In this working paper, we compare a German lemma list based on DEReKO and a basic vocabulary list for German with entries from the German Wiktionary in order to get a more detailed idea of the database we are dealing with. The entries in the German Wiktionary were used for the analyses in our article “Observing online dictionary users: studies using Wiktionary log files” by Carolin Müller-Spitzer, Sascha Wolfer and Alexander Koplenig, to appear in the International Journal of Lexicography (IJL). A Wiktionary entry is included in the present comparison if (1) it is marked as German in the German language edition of Wiktionary (with “(Deutsch)” directly after the headword), (2) we found frequency information for the specific entry and (3) it received more than one visit in one million visits in 2013 (based on the log files provided by the Wikimedia Foundation). This is the large-scale, maximalist point of view for comparison. On the other hand, we also adopted a small-scale “minimalist” perspective comparing the selected Wiktionary entries with a lemma list based on a language learner’s point of view.

Some parts of this working paper overlap with the IJL paper. At several points, however, we go into greater detail.

1.1 Maximalist perspective
The basis of comparison for the maximalist perspective is a lemma list based on DEReKO. This list is based on the same German corpus we also extracted the frequency measures for the Wiktionary headwords from and thus provides an adequate basis of comparison. The list contains frequency information for 326,949 German lemmas. There are 70,891 articles in our selected entries (see figure 1 in section 2.3 of our article). By definition, the lemma list does not contain any inflected forms, but the selected entries do. So, it is not surprising, that there are 27,574 headwords (38.9%) in the selected entries that are not in the lemma list. From these 27,574 “missing” entries, almost half (12,046 entries, 43.7%) are inflected forms. The other very large group (11,105 entries, 40.3%) consists of nouns. These nouns mainly comprise many geographic and proper names and terminological terms (e.g., Tetrahydrofolsäure, engl. "Tetrahydrofolic acid"; Stangenstromabnehmer, engl. "arm current collector") that are not covered in the lemma list. We did not extract the exact numbers of geographic names and terminological words within the nouns not found in the lemma list. Of course, the list also contains words that do not fall into one of these two groups. Another 2,003 (7.3%) entries are abbreviations, 658 (2.4%) adjectives, 348 (1.3%) affixes and 277 (1.0%) verbs. The remaining 1,137 entries are comprised of adverbs, participles, extended infinitives, interjections and proper names. So obviously, the Wiktionary articles we selected are not a real subset of the lemma list compiled from DEReKO.

As a next step, we can investigate how the DEReKO lemma list and the selected German entries relate in terms of frequency. Figure 1 gives an impression of the frequency distributions.
We are plotting probability of occurrence (and not raw numbers) on the y axis to see how probable it is for a certain frequency to occur, given the distribution of frequencies it stems from. Figure 1 suggests that selected entries from Wiktionary occur rather frequently in comparison to all lemmas of the German language as represented in the DeReKo corpus. The near to normal distribution of frequencies for the selected entries could be the result of two distinct principles. (1) Entries for low frequency headwords were excluded because they were not visited often enough to reach the threshold of one visit per one million visits. (2) The fact that highly frequent headwords are not so probable as to be described in Wiktionary is simply due to the fact that there are so few high frequency lemmas in every natural language. From the maximalist perspective, we can conclude that the selected Wiktionary articles represent a mainly mid- to highly frequent proportion of the German language (as represented by the DeReKo corpus). Furthermore, Wiktionary also contains entries that are not captured by the lemma list. These are mainly entries for inflected forms, geographic and proper names and terminological nouns.

1.2 Minimalist perspective

The word list based on levels C1 through B1 from the German CEFR (please see our IJL article for further information about the CEFR) can be considered the minimalist perspective for comparison. It contains all words, which are thought to be important for "Threshold Independent Users" (level B1, cf. Council of Europe, 2001, p. 23) of German. This is the third of six competence levels that a learner of German can achieve. It is also the highest level the Goethe Institute\(^1\) provides an explicit word list for. Henceforward we will call this list “CEFR words”. The list contains 3,405 words. This is only 2.1 % of all German words in Wiktionary and 4.8 % of all German words we have frequency information for and that are visited more than once in one million visits. Considering that the CEFR words make up

---

\(^1\) The Goethe Institute is the institution, which is responsible for CEFR-related activities in Germany.
the basic vocabulary a learner of German should know, we expected that all words would be
described in the German Wiktionary. As it turns out, this is not the case. 161 words (4.7 % of all CEFR
words) are not contained in the German entries (161 words, or 4.8 %, for entries with frequency
information and 177 words, 5.2 %, for the selected entries). Even when the 20 feminine occupational
listener”) that are described as the masculine counterparts2 (“Reporter”, “Zuhörer”) are excluded,
there are still some CEFR words that cannot be found in Wiktionary. Five of them are numerals like
“neunzehnhundertneunundneunzig” (the graphemic representation of 1999). Some Anglicisms (e.g.,
“Mailbox”) in the CEFR words are tagged as English in the German Wiktionary and were hence
excluded from our Wiktionary database. However, we are still left with CEFR words where there does
not seem to be an apparent reason why they are not described in the German Wiktionary. A few
examples are “allerseits” (engl. “on all sides”), “ärgerlich” (“annoying”), “vegetarisch” (“vegetarian”,
“further education”), “Rundfahrt” (“tour” / “round trip”), or “Zahncreme” (“tooth paste”). To be
precise, 19 of the missing words (but none of the examples) were described until the end of August
2013 (e.g., “Realschule” on 03/03/2014 or “Durchsage” on 06/04/2014) but were not described early
enough to be included in the database of the IJL paper. Also, for some cases, very close synonyms
like “Zahnpasta” for “Zahncreme” or closely connected forms from other parts-of-speech
(“Vegetarier” and “Vegetarierin” as personal descriptions for “vegetarisch”) are described in the
German Wiktionary. The list of missing CEFR words would provide a good mean to explore potential
inconsistencies in a collaborative, crowd-sourced dictionary like Wiktionary.

2 Not every feminine occupational title or person description is described in the respective entry for the
masculine counterpart. The “policy” of the crowd-sourced Wiktionary is not consistent here.