Using Google Books Unigrams to Improve the Update of Large Monolingual Reference Dictionaries

Alexander Geyken & Lothar Lemnitzer

Keywords: *practical lexicography*, *computational linguistics*, *corpus statistics*, *lemma list*.

Abstract

This paper describes ongoing work to extend a traditional dictionary using a large opportunistic corpus in combination with a unigram list from the Google Books project. This approach was applied to German with the following resources: the *Wörterbuch der Deutschen Gegenwartssprache* (WDG, 1961-1977), the German unigram-list of Google Books and the DWDS-E corpus. Both corpus resources were normalized. The subsequent analysis shows that the normalized unigram list has clear complementary information to offer with respect to DWDS-E and that a comparatively small amount of manual work is sufficient to detect a fairly large number of new and relevant dictionary entry candidates.

1. Introduction

The "Digitales Wörterbuch der Deutschen Sprache" (DWDS, cf. Klein and Geyken 2010) is a project of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). Its goal is to implement a digital lexical system that combines dictionaries, text corpora and language statistics (www.dwds.de). At the core of this system is the DWDS dictionary, a monolingual general language dictionary for German that comprises rich lexicographical information for approx. 90,000 headwords and minimal (semantic) information for another 30,000 compounds. The DWDS dictionary is based on the digitized version of the six-volume „Wörterbuch der deutschen Gegenwartssprache" ([WDG]). The WDG was published between 1961 and 1977 in a traditional way from A-Z, hence the vocabulary after 1975 (depending on the letter even earlier) is not covered by the dictionary. The project plan of the DWDS project foresees to update the lexical substance of the DWDS dictionary in the years between 2013 and 2018. The revision process will include the addition of approximately 25,000 new lexical units as full entries with rich lexicographical information. Furthermore, another 20,000 lexical entries will be provided with minimal information, basically grammatical information and corpus citations. The first step of this revision process is the selection of an appropriate lemma list of 45,000 lexical units. In the remainder of this paper we will briefly describe the two corpus resources used for the lemma selection process, namely the DWDS-E corpus, an opportunistic corpus of contemporary German, and the German unigram list of the Google Books project (section 2). In section 3 and 4 we will present the methods we employed to make both corpora comparable and thus suitable for a common analysis. We will then present the results of the extraction process (section 5) and we will draw some conclusions (section 6).

2. Resources

The resources used for the lemma extraction process are two corpora, more precisely, the annotated word lists extracted from DWDS-E and the Google Books unigram list, henceforth GB-unigram list. We consider the DWDS-E word type list to be our base list and the other lists, including the GB-unigram list, to be control lists. From the DWDS-E list we have gleaned entry candidates for which we can easily get a significant amount of corpus citations.

The major purpose of the control lists is to narrow down the search space of the DWDS-E list to a manageable size and to draw the lexicographer's attention to lexical units which are more widely used (as made evident by the frequencies in the control lists) than it is reflected in the DWDS-E corpus list.

DWDS-E is a large opportunistic corpus collection containing texts from 1900 up to 2010. It is continually updated and currently consists of approximately 2.8 billion tokens. For our study we have extracted all texts from 1980-2009 with a total of 2.5 billion tokens. The major part of the collection consists of influential daily and weekly national newspapers and magazines: Bild, Frankfurter Allgemeine Zeitung, Neue Zürcher Zeitung, Süddeutsche Zeitung, Spiegel, WELT, and ZEIT. Furthermore three Berlin-based newspapers were added: Berliner Zeitung (East Berlin), die tageszeitung, Tagesspiegel (both West Berlin). In addition to newspapers and magazines, DWDS-E contains three other text genres derived from the DWDS-Kernkorpus (Geyken 2007), i.e. fiction, functional texts and scientific prose. This subcorpus contains carefully selected works of important authors such as Martin Walser, Hans-Magnus Enzensberger, Franz Kafka, Martin Suter or Jürgen Habermas. The total amount of this subcorpus is 100 million tokens.

The second corpus used for our work is the German part of the corpus of the Google Books project (Michel et al. 2010). More precisely, our study draws on the type list of this corpus, the GB- unigram list. The published list of about 5 GBytes contains for each word (type) several kinds of frequencies: type frequency per year, normalized type frequency (per million) as well as document frequency. The published list is not a full account of all types in the list: "In order to ensure that ngrams could not be easily used to identify individual text sources, we did not report counts for any n-grams that appeared fewer than 40 times in the corpus. Of course, the most robust historical trends are associated with frequent n-grams, so our ability to discern these trends was not compromised by this approach." (Michel et al. 2010). For reasons of compatibility with DWDS-E corpus we derived only those word types and type frequencies from the list which are based on documents published 1980 or later. The resulting sum of tokens corresponds to approximately 37.5 billion tokens.

3. Normalization of the data sets

In order to obtain compatible data sets, both the DWDS-E corpus and the GB-unigram list were transformed into the same format.

