numbering
  - without text
- **Other**, optional, any other comment (free text)

The annotations of the **course**, **scenario**, **phase**, **learning activity** and **task** can be used to locate the segment in the course, e.g. to analyze progression in the material.

The annotation of the type of **segment** follows the inherent structure of the material and allows to distinguish between different kinds of text that a learner encounters. For example, the *editing field* and *media* typically contain the primary learning content of level A1 about a certain topic, for example a cloze text about groceries. The *instructions* or *task description* in turn may contain vocabulary or grammatical constructions that are not targeted with the current task but which are needed to convey the task. Other categories like *description of grammar* or *description of learning content* consist of even more abstract language, e.g. about grammatical categories like *unbestimmter Artikel im Nominativ* ('indefinite article in nominative case').

The **medium** informs about the mode of presentation (*text* vs. *audio* vs. *video*). If text is presented as a table, its layout is not preserved in our corpus but it is annotated with the category *table* so that this information can be considered.

Under **text unit** it is stored whether the segment consists of whole sentences or a dialogue, which can be analyzed syntactically, or if it is only a list of words or phrases, where e.g. dependency parsing may not be meaningful.

We defined a set of annotations stored under **comments** which can be taken into account when further processing the corpus. For example, an *incomplete cloze text* (with gaps such as *Wir ___ viel Spaß!* 'We ___ a lot of fun!') or a multiple choice cloze text (such as *Ich arbeite seit drei Jahren / vor einem Jahr als Ingenieur* 'I have been working as an engineer for three years / one year ago') may present challenges for parsing as the sentences are incomplete or contain superfluous words. There are also sorting tasks which present learners with a wrong word order. Such sentences should not be taken into account when e.g. analyzing co-occurrences of words. When there is no transcription for audio material within a listening task, we include this segment in the corpus to retain the course structure. In this case, it does not contain any tokens but only metadata, in particular the *without text* comment.

4. Automatic Linguistic Annotation

4.1. Gold Standard Annotation

We created an evaluation set with manual gold standard annotations for lemmas and POS tags based on the STTS tagset (Schiller et al., 1999) using the annotation platform INCEpTION (Klie et al., 2018). We aimed at making the evaluation set as diverse as possible, covering all categories for *segment*, *medium*, and *text unit*. We included a wide range of different kinds of learning material from different chapters, such as cloze texts with gaps, word lists both with and without articles, words annotated with grammatical categories or suffixes like *Sg.* (Singular) or *-e*, phrases or isolated words without context as well as dialogues, both extended ones and those consisting of only two sentences. Our evaluation set consists of 843 tokens. The first two authors of this paper independently annotated the texts and subsequently discussed diverging cases for reaching a gold standard. Most cases were clear, only a few tags were debatable, for example *Arbeiten im Semester?*, where *Arbeiten* could be read as an infinitive verb ('Working during the semester') or as a plural noun ('Work during the semester').

4.2. Models

We compare three different models for automatic POS tagging and lemmatization, the **small** and **medium** German models of **spaCy** (Honnibal et al., 2020)[4] and **Stanza** (Qi et al., 2020)[5]. In order to avoid alignment issues due

[4] spaCy v3.8.4, models de_core_news_sm v.3.8.0 and de_core_news_md v.3.8.0

[5] Stanza version 1.10.1

Model	Accuracy
spaCy small	0.886
spaCy medium	0.902
Stanza	0.896
Stanza + APPRART	0.910

Table 1: Overall accuracy for POS tagging.

to differences in tokenization, we use the tokenization from the INCEpTION platform for the evaluation set for all models. Differences in tokenization arise, for instance, because Stanza, trained on Universal Dependencies treebanks, tokenizes contractions of prepositions and definite articles such as *im* (= *in dem* 'in the') as two tokens *in* and *dem*, while spaCy treats them as a single token. For the evaluation, we measure accuracy, precision, and recall on the evaluation set using *scikit-learn* (Pedregosa et al., 2011).

