GerVLPro: A CEFR-Graded Vocabulary List of L2 Learners' Productive Vocabulary in German

Noah-Manuel Michael, Anna Hülsing, Andrea Horbach
Kiel University, Germany
Leibniz Institute for Science and Mathematics Education, Kiel, Germany
{michael,huelsing,horbach}@ipn.uni-kiel.de

Abstract

CEFR-graded vocabulary lists are a valuable tool for second-language (L2) learners as they provide guidance on the order in which to acquire vocabulary items. Thus, they are essential for informing computer-assisted language learning solutions that target vocabulary development in learners. However, the vast majority of GVLs are prescriptive in that they determine which items learners should learn at each level, and they provide little information about which items learners actually know. Moreover, in the case of German, almost all established GVLs focus exclusively on learners' receptive vocabulary. To remedy this, we introduce **GerVLPro**: A CEFR-**G**raded **V**ocabulary **L**ist of L2 learners' **Pro**ductive vocabulary in **Ger**man. We derived GerVLPro from a comprehensive aggregation of available CEFR-annotated German L2 learner corpora to represent a wide range of learners and contexts. The resulting list comprises 4,015 lemma–POS entries (A1: 611; A2: 1,134; B1: 903; B2: 1,103; C1: 249; C2: 15), assigned via a normalized share-based method. We then conducted a large-scale cross-evaluation against seven established GVLs and six prominent frequency lists. Despite sizable lexical overlap among resources, we found only weak to moderate alignment with GerVLPro. Finally, we investigated whether Gpt-4o and Gpt-5 can reliably grade the productive vocabulary items in GerVLPro. Although both models exhibit roughly similar predictive capacity, they underperform most of the established GVLs on alignment and do not accurately capture productive difficulty. Overall, our findings suggest that established GVLs, frequency lists, and LLM grading insufficiently reflect the trajectory of learners' productive vocabulary, underscoring the need for descriptive, learner-based resources such as GerVLPro.

Keywords: graded vocabulary lists, CEFR, German, productive lexicon, cross-resource evaluation

1. Introduction

Vocabulary acquisition plays an essential role in mastering a second language (Schmitt, 2008). Consequently, NLP applications in the field of computer-assisted language learning (CALL), such as automatic essay scoring, text simplification, and exercise generation, depend on CEFR[1]-graded vocabulary lists (GVLs) (Stowe et al., 2022; Bannò et al., 2025; Li et al., 2025). This is because GVLs provide a basis for determining the sequence in which vocabulary should be acquired and, as such, are vital in informing applications that target the development and evaluation of learners' vocabulary.

However, most established GVLs are derived from curated instructional materials for L2 learners and are therefore prescriptive in nature, specifying which vocabulary items learners are supposed to learn at each proficiency level. Ehara et al. (2012) note that the vast majority of studies focus either on developing methods for measuring the size of learners' L2 vocabulary or on determining the vocabulary items learners are supposed to learn, while few studies examine which words learners actually know and whether they do indeed acquire vocabulary in the order second-language acquisition (SLA) experts recommend. Despite recent efforts for English (Schmitt, 2024), to date, the need for descriptive resources that capture the development of German learners' vocabulary remains unmet.

Moreover, with the exception of the productive Profile deutsch list, established German GVLs focus exclusively on receptive vocabulary. As a result, they specify what learners should understand at each proficiency level, while offering little insight into what learners can actually produce. From a descriptive perspective, it therefore remains an open question whether learners learn to produce vocabulary in the order suggested by GVLs. This question directly affects the development of CALL solutions that target productive vocabulary development.

In an attempt to remedy this lack of a descriptive GVL of German learners' productive vocabulary, we introduce **GerVLPro**: A CEFR-**G**raded **V**ocabulary **L**ist of L2 learners' **Pro**ductive vocabulary in **Ger**man. GerVLPro comprises 4,015 lemma–POS[2] entries (A1: 611; A2: 1,134; B1: 903; B2: 1,103; C1: 249; C2: 15). We derived GerVLPro from a large, aggregated collection of CEFR-annotated German learner texts, totaling approximately 1,390,000 tokens and 5,631 individual texts, and we verified each entry manually. To the best of our knowledge, this is the first productive German GVL derived from such a large dataset.

In addition, we conducted a large-scale cross-

[1] Common European Framework of Reference.

[2] Part-of-speech.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11942–11959
11–16 May 2026.

evaluation of GerVLPro, seven established GVLs, and six prominent frequency lists, and found that none of the existing resources manage to model the development of German learners' productive vocabulary adequately. Finally, we investigated whether Gpt-4o and Gpt-5 can accurately classify the items contained in GerVLPro, thereby assessing large language models' (LLMs) potential as automated tools for estimating word difficulty in productive use, and we found only weak predictive capacity.

Our code and resource are available at https://github.com/noahmanu/gervlpro.

2. Background & Related Work

We now review existing approaches to GVL construction and related research on receptive and productive vocabulary, lexical size estimation, and recent attempts to use LLMs for GVL construction.

2.1. Pragmatic GVL Construction

According to Möhring and Wallner (2013), two strategies for vocabulary selection in GVLs can be distinguished: pragmatic and frequency-based approaches. Pragmatic approaches prioritize words linked to learner-relevant scenarios and are based on the recommendations of SLA experts. Möhring and Wallner (2013) report critiques that in pragmatic approaches, the choice and definition of thematic categories, as well as the allocation of vocabulary to specific topics and proficiency levels, are often insufficiently transparent and lack empirical grounding. In a similar vein, Tschirner (2019) also highlights limitations of pragmatically motivated GVLs. He shows that their inventories contain many low-frequency, concrete words such as *Abfalleimer* "garbage can", which learners seldom have the opportunity to encounter, let alone to use. Furthermore, he points out that pragmatically motivated GVLs also often omit frequent, abstract vocabulary that is indispensable for text comprehension, such as the word *Grundlage* "basis". Tschirner hence argues that pragmatic GVLs are primarily suitable for productive, everyday communication.

2.2. Frequency-Based GVL Construction

Frequency-based approaches emphasize words that appear most often in general language use. Tschirner (2019) therefore argues that frequency-based GVLs are essential for building receptive competence. This approach intuitively seems to make sense, as word frequency strongly predicts word difficulty (Ellis, 2002; Johannsen et al., 2012; Ehara et al., 2012). However, Johannsen et al. admit that this relationship may not always hold in L2 contexts or for people with low literacy. Crucially, the usefulness of frequency-based GVLs depends on the choice of the underlying background corpus. Many traditional frequency lists rely on publicly available corpora that consist predominantly of newspaper texts and other forms of standard written language, which are poorly aligned with learner language and communicative needs. To mitigate this limitation, frequency lists can instead be derived from alternative sources such as movie subtitles or mixed-genre corpora, with the aim of better approximating everyday language use (Brysbaert and New, 2009; Nohejl et al., 2025). Nevertheless, regardless of their source, frequency lists derived from anything other than learner language itself are unlikely to accurately reflect its lexicon.

In addition, numerous studies have proposed more refined metrics than frequency as proxies for word difficulty, such as contextual diversity and dispersion (Adelman et al., 2006; Chen and Meurers, 2016; Gries, 2021). However, no existing German vocabulary lists are based on these measures. Consequently, the frequency lists we consider in our cross-evaluation rely on simple word frequency.

2.3. Receptive vs. Productive Vocabulary

Existing resources and studies disproportionately target receptive vocabulary (see Table 1; François et al., 2014, 2016; Dürlich and François, 2018; Tack et al., 2018), leaving the development of learners' productive vocabulary underexplored beyond the broad notion that comprehension precedes production (Laufer, 1998; Laufer and Paribakht, 1998; Pignot-Shahov, 2012). This is especially true for German, where Profile deutsch is the only established GVL that provides both receptive and productive CEFR grades. This notably stands in contrast to Tschirner (2019)'s observation that pragmatically motivated GVLs are particularly valuable for productive language use. All but one established GVL is pragmatically motivated, yet only one explicitly targets productive vocabulary.

An important step toward exploring learners' productive vocabulary was made by Volodina et al. (2016), who, similarly to our approach, derived word frequency information from a corpus of Swedish L2 learner essays. Unlike a GVL, however, their resource reports word usage across proficiency levels without assigning CEFR grades. To the best of our knowledge, no similar work has been attempted for German. By pooling together as many CEFR-graded German L2 corpora as possible, we aimed to create a GVL that provides a broader representation of learners and writing contexts than could be captured in a single corpus.

2.4. Size of L2 Learners' Lexicons

Despite the abundance of studies that have investigated and attempted to estimate the size of L2

Resource	Modality	Approach	Medium	#Entries						
				A1	A2	B1	B2	C1	C2	Total
DaFLex	Receptive	Frequency	GUI	3,219	3,755	5,667	8,359	7,803	6,059	34,862
Lingster Academy	N/A	Pragmatic	PDF	464	808	1,317	735	32	2	3,358
Aspekte neu	N/A	Pragmatic	PDF	–	–	2,789	2,125	2,054	–	6,968
Profile deutsch	Receptive	Pragmatic	CD	815	1,106	1,243	935	–	–	4,099
	Productive	Pragmatic	CD	464	618	1,244	1,268	–	–	3,594
Goethe Zertifikat	Receptive	Pragmatic	PDF	653	612	1,756	–	–	–	3,021
Telc	Receptive	Pragmatic	PDF	1,103	1,078	1,034	–	–	–	3,215

Table 1: Overview of established German CEFR-graded vocabulary lists. *N/A* indicates that modality is not specified by the source. *#Entries* measured after filtering. References in Appendix A.