All tokens in the DWDS-E were lemmatized using the TAGH morphology, a German morphological analyzer that takes into account composition and derivation (Geyken and Hanneforth 2006), and part-of-speech-tagged (Jurish 2003). In cases of ambiguous lemmas two heuristics were used: first, the most probable lemma was taken according to its POS information. Thus, a word like *heute* (engl. today) is (in most of the cases) correctly associated with the adverb and not with the past tense form of the verb *heuen* (to hay). Second, in ambiguous cases with the same tag, the inflected form is associated with the most frequent lemma. Additionally, tokens unknown to the TAGH morphology were added to the list of potential lemmas. As a result of the morphological analysis, we obtained approximately 9 million "lemmas" (out of a total of 16.5 million types) including hapax legomena, but only 6 million of them were recognized by the morphology. The rest are either Named-Entities, spelling or keyboarding errors, special tokens (such as numbers or combinations of numbers and characters), foreign words or errors of the tokenizing process. Each lemma in the DWDS-E list was provided with absolute frequency information per year (drawing on the publication date of the underlying document) and with a normalized relative frequency (number of occurrences per million, henceforth opm).

The GB-unigram list contains approximately 6 million types. After processing this list with the TAGH morphology, the GB-unigram list is reduced to 2.9 million lemmas. A different heuristic was used here in cases of ambiguous morphological analysis. We used the most probable lemmatization adopted on the basis of the lemmatization process of the DWDS-E corpus. Thus, erroneous lemmatization results such as the above-mentioned *heute* as *heuen* are avoided as well. It has to be noted here that these much smaller figures in the GB-unigram list are due to the fact that types with a frequency of less than 40 were not taken into consideration by the unigram-list (cf. section 2 above) whereas the DWDS-E lemma list gives a complete account of all types including hapax legomena.

In the resulting data structure for both the DWDS-E list and the GB-unigram list we represent a) the canonical form of each lexical unit (the lemma) b) the summed up total frequencies of the types which have been subsumed under this lemma for the time span between 1980 and 2009, c) the year d) the frequency in the normalized form of a relative frequency, expressed as occurrences per million.

4. Reducing the data size

We used the lemma list of the DWDS-E as a starting point. From this list we removed the 120 000 headwords (and spelling variants) from the DWDS dictionary (cf. section 1). For the resulting DWDS-E list, we took only those lemmas into consideration that had a normalized frequency of more than 0.3 opm. This was the case for 46,000 lemmas. This list was manually analyzed. Named entities, foreign language material, spelling errors and special tokens were discarded. The remaining list of 13,000 candidates was subdivided into a candidate list for full entries as well as a candidate list for minimal entries.[1]

Similarly, we removed from the GB-unigram list all lemmas that were in the DWDS dictionary as well as all lemmas in the DWDS-E list that were above the threshold of 0.3 opm. Additionally, we took only those lemmata into consideration that were in the intersection of DWDS-E and the GB- unigram list. This assumption reduces the number of lemmas to be examined in the GB-unigram list. A manual inspection of the top 5,000 candidates (ordered by frequency) showed that this assumption was safe in the sense that words that do occur in the GB-unigram list but do not occur in DWDS-E at all are not suitable candidates for a list of lexicographically relevant lemmas. This resulted in a reduced list of approximately 1.4 million lemmas that are in the intersection of GB-unigram and DWDS-E. We subdivided this list into four different data sets. Set 1 consists of all entries in the GB- unigram list with a normalized frequency of above 0.3 opm. Additionally we introduced a threshold on the DWDS-E frequencies of 0.001 opm (that corresponds to an absolute frequency of 3). Set 2 consists of entries with a normalized frequency below 0.3 opm that are underrepresented in DWDS-E in comparison with the GB-unigram list. We modeled this by considering all entries in the GB-unigram list where the ratio of opm (DWDS-E) to opm(GB-unigram) is less than 0.5 Set 3 consists of all entries that are overrepresented in DWDS-E with respect to the GB-unigram list (ratio larger than 1.5), and finally, set 4 contains all lemmas with a ratio between 0.5 and 1.5). Since the corresponding data sets amount to a total of 173,575 lemmas, we introduced thresholds for GB-unigram list such that the resulting number of lemmas is around 30,000, thus being manageable for manual investigation.

After discarding Named-entities, misspellings, special forms and foreign language material in these data sets we arrive at additional 14,869 entry candidates (cf. table 1).

Table 1.

Name	Thresholds for	lemmas with (without) threshold	Entry candidates

	GB;DWDS		
Set 1	0.3; 0.001	10443 (23,960)	8,407
Set 2	0.2; 0.01	8,111 (84,785)	2,003
Set 3	0.05;0.05	5,680 (12,827)	2,561
Set 4	0.15;0.15	3,517 (52,003)	1,898
Total		29,996 (173,575)	14,869

5. Results

In this section we discuss the relevance of the pre-selected list of 14,869 entry candidates from the GB- unigram list for the selection of new lemma candidates for the DWDS dictionary. Table 1 clearly shows that GB-unigrams has complementary information to offer with respect to DWDS-E. Set 1 above (s. table 1) illustrates this fact. Indeed, 8,407 new candidates were obtained by simply going through GB-unigrams with a normalized frequency of more than 0.3 opm. The comparison with DWDS-E shows that many of those entry candidates are comparatively infrequent in DWDS-E. Indeed, 4,905 out of the 8,407 entry candidates in data set1 have a normalized frequency of less than 0.1 opm in DWDS-E, and 661 entry candidates have a normalized frequency of less than 0.01 opm. A systematic investigation of lemmas in DWDS-E with an opm of less than 0.01 would imply to investigate a list of almost 1 million entries (frequency rank of lemmas with opm > 0.01).