4.3. Evaluation of POS Tagging

The upper part of Table 1 shows the overall accuracy of the three models across all POS tags. Their performance just around .90 is only slightly worse than what Ortmann et al. (2019) report for other registers. While numerically, spaCy (medium) and Stanza perform almost on par, looking at the tagging errors that each model makes reveals some important qualitative differences. We restrict the following discussion to Stanza and the spaCy medium model, which performed slightly better than the small model.

4.3.1. Confusion of POS Tags

For spaCy medium, there are 83 tagging errors in total. The most frequent confusion (12 times) is to tag a normal noun (NN) as a proper noun (NE). Stanza has 88 tagging errors and the confusion of NE for NN only happened seven times. This indicates that Stanza has a broader lexicon because normal nouns such as *HNO* ('ENT specialist'), *Nachhilfelehrer* ('tutor'), *Babysitter* ('babysitter') and *Animateur* ('entertainer') were treated as proper nouns by spaCy, a frequent fallback tag for unknown words, but recognized correctly by Stanza.

The most frequent confusion of Stanza (19 times) is to tag non-words (XY) as punctuation marks ($(). Firstly, this concerns emojis and secondly, we used the tag XY to tag lines indicating gaps in a cloze text. It is very plausible to tag such cases as punctuation marks instead and while it affects Stanza's overall accuracy in this evaluation, in practical applications, these word classes should only play a minor role. spaCy, on the other hand, tags them partly as foreign words (FM) and partly as adjectives (ADJA). This could impact analyses of the learning material in an undesirable way.

4.3.2. Creation of a Combined Model

Based on the previous observations, we decided for Stanza to be more appropriate than spaCy for tagging the DUO_DE A1 corpus. However, as addressed in Sec. 4.2, Stanza treats contracted prepositions and articles such as *im* as two tokens. We find this undesirable because there are differences in meaning between the contracted forms and the split forms and they cannot be used interchangeably (see e.g. Cieschinger, 2016), hence keeping the original form is important. Contracted prepositions and articles always get the POS tag APPRART in the gold standard, which Stanza never assigns because it was not present in its training data. Instead, Stanza would assign the tags for article (ART, 11 times) or preposition (APPR, 1 time) in our evaluation set. Therefore, we decided to create a combined POS tagging model which uses Stanza's POS tags except for tokens where spaCy assigns APPRART, which then overwrites the Stanza tag. This combined model achieves a higher accuracy (0.91) than the individual models, see Table 1.

Figure 1 shows the confusion matrix of its remaining tagging errors and Table 2 the precision, recall and F1 score per POS tag. We can see that besides NE, XY and $(that were discussed above, most of the tags with an F1 score $<$ .90 occurred less than 10 times: ITJ (5), PDAT (1), PDS (1), PIAT (8), PIS (1), PTKANT (2), PWAV (6), VVIMP (1). Adverbs (ADV) only have an F1 score of .81 partly because Stanza tags answer particles (PTKANT) as adverbs, and infinitive verbs (VVINF) are challenging (F1 = .86) because they are of-

XPOS	precision	recall	F1	# toks.
$(	.82	.94	.88	94
$,	1	1	1	24
$.	.97	1	.98	85
ADJA	1	.85	.92	13
ADJD	.92	1	.96	12
ADV	.73	.92	.81	12
APPR	1	.97	.99	37
APPRART	1	1	1	12
ART	1	.95	.97	60
CARD	1	1	1	34
ITJ	0	0	0	5
KON	.94	.94	.94	17
NE	.72	.91	.80	43
NN	.94	.94	.94	160
PAV	1	1	1	1
PDAT	0	0	0	1
PDS	.33	1	.50	1
PIAT	.71	.62	.67	8
PIS	.20	1	.33	1
PPER	.95	1	.97	55
PPOSAT	1	.85	.92	20
PTKANT	0	0	0	2
PTKNEG	1	1	1	2
PTKVZ	1	1	1	1
PWAT	1	1	1	1
PWAV	1	.33	.5	6
PWS	1	1	1	7
VAFIN	1	.87	.93	31
VMFIN	1	1	1	9
VMINF	1	1	1	1
VVFIN	.88	.97	.92	38
VVIMP	0	0	0	1
VVINF	.94	.79	.86	19
VVPP	.88	1	.93	7
XY	1	.09	.16	23
macro avg	.76	.76	.73	843
weighted avg	.92	.91	.90	843
accuracy		.91		843

Table 2: By-tag performance of the combined model of Stanza + spaCy's APPRART.

ten confused with finite verbs which share the same word form.

