data

MDPI

Data Descriptor

Introducing DeReKoGram: A Novel Frequency Dataset with Lemma and Part-of-Speech Information for German

Sascha Wolfer *, Alexander Koplenig, Marc Kupietz and Carolin Müller-Spitzer

Leibniz Institute for the German Language (IDS), 68161 Mannheim, Germany
* Correspondence: wolfer@ids-mannheim.de

Abstract: We introduce DeReKoGram, a novel frequency dataset containing lemma and part-of-speech (POS) information for 1-, 2-, and 3-grams from the German Reference Corpus. The dataset contains information based on a corpus of 43.2 billion tokens and is divided into 16 parts based on 16 corpus folds. We describe how the dataset was created and structured. By evaluating the distribution over the 16 folds, we show that it is possible to work with a subset of the folds in many use cases (e.g., to save computational resources). In a case study, we investigate the growth of vocabulary (as well as the number of hapax legomena) as an increasing number of folds are included in the analysis. We cross-combine this with the various cleaning stages of the dataset. We also give some guidance in the form of Python, R, and Stata markdown scripts on how to work with the resource.

Dataset: https://www.owid.de/plus/derekogram/ (along with information and sample code).

Dataset License: DeReKo license (non-commercial, academic).

Keywords: language; n-grams; corpus frequency; dataset; German; vocabulary growth

Citation: Wolfer, S.; Koplenig, A.; Kupietz, M.; Müller-Spitzer, C. Introducing DeReKoGram: A Novel Frequency Dataset with Lemma and Part-of-Speech Information for German. *Data* **2023**, *8*, 170. https://doi.org/10.3390/data8110170

Academic Editor: Francesco M. Donini

Received: 27 September 2023
Revised: 22 October 2023
Accepted: 6 November 2023
Published: 10 November 2023

1. Introduction

In corpus linguistics, we are interested in systematic analyses of large collections of texts (corpora) to gain insights into language usage patterns, structure, and meaning. Here, full-text corpora (i.e., collections where the sequential ordering of words and their distribution over documents are preserved) are the most general resource, in the sense that all the information that might become relevant during a research endeavor can be derived, especially all measures that rely on sequential information. However, a lot of interesting language resources that we want to include in corpora are subject to restrictions under national copyright law or licenses. The latter is (often) a matter of negotiation between licensees and licensors and might even differ between text sources [1]. This not only complicates things considerably when compiling corpora but is also an obstacle for efforts regarding standards of open science [2]. In other words, in an ideal world shaped for corpus-linguistic research, we would be able to distribute all corpus resources freely for everyone who wants to conduct original research or replicate findings based on these resources.

Since, unfortunately, we do not live in this ideal world, the corpus-linguistic community often has to compromise when distributing corpora. For example, we might prevent access to the full-text corpora themselves and give users the opportunity to access them via corpus platforms that are designed to search for specific patterns and allow for a specified set of analyses and the extraction of small parts of the corpus (e.g., keyword-in-context outputs). Several such corpus platforms are available for the German language. COSMAS II and KorAP (COSMAS II is available via https://cosmas2.ids-mannheim.de. KorAP is available via https://korap.ids-mannheim.de (accessed on 7 November 2023)) and provide access to the corpus on which the dataset presented here is based. Access to another German

corpus resource is offered by the corpus search of the Digital Dictionary of the German Language (DWDS) (the corpus search of the DWDS is available via https://www.dwds.de/r (accessed on 7 November 2023)). While opening many linguistic research avenues, there are, however, also many large-scale corpus-linguistic procedures (e.g., calculations of transition probabilities for all 2-grams, i.e., two-word sequences, in a corpus) where access paths through corpus platforms, and this includes the ones mentioned above, are not sufficient.

For such applications, *n*-gram (*n*-grams are adjacent sequences of words where *n* encodes the window size: 1-grams (unigrams) are single words, 2-grams (bigrams) are adjacent sequences of two words, and so on) frequency lists are another possibility for allowing researchers to leverage large-scale corpora. Maybe the best-known of such lists are the Google Books corpora [3], which are also available for German but come with their own restrictions, like a frequency threshold and other issues [4,5]. Frequency lists enable researchers to devise a range of measures (for example, the transitional probabilities mentioned above or more sophisticated language models based on *n*-grams) and can be used to train word-level machine learning models with a fixed context [6]. Basically, frequency lists become relevant whenever all items contained in the corpus need to be factored into the analyses. Frequency lists can also inform joint research with other linguistic disciplines. Word frequencies (or derived measures) might become relevant in choosing stimuli for psycholinguistic studies (e.g., if the corpus frequency of experimental items should be held constant or manipulated in an experimental setting). They are also used as covariates or predictors for behavioral data, such as eye movements during reading [7,8] or ERP data [9,10].