Likewise set 2 and set 3 show the advantage of focusing on lemmas that are either over-or underrepresented in one of both corpora. An additional 4,564 entry candidates could be found by investigating another 13,791 entries. Finally set 4 shows that the threshold of 0.3, initially chosen for DWDS-E, is somewhat arbitrary since the manual analysis of lemmas with a threshold above 0.15 still yields relevant entry candidates.

As stated above, not all of these extracted lemmas are candidates for full entries. The manual analysis of these 14,869 candidates provided by the four data sets described above results in 11,961 candidates for full entries, another 2,905 entries that are suitable for minimal entries. The overwhelming majority of entry candidates are nouns (more than 12,000), followed by adjectives (1,200). Only 200 new verbs (many of them are anglicisms such as traden (to trade) or faken (to fake)) and less than 50 adverbs could be found. The high proportion of nouns is due to the potential of compounding in German, where even complex compounds are widely used. Commonly used verb or adjective compounds are much rarer.

Finally, the analysis of GB-unigrams shows that the candidates are distributed over a variety of domains such as mathematics/statistics, technology, medicine, social sciences and linguistics as well as economy and law. In general and due to the base of the GB unigram list, more special language terms can be found in the higher frequency ranks.

6. Summary and future work

We have presented a method of preparing a lemma list which can be useful in general for projects planning to update a legacy lexical resource. The required resources are a lemma list of the legacy resource as well as the unigram lists of the Google Books texts that serve as a control resource to a project specific large reference corpus or any other large opportunistically collected corpus. Unigram lists of the Google Books project are available for

many languages. The lemma list of the control corpus helps to control the amount of work which has to be spent in reducing the amount of data in the reference corpus word list (DWDS-E, in our case).

There are still some issues which need further investigation: a) the selection of the respective thresholds (0.3 opm) is arbitrary and lacks independent justification. We have tried to overcome this problem by the addition of supplementary entry candidates from lemmata that are either under-or overrepresented in both corpora. Much depends on the number of lemma candidates which one wants to obtain at the end of the process; b) we realized that even with a combination of reference corpus and Google Books data some aspects of daily life are still not well represented lexically. These are the aspects which are typically not to be found in newspapers, scientific texts etc, but could be covered with a larger proportion of literary texts or other non-fictional materials. It might be worth to consider harvesting texts of such genres from the web with the help of appropriate seed words (cf. Baroni and Bernardini 2004, Sharoff 2006).

Note

[1] It is worth mentioning that another approx. 13,000 lemmas have been selected from other lexical resources (Wiktionary, GermaNet, the WDG list of minimal compound entries), but this is less relevant for the topic of this paper.

References

A. Dictionaries

Klappenbach, R. und W. Steinitz 1961-1977. *Wörterbuch der deutschen Gegenwartssprache* 1-6. Berlin: Akademie-Verlag.

B. Other literature

Baroni, M. and S. Bernardini 2004. 'BootCaT: Bootstrapping corpora and terms from the web.' In M. T. Lino et al. (eds.), *Proceedings of the Fourth Language Resources and Evaluation Conference*. Lisbon/Paris: ELRA, 1313–1316.

Geyken, A. 2007. 'The DWDS Corpus: a Reference Corpus for the German Language of the 20th Century.' In C. Fellbaum (ed.), *Collocations and Idioms: Linguistic, lexicographic, and computational aspects*. London: Continuum Press, 23–41.

Geyken, A. and T. Hanneforth 2006. 'TAGH: a Complete Morphology for German Based on Weighted Finite State Automata.' In A. Yli-Jyrä et al. (eds.), *Finite-state methods and natural language processing, 5th international workshop, FSMNLP 2005, Helsinki, Finland, Revised Papers*. Lecture Notes in Artificial Intelligence 4002. Berlin/Heidelberg: Springer, 55–66.

Jurish, B. 2003. *A Hybrid Approach to Part-of-Speech Tagging*. Final report. Project 'Kollokationen im Wörterbuch'. Berlin-Brandenburgische Akademie der Wissenschaften, 31 Oct. 2011. http://www.ling.uni-potsdam.de/~moocow/pubs/dwdst-report.pdf.

Klein, W. and A. Geyken 2010. 'Das Digitale Wörterbuch der Deutschen Sprache (DWDS).' *Lexicographica: International Annual for Lexicography* 26: 79–96.

Sharoff, S. 2006. 'Open-source corpora: using the net to fish for linguistic data.' *International Journal of Corpus Linguistics* 11.4: 435–462.