4.3.3. Impact of Type of Text

Table 3 shows the performance of the models across the different categories for *text unit*, *medium* and *segment*. We hypothesized that the taggers would struggle most with fragmented text segments such as tables or word lists where no complete sentence context is given. When comparing the results for the different *text units*, we see that numerically, all models indeed perform better on sentences/ dialogues and phrases than on isolated words. However, the worse performance for words vs. sentences is only statistically significant for spaCy ($\chi^2(1) = 7.50, p < .01$, one-sided) but not for Stanza ($\chi^2(1) = 0.91, p > .17$, one-sided) nor the combined model ($\chi^2(1) = 2.18, p > .06$, one-sided).

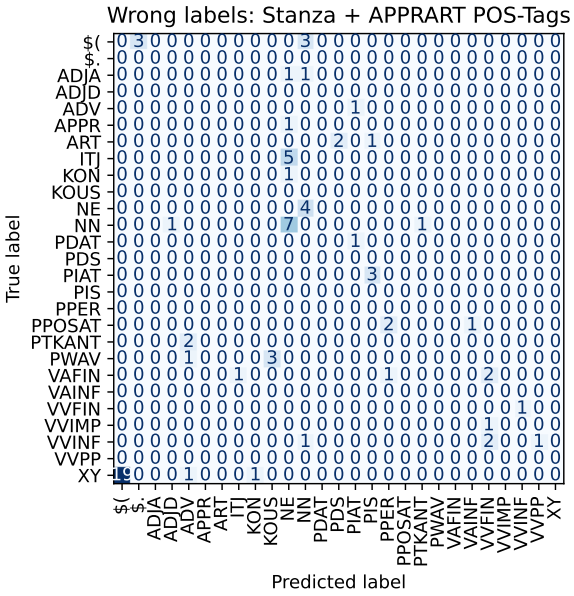

Figure 1: Confusion matrix of POS tags for the combined model of Stanza + spaCy's APPRART. Only wrong assignments are counted.

When comparing the different types of text segments, we see no large differences between them for either of the models, except for a worse performance on the category *further information*. A closer inspection revealed that this was largely caused by presenting isolated word forms in an inflectional paradigm, which are ambiguous without sentence context.

4.4. Evaluation of Lemmatization

For the evaluation of the lemmatization, we focus only on nouns, verbs and adjectives (39.6% of the evaluation set). Other word classes either do not inflect or, e.g. in case of pronouns or determiners, their lemma is usually not of primary interest but rather their grammatical properties. Table 4 shows the accuracy of the three models for nouns, verbs and adjectives.

text unit				
	spaCy	**Stanza**	**Mix**	**#toks**
dialogue	.93	.90	.93	73
sentences	.92	.90	.92	382
phrases	.90	.91	.91	284
words	.82	.86	.87	104
medium				
	spaCy	**Stanza**	**Mix**	**#toks**
audio	.87	.91	.91	23
table	.81	.86	.86	43
text	.91	.90	.91	777
segment				
	spaCy	**Stanza**	**Mix**	**#toks**
title	1.0	.90	.90	10
situation & instructions	.95	.95	1.0	22
task description	.94	.97	.97	35
media	.96	.87	.96	46
editing field	.89	.89	.90	375
hint	.68	.75	.76	76
model solution	.98	.99	.99	102
descr. of grammar	1.0	.95	.95	22
descr. of learning content	.95	.91	.93	94
list of new expressions	.92	.87	.87	61

Table 3: Accuracy of POS tagging per text unit, medium and segment type for spaCy (md), Stanza and the combined model (*Mix*).

POS	**spaCy sm**	**spaCy md**	**Stanza**	**#toks**
Noun	.92	.93	.98	203
Verb	.89	.90	.96	106
Adj.	.84	.92	.88	25
total	.90	.92	.96	334

Table 4: Lemmatization accuracy for nouns, verbs, adjectives and all three word classes.