Here, we introduce new (up to 3-gram) frequency lists based on the German Reference Corpus "DeReKo" [1] that contain lemma, part-of-speech (POS), and frequency information from a corpus of around 43.2 billion tokens. We will first describe which parts of DeReKo provide the basis for the frequency list and introduce the data structure. We will then evaluate the distribution of data over the 16 parts (henceforth "folds", Section 3) and present a case study (Section 4) on vocabulary growth utilizing several cleaning stages in the dataset. Further details and sample code for using the dataset are available at https://www.owid.de/plus/derekogram/ (accessed on 7 November 2023) (Supplementary Materials). With this accompanying webpage, we would like to facilitate work with DeReKoGram for as many researchers as possible. We give pointers on how to back-translate the integer codes (see Section 2 for an explanation) to human-readable wordforms and lemmas, aggregate, lower (i.e., transform all characters to lower-case), and clean the dataset and search for specific patterns based on a linguistic example. For Python, we also show how to train smoothed *n*-gram language models with DeReKoGram. For Stata and R code for another linguistic application please see the supplementary material of a previous study [11].

2. Data Description

DeReKo, the corpus DeReKoGram is derived from, currently contains around 50 billion tokens and comprises a multitude of genres, such as (for the most part) newspaper texts (newspapers from most regions of German-speaking countries are included; in addition, the included newspapers vary in the size of their distribution area and their political orientation), fiction, or specialized texts. The corpus currently grows by around 3 billion words per year [1].

2.1. Data Selection

For DeReKoGram, the dataset we are introducing here, a large subset of DeReKo was selected, tokenized with the KorAP tokenizer [12], and tagged with the TreeTagger [13]. The basis of the sample is a 2020 release of DeReKo, from which we excluded sub-corpora with high proportions of foreign-language passages and high proportions of non-redacted texts (e.g., material taken from Wikipedia discussions), as well as some material where publication in this format would be legally problematic. We provide 1-, 2-, and 3-gram frequency lists with and without punctuation (as identified by the TreeTagger). Note that

excluding punctuation increases the number of lexical items contained, for example, in some 3-grams. The original corpus sequence *ich glaube , dass* (Engl. "I believe that") counts as a 4-gram and is thus not included in the dataset with punctuation. In the dataset without punctuation, however, the 3-gram *ich glaube dass* is included. This might be beneficial for all analyses where punctuation is not important. We included begin- and end-of-document markers *«START»* and *«END»*. Please refer to Koplenig et al. [11] for an explanation. Note that *n*-grams crossing sentence boundaries can be excluded by deleting *n*-grams based on the POS tag *$.*, which identifies the end of a sentence. Hence, *n*-grams crossing sentence boundaries cannot be excluded from the dataset without punctuation.

2.2. Data Structure

One row in each dataset consists of integer IDs for the wordform and lemma, POS tag, and the associated corpus frequency. Hence, 1-gram datasets have 4, 2-gram datasets 7, and 3-gram datasets 10 columns. Table 1 shows an excerpt from the bigram data with punctuation for fold 12. No frequency threshold or stop word list was employed regarding the inclusion of *n*-grams. The dataset was divided into 16 folds (16 folds have the advantage (e.g., over 10 folds) that one can create 5 datasets where the size is doubled in each step by using 1, 2, 4, 8, and 16 folds) where each fold consists of 1/16 of all corpus documents (randomly sampled without replacement). The data are stored using integer IDs [14]; i.e., each wordform and lemma is independently mapped to a unique integer. This drastically reduces file sizes and allows for faster and more resource-efficient processing for research questions where the actual word form or lemma is not required. We also provide the dictionaries necessary to map integer IDs to their respective wordforms and lemmas and routines for Python, R, and Stata to use these dictionaries. To enable a resource-efficient approach, we also provide dictionaries for each of the 16 folds separately.

Table 1. Sample bigram data for fold 12 with punctuation (rows 38 to 42 of the respective file). Column 1: wordform code for the first element of the bigram, column 2: lemma code for the first element of the bigram, column 3: part-of-speech tag for the first element of the bigram, columns 4 to 6: respective information for the second element of the bigram, column 7: frequency of the bigram in fold 12. For convenience, columns 8 to 11 show the back-translated ("clear") wordforms and lemmas. Note that only columns 1 to 7 are included in the dataset. Columns 8 to 11 were back-translated with the lemma and wordform dictionaries for fold 12, which are also available for download.

Form 1	Lem. 1	POS 1	Form 2	Lem. 2	POS 2	Freq.	Form 1 (Clear)	Lem. 1 (Clear)	Form 2 (Clear)	Lem. 2 (Clear)
9	12	APPR	2	0	ART	1,775,493	mit	mit	der	die
1	2	$,	6	4	APPR	1,762,655	,	,	in	in
15	17	APPR	7	0	ART	1,761,184	für	für	den	die
1	2	$,	50	39	KOUS	1,760,171	,	,	wie	wie
1	2	$,	58	41	ADV	1,721,500	,	,	so	so

3. Evaluation of Fold Distribution

In what follows, we evaluate the underlying sampling process, which was carried out using the cryptographic hash function BLAKE2b [15] to create the 16 folds by comparing several frequency parameters over all folds. We did so for a corpus that had not been cleaned in any way and for raw and lowered, i.e., where all characters had been converted to lower case, datasets separately.