Resource	#Entries
dlexDB	1,321,427
DeReWo	317,835
Subtlex-de	135,634
Leipzig Corpora Frequencies Top 100k	56,293
Google Ngrams Top 10k	6,549
Routledge Frequency Dictionary	4,953

Table 2: Overview of prominent German word frequency lists. *#Entries* measured after filtering. References in Appendix A.

learners' lexicons at different CEFR levels, there is no general consensus. Results vary considerably both across and within languages, depending on the setting in which the languages were studied and the learners' first languages, among other factors (Nation and Waring, 1997; Nation, 2006; Milton and Alexiou, 2009; Laufer and Ravenhorst-Kalovski, 2010; Milton, 2010; Tschirner et al., 2020). For German, Tschirner (2019) reports a receptive lexicon size of 4,049 words at the B2 level. We are unaware of any studies quantifying the size of German learners' productive lexicons.

2.5. GVLs Beyond the B2 Level

Apart from the issue of undefined level bands, Glaboniat et al. (2017) call the validity of GVLs beyond the B2 level into question. This is because it is difficult to determine whether vocabulary beyond a certain threshold can still be considered indicative of a proficiency level rather than of individual interests and specialization. Consequently, vocabulary lists for advanced learners (C1–C2) often specialize in certain domains, and it is questionable whether the notion of a general-purpose GVL for levels beyond B2 is meaningful. Hence, in our evaluation, we only consider the levels A1–B2.

2.6. GVLs and LLMs

Recent work has shown that LLMs can perform well on tasks related to lexical difficulty estimation. In the MLSP Shared Task (Shardlow et al., 2024), a Gpt-4-based model achieved the best performance in predicting lexical complexity scores on a Likert scale from 1 to 5 (Enomoto et al., 2024), a task closely related to the classification of vocabulary items by CEFR level. Similarly, Alfter (2024) explored the performance of generative LLMs in a zero-shot setting in predicting the CEFR level of vocabulary items and in automatically generating GVLs. He found that while LLMs perform reasonably well in classifying tokens according to the CEFR level, the automatic generation of GVLs works poorly for languages other than English. Although his analysis includes five common European languages, it does not include German. We build on Alfter's findings by investigating the performance of Gpt-4o and Gpt-5 in predicting CEFR grades for the productive vocabulary items contained in GerVLPro.

3. Survey of Existing Resources

In the following section, we provide a brief survey of established German GVLs and of a number of prominent German frequency lists. We then proceed to provide a comprehensive overview of currently existing and publicly available CEFR-annotated German learner corpora.

Established German GVLs. Table 1 provides an extensive overview of established German general-purpose GVLs. Currently, Profile deutsch is the only established GVL that assigns both receptive and productive CEFR levels. With the exception of DaFLex, all established GVLs follow the pragmatic construction approach. However, DaFLex does not constitute a GVL in the conventional sense, as it does not assign definitive CEFR levels to individual vocabulary items but instead provides information on their distribution across all levels based on a corpus of curated instructional materials. Thus, DaFLex can technically be classified as a hybrid of both approaches to GVL construction. Because of the lack of definitive level assignments, we applied the same level assignment method to DaFLex as we did to GerVLPro (see Section 4.2).

Corpus	Source	CEFR	L1s (>1%)	#Texts	#Tokens
BELDEKO	Strobl and Wedig (2023)	B2	nld	301	71k
CLEG13	Maden-Weinberger (2013)	B1–C1	eng	729	285k
DISKO-L2	Wisniewski et al. (2022)	B2–C1	ara, bul, ces, + 43 others	510	240k
DISKO-WEBTESTDAF	Wisniewski et al. (2022)	A2–C1	ara, bul, fas, + 26 others	479	91k
DULKO	Beeh et al. (2021)	B2–C1	hun	64	21k
FALKOESSAYL2	Lüdeling et al. (2008)	B2–C2	afr, ces, dan, + 34 others	248	145k
FALKOSUMMARYL2	Lüdeling et al. (2008)	C1	bul, eng, fas, + 21 others	106	41k
FALKOWHIGL2	Reznicek et al. (2012)	B2–C2	eng, fra, pol, + 7 others	196	131k
GERSUMCO	Wedig and Strobl (2024)	B2	ara, dan, eng, + 19 others	108	27k
KANDEL	Vyatkina (2016)	A2	eng	688	122k
KOBALT-DAF	Zinsmeister et al. (2012)	B2	bel, rus, swe, zho	51	33k
KOLIPSI-1	Glaznieks et al. (2023)	A1–C1	ita, lld, + others	523	87k
KOLIPSI-2	Glaznieks et al. (2023)	A1–C1	ita, lld, + others	700	106k
MERLIN	Boyd et al. (2014)	A1–C2	ara, eng, + min. 11 others	1,033	154k
Total				5,736	1,554k

Table 3: Overview of available German learner corpora with CEFR annotations. Language codes follow ISO 639-2. *#Texts* and *#Tokens* measured before filtering.

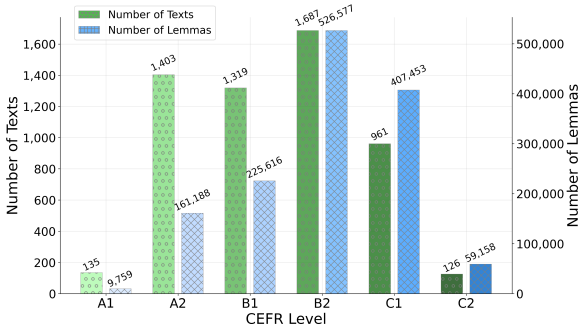

Figure 1: Number of texts and lemma tokens per CEFR level in the aggregated learner corpus after filtering.

Furthermore, Glaboniat et al. (2017)'s critique of GVLs beyond the B2 level is mirrored in the established resources, as only ASPEKTE NEU assigns a noteworthy number of vocabulary items to level C1 (see Section 2.5).[3] See Appendix B for an overview of the general and GVL-specific preprocessing steps we performed in order to enable the large-scale cross-evaluation.

Prominent German Word Frequency Lists. Table 2 presents a selection of prominent German word frequency lists. We processed the entries of the frequency lists in the same way as the entries of the established GVLs.

CEFR-Graded German Learner Corpora. Finally, Table 3 provides a comprehensive overview of German L2 learner corpora annotated with CEFR levels (see Appendix D, Table 4 for a more detailed version). We included only those corpora that either provide CEFR labels at the text level, provide reliable learner-level CEFR information, or allow for their derivation through conversion (e.g., from C-tests). After filtering out texts with missing CEFR annotations and removing punctuation tokens, we were left with a final aggregated learner corpus comprising 5,631 texts and 1,389,751 lemma tokens. Figure 1 summarizes the number of texts and lemmas contained within each CEFR level in the aggregated corpus. As expected, texts at higher CEFR levels are significantly longer and therefore contain more tokens than at lower levels.

4. Construction of GERVLPro from the Aggregated Learner Corpus

From the aggregated corpus, we constructed GERVLPro, using the lemmas and POS tags returned by STANZA.[4] To obtain the final set of vocabulary items to be included in GERVLPro, we further narrowed down the 1.39 million lemma tokens by applying additional filtering criteria. In what follows, we describe this selection procedure and the method we developed to assign CEFR levels to the final set of vocabulary items.

4.1. Vocabulary Filtering

By pairing lemma tokens with their respective POS tags, we were able to collapse the 1.39 million lemma tokens contained within the aggregated corpus into a set of 51,199 unique lemma–POS pairs

[3] We do not count DAFLEX here because the level distribution we report is based on our own level assignment method.

[4] https://stanfordnlp.github.io/stanza/, last accessed 2026/03/14.

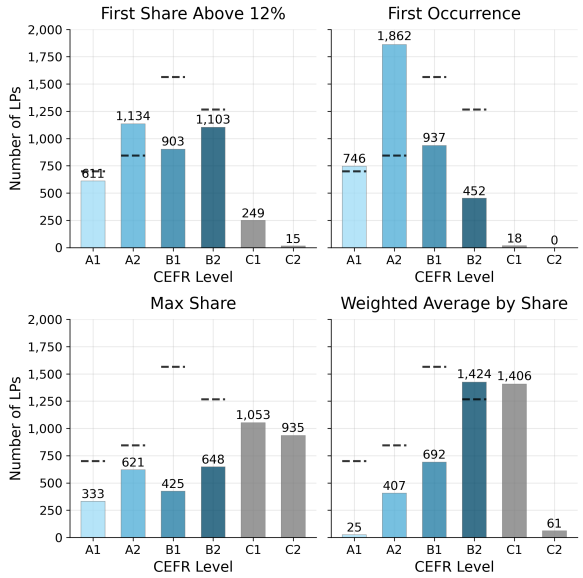

Figure 2: Number of vocabulary items at each CEFR level in GERVLPro using different level assignment methods. Reference distribution D_A represented by dashed lines.

(LPs). However, dealing with noisy learner language implies that not all of the 51,199 LPs were valid candidates for inclusion in the final vocabulary set. Many of them contained spelling errors, prompt-leaked material, foreign material, personal names, or very infrequent material, the latter of which would also indicate either potentially erroneous material or learner-specific vocabulary acquired through personal interest or individual domain knowledge.

Our aim was to extract a core vocabulary shared among a significant number of learners at each CEFR level. Thus, we employed the following filtering strategies to ensure that at each level, we only captured those vocabulary items that learners are demonstrably able to produce actively, independently, repeatedly, and correctly:

Document Frequency-Based Filtering. First, we restricted the lexicon to LPs that occur in at least 10 different learner texts. We prioritized document frequency over absolute frequency to ensure that we only retained vocabulary that is attested across multiple learners. This ensures that the captured vocabulary forms part of the shared core vocabulary at a given CEFR level, rather than reflecting individual knowledge. Additionally, this approach reduces noise from repeated spelling errors in single productions, which affect absolute word frequency but not document frequency. Document frequency-based filtering reduced the candidate vocabulary from 51,199 LPs to 4,878 LPs.

