Managing indirect user contributions for the DWDS project

Alexander Geyken[1], Frank Wiegand[1]

[1] Berlin-Brandenburg Academy of Sciences, Jägerstr. 22/23, D-10117 Berlin, Germany
E-mail: {geyken,wiegand}@bbaw.de

Abstract

Keywords: on line dictionaries, quality control, indirect user contributions

1. Introduction and problem statement

The Digital Dictionary of the German Language (DWDS, Digitales Wörterbuch der deutschen Sprache) is a long term project of the Berlin-Brandenburg Academy of Sciences and the Humanities (BBAW, Berlin-Brandenburgische Akademie der Wissenschaften). The goal of the DWDS project is to compile a large aggregated word information system based on large legacy dictionaries, large corpora, word statistics and automated methods to speed up the compilation process (Geyken 2014). The platform integrates an automatic collocation extractor and a good example finder (Didakowski and Geyken 2012, Didakowski et al 2012). Finally, the DWDS draws on large corpora with a size of 8 billion running words (as of January 2016) that cover the period between 1600 to the present. The DWDS-website with all the data and functions described in the article can be found under its beta-version http://zwei.dwds.de/. The dictionary component of the DWDS draws mainly on two legacy dictionaries: the Dictionary of the German Contemporary Language (WDG, Wörterbuch der deutschen Gegenwartssprache, WDG [1961-1977]), a synchronic dictionary of 4,800 pages in 6 volumens with 120,000 keywords, compiled between 1961 and 1977 at the GDR Academy of Sciences, and second, a subset of about 70,000 articles of the Duden GWDS (Duden-GWDS 1999), the largest printed dictionary of contemporary German. Articles from Duden were chosen for cases where the WDG articles are missing, incomplete or outdated. In addition to these entries in WDG and Duden, another 45,000 entries were selected by corpus-based methods (Geyken & Lemnitzer 2012) and integrated as entries with minimal morphological information into the DWDS on line dictionary. Since 2013, a team of six (FTE) lexicographers edits new articles and revises the existing entries in the lexicographic substances. The ultimate goal of the DWDS project is to obtain a coherent and up-to-date lexicographic description of the present German language at the end of the project in 2025. In the mean time the project's goal is twofold: it should provide reliable information for all types of lexicographic information (spelling, morphology, sense, collocation, phraseology) unless the entry is explicitely marked as being outdated or incomplete on some 'zone' of lexicographic information. This form of

quality control requires a check of all dictionary entries for their correctness and up-to-dateness on all the above-mentioned lexicographic levels. This process is feasible only by a distributed effort, and it goes without saying that this revision process is too complex to be done without digital assistance. In the next section we present the management system that is currently been used to store and to process “tickets” submitted either by the team or external users of the DWDS on line platform. This issue management can be qualified as indirect user contribution (Abel & Mayer, 2013).

2. The DWDS issue management

In order to maintain the consistency and the up-to-dateness of the lexicographic entries, we use MantisBT (https://www.mantisbt.org/, henceforth we use Mantis as a shortform), an open source, web-based issue tracker that is easy to install and requires only little time for the users to familiarize with the system. There are more powerful OpenSource issue tracking systems such as Redmine or Trac. However, for the needs of our projects, Mantis proved to be satisfactory since lexicographic issues do not yield many dependencies and the management does not require elaborated reporting functions - just to name two differences between Mantis and Redmine. Even though MantisBT is typically used for bug tracking in small to medium sized software projects, it can easily be customized to other project types. In order to customize Mantis to the DWDS project we have redefined some of the fields of Mantis. However, the majority of the fields with closed values could be reused without modification. In the case of the DWDS an issue consists of this following fields: **severity** (minor, major). An issue can have several **status** values: new, feedback, acknowledged, assigned, won’t fix, no change required, resolved or closed. The field **steps to reproduce** points to the persistent url of the dictionary entry: http://beta.dwds.de/wb/<entry>. **Category** is a field with an open set of values. We have customized it for the DWDS project by using lexicographic and functional categories. The following values can be assigned for the field category: entry missing, meaning is missing/incomplete, image/pictogram missing, form-part is incomplete, web-css of entry is wrong, frequency time line is not plausible, collocation information is wrong, examples by the good example extractor is wrong. And word segmentation is wrong. Furthermore, we use the field **Tags** to provide the reported issue with additional workflow information such as ‘for this word, a basic entry is sufficient’, ‘provide definition only’, ‘word should become a full entry’. Those Tag values can be used as a flag to be displayed on the DWDS web-platform. Currently we don’t use hierarchical issues because the overwhelming number of issues point to one dictionary entry, even though possibly to several lexicographic zones. For example, an entry can have a wrong form part and a missing meaning as well.