We chose five "direct" measures: (1) number of different wordforms, i.e. the number of wordform types. One may wonder why we do not evaluate the number of wordform tokens. However, this would be somewhat trivial, since the sampling process randomly assigned corpus documents to folds (see Section 2). The length of the documents (measured in tokens) may therefore differ only minimally between the folds. With the evaluations presented here, we aim to show that this similarity also holds for somewhat more complex measures. (2) number of hapax legomena, (3) number of different wordforms tagged

as normal nouns, (4) function words (part-of-speech tags are taken from the automatic (and non-corrected) classification of the TreeTagger, which uses the Stuttgart-Tübingen-Tagset [16]. The following tags were classified as function words: APPO, APPR, APPRART, APZR, ART, KOKOM, KON, KOUI, KOUS, PAV, PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS, PRF, PTKA, PTKNEG, PTKZU, PWAT, PWAV, PWS) as well as (5) the ratio between the number of name/noun tokens and finite verb tokens. We also report two "derived" measures: (6) type-token ratio (where wordforms are treated as types) and (7) the entropy (entropy h is defined as $h = -\sum_{i=1}^{N} p_i * \log_2\ p_i$, where p_i is the relative frequency of wordform i and N is the number of wordform types in the corpus fold) of the complete frequency distribution for wordforms. For all these measures, we calculated the difference in percentages between the minimum and maximum value over the 16 folds: Δ1-gram = (max (x_1, x_2, …, x_{16}) − min(x_1, x_2, …, x_{16}))/max(x_1, x_2, …, x_{16}); where x_n is the value of the respective measure in the nth fold (accordingly for 2-grams). Ideally, the difference percentages should be close to zero, which would indicate no difference between the corpus folds. We also calculated coefficients of variation (CV, also known as relative standard deviation, RSD, also reported as percentages) defined as $CV = \sigma_x/\mu_x$, where σ_x is the standard deviation and μ_x the mean of the respective measure over the 16 folds.

Table 2 shows that the highest difference percentages for 1-grams and the highest coefficients of variation can be observed for the number of different wordforms tagged as function words. This is presumably because the number of different wordforms of function words is indeed much smaller than, e.g., nouns. For example, for the lowered folds, the number of different wordforms ranges from 764 to 772. On the 2-gram level, the difference percentages and the coefficients of variation are also very low.

Table 2. Difference percentages on the 1-gram level (Δ1-gram, column 2 and 3) and coefficients of variation (CV, column 4 and 5) for a number of frequency measures for the 16 raw and lowered 1-gram folds. Column 6 and 7 hold information for raw 2-gram folds (column 6: difference percentages; column 7: coefficients of variation). All values are reported as percentages.

Measure	Δ1-Gram		CV		2-Gram (Raw Only)	
	Raw	Low	Raw	Low	Δ2-Gram	CV
Number of wordforms/2-grams	0.22	0.23	0.05	0.06	0.21	0.05
Number of hapax legomena	0.26	0.29	0.07	0.07	0.23	0.06
Number of different wordforms tagged as nouns	0.20	0.22	0.05	0.06	-	
Number of different wordforms tagged as function words	1.85	1.04	0.5	0.3	-	
NE + NN tokens/fin. verb tokens	0.10		0.02		-	
Entropy	0.01	0.01	<0.01	<0.01	0.01	<0.01
Type-token ratio	0.17	0.17	0.05	0.05	0.13	0.03

Another way to compare the folds on the 1-gram level is to compare the frequency ranks of wordforms over all 16 corpus folds. Since each wordform has been assigned a numerical code based on its token frequency rank in the overall corpus, we can correlate the codes of the first n wordforms (ranked by their frequency in each fold) with the respective first n codes in every other fold. This yields a 16-by-16 correlation matrix with Spearman's correlation coefficients ρ. Ideally, all ρs should be very close to 1, which would indicate a perfect match of frequency ranks over all corpus folds. We calculated correlation matrices with increasing values of n for raw corpus folds without any cleaning and extracted the lowest ρ ($\rho_{\min}$) for each value of n.

Table 3 shows that all values for $\rho_{\min}$ are very close to 1 but are steadily decreasing as n increases. This is because the token frequencies in higher ranks (= lower frequencies) tend to produce more ties; i.e., wordforms with equal frequencies. In other words, wordform frequencies lose their distinguishing power for higher values of n. For example, in fold 4, the token frequency for all wordforms with the in-fold token frequency ranks of 995,506

to 1,018,394 (22,889 wordforms) is 28. To demonstrate that a similar pattern can also be observed for other folds: in fold 13, all tokens with the in-fold frequency ranks 994,655 to 1,017,581 (22,927 tokens) share a frequency of 28. This lack of distinguishing power in the lower frequency ranges consequently leads to lower values of ρ_{min}.

Table 3. Lowest spearman's correlation coefficients ρ for the first n wordform codes ranked by their in-fold frequency from 16-by-16 correlation matrices correlating all folds.