We can see that the overall lemmatization accuracy is very high, especially for Stanza, where the accuracy of 96% corresponds to only 12 wrongly assigned lemmas. This includes cases where Stanza uses the old spelling (*Schloß* for *Schloss* 'lock') and cases where it indicates an ambiguous lemma (e.g. *Dosis/Dose* 'dose/can' for *Dosen*). spaCy, in contrast, often fails to provide a lemmatized form, sticking with the inflected word form (e.g. *warst* '(you) were' or *Würste* 'sausages') or it seems to apply rules that lead to non-existent word forms, such as **isen* instead of *essen* for the (irregularly inflected) verb form *isst* '(you) eat' or **Wetterlag* for *Wetterlage* 'weather conditions'.

5. Corpus Analysis

Following the results on the evaluation set, the whole corpus was tokenized and split into sentences with spaCy (medium) because it keeps contractions of prepositions and articles as one token. We lemmatized the corpus with Stanza and POS tagged it with the combined model based on Stanza with spaCy's APPRART tags. In total, the DUO_DE A1 corpus consists of 126,142 tokens across 4,084 segments.[6] This number includes punctuation marks and whitespace tokens such as tabs and newlines, which are preserved for some layout information. The following analyses are based on pure words, which we define as tokens that consist only of alphabetic characters and potentially hyphens. The corpus contains a total of 85,680 words. The ten most frequent lemmas are shown in the first column of Table 5.

Table 6 shows how many words belong to each kind of *text unit*, *medium* and *segment*. We can see that coherent text, i.e. sentences and dialogue only make up about 80% of the corpus. While phrases and isolated words showed high POS tagging accuracy as well, this part of the corpus may not be suitable for subsequent syntactic analyses like dependency parsing.

Using the *segment* annotation, we can approximate a distinction of text containing the primary learning content vs. instructional or metalinguistic content, which in the following we refer to as meta language. Meta language, when defined as comprising the segments *task description*, *task type*, *situation and instructions*, *description of learning content* and *description of grammar*, makes up about 35% of the words in the course. The language of the primary learning content (= all other segments) differs from the meta language as shown in Table 5. The table contains the ten most frequent noun, verb and adjective lemmas (according to the automatic annotation) for learning content vs. meta language. For this analysis, we leave out the *task type* because it consists of rather standardized task descriptions such as

[6] Six of these segments do not contain any tokens but were only included for structural reasons, see the *without text* annotation in Sec. 3.3.

overall		nouns				verbs				adjectives			
		content		meta		content		meta		content		meta	
der *the*	7,157	Uhr *clock*	315	Übung *exercise/task*	212	sein *be*	1,702	lesen *read*	379	gut *good*	426	anderer *other*	84
sie *she/you*	4,539	Zeit *time*	139	Tip *hint*	151	haben *have*	844	sein *be*	377	neu *new*	98	gut *good*	72
ich *ich*	2,258	Tag *day*	134	Dialog *dialogue*	139	können *can*	508	hören *listen*	258	alt *old*	85	neu *new*	45
sein *be*	2,179	Jahr *year*	120	Frage *question*	93	gehen *go*	404	passen *pass*	186	schön *pretty*	66	richtig *right*	35
und *and*	2,031	Frau *woman/Ms.*	114	Verb *verb*	75	machen *make*	344	machen *make*	182	schwarz *black*	54	passend *fitting*	20
in *in*	1,668	Hose *pants*	101	Person *person*	75	kommen *come*	267	haben *have*	157	anderer *other*	45	verschieden *different*	16
ein *a*	1,665	Haus *house*	100	Studierende *student*	72	müssen *must*	209	sehen *see*	153	super *super*	43	wichtig *important*	14
haben *have*	1,002	Freund *friend*	86	Bild *picture*	72	wollen *want*	197	sagen *say*	152	klein *small*	42	falsch *wrong*	14
was *what*	998	Zimmer *room*	83	Satz *sentence*	61	finden *find*	191	lernen *learn*	134	kalt *cold*	41	international *international*	12
mit *with*	816	Wochenende *weekend*	79	Gespräch *conversation*	50	mögen *like*	183	können *can*	127	warm *warm*	41	spät *late*	11