POS-Based Filtering. Second, we excluded any LPs tagged as `NUM` or `X`, i.e., numbers, abbreviations, foreign words, or unrecognizable strings. This reduced the remaining LPs to 4,661.

Writing Prompt-Based Filtering. Third, we grouped texts elicited by the same writing prompts into composite documents. On these, we computed c-TF-IDF scores over content word lemmas (`NOUN`, `PROPN`, `VERB`, `ADJ`, `ADV`).[5] To minimize the influence of prompt-driven vocabulary, we pooled LPs into two bins across prompt groups: a top 20% bin, and a bottom 80% bin. The top bin is an aggregation of the highest-ranked 20% of LPs within each composite document; the bottom bin aggregates all remaining LPs. All items from the bottom bin were retained, as they either do not play a key role in the respective learner texts or appear frequently among several texts. Items that occured only in the top-20% bin were excluded: They are likely to reflect lexical choices triggered directly by prompts or source texts (Crossley et al., 2013; Wu, 2013; van Weijen et al., 2019). If an LP appeared in both bins, i.e., top-ranked in one group but bottom-ranked in another, we retained it, since its wider distribution suggests availability beyond prompt effects. Prompt-based filtering further reduced the remaining candidate vocabulary to 4,366 LPs.

Manual Filtering. Lastly, we manually removed any remaining tokens from the candidate vocabulary that contained spelling errors, personal names, or foreign-language material, reducing the remaining 4,366 candidate LPs to a final set of 4,015 LPs to be included in GERVLPro. This is in line with the general lexicon sizes determined in previous research (see Section 2.4). Manual filtering ensured that our dataset is of very high quality.

4.2. Level Assignment Method

Let V_C ($|V_C| = 4,015$) be our final set of ungraded LPs. For each LP, we report absolute and relative frequencies in the entire corpus, absolute and relative frequencies within each CEFR level, and a *normalized share* across levels, which quantifies how characteristic an LP is at each level.

As explained in Section 2.4, there is no consensus on how many items a learner typically knows at each CEFR level. Thus, a level assignment based on simple word frequency is not possible. Assigning items to the level of first occurrence is also inadequate, as it does not ensure that a word forms part of the core vocabulary shared by a broader group of learners. To align the unannotated candidate vocabulary list V_C with a target CEFR distribution, we therefore induce levels for words in V_C by

[5] We used `scikit-learn`'s TF-IDF implementation.

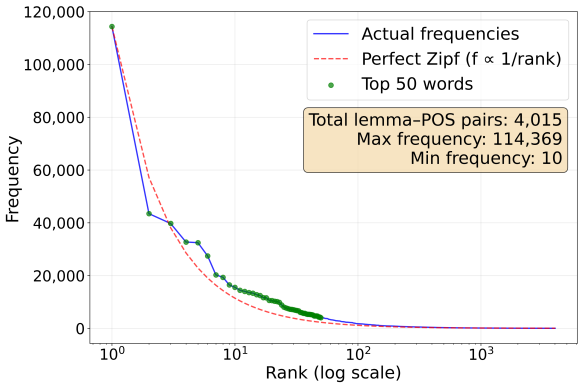

Figure 3: GERVLPro's LP frequencies against the Zipfian curve.

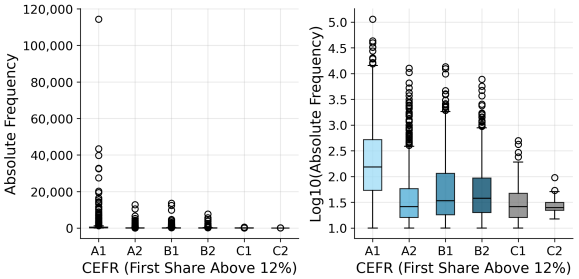

Figure 4: GERVLPro's LP frequencies against CEFR levels.

thresholding their *normalized share* across levels. For a lemma–POS pair LP and level ℓ, let $f(LP, \ell)$ denote its relative frequency at ℓ; the normalized share is

$$s(LP, \ell) = \frac{f(LP, \ell)}{\sum_{\ell'} f(LP, \ell')}.$$

A word is assigned to the *lowest* level ℓ such that $s(LP, \ell) \geq x$, where x is a tunable threshold. For the lack of a better approximation, the reference distribution D_A is constructed as the average of the level distributions within the established GVLs introduced in Table 1 $\{D_1, \ldots, D_6\}$, excluding DAFLEX:[6]

$$D_A(\ell) = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} D_i(\ell),$$

where n_ℓ is the number of lists that provide annotations for level ℓ (i.e., only lists with data for the respective level are included). The size of $|D_A| = 4,374$ is very similar to the lexicon sizes identified in previous research and the number of candidate LPs in V_C (see Section 2.4), supporting the validity of our filtering approach. Given the induced distribution $D^x_{V_C}$ over levels in V_C, we quantify its deviation from D_A by summing absolute differences:

$$P(x) = \sum_{\ell \in \{A1, \ldots, B2\}} \Big| |D^x_{V_C}(\ell)| - |D_A(\ell)| \Big| + \Big| |D^x_{V_C}| - |V_C| \Big|,$$

where the second term penalizes discrepancies in the overall number of items, since higher thresholds may leave words unassigned.

Figure 2 showcases the level distributions resulting from different potential level assignment strategies. It illustrates that, when assigning the level at which an LP first occurs, no items are assigned to C2, and only 18 to C1. This is despite the number of lemma tokens contained within the C1 portion of our aggregated corpus exceeding the number of tokens contained in the A1–B1 levels combined. In practice, this means that neither of the two C-levels introduces any new vocabulary. Taken together with the questionable validity of GVLs beyond B2 (see Section 2.5), we aim to situate all candidate LPs within the levels A1–B2.

The optimal threshold is thus given by

$$x^* = \arg\min_x P(x).$$

Empirically, $x^* = 0.12$ (12%) yields the lowest penalty, producing the closest match between the induced distribution of V_C and the averaged reference distribution, while including as many LPs as possible. The resulting GVL is GERVLPro.

Finally, the frequency distribution of the vocabulary items in GERVLPro follows the expected Zipfian curve relatively closely (see Figure 3), indicating that the resulting lexicon exhibits the expected distributional properties of natural language vocabularies. Moreover, Figure 4 illustrates that, as expected, our level assignment method assigns the most frequent LPs to A1, with frequency scores generally declining gradually as the CEFR level increases.

5. Validation Study

To assess whether established GVLs and prominent frequency lists can model the productive vocabulary development of learners as represented by GERVLPro, and how well individual lists align within and across these two resource types, we carried out a large-scale cross-evaluation of GERVLPro, the seven established GVLs, and the six prominent frequency lists introduced in Section 3. In what follows, we briefly introduce the similarity and alignment measures we report. We computed all metrics on the intersecting LPs within the A1–B2 levels of the respective resource pairs. Additionally, we explain the setup of our exploration of LLMs as a tool for grading productive GVLs.

[6] We excluded DAFLEX because it does not constitute a GVL in the conventional sense.

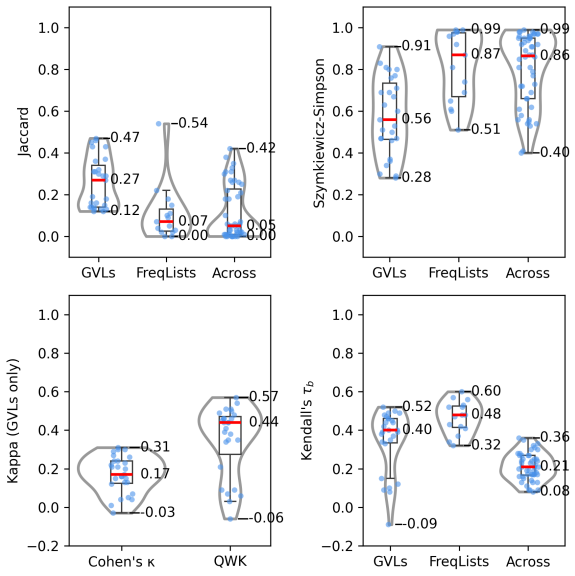

Figure 5: Visualization of pairwise similarity and alignment results, aggregated by resource types (see Appendix F, Tables 5, 6, 7). We excluded the PROFILE DEUTSCH R. and P. pairs from the visualization because the two lists are interdependent.

5.1. Similarity Measures

We use the following measures to assess resource similarity across and within GERVLPro, established GVLs, and prominent frequency lists:

Jaccard Index. We report the Jaccard index, defined as the ratio of the intersection to the union of two sets, to measure overall resource similarity (Jaccard, 1901).

Szymkiewicz–Simpson Coefficient. We also report the related Szymkiewicz–Simpson coefficient, defined as the ratio of the intersection to the smaller of the two set sizes (Vijaymeena and Kavitha, 2016). This provides a complementary perspective to the Jaccard index, which is especially relevant for comparing resources of different sizes.

5.2. Alignment Measures

We use the following measures to assess alignment within and across GERVLPro, established GVLs, and prominent frequency lists, as well as to evaluate the performance of GPT-4o and GPT-5 in grading the vocabulary items contained in GERVLPro:

Cohen's κ. For established GVLs and LLM predictions, which assign explicit CEFR levels rather than frequency ranks, we report Cohen's κ to quantify exact-level agreement, treating CEFR levels as unordered categories (Cohen, 1960).