First *n* Wordforms	Spearman's ρ_{min}
100	0.9999
1000	0.9999
10,000	0.9997
100,000	0.9975
1,000,000	0.9660

Furthermore, we Spearman-correlated all POS frequency distributions in the 16 folds with all other folds, again yielding a 16-by-16 correlation matrix. Here, all ρs were larger than 0.9999.

Finally, we checked whether the frequency distributions as captured by the Zipf-Mandelbrot power law [17,18] yielded comparable parameters over all folds. Using the R package *zipfR* [19], we fitted large number of rare events (LNRE) models to the frequency distributions of (unigram) wordforms for each fold. Indeed, the parameter ranges were very narrow, with the exponent parameter α ranging from 0.6316296 to 0.6316336 and the second free parameter β ranging from 0.00043567 to 0.00043572.

Given these comparisons, we concluded that the sampling process produced 16 homogeneous folds. Methodologically, this means that many analyses performed using all 16 parts should produce very similar results when performed using only one fold.

4. Case Study: Vocabulary Growth

First, it must be noted that we do not understand vocabulary growth here in the sense of language acquisition research. Here, we investigate how vocabulary develops as the corpus grows [20]. Of course, as corpus size increases, we would expect the vocabulary to grow. However, vocabulary growth curves should differ according to the amount and type of cleaning we apply to the frequency lists, because an increasing number of wordforms are being excluded from the dataset. For this case study, we applied various cleaning stages:

A. no cleaning at all;
B. exclusion of punctuation, names, and start-end-symbols (all identified via their respective POS tags), URLs, and wordforms only consisting of numbers (both identified by regular expressions);
C. exclusion of wordforms containing numbers;
D. exclusion of wordforms that contain upper-case letters following lower-case letters (in this cleaning stage, we exclude wordforms with non-conventional capitalization (e.g., *dEr*) while, at the same time, keeping capitalized abbreviations (e.g., *NATO*). For this cleaning stage, we only use the raw version because there cannot be any difference to the lowered version);
E. exclusion of wordforms where the TreeTagger could not assign a lemma;
F. selection of wordforms that are themselves (or the associated lemma (note that this means that, for example, the inflected wordform *Weihnachtsmannes* is also included, although it is not on the BLL itself, but the associated lemma *Weihnachtsmann* is. Another example is the wordform *u.*, which is shorthand for the lemma *und*)) on a basic lemma list (BLL) of New High German standard language to identify a set of conventionalized word forms [21]. For more information regarding this basic lemma list, please refer to Koplenig et al. [11].

Cleaning stages A through D are cumulative. For example, cleaning stage D incorporates stages B and C. Stages E and F, however, both rely on stage D, because they can

be understood as being equivalent regarding their aim: identifying "true" lemmas and wordforms of German. We chose these cleaning stages because they represent very general selections/exclusions in potential research projects in corpus or computational linguistics. One could also think of these cleaning stages as becoming more rigorous towards the core vocabulary of the language in each step. Of course, the datasets provided can also be used to test other selections adapted to specific research questions (e.g., only selecting certain POS or applying frequency thresholds).

4.1. Number of Wordform Types

We will first examine how the number of wordform types develops when including an increasing number of the 16 corpus folds. Figure 1 shows that the first four cleaning stages (panels A through D) exhibit roughly the same overall pattern for raw and lowered versions: the vocabulary growth curves do not show clear signs of approaching a ceiling value. This finding replicates several studies for English language corpora that were summarized by [22], who also "failed to find any flattening of the predicted linear curve, indicating that the pool of possible word types was still far from exhausted". There is, in other words, "no indication of a stop to the growth", which is an instantiation of Herdan's [23] or Heaps' [24] law.