Table 5: Most frequent lemmata for primary learning content (*content*) and instructional and metalinguistic content (*meta*).

text unit		
	#words	**%**
sentences	83,146	65.9
dialogue	20,569	16.3
phrases	17,820	14.1
words	4,450	3.5
other	157	0.1
medium		
	#words	**%**
text	71,104	83.0
table	10,137	11.8
audio	4,355	5.1
video	84	0.1
segment		
	#words	**%**
editing field	32,715	38.2
media	14,126	16.5
task description	13,087	15.3
task type	8,528	10.0
situation and instructions	6,605	7.7
hint	4,044	4.7
list of new expressions	2,265	2.6
model solution	1,944	2.3
description of learning content	1,768	2.1
title	361	0.4
description of grammar	179	0.2
additional information	58	0.1

Table 6: Distribution of categories for *text unit*, *medium* and *segment* based on words in the whole corpus.

Verbinden Sie die Elemente 'Connect the elements'. While the *task type* segments consist of 8,528 words, one can find only 131 different lemma types and these would skew the analysis.

6. JSON Format

The corpus is stored in Tabular JSON format (Roussel, 2024). Each segment is represented as one JSON object. Each file represents one chapter/scenario with an array of all JSON objects belonging to this chapter. On the top level, each JSON object consists of an `id`, a `metadata` object, as well as a `tokens` and `sentences` array. The `metadata` object includes the information described in Sec. 3.3, e.g. what course and scenario the text is from or what units of text it contains. Additionally, it indicates how the data was tokenized and annotated with POS and lemmas. The `tokens` array consists of one object per token, which contains the token's id, word form, lemma, and POS tag. The `sentences` array describes the span of each sentence, i.e. from which token to which token each sentence within the current segment spans. The JSON format is made publicly available with the exception of the word forms and lemmas. In this derived text format, copyright is respected as the original

texts cannot be reconstructed from our annotations. A complete example for one segment can be found in Appendix A.

7. Conclusion and Future Work

We presented the creation of DUO_DE A1, a corpus compiled from A1-level learning material from the Deutsch-Uni Online GFL online course. The corpus is enriched with fine-grained metadata characterizing different kinds of text segments and was tokenized, POS tagged and lemmatized with high accuracy. While the texts are copyrighted, we publish all structural features, metadata and POS tags in a structured JSON format in order to contribute to transparency in corpus linguistic research on GFL learning material. The corpus can be used, for example, to analyse the structural composition of the learning material or the progression of POS distributions. In future work, we want to enrich the corpus with further linguistic annotations like syntactic dependency parsing and morphological analysis. We are also planning to process material from other CEFR levels from the Deutsch-Uni Online course in a similar way.

8. Acknowledgments

We gratefully acknowledge the Society for Academic Study Preparation and Test Development (g.a.s.t. e.V.) for providing access to the learning materials from German University Online (DUO) and for supporting this research. Furthermore, we thank the anonymous reviewers for their helpful comments.

9. Bibliographical References

Maria Rosario Bautista Zambrana. 2018. Corpus analysis of phraseology in an A1 level textbook of German as a foreign language. *Quaderns de Filologia - Estudis Lingüístics*, 22:13–32.

Lars Behnke. 2023. Korpuslinguistische Betrachtungen zum grammatischen Wandel in DaF-Lehrwerken. Zwischen Authentizität und Lernbarkeit. *AUC PHILOLOGICA*, 2022(3):11–39.

Maria Cieschinger. 2016. *The contraction of preposition and definite article in German. Semantic and pragmatic constraints*. PhD Thesis, University of Osnabrück.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

Thomas François, Patricia Kerres, Damien De Meyere, and Ferran Suñer Muñoz. 2021. DAFLex: A CEFR-graded lexical resource for German as a foreign language. Presentation at the first Workshop on Building CEFR-graded resources for second and foreign language learning (GR4L2), Louvain-la-Neuve, Belgium.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, USA. Association for Computational Linguistics.