Quadratically Weighted κ. We also report QWK to assess whether level shifts primarily seem to involve adjacent levels or whether larger jumps are common (Cohen, 1968).

Kendall's τ_b. For all resource types and the LLM predictions, we report Kendall's τ_b to assess alignment in overall low-to-high ordering of LPs across resources (Kendall, 1938). We chose τ_b over other correlation metrics because it captures pairwise ordering consistency with explicit tie correction, which is required when working with only a few categories such as the CEFR levels in GVLs. For frequency lists, we used the provided frequency bands or rank categories where available, and we fell back on absolute frequency ranks otherwise.

5.3. LLM Grading

As outlined in Section 2.6, Alfter (2024) demonstrated that several LLMs perform reasonably well in assigning receptive CEFR levels to lexical items using zero-shot prompting across English, Spanish, French, Swedish, and Dutch. We extend this work by evaluating the predictions of GPT-4o, the best-performing model reported by Alfter, and GPT-5, for productive CEFR levels in German (see Appendix E for our adapted prompt; the original prompt did not specify whether to assign receptive or productive levels). Following Alfter, we used default hyperparameters. We report means over three runs.

6. Results

This section presents the results of our cross-evaluation of the GVLs and frequency lists introduced in Section 3, their similarity and alignment with GERVLPro in particular, and the exploration of LLMs as tools for grading the productive vocabulary items contained in GERVLPro.

6.1. Similarity Across Resource Types

Pairwise Similarity. Across resources, the two similarity measures reveal systematically different patterns. Using Jaccard similarity, established GVLs generally exhibit higher mutual similarity than frequency lists, indicating that in absolute terms, GVLs tend to share a larger proportion of lexical items with one another. In contrast, Szymkiewicz–Simpson coefficients are highest for comparisons involving frequency lists, both within the group of frequency lists and in cross-type comparisons with GVLs, reflecting the broader lexical coverage of these resources. As a result, cross-type comparisons often show high containment of GVL vocabularies by frequency lists despite relatively low Jaccard similarity. ASPEKTE NEU differs

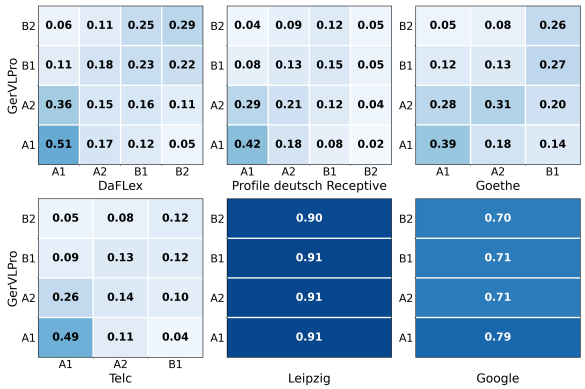

Figure 6: Pairwise Szymkiewicz–Simpson coefficient per level set in GERVLPro and a selection of established GVLs and prominent frequency lists.

the most from the other GVLs, which is unsurprising as, unlike all other GVLs, it does not hold A-level items. Among frequency lists, the highest Jaccard similarity is observed for ROUTLEDGE and GOOGLE, both nearly an order of magnitude smaller than the next smallest list. The distributional patterns are illustrated in Figure 5 and correspond to the pairwise values reported in Appendix F, Table 5.

Per-Resource Similarity. To statistically compare similarity patterns across resource groups, we derive similarity scores at the individual resource level (pairwise comparisons are not independent; each resource participates in multiple pairs). To this end, we average each resource's similarity to others within its group and across groups (see Appendix F, Table 8). These per-resource averages corroborate the pairwise patterns. For Jaccard, within-GVL similarity is substantially higher than cross-type similarity (μ within-GVL = 0.26 vs. μ cross-type = 0.11; Wilcoxon signed-rank test on per-resource means: p_{Holm} = 0.0234, RBC = 1.00). Within-GVL similarity is also higher than within-frequency-list similarity (μ within-frequency-list = 0.11; Mann–Whitney U: p_{Holm} = 0.0053, Cliff's δ = 0.92). For Szymkiewicz–Simpson, cross-type similarity is higher than within-GVL similarity (μ within-GVL = 0.58 vs. μ cross-type = 0.80; Wilcoxon signed-rank test: p_{Holm} = 0.0313, RBC = -0.94). This pattern is expected given typical size differences between GVLs and frequency lists.

6.2. Alignment Across Resource Types

Pairwise Alignment. To assess alignment in ordering across all resources, we report Kendall's τ_b (Appendix F, Table 7; Figure 5). Within the set of established GVLs, pairwise correlations range from weak to moderate (τ_b range: -0.09–0.52).[7] Most pairs involving LINGSTER, DAFLEX, PROFILE DEUTSCH (receptive and productive), GOETHE, and TELC reach moderate correlations of typically > 0.40, whereas combinations involving ASPEKTE NEU exhibit weak or near-zero alignment.

Looking specifically at agreement in discrete level assignments among established GVLs, Cohen's κ is consistently weak ($\kappa \leq 0.31$; Appendix F, Table 6; Figure 5). When the magnitude of level shifts is taken into account, QWK increases for the more closely related GVLs identified above (QWK ≤ 0.57), while comparisons involving ASPEKTE NEU remain close to zero. These results indicate that, while some established GVLs exhibit similar ordering tendencies, there is no uniform level ordering across GVLs. Even when level distances are considered, agreement remains moderate at best.

Within-frequency-list correlations are generally moderate to strong (τ_b range: 0.32–0.60). In contrast, cross-type alignment between GVLs and frequency lists is consistently weaker (τ_b range: 0.08–0.36). Overall, these correlation patterns show that ordering agreement is highest within frequency lists, lower within GVLs, and weakest in cross-type comparisons, indicating that level assignments in GVLs are not primarily frequency-driven.

Per-Resource Alignment. Resource-level alignment also reflects the fact that in their ordering, GVLs are more similar to each other than to frequency lists (μ within-GVL τ_b = 0.35 vs. μ cross-type τ_b = 0.21; Wilcoxon signed-rank test on per-resource means: p_{Holm} = 0.0313, RBC = 0.944), but the absolute level of alignment within GVLs remains only moderate. In other words, there is no single shared ordering across GVLs, and frequency-based ordering does not closely align with pragmatically motivated level assignments.

6.3. Relationship to GERVLPro

Similarity. In terms of Jaccard, GERVLPro tends to be more similar to other GVLs than to frequency lists (μ $\times$GVLs = 0.26 vs. μ $\times$frequency lists = 0.15), although given the small sample size, this difference does not reach statistical significance (Mann–Whitney U: p_{Holm} = 0.2949; Cliff's δ = 0.38). Among GVLs, GOETHE, LINGSTER, and the productive version of PROFILE DEUTSCH reach scores ≥ 0.30. The pairwise score distribution for frequency lists ranges from 0.00 to 0.42, with GOOGLE and ROUTLEDGE, the two smaller lists, unsurprisingly being the most similar.

In terms of Szymkiewicz-Simpson, GERVLPro shows substantially higher overlap with frequency

[7] Excluding the values for the interdependent PROFILE DEUTSCH R. and P. lists.

lists than with other GVLs (μ ×GVLs = 0.83 vs. μ ×frequency lists = 0.52; Mann–Whitney U: p_{Holm} = 0.0234; Cliff's δ = -0.91). This indicates that frequency lists typically include a large share of the lexical items in GERVLPro. However, because this measure normalizes by the smaller set size, high values are expected when comparing a smaller GVL to much larger frequency lists. Figure 6 further illustrates that overlap between GERVLPro and other resources is typically concentrated at the lower proficiency levels, with A1 showing the highest degree of overlap, while coverage generally decreases at higher levels.

Alignment. GERVLPro generally exhibits moderate alignment with other GVLs (μ ×GVLs τ_b = 0.33; range: 0.09–0.40), with the only notable downward outlier being ASPEKTE NEU, but only weak alignment with frequency lists (μ ×frequency lists τ_b = 0.14; range: 0.08–0.27), with SUBTLEX-DE being the only list that reaches a score > 0.20. The difference between resource types is statistically supported (Mann–Whitney U: p_{Holm} = 0.0438; Cliff's δ = 0.79). Agreement in exact-level assignments between GERVLPro and other GVLs is consistently weak ($\kappa \leq 0.17$), while QWK reaches moderate values for some comparisons (QWK ≤ 0.46). Thus, while other GVLs model GERVLPro more closely than frequency lists in terms of ordering, the observed agreement remains only moderate. Overall, with the exception of ASPEKTE NEU, GERVLPro shows the weakest alignment with all other GVLs and across all alignment measures.

Summary. In sum, neither established GVLs nor frequency lists adequately model GERVLPro. Frequency lists tend to cover a large proportion of GERVLPro's vocabulary, but they show weak ordering alignment with GERVLPro, suggesting that frequency-based ordering does not capture the level progression observed in productive learner vocabulary. Established GVLs show higher ordering alignment with GERVLPro than frequency lists, but that alignment is still only moderate. The same goes for exact and weighted agreement in discrete level assignments. Overall, existing resource types provide partial but incomplete approximations: Frequency lists primarily offer broad lexical coverage, whereas GVLs provide closer but still rather weak ordering alignment.

6.4. LLM Grading

An examination of the scores returned by GPT-4o and GPT-5 shows that exact-level matches are similarly rare to those obtained in the comparisons with the established GVLs (Appendix F, Table 6; κ range: 0.06–0.11, SD range: 0.00–0.04). Taking into account the size of level shifts through QWK again improves agreement, though not to the same extent as observed for the established GVLs (QWK range: 0.24–0.31, SD range: 0.00–0.05).