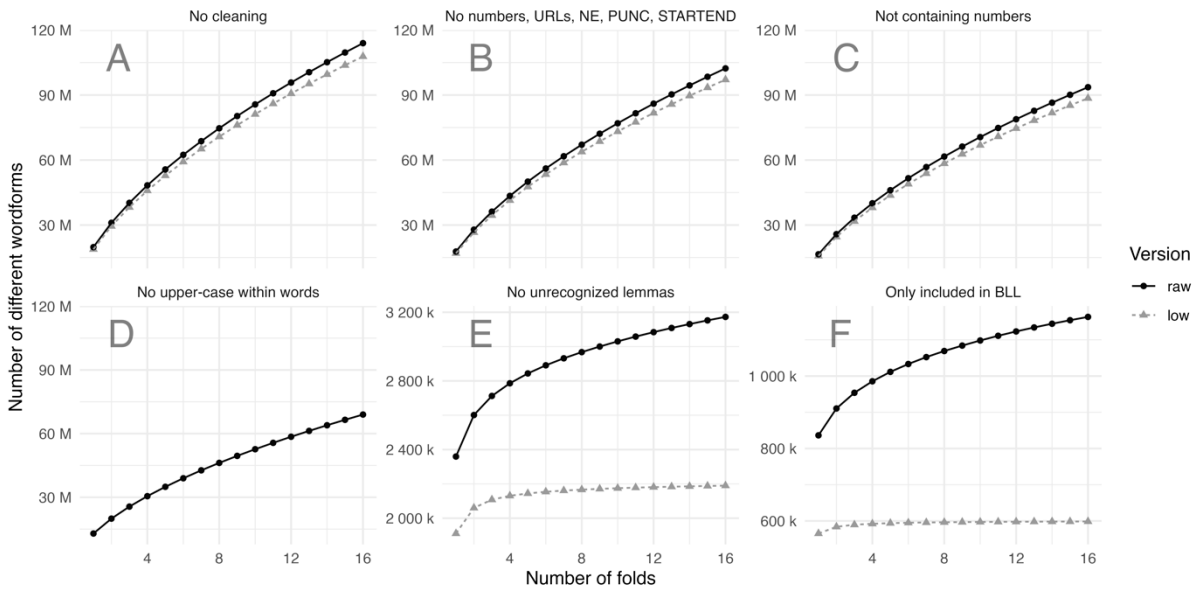

Figure 1. Number of different wordforms (M: million; k: thousand) in relationship to the number of included folds (*x*-axis), cleaning stage (panels (**A**–**F**), see panel title), and lowered (grey lines) vs. raw corpus (black lines). Note the varying *y*-axes between panels.

The final two cleaning stages (panels E and F) quickly showed asymptotical behavior, but only in the vocabulary growth curves for the lowered dataset (grey lines). This is especially true for cleaning stage F, where we restricted the corpus to a fixed set of wordforms. For the raw corpus version, many new forms were still observed, even approaching the full corpus.

The same data can also be visualized as percentage increases as increasing folds are included (Figure 2). There is virtually no difference between the raw vs. lowered versions for the first three cleaning stages (panels A through C), and the number of observed wordforms still increases in the last step (15 to 16 folds), by approx. 4%. This is remarkably close to the figure reported by [25], who used a corpus of ten introductory psychology textbooks and investigated the growing lexical diversity when adding whole textbooks to their sample one after another. However, this similarity might be coincidental, because Miller and Biber used lemmas (not wordforms) and a much more thematically restricted corpus in their study. Moreover, their overall corpus was much smaller.

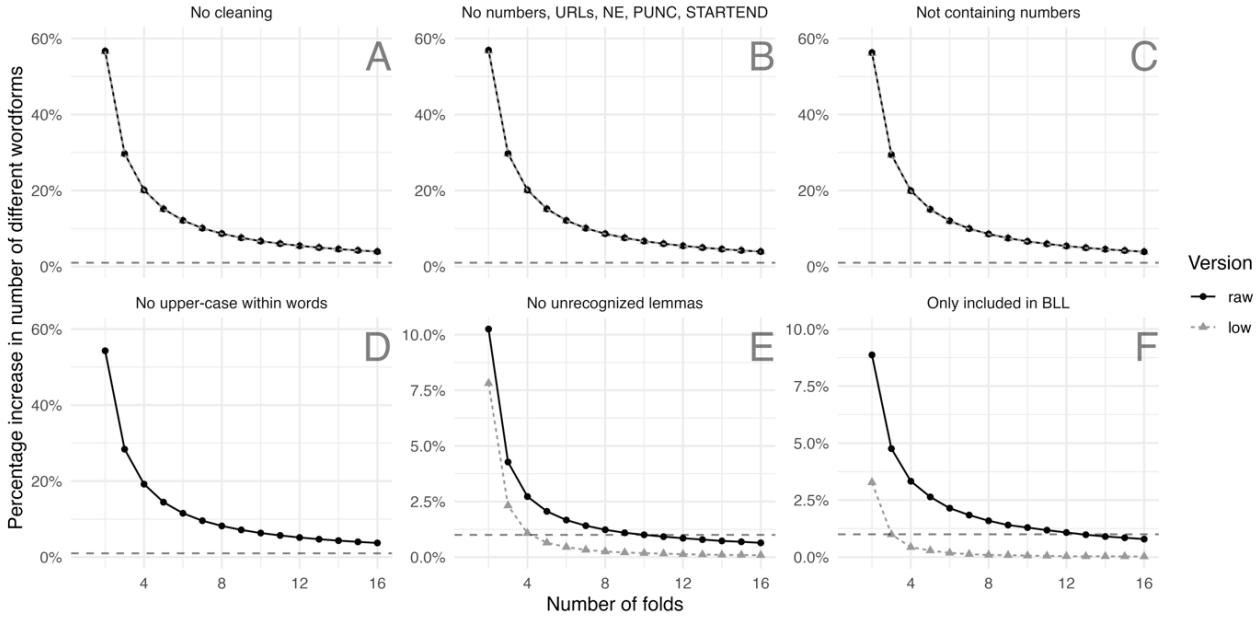

Figure 2. Percentage increase in the number of different wordforms in relationship to the number of included folds (*x*-axis), cleaning stage (panels (**A**–**F**), see panel title), and lowered (grey lines) vs. raw corpus (black lines). Note the varying *y*-axes between panels. The dashed horizontal lines mark a percentage increase of 1%.