Elen Le Foll. 2023. *Textbook English: A corpus-based analysis of the language of EFL textbooks used in secondary schools in France, Germany and Spain*. PhD Thesis, University of Osnabrück. Publisher: Universität Osnabrück.

Fanny Meunier and Céline Gouverneur. 2009. New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In Karin Aijmer, editor, *Studies in Corpus Linguistics*, volume 33, pages 179–201. John Benjamins Publishing Company, Amsterdam.

Katrin Ortmann, Adam Roussel, and Stefanie Dipper. 2019. Evaluating off-the-shelf NLP tools for German. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 212–222, Erlangen, Germany.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake. Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Adam Roussel. 2024. Tabular JSON: A proposal for a pragmatic linguistic data format. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 166–172, Vienna, Austria. Association for Computational Linguistics.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Universität Stuttgart, Universität Tübingen.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: A pedagogically annotated corpus of coursebooks for Swedish as a second language. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 128–144, Uppsala, Sweden. LiU Electronic Press.

10. Language Resource References

Thomas François and Patricia Kerres and Damien De Meyere and Ferran Suñer Muñoz. 2021. *DAFlex: A CEFR-graded lexical resource for German as a foreign language*. Centre de Traitement automatique du Langage (CENTAL). PID https://cental.uclouvain.be/cefrlex/daflex/.

Laâguidi, Jammila and Ruban, Vitaliia and Laarmann-Quante, Ronja and Drackert, Anastasia. 2026. *DUO_DE A1: An Annotated Corpus of A1-Level Learning Material from the Deutsch-Uni Online Courses for German as a Foreign Language*. Zenodo.

A. Appendix

The following shows a complete example of an annotated segment in JSON format. The segment reads *Günstig oder teuer? Wie ist Ihre Meinung?* ('Cheap or expensive? What is your opinion?'). In the published corpus, the lemma and word form (marked in red) are removed for copyright reasons.


```
{
    "id": "seg324",
    "metadata": {
        "course": "A1.1",
        "scenario": "3",
        "phase": "6",
        "learning_activity": "1",
        "task": "4",
        "medium": "text",
        "segment": "task description",
        "text_unit": "sentences",
        "comments": null,
        "other": null,
        "annotations": {
            "pos": {
                "use": "pos_xpos"
            },
            "token": {
                "type": "property",
                "model": "de_core_news_md
                    ",
                "source": "spaCy"
            },
            "pos_xpos": {
                "type": "property",
                "source": "stanza"
            },
            "lemma": {
                "type": "property",
                "model": "de",
                "source": "stanza"
            }
        }
    },
    "tokens": [
        {
            "id": "t1",
            "form": "Günstig",
            "lemma": "günstig",
            "pos_xpos": "ADJD"
        },
        {
            "id": "t2",
            "form": "oder",
            "lemma": "oder",
            "pos_xpos": "KON"
        },
        {
            "id": "t3",
            "form": "teuer",
            "lemma": "teuer",
            "pos_xpos": "ADJD"
        },
        {
            "id": "t4",
            "form": "?",
            "lemma": "?",
            "pos_xpos": "$."
        },
        {
            "id": "t5",
            "form": "Wie",
            "lemma": "wie",
            "pos_xpos": "PWAV"
        },
        {
            "id": "t6",
            "form": "ist",
            "lemma": "sein",
            "pos_xpos": "VAFIN"
        },
        {
            "id": "t7",
            "form": "Ihre",
            "lemma": "ihr",
            "pos_xpos": "PPOSAT"
        },
        {
            "id": "t8",
            "form": "Meinung",
            "lemma": "Meinung",
            "pos_xpos": "NN"
        },
        {
            "id": "t9",
            "form": "?",
            "lemma": "?",
            "pos_xpos": "$."
        }
    ],
```

```
    "sentences": [
        {
            "id": 1,
            "begin": 1,
            "end": 4
        },
        {
            "id": 2,
            "begin": 5,
            "end": 9
        }
    ]
}
```