In terms of correlations, GPT-4o's and GPT-5's predictions show only moderate alignment with GERVLPro. All values remain below those of the established GVLs (Appendix F, Table 7; τ_b = 0.29 for all settings, SD range: 0.00–0.02), with the exception of ASPEKTE NEU. This indicates that even the most advanced LLMs currently available are not able to accurately grade the productive vocabulary items contained in GERVLPro. Moreover, explicitly instructing the LLMs to differentiate between receptive and productive scores does not considerably improve performance.

7. Conclusion

We have introduced GERVLPro: The first descriptive, CEFR-graded list of German L2 learners' productive vocabulary, built from the largest known aggregation of CEFR-annotated German learner corpora to date. Through stringent filtering and a normalized share-based level assignment method, we derived a lexicon of 4,015 lemma–POS pairs, primarily assigned to the A1–B2 levels. Our cross-evaluation against seven established GVLs and six prominent frequency lists shows sizable lexical overlap but weak alignment between resources. Among established GVLs, LINGSTER, PROFILE DEUTSCH, GOETHE, and TELC align most closely with each other, suggesting a potential relationship between these resources and, to a weaker extent, with GERVLPro. This might be an artifact of the fact that both GOETHE and TELC are the main providers of German language proficiency tests, which incentivizes learners to study the vocabulary items featured in their lists as part of exam preparation.

Among frequency lists, SUBTLEX-DE relates more closely to learner behavior than lists based on corpora of written standard language, yet none of the aforementioned resources manages to adequately model productive vocabulary development, especially beyond A2–B1. Level set-wise overlap with GERVLPro concentrates at A1 and deteriorates at higher levels, supporting doubts about the usefulness of general-purpose GVLs above B2. Finally, GPT-4o's and GPT-5's predictions yield only modest correlations with GERVLPro, underperforming curated GVLs on alignment. Taken together, these findings indicate that established GVLs and frequency lists do not sufficiently reflect the progression of productive lexical competence, particularly beyond beginner levels. GERVLPro provides the first data-driven baseline for informing applications and research on productive vocabulary development in learners of German.

8. Limitations

Lastly, we briefly outline a number of methodological limitations and design decisions that affect how GERVLPro was constructed and how it should be interpreted.

Unlike Volodina et al. (2016), we decided not to apply grammatical error correction as a first step in preprocessing the learner texts we derived GERVLPro from. Learner productions, especially at lower proficiency levels, often permit multiple valid target hypotheses and human agreement is limited (Tetreault and Chodorow, 2008); therefore, automatic normalization risks introducing bias. Instead, we relied on the filtering procedures described in Section 4 and a final manual check to remove erroneous lemmas.

Our preprocessing procedure also fragmented multiword expressions. However, learners acquire expressions such as *auf Wiedersehen* "goodbye" as a unit and do not necessarily understand each lexeme individually (Underwood et al., 2008). A separate modeling of multiword expression acquisition order is left for future work.

Lemmatization further poses challenges. Established GVLs and frequency lists were lemmatized as isolated tokens without context, whereas vocabulary items from learner essays were lemmatized with context, which can lead to more POS options for the same lemma. Moreover, lemmatization is likely to be affected by learner language noise. Additionally, lemma–POS information can be insufficient to disambiguate meaning, as a single lemma–POS pair may correspond to multiple senses. Given that most GVLs and frequency lists provide their entries without context, we accepted this non-optimal compromise.

A further limitation concerns the reliability and comparability of CEFR labels across the aggregated learner corpora. CEFR annotations may stem from different assessment procedures (e.g., standardized tests, course placement, or instructor judgment), which can introduce inconsistencies across datasets. This potential label noise may affect the distribution of vocabulary items across CEFR levels and, consequently, the level assignments in GERVLPro.

Finally, using the average number of vocabulary items contained at each level in the existing GVLs introduced in Section 3 as the reference distribution D_A, with which we aimed to align the distribution of vocabulary items per CEFR level in GERVLPro, is arguably not optimal. However, since there is no consensus on how many vocabulary items L2 German learners typically can produce at each CEFR level, this approach presented itself as the most pragmatic one. In the absence of a more robust alternative for assigning vocabulary items to discrete CEFR levels, which is necessary for constructing a *graded* vocabulary list, the resulting CEFR-level distribution in GERVLPro should be understood as a rough calibration rather than a definitive prescriptive target distribution. The main contribution of this work lies in extracting a descriptive inventory of the productive vocabulary of L2 German learners, whereas the precise level assignments may remain open to discussion. Future studies on lexical size estimation may enable the determination of a more robust threshold for the normalized share we used for assigning CEFR levels.

9. Ethics Statement

We do not see any major ethical concerns in the context of this work. However, GERVLPro can only be representative of the learners present in the data. The learner corpora aggregated for its construction differ in terms of L1 backgrounds, age groups, task types, and annotation practices, and therefore inevitably introduce biases related to the populations and settings represented in the data. Consequently, many demographic and contextual variables are not explicitly controlled for or may be underrepresented. For this reason, we do not claim that GERVLPro captures the "average" German learner in absolute terms. Rather, it models the "average" learner as reflected in the aggregated learner corpus used in this study. Expanding the dataset with additional learner corpora representing more diverse backgrounds, learning contexts, and production settings would further improve the representativeness of the resource. Moreover, future work may investigate stratified analyses by corpus, L1 background, or task type to better understand how such factors influence the distribution of productive vocabulary.

10. Acknowledgments

We would like to thank everyone who made their corpus data available to us in a machine-readable format. Special thanks go to our student assistant Jette Sönnichsen for her help in meticulously extracting vocabulary from PDFs and other data formats and in the manual filtering of the final set of vocabulary items in GERVLPro.

11. Bibliographical References

James S. Adelman, Gordon D.A. Brown, and José F. Quesada. 2006. Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–19, Rennes, France. LiU Electronic Press.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Exploiting the English vocabulary profile for L2 word-level vocabulary assessment with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 632–646, Vienna, Austria. Association for Computational Linguistics.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–94, San Diego, CA. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Scott A. Crossley, Laura K. Varner, and Danielle S. McNamara. 2013. Cohesion-based prompt effects in argumentative writing. In *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*, pages 202–207. Grantee Submission, non-journal.

Luise Dürlich and Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of COLING 2012*, pages 799–814, Mumbai, India. The COLING 2012 Organizing Committee.

Nick C Ellis. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, 24(2):143–188.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598, Mexico City, Mexico. Association for Computational Linguistics.

Thomas François, Nùria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3766–3773, Reykjavik, Iceland. European Language Resources Association (ELRA).

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 213–219, Portorož, Slovenia. European Language Resources Association (ELRA).

Stefan Gries. 2021. What do (most of) our dispersion measures measure (most)? dispersion? *Journal of Second Language Studies*, 5.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles*, 37(142):547–579.

Anders Johannsen, Héctor Martínez, Sigrid Klerke, and Anders Søgaard. 2012. EMNLP@CPH: Is frequency all there is to simplicity? In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 408–412, Montréal, Canada. Association for Computational Linguistics.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Batia Laufer. 1998. The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, 19(2):255–271.

Batia Laufer and T. Sima Paribakht. 1998. The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, 48(3):365–391.

Batia Laufer and Geke C. Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1):15–30.

Guanlin Li, Yuki Arase, and Noel Crespi. 2025. Aligning sentence simplification with ESL learner's proficiency for language acquisition. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 492–507, Albuquerque, New Mexico. Association for Computational Linguistics.

James Milton. 2010. The development of vocabulary breadth across the cefr levels. In Ineke Vedder, Inge Bartning, and Merete Martin, editors, *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, volume 1 of *Monograph Series*, pages 211–232. Second Language Acquisition and Testing in Europe (SLATE), Rome.

James Milton and Thomaï Alexiou. 2009. Vocabulary size and the common european framework of reference for languages. In Michael H. Daller, James Milton, and Jeanine Treffers-Daller, editors, *Vocabulary Studies in First and Second Language Acquisition*, pages 194–211. Palgrave Macmillan, London.

Jupp Möhring and Franziska Wallner. 2013. Wortschatzlisten auf dem Prüfstand. In Hana Bergerová, Marek Schmidt, and Georg Schuppener, editors, *Aussiger Beiträge, Band 7*, pages 119–133. Univerzita J. E. Purkyně, Filozofická fakulta, Ústí nad Labem.

Paul Nation. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1):59–82.

Paul Nation and Robert Waring. 1997. Vocabulary size, text coverage and word lists. In Norbert Schmitt and Michael McCarthy, editors, *Vocabulary: Description, Acquisition and Pedagogy*, pages 6–19. Cambridge University Press, Cambridge.

Adam Nohejl, Frederikus Hudi, Eunike Andriani Kardinata, Shintaro Ozaki, Maria Angelica Riera Machin, Hongyu Sun, Justin Vasselli, and Taro Watanabe. 2025. Beyond film subtitles: Is YouTube the best approximation of spoken vocabulary? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9566–9585, Abu Dhabi, UAE. Association for Computational Linguistics.

Virginie Pignot-Shahov. 2012. Measuring L2 Receptive and Productive Vocabulary Knowledge. *Language Studies Working Papers*, 4:37–45.

Norbert Schmitt. 2008. Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12:329–363.

Norbert Schmitt. 2024. *Knowledge-Based Vocabulary Lists*. Equinox Publishing Ltd.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Kevin Stowe, Debanjan Ghosh, and Mengxuan Zhao. 2022. Controlled language generation for language learning items. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 294–305, Abu Dhabi, UAE. Association for Computational Linguistics.