The results for cleaning stages E and F are different: the fourth step (4 to 5 folds) in panel E shows a percentage increase of below 1% for the lowered dataset. In panel F, it is already the second step (2 to 3 folds). The lowest percentage increase is observed for the last step of the lowered corpora for cleaning stage F: 0.03% (or, in absolute numbers, 157 newly observed wordform). One of these wordforms (*habilitationsprofessur*) appears 5 times, and 14 wordforms appeared twice. Note that these numbers and the wordforms themselves depend on which fold is added last to the growing dataset. Here, we simply included folds 1 to 16 subsequently. So, for cleaning stages E and F, we can conclude that the boundless growth in vocabulary is far less pronounced than in the previous cleaning stages. This makes sense given that E and F are the cleaning stages where we tried to identify a basic set of German wordforms.

4.2. Number of Hapax Legomena

Hapax legomena (henceforth: HL), items appearing only once in a frequency list, are often used in calculations of quantitative linguistic measures, for example in analyses concerning productivity [26]. It is therefore interesting to see how the number of HL is influenced by corpus size (=number of folds), cleaning stage, and lowering of the dataset.

In terms of frequency, we would expect more HL in larger corpora, especially when no or rather light (cleaning stages A through D) cleaning is performed, because we observe more and more non-canonical wordforms as the corpus becomes larger. However, it is hard to hypothesize what the pattern would look like for the last two cleaning stages E and F.

There are no considerable differences between the first four cleaning stages (panels A through D in Figure 3) and, indeed, none of the curves show any signs of approaching a point where no new HL are being added after a specific corpus size, which would be indicated by the curve approaching a horizontal asymptote. The highest number of HL was observed for the full raw corpus without any cleaning in panel A of Figure 3 (66,149,313 wordforms, 58.0% of all wordforms).

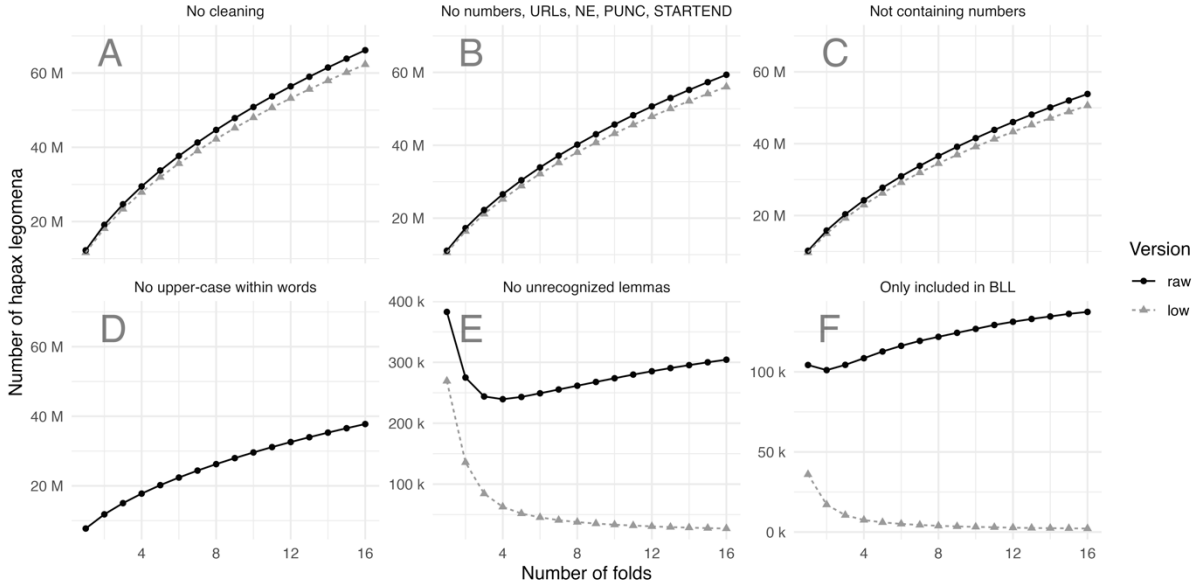

Figure 3. Number of hapax legomena (wordforms) in relationship to the number of included folds (*x*-axis), cleaning stage (panels (**A**–**F**), see panel title), and lowered vs. raw corpus (line color). Note the varying *y*-axes between panels.

The last two cleaning stages (panels E and F), which we consider quite "strict", show a different pattern compared to the first four stages. In these cleaning stages, the trajectory of the curves for raw and lowered datasets differs. For the lowered dataset, the number of HL steadily decreases, which is what we would expect, given datasets where only recognized lemmas (stage E) or elements from a well-defined word list (stage F) are allowed. Consequently, the lowest number of HL was observed for the full corpus in cleaning stage F for a lowered dataset (2205 wordforms), which account for 0.37% of all wordforms at this point.