Anaïs Tack, Thomas François, Piet Desmet, and Cédrick Fairon. 2018. NT2Lex: A CEFR-graded lexical resource for Dutch as a foreign language linked to open Dutch WordNet. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146, New Orleans, Louisiana. Association for Computational Linguistics.

Joel Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: Experiments in preposition error detection. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 24–32, Manchester, UK. Coling 2008 Organizing Committee.

Erwin Tschirner. 2019. Der rezeptive Wortschatzbedarf im Deutschen als Fremdsprache. In *IDT 2017: Vielfalt und Einheit in der Deutschdidaktik.*

Band 1, pages 98–111. Erich Schmidt Verlag, Berlin.

Erwin Tschirner, Jane Hacking, and Fernando Rubio. 2020. The relationship between reading proficiency and vocabulary size: An empirical investigation. In Peter Ecke and Susanne Rott, editors, *Understanding Vocabulary Learning and Teaching: Implications for Language Program Development*, AAUSC Issues in Language Program Direction, pages 58–77. Cengage, Boston, MA. AAUSC 2018 Volume.

Geoffrey Underwood, Norbert Schmitt, and Adam Galpin. 2008. The eyes have it: An eye-movement study into the processing of formulaic sequences. In *Formulaic sequences: Acquisition, processing and use*, pages 153–172. John Benjamins Publishing Company.

Daphne van Weijen, Gert Rijlaarsdam, and Huub van den Bergh. 2019. Source use and argumentation behavior in L1 and L2 writing: a within-writer comparison. *Reading and Writing*, 32(6):1635–1655.

M. K. Vijaymeena and K. Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(1):19–28.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016. SweLLex: Second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 76–84, Umeå, Sweden. LiU Electronic Press.

Ruey-Jiuan Regina Wu. 2013. Native and non-native students' interaction with a text-based prompt. *Assessing Writing*, 18(3):202–217.

12. Language Resource References

Christoph Beeh, Ewa Drewnowska-Vargáné, Péter Kappel, Bernadett Modrián-Horváth, Andreas Nolda, Orsolya Rauzs, and György Scheibl. 2021. Dulko-Handbuch. Aufbau und Annotationsverfahren des deutsch-ungarischen Lernerkorpus. Version 1.0.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. pages 1281–1288.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bolte, and Andrea Bohl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in german. *Experimental Psychology*, 58(5):412–424.

Manuela Glaboniat, Martin Müller, Paul Rusch, Helen Schmitz, and Lukas Wertenschlag. 2017. Profile deutsch: Lernzielbestimmungen, kannbeschreibungen, kommunikative mittel. niveau a1–a2, b1–b2, c1–c2. CD-ROM Version 2.0 with accompanying book; originally published 2005 by Langenscheidt KG, München; 8th printing 2022.

Andrius Glaznieks, Jérôme-C Frey, Andrea Abel, Laure Nicolas, and Chiara Vettori. 2023. The Kolipsi Corpus Family: Resources for learner corpus research in Italian and German. *Italian Journal of Computational Linguistics*, 9(2).

Julian Heister, Kay-Michael Würzner, Johannes Bubenzer, Eva Pohl, Thomas Hanneforth, Alexander Geyken, and Reinhold Kliegl. 2011. dlexdb—eine lexikalische datenbank für die psychologische und linguistische forschung. *Psychologische Rundschau*, 62(1):10–20.

Anke Lüdeling, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt, and Maik Walter. 2008. Das lernerkorpus falko. *Deutsch als Fremdsprache*, (2):67–73.

Ursula Maden-Weinberger. 2013. CLEG13: Corpus of Learner German (version 07-19-2013).

Marc Reznicek, Anke Lüdeling, Cedric Krummes, Franziska Schwantuschke, Maik Walter, Karin Schmidt, Hagen Hirschmann, and Torsten Andreas. 2012. Das Falko-Handbuch: Korpusaufbau und Annotationen. Version 2.01.

Carola Strobl and Helena Wedig. 2023. Beldeko Summary Corpus v1.1.0. Eurac Research CLARIN Centre.

Erwin Tschirner and Jupp Möhring. 2019. *A frequency dictionary of German*, 2 edition. Routledge Frequency Dictionaries. Routledge, London, England.

Nina Vyatkina. 2016. The Kansas Developmental Learner Corpus (KANDEL): A developmental corpus of learner German. *International Journal of Learner Corpus Research*, 2(1):101–119.

Helena Wedig and Carola Strobl. 2024. German Summary Corpus (GerSumCo) v1.0.0. Eurac Research CLARIN Centre.

Katrin Wisniewski, Eva Muntschick, and Angelika Portmann. 2022. Schreiben in der Studiersprache Deutsch: Das Lernerkorpus DISKO. pages 283–304.

Heike Zinsmeister, Marc Reznicek, Julia Ricart Brede, Christina Rosén, and Dirk Skiba. 2012. Das Wissenschaftliche Netzwerk "Kobalt-DaF". *Zeitschrift für Germanistische Linguistik*, 40(3):457–458.

Appendix A. Resource References.

DaFLex: https://cental.uclouvain.be/cefrlex/daflex/, last accessed 2026/03/14.

Lingster Academy: https://lingster.de/wp-content/uploads/2023/03/Der-deutsche-Wortschatz-von-A1-bis-B2-Lingster-Academy.pdf, last accessed 2026/03/14.

Aspekte neu: https://www.klett-sprachen.de/downloads/5376/alphabetische-wortliste/pdf; https://www.klett-sprachen.de/downloads/7058/alphabetische-wortliste/pdf; https://www.klett-sprachen.de/downloads/10201/alphabetische-wortliste/pdf, last accessed 2026/03/14.

Profile deutsch: https://www.klett-sprachen.de/profile-deutsch/t-1/9783126065184, last accessed 2026/03/14.

Goethe Zertifikat: https://www.goethe.de/pro/relaunch/prf/de/A1_SD1_Wortliste_02.pdf; https://www.goethe.de/pro/relaunch/prf/de/Goethe-Zertifikat_A2_Wortliste.pdf; https://www.goethe.de/pro/relaunch/prf/en/Goethe-Zertifikat_B1_Wortliste.pdf, last accessed 2026/03/14.

Telc: https://www.telc.net/fileadmin/user_upload/Downloads_Verlag/Einfach_gut/Wortschatzlisten/Einfach_gut_A1_Wortschatzliste_alphabetisch.pdf; https://www.telc.net/fileadmin/user_upload/Downloads_Verlag/Einfach_gut/Wortschatzlisten/Einfach_gut_A2_Wortschatzliste_alphabetisch.pdf; https://www.telc.net/fileadmin/user_upload/Downloads_Verlag/Einfach_gut/Wortschatzlisten/Einfach_gut_B1_Wortschatzliste_alphabetisch.pdf, last accessed 2026/03/14.

dlexDB: Heister et al. (2011).

DeReWo: https://www.ids-mannheim.de/digspra/pb-s1/projekte/methoden/derewo/, last accessed 2026/03/14.

Subtlex-de: Brysbaert et al. (2011).

Leipzig Corpora Frequencies Top 100k: https://api.wortschatz-leipzig.de/ws/swagger-ui/index.html, last accessed 2026/03/14; corpus: deu_news_2012_3M.

Google Ngrams Top 10k: https://github.com/orgtre/google-books-ngram-frequency/blob/main/ngrams/1grams_german.csv, last accessed 2026/03/14.

Routledge Frequency Dictionary: Tschirner and Möhring (2019).

Appendix B. GVL-Specific Preprocessing Steps.

For all established GVLs, we applied the following general preprocessing steps: If the GVL was not already in a machine-readable format, we extracted its entries from PDFs using Gemini and manually checked its output (see Appendix C for the prompt we use).[8] We cleaned punctuation such as brackets and commas as they are often used to provide additional information or introduce inflected forms and split entries on whitespace to naively deconstruct multiword expressions. We then processed the items with stanza[9] and filtered out any items that returned multiple POS tags[10] or were tagged as either NUM, X, or PUNCT. For items passing this filter, we extracted the lemma and POS tag. We filtered out any non-purely alphabetic entries and ensured that the length of every entry exceeded a single character. Finally, we created a unique

[8] https://gemini.google.com, last accessed 2026/03/14.

[9] We chose stanza over spaCy because, compared to stanza, spaCy produces ∼13% more unique lemmas and ∼18% more unique LPs on the aggregrated learner corpus data overall. It also yields far more system-exclusive forms: ∼61% more lemmas and ∼48% more LPs than the number of exclusive forms in stanza. Furthermore, spaCy assigns multiple POS tags to ∼58% more lemmas than does stanza. This suggests that stanza with its neural sequence-to-sequence architecture yields a more stable and consistent lemma inventory, reducing spurious variants and POS ambiguities. Such consistency is especially beneficial when dealing with inflection and learner language errors.

[10] This effectively filters out entries that are not standalone words but affixes such as *auf-*, as stanza returns two POS tags for hyphenated items.

identifier that combines the lemma and POS tag into a `lemma_POS` pair (LP), removed any duplicate entries based on this identifier, and kept only the entry with the lowest level assignment.

In addition to this general preprocessing pipeline, we applied the following source-specific preprocessing steps:

Goethe. Vocabulary items containing forward slashes were normalized by replacing slashes with spaces to account for notational variants (e.g., *ec-Karte/EC-Karte*).

Lingster. CEFR level annotations were normalized by mapping *B1+*→*B1* and *B2+*→*B2*. Vocabulary items were split on both spaces and commas to handle notational variants (e.g., *nächster,nächste,nächstes*). Parentheses were removed, as they typically indicate optional or inflected forms (e.g., *Kilogramm (kilo)*).