For the raw version, it is especially noteworthy that the number of HL first decreased and then increased again. In stage E, the lowest point of the curve was reached for 4 folds (239,461 HL, 8.6% of all wordforms) with steadily rising counts until the corpus was complete (304,293 HL, 9.6%). To get an idea of where this effect comes from, we can look at the HL in the complete 16-fold dataset that were not observed in the 4-fold dataset (247,607 wordforms). These 228,156 HL had to be added somewhere between the inclusion of folds 5 to 16. A total of 51.8% of these wordforms (n = 118,076) consisted of upper-case letters only (in stage D, we only excluded wordforms where upper-case letters follow at least one lower-case letter). Another 7.9% (n = 18,111) begin with more than one upper-case letter (e.g., *COusinen* or *HERZstiche*), also indicating irregular capitalizations of wordforms that would not be hapax legomena if capitalized in a regular way. Thus, irregular capitalization seemed to play a large role in the increasing number of hapax legomena after the few initial folds for the final cleaning stages of raw corpora.

A total of 35,014 HL (15,3%) had an upper-case letter only in the word-initial position. Most of these wordforms turned out to be sentence-initial or nominalized forms of adjectives (*Flauschigsten*, *Abwischbarer*) and verbs (*Durchtauchte, Lullten*) or compound nouns (*Waldgesundheitsprogramm*, *Dampflokomotivkessel*). We made this observation by sampling 200 of these wordforms: 128 are adjectives or verbs with a word-initial upper-case letter; 62 are compound nouns (41 with two, 18 with three, and 3 with four constituent parts). In addition, several forms show alternative spellings of umlauts (e.g., *Gegenstaendlich* instead of *Gegenständlich*) or ß/ss alternation (e.g., *Fussten* or *Verkehrskompaß*) or simply typos (*Unwahrschienlich* instead of *Unwahrscheinlich*). Since all the effects reported above involve capitalization, they were not observed for lowered corpora. Hence, the diverging patterns for the respective curves in panels E and F in Figure 3.

5. Summary

We presented a novel frequency list dataset for German, DeReKoGram, containing all 1-, 2-, and 3-grams in a large subset (43.2 billion tokens) of the German Reference Corpus DeReKo with lemma and POS information. We evaluated the distribution of the corpus material over the 16 folds and concluded that, at least for a number of analyses, results obtained using one fold are generalizable to the whole corpus. This should make it more convenient for people with fewer computational resources to work with the frequency lists. In a case study, we showed how the size of the wordform vocabulary develops if an increasing number of corpus folds are included. We cross-combined this with five different cleaning stages and a raw vs. lowered version of the dataset. We paid particular attention to the number of hapax legomena that were found in the growing datasets. To download the dataset and get started working with DeReKoGram, you can visit the OWIDplus page at https://www.owid.de/plus/derekogram/ (accessed on 7 November 2023).

Supplementary Materials: Sample code for Python, R, and Stata is available via https://www.owid.de/plus/derekogram (accessed on 7 November 2023) and https://github.com/saschawo/DeReKoGram (accessed on 7 November 2023).

Author Contributions: Conceptualization, all authors; methodology, S.W., A.K., and M.K.; software, S.W., A.K., and M.K.; validation, S.W., A.K., and M.K.; formal analysis, S.W., and A.K.; investigation, S.W. and A.K.; resources, S.W., A.K., and M.K.; data curation, S.W., A.K., and M.K.; writing—original draft preparation, S.W.; writing—review and editing, all authors; visualization, S.W.; supervision, C.M.-S.; All authors have read and agreed to the published version of the manuscript.

Funding: The publication of this article was funded by the Open Access Fund of the Leibniz Association.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset generated and analyzed in this contribution is available from a persistent research data repository of the Leibniz Institute for the German Language. You can access the data via https://www.owid.de/plus/derekogram (accessed on 7 November 2023) after signing an EULA with an academic, non-commercial license. On this page, you can also find sample code in markdown documents for Python, R, and Stata. The data can also be accessed via http://hdl.handle.net/10932/00-0630-C8F8-FA80-F401-4 (accessed on 7 November 2023).

Acknowledgments: We thank our colleague Peter Fankhauser for many fruitful discussions during the initial stages of the project. We also thank Pawel Kamocki, who assisted us in clarifying the legal conditions of publishing the dataset, and Denis Arnold for publishing the datasets in the IDS repository, as well as Frank Michaelis for setting up the OWIDplus page.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kupietz, M.; Lüngen, H.; Kamocki, P.; Witt, A. The German Reference Corpus DeReKo: New Developments—New Opportunities. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7 May 2018; Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., et al., Eds.; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.
2. Marsden, E. Open Science and Transparency in Applied Linguistics Research. In *The Encyclopedia of Applied Linguistics*; Chapelle, C.A., Ed.; Wiley: Hoboken, NJ, USA, 2019; pp. 1–10. ISBN 978-1-4051-9473-0.
3. Michel, J.-B.; Shen, Y.K.; Aiden, A.P.; Verses, A.; Gray, M.K.; The Google Books Team; Pickett, J.P.; Hoiberg, D.; Clancy, D.; Norvig, P.; et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science* **2011**, *331*, 176–182. [CrossRef] [PubMed]
4. Pechenick, E.A.; Danforth, C.M.; Dodds, P.S. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLoS ONE* **2015**, *10*, e0137041. [CrossRef] [PubMed]
5. Schmidt, B.; Piantadosi, S.T.; Mahowald, K. Uncontrolled Corpus Composition Drives an Apparent Surge in Cognitive Distortions. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2115010118. [CrossRef] [PubMed]
6. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 3rd ed.; 2023. Available online: https://web.stanford.edu/~jurafsky/slp3/ (accessed on 7 November 2023).