Profile Deutsch. Entries lacking both receptive and productive level assignments were excluded. During deduplication, entries were prioritized first by receptive level and then by productive level, retaining the entry with the lowest receptive level assignment, as comprehension precedes production.

DaFLex. Forward slashes were replaced by the pipe character (`|`), which DaFLex uses to indicate notational variants (e.g., *Würfel|würfeln*). Entries were then split on the latter. CEFR levels were assigned according to the level assignment method introduced in Section 4.2.

Appendix C. Gemini Prompt Used for Vocabulary Extraction.

We used the following prompt to extract vocabulary from PDFs and other data formats (content input via copy and paste):

> "Please provide me with only the German entries, every entry on a new line, normalized in the sense that examples and specifiers after the original entry get removed, so we keep only the basic word form. The word and its associated level are tab-separated so I can copy and paste them into Excel."

We included the second sentence only when level annotations were present within the source file; when the file contained vocabulary from a single level only, the sentence was omitted.

Appendix D. Overview of German Learner Corpora.

Table 4 provides an extensive overview of the source corpora we used to derive GerVLPro.

Some corpora capture more than one L1 for each learner, if applicable. For the sake of simplicity, we only count the first L1 saved for each learner.

The DISKO_L2 corpus employs a rating scale from 2 to 5 for evaluating learner texts, where 2 corresponds to a proficiency level "below B2" and 5 to "C1 or higher". To align this scale with the CEFR levels used in this study, we map the scores as follows: *2*→*B1*, *3*→*B2*, *4*→*C1*, and *5*→*C2*.
KanDeL is searchable via ANNIS. However, the full texts are not available in the `Visualizer`. We received the full texts directly from the corpus's creator.

In MERLIN, six of the learner texts are annotated with German as the author's L1. We excluded these texts from our analyses.

Appendix E. Prompt for LLM Grade Generation.

We used the following prompt, adapted from Alfter (2024), to elicit productive CEFR ratings from LLMs for all vocabulary items in our experiments:

> "You are an experienced teacher of German as a second language. You can easily assess the receptive and productive difficulty of words in German for learners. You assess words on a scale from 0 (easiest) to 4 (hardest). You only answer with two numbers, one for the receptive difficulty and one for the productive difficulty. Assess: <lemma>, <pos>"

We asked the models for both receptive and productive grades in order to steer them towards assigning different scores for both dimensions. We utilized a numerical scale because, according to Alfter, it has shown improved performance over the corresponding CEFR scale.

Appendix F. Result Tables.

Corpus	Avail.	Source	Size	# Texts	Text Types	Setting	CEFR	L1s (>1%)*
Beldeko	Academic	Strobl and Wedig (2023)	71K	301	Summaries (source-based)	C	B2	nld
CLEG13	ANNIS	Maden-Weinberger (2013)	285K	729	Argumentative essays (independent); summaries, commentaries (source-based)	CEH	B1–C1	eng
DISKO_L2	Academic	Wisniewski et al. (2022)	240K	510	Argumentative essays (source-based)	E	B2–C1*	ara, bul, ces, eng, fas, fra, hun, ind, ita, kor, kur, pol, rus, spa, sqi, srp, ukr, vie, zho, + 27 others
DISKO_WebTestDaF	Academic	Wisniewski et al. (2022)	91K	479	Argumentative essays (independent); summaries (source-based)	E	A2–C1	ara, bul, fas, ita, jpn, kat, kor, kur, pol, por, ron, rus, spa, sqi, tur, ukr, vie, zho, + 11 others
Dulko	ANNIS	Beeh et al. (2021)	21K	64	Argumentative essays (independent); translations (source-based)	C	B2–C1	hun
FalkoEssayL2	Public	Lüdeling et al. (2008)	145K	248	Argumentative essays (independent)	E	B2–C2	afr, ces, dan, ell, eng, fin, fra, ita, jpn, kik, lua, nld, nor, pol, ron, rus, spa, swe, tur, ukr, uzb, zho, + 15 others
FalkoSummaryL2	Public	Lüdeling et al. (2008)	41K	106	Summaries (source-based)	E	C1	bul, eng, fas, jpn, kat, kor, lit, mon, pol, por, rus, ukr, zho, + 11 others
FalkoWHiGL2	Public	Reznicek et al. (2012)	131K	196	Argumentative essays (independent)	E	B2–C2	eng, fra, pol, rus, + 6 others
GerSumCo	Academic	Wedig and Strobl (2024)	27K	108	Summaries (source-based)	CH	B2	ara, dan, eng, fas, fra, ita, kor, pol, por, rus, spa, sqi, tur, ukr, zho, + 7 others
KanDeL	ANNIS*	Vyatkina (2016)	122K	688	Short narrative/argumentative essays (independent)	CH	A2	eng
Kobalt-DaF	ANNIS	Zinsmeister et al. (2012)	33K	51	Argumentative essays (independent)	E	B2	bel, rus, swe, zho
Kolipsi-1	Academic	Glaznieks et al. (2023)	87K	523	Narrative emails (source-based); argumentative letters (independent)	E	A1–C1	ita, lld, + others
Kolipsi-2	Academic	Glaznieks et al. (2023)	106K	700	Narrative emails (source-based); argumentative emails (independent)	E	A1–C1	ita, lld, + others
MERLIN	Public	Boyd et al. (2014)	154K	1,033	(In)formal letters/emails, argumentative essays, articles, reports (independent)	E	A1–C2	ara, eng, fra, hun, ita, pol, por, rus, spa, tur, + min. 3 others*

Table 4: Overview of available German learner corpora with CEFR annotations. Sizes in tokens (rounded to the nearest thousand). Language codes follow ISO 639-2. C: In-**C**lass assignment, E: **E**xam, H: **H**omework.

	GerVLPro	DaFLex	Lingster	Aspekte	Profile R.	Profile P.	Goethe	Telc	dlexDB	DeReWo	Subtlex	Leipzig	Google	Routledge
GerVLPro	1.00	.76 | .13	.50 | .31	.37 | **.19**	.47 | .29	.47 | .30	.60 | .36	.47 | .27	.89 | .00	.86 | .01	.91 | .03	.91 | .06	.72 | **.35**	.69 | **.42**
DaFLex	**.76** | .13	1.00	**.81** | .12	**.80** | .18	**.77** | .14	**.80** | .13	**.91** | .13	**.83** | .12	.81 | **.01**	.81 | **.05**	.66 | **.10**	.56 | **.18**	.66 | .19	**.79** | .18
Lingster	.50 | .31	.81 | .12	1.00	.34 | .16	.69 | **.45**	.67 | **.47**	.63 | .43	.46 | .29	**.98** | .00	**.99** | .01	**.96** | .02	.92 | .05	.62 | .27	.61 | .33
Aspekte	.37 | .19	.80 | **.18**	.34 | .16	1.00	.28 | .15	.29 | .14	.36 | .16	.30 | .14	.92 | .00	.95 | .01	.85 | .03	.77 | .07	.41 | .21	.40 | .25
Profile R.	.47 | .29	.77 | .14	.69 | .45	.28 | .15	1.00	1.00 | .88	.71 | .43	.53 | .31	.96 | .00	.97 | .01	.94 | .03	.86 | .06	.54 | .26	.53 | .31
Profile P.	.47 | .30	.80 | .13	.67 | **.47**	.29 | .14	1.00 | .88	1.00	.69 | **.46**	.51 | .32	.97 | .00	.97 | .01	.95 | .03	.87 | .06	.58 | .26	.56 | .31
Goethe	.60 | **.36**	**.91** | .13	.63 | .43	.36 | .16	.71 | .43	.69 | .46	1.00	.56 | **.37**	**.98** | .00	.97 | .01	**.96** | .02	**.95** | .05	**.73** | .30	.72 | .38
Telc	.47 | .27	.83 | .12	.46 | .29	.30 | .14	.53 | .31	.51 | .32	.56 | .37	1.00	.95 | .00	.98 | .01	.93 | .02	.87 | .05	.55 | .22	.54 | .27
dlexDB	.89 | .00	**.81** | .01	.98 | .00	.92 | .00	.96 | .00	**.97** | .00	**.98** | .00	.95 | .00	1.00	.65 | **.15**	.69 | .07	.74 | .03	.92 | .00	**.99** | .00
DeReWo	.86 | .01	**.81** | .05	**.99** | .01	**.95** | .01	**.97** | .01	**.97** | .01	.97 | .01	**.98** | .01	.65 | **.15**	1.00	.51 | .18	.60 | .10	.87 | .02	.98 | .02
Subtlex	**.91** | .03	.66 | .10	.96 | .02	.85 | .03	.94 | .03	.95 | .03	.96 | .02	.93 | .02	.69 | .07	.51 | **.18**	1.00	.61 | **.22**	.96 | .05	.98 | .04
Leipzig	**.91** | .06	.56 | .18	.92 | .05	.77 | .07	.86 | .06	.87 | .06	.95 | .05	.87 | .05	.74 | .03	.60 | .10	.61 | **.22**	1.00	**.97** | .11	**.99** | .09
Google	.72 | .35	.66 | **.19**	.62 | .27	.41 | .21	.54 | .26	.58 | .26	.73 | .30	.55 | .22	.92 | .00	.87 | .02	.96 | .05	.97 | .11	1.00	.81 | **.54**
Routledge	.69 | **.42**	.79 | .18	.61 | **.33**	.40 | **.25**	.53 | **.31**	.56 | **.31**	.72 | **.38**	.54 | **.27**	**.99** | .00	**.98** | .02	**.98** | .04	**.99** | .09	.81 | **.54**	1.00

Table 5: Pairwise Szymkiewicz–Simpson coefficient (left) and Jaccard similarity (right) between established graded vocabulary lists and frequency lists.

	GerVLPro	DaFLex	Lingster	Aspekte	Profile R.	Profile P.	Goethe	Telc
GerVLPro	1.00	.15 | .46	.17 | .44	.01 | .07	.15 | .38	.12 | .39	.16 | .41	.13 | .35
DaFLex	.15 | **.46**	1.00	.20 | .46	**.30** | **.21**	.21 | .45	.16 | .35	.24 | .46	.27 | .48
Lingster	**.17** | .44	.20 | .46	1.00	.04 | .06	.26 | .51	.21 | **.51**	**.31** | .54	.22 | .44
Aspekte	.01 | .07	**.30** | .21	.04 | .06	1.00	.05 | .09	.07 | .09	.04 | -.06	-.03 | .03
Profile R.	.15 | .38	.21 | .45	.26 | .51	.05 | .09	1.00	.34 | .72	.30 | **.57**	.24 | .44
Profile P.	.12 | .39	.16 | .35	.21 | .51	.07 | .09	.34 | .72	1.00	.23 | .50	.13 | .34
Goethe	.16 | .41	.24 | .46	**.31** | **.54**	.04 | -.06	**.30** | **.57**	**.23** | .50	1.00	**.29** | **.49**
Telc	.13 | .35	.27 | **.48**	.22 | .44	-.03 | .03	.24 | .44	.13 | .34	.29 | .49	1.00
Gpt-5 R.	.06 ± .04 | .24 ± .05							
Gpt-5 P.	.11 ± .01 | .31 ± .01							
Gpt-4o R.	.11 ± .01 | .29 ± .00							
Gpt-4o P.	.10 ± .00 | .29 ± .01							

Table 6: Pairwise Cohen's κ agreement (left) and QWK (right) between established graded vocabulary lists and LLM predictions (mean over three runs).

	GerVLPro	DaFLex	Lingster	Aspekte	Profile R.	Profile P.	Goethe	Telc	dlexDB	DeReWo	Subtlex	Leipzig	Google	Routledge
GerVLPro	1.00	2,857 | .40	1,671 | .39	1,398 | .09	1,745 | .35	1,684 | .34	1,809 | .39	1,500 | .37	3,345 | .08	3,230 | .14	3,397 | .27	3,399 | .09	2,693 | .11	2,585 | .13
DaFLex	2,857 | **.40**	1.00	2,694 | .42	3,951 | **.33**	3,145 | .40	2,881 | .38	2,737 | .45	2,667 | .46	17,104 | .09	16,952 | .13	13,857 | .17	11,689 | .09	4,311 | .13	3,910 | .17
Lingster	1,671 | .39	2,694 | .42	1.00	1,135 | .08	2,292 | .47	2,220 | .48	1,899 | .49	1,467 | .46	3,248 | .19	3,286 | .25	3,185 | .33	3,056 | .18	2,070 | .17	2,030 | .22
Aspekte	1,398 | .09	3,951 | .33	1,135 | .08	1.00	1,155 | .15	1,056 | .12	1,089 | -.09	979 | .10	4,535 | .16	4,679 | .16	4,199 | .17	3,787 | .15	2,027 | .19	1,947 | .21
Profile R.	1,745 | .35	3,145 | .40	2,292 | .47	1,155 | .15	1.00	3,594 | .77	2,135 | **.52**	1,710 | .44	3,955 | **.27**	3,987 | **.31**	3,870 | **.36**	3,505 | **.24**	2,228 | **.28**	2,153 | **.30**
Profile P.	1,684 | .34	2,881 | .38	2,220 | .48	1,056 | .12	3,594 | .77	1.00	2,087 | .50	1,642 | .44	3,470 | **.27**	3,495 | .30	3,417 | .35	3,134 | **.24**	2,081 | .25	2,026 | .28
Goethe	1,809 | .39	2,737 | .45	1,899 | **.49**	1,089 | -.09	2,135 | **.52**	2,087 | **.50**	1.00	1,684 | **.49**	2,950 | .19	2,941 | .23	2,913 | .32	2,865 | .17	2,195 | .21	2,185 | .24
Telc	1,500 | .37	2,667 | **.46**	1,467 | .46	979 | .10	1,710 | .44	1,642 | .44	1,684 | .49	1.00	3,068 | .20	3,142 | .21	2,982 | .27	2,795 | .17	1,765 | .23	1,732 | .25
dlexDB	3,345 | .08	17,104 | .09	3,248 | .19	4,535 | .16	3,955 | .27	3,470 | .27	2,950 | .19	3,068 | .20	1.00	208,093 | .43	93,083 | .37	41,397 | .34	5,994 | .52	4,886 | .53
DeReWo	3,230 | .14	16,952 | .13	3,286 | .25	4,679 | .16	3,987 | .31	3,495 | .30	2,941 | .23	3,142 | .21	208,093 | .43	1.00	69,066 | **.48**	33,895 | **.56**	5,694 | .52	4,834 | **.60**
Subtlex	3,397 | **.27**	13,857 | **.17**	3,185 | **.33**	4,199 | .17	3,870 | **.36**	3,417 | **.35**	2,913 | **.32**	2,982 | **.27**	93,083 | .37	69,066 | .48	1.00	34,321 | .32	6,311 | .41	4,853 | .42
Leipzig	3,399 | .09	11,689 | .09	3,056 | .18	3,787 | .15	3,505 | .24	3,134 | .24	2,865 | .17	2,795 | .17	41,397 | .34	33,895 | .56	34,321 | .32	1.00	6,359 | .47	4,887 | .48
Google	2,693 | .11	4,311 | .13	2,070 | .17	2,027 | .19	2,228 | .28	2,081 | .25	2,195 | .21	1,765 | .23	5,994 | .52	5,694 | .52	6,311 | .41	6,359 | .47	1.00	4,030 | .57
Routledge	2,585 | .13	3,910 | **.17**	2,030 | .22	1,947 | **.21**	2,153 | .30	2,026 | .28	2,185 | .24	1,732 | .25	4,886 | **.53**	4,834 | **.60**	4,853 | .42	4,887 | .48	4,030 | **.57**	1.00
Gpt-5 R.	.29 ± .01													
Gpt-5 P.	.29 ± .02													
Gpt-4o R.	.29 ± .00													
Gpt-4o P.	.29 ± .01													

Table 7: Intersection size (left; if applicable) and pairwise Kendall's τ_b rank correlation (right) between established graded vocabulary lists, frequency lists, and LLM predictions (mean over three runs).

Resource	Jacc_within	Jacc_across	Szym_within	Szym_across	Kendall_within	Kendall_across
GerVLPro	0.26	0.15	0.52	0.83	0.33	0.14
DaFLex	0.14	0.12	0.81	0.72	0.41	0.13
Lingster	0.32	0.11	0.59	0.85	0.40	0.22
Aspekte	0.16	0.10	0.39	0.72	0.11	0.17
Profile R.	0.30	0.11	0.58	0.80	0.39	0.29
Profile P.	0.30	0.11	0.57	0.82	0.38	0.28
Goethe	0.33	0.13	0.64	0.89	0.39	0.23
Telc	0.26	0.10	0.52	0.80	0.39	0.22
Mean	0.26	0.11	0.58	0.80	0.35	0.21
dlexDB	0.05	0.00	0.80	0.93	0.44	0.18
DeReWo	0.09	0.02	0.72	0.94	0.52	0.22
Subtlex	0.11	0.04	0.75	0.90	0.40	0.28
Leipzig	0.11	0.07	0.78	0.84	0.43	0.17
Google	0.14	0.26	0.91	0.60	0.50	0.20
Routledge	0.14	0.31	0.95	0.61	0.52	0.23
Mean	0.11	0.12	0.82	0.80	0.47	0.21

Table 8: Per-resource mean similarity and alignment values. Each value represents the average similarity/alignment of a resource to all other resources within the same type (within) or to all resources of the other type (across). The within values for Profile deutsch R. and P. exclude the direct comparison between the two resources because the two lists are interdependent. These per-resource means constitute the observational units used in the resource-level statistical analyses.

Metric	Comparison	p	p_{Holm}	Effect size
Jaccard	within-GVL vs. cross-type	0.0078	0.0234	RBC = 1.000
	within-frequency-list vs. cross-type	1.0000	1.0000	RBC = -0.048
	within-GVL vs. within-frequency-list	0.0027	0.0053	δ = 0.917
	GerVLPro ($\times$GVLs vs $\times$frequency lists)	0.2949	0.2949	δ = 0.381
Szymkiewicz–Simpson	within-GVL vs. cross-type	0.0156	0.0313	RBC = -0.944
	within-frequency-list vs. cross-type	1.0000	1.0000	RBC = 0.048
	within-GVL vs. within-frequency-list	0.0080	0.0080	δ = -0.833
	GerVLPro ($\times$GVLs vs $\times$frequency lists)	0.0078	0.0234	δ = -0.905
Kendall's τ_b	within-GVL vs. cross-type	0.0156	0.0313	RBC = 0.944
	within-frequency-list vs. cross-type	0.0313	0.0938	RBC = 1.000
	within-GVL vs. within-frequency-list	0.0013	0.0040	δ = -0.958
	GerVLPro ($\times$GVLs vs $\times$frequency lists)	0.0219	0.0438	δ = 0.786

Table 9: Non-parametric statistical comparisons of similarity and alignment metrics. Within–across comparisons were tested using Wilcoxon signed-rank tests; between-group comparisons used Mann–Whitney U tests. Holm correction was applied across metrics within each comparison family. Effect sizes are reported as rank-biserial correlation (RBC) for Wilcoxon tests and Cliff's δ for Mann–Whitney tests.