7. Frisson, S.; Rayner, K.; Pickering, M.J. Effects of Contextual Predictability and Transitional Probability on Eye Movements During Reading. *J. Exp. Psychol. Learn. Mem. Cogn.* **2005**, *31*, 862–877. [CrossRef] [PubMed]
8. Kliegl, R.; Grabner, E.; Rolfs, M.; Engbert, R. Length, Frequency, and Predictability Effects of Words on Eye Movements in Reading. *Eur. J. Cogn. Psychol.* **2004**, *16*, 262–284. [CrossRef]
9. Hauk, O.; Pulvermüller, F. Effects of Word Length and Frequency on the Human Event-Related Potential. *Clin. Neurophysiol.* **2004**, *115*, 1090–1103. [CrossRef] [PubMed]
10. Hendrix, P.; Bolger, P.; Baayen, H. Distinct ERP Signatures of Word Frequency, Phrase Frequency, and Prototypicality in Speech Production. *J. Exp. Psychol. Learn. Mem. Cogn.* **2017**, *43*, 128–149. [CrossRef] [PubMed]
11. Koplenig, A.; Kupietz, M.; Wolfer, S. Testing the Relationship between Word Length, Frequency, and Predictability Based on the German Reference Corpus. *Cogn. Sci.* **2022**, *46*, e13090. [CrossRef] [PubMed]
12. Diewald, N.; Kupietz, M.; Lüngen, H. Tokenizing on Scale. Preprocessing Large Text Corpora on the Lexical and Sentence Level. In Proceedings of the Dictionaries and Society, Proceedings of the XX EURALEX International Congress, Mannheim, Germany, 12–16 July 2022; Klosa-Kückelhaus, A., Engelberg, S., Möhrs, C., Storjohann, P., Eds.; IDS-Verlag: Mannheim, Germany, 2022; pp. 208–221.
13. Schmid, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, 6–8 July 1994.
14. Brants, T.; Popat, A.C.; Xu, P.; Och, F.J.; Dean, J. Large Language Models in Machine Translation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007; Association for Computational Linguistics: Stroudsburg, PA, USA, 2007; pp. 858–867.
15. Aumasson, J.-P.; Meier, W.; Phan, R.C.-W.; Henzen, L. BLAKE2. In *The Hash Function BLAKE*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 165–183.
16. Schiller, A.; Teufel, S.; Stöckert, C.; Thielen, C. *Guidelines für das Tagging Deutscher Textcorpora mit STTS*; Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart: Stuttgart, Germany, 1999.
17. Mandelbrot, B. An Informational Theory of the Statistical Structure of Language. In *Communication Theory*; Jackson, W., Ed.; Butterworths Scientific Publications: London, UK, 1953; pp. 468–502.
18. Zipf, G.K. *The Psycho-Biology of Language*; Houghton, Mifflin: Oxford, UK, 1935.
19. Evert, S.; Baroni, M. zipfR: Word Frequency Distributions in R. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions, Prague, Czech Republic, 25–27 June 2007; pp. 29–32.
20. Baayen, R.H. The Effects of Lexical Specialization on the Growth Curve of the Vocabulary. *Comput. Linguist.* **1996**, *22*, 455–480.
21. Stadler, H. *Die Erstellung der Basislemmaliste der Neuhochdeutschen Standardsprache aus Mehrfach Linguistisch Annotierten Korpora*; Blühdorn, H., Elstermann, M., Klosa, A., Eds.; Institut für Deutsche Sprache: Mannheim, Germany, 2014.
22. Brysbaert, M.; Stevens, M.; Mandera, P.; Keuleers, E. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Front. Psychol.* **2016**, *7*, 1116. [CrossRef] [PubMed]
23. Herdan, G. *Quantitative Linguistics*; Butterworths: London, UK, 1964.
24. Heaps, H.S. *Information Retrieval, Computational and Theoretical Aspects*; Library and Information Science; Academic Press: New York, NY, USA, 1978; ISBN 978-0-12-335750-2.
25. Miller, D.; Biber, D. Evaluating Reliability in Quantitative Vocabulary Studies: The Influence of Corpus Design and Composition. *Int. J. Corpus Linguist.* **2015**, *20*, 30–53. [CrossRef]
26. Baayen, R.H. Productivity in Language Production. *Lang. Cogn. Process.* **1994**, *9*, 447–469. [CrossRef]