Users submit issues not via Mantis but via a web-form on the DWDS website. The idea is that it is much easier for the user to mark up a problematic entry when being on the website than to open Mantis at the same time. The web-form has the advantage

that some of the required fields in Mantis are filled in with default values. For example the field reporter is always given the default value dwdsweb, categories can be selected via a drop-down menu, and the field assigned to is given the value dwds-issue-manager. The web-form contains also a mandatory comment field that encourages the user to justify why, for example, a meaning is missing, or why a new entry is lexicographically relevant. After having entered all mandatory information on the web-form it can be sent to Mantis via the Mantis API.

Roles in the DWDS issue management are important since only developers (lexicographers) have more access rights than external users. They can process the issue (i.e. change status values), the can modify the severity, i.e. change it from minor to major. And finally, they can use tags. ‘display: definition is missing’; display: entry is missing’

The following example gives a typical way how the issue management is currently used. The user looks up a word on the DWDS website and finds that an existing meaning of that word does not exist. The user opens the web-form, selects the category ‘missing meaning’ and presses enter to send it to Mantis. All the other field are given default values, e.g. severity, status, reporter, assigned to. In the next step the issue manager goes through all new issues and assigns each issue to a field expert (e.g. morphology is different from definition expert). The expert to whom the issue is assigned can then decide to ‘upgrade’ the issue (minor -> major) and to use a tag, e.g. display: definition is missing’. This ‘upgrade’ by the expert would create a message for this entry on the website: “this entry is not up-to-date: there is a sense description missing”. If the severity is marked major, the message text will be followed by “The DWDS team will update this entry in the coming weeks“. If the severity is unmarked, the message is: “The DWDS team will update this entry in the present project phase“.

3. Results (Work in progress)

The issue management system is used since August 2015. As of March, 21st more than 11,500 issues have been submitted by a group of 20 users, all employees of the BBAW but not staff of the DWDS project. According to the summary page of the MantisBT the top 3 issues are: missing entry (7502), missing/wrong meaning (2930), and wrong form parts: 438. Currently, 5300 issues were (partially) fixed, i.e. form errors were corrected and missing entries were provided with minimal information (i.e. form information). However, most of the missing entries were reported to be described as basic entries (60%) or full entries (40%). Therefore, the longer part of the work will consist in providing those minimal entries with good corpus examples, collocations and a sense description. Thus, the informational depth of the entries will be gradually accomplished.

To sum up. Distributed issue management has proven to be very useful for our

dictionary project. The current bottle-neck does not consist in the number of reporting users but rather in the work-load our developers (lexicographers) can spend for issue solving.

4. References

Abel, Andrea, Meyer, Christian, M. (2013). The dynamics outside the paper: user contributions to online dictionaries. In: Proceedings of elex 2013, Tallinn, p. 179-194.

Didakowski, J. and Geyken, A. (2012). From DWDS corpora to a German Word Profile – methodological problems and solutions. In Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. Arbeiten zur Linguistik 2/2012 (OPAL), Mannheim: Institut für Deutsche Sprache, pp. 43–52.

Didakowski, J., Geyken, A. and L. Lemnitzer (2012). Automatic example sentence extraction for a contemporary German dictionary. In: Proceedings of EURALEX, Oslo, p. 343-349.

Duden-GWDS [1999]. Das große Wörterbuch der deutschen Sprache. 10 volumes. Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

Geyken, Alexander (2014). Methoden bei der Wörterbuchplanung in Zeiten der Internetlexikographie. In: Heid, Ulrich/Schierholz, Stefan/Schweickard, Wolfgang/Wiegand, Herbert Ernst/Gouws, Rufus H./Wolski, Werner (Hg.). Lexicographica. Berlin / Boston: de Gruyter, S. 77-112.

Geyken, Alexander, and Lemnitzer, Lothar. Using Google Books Unigrams to Improve the Update of Large Monolingual Reference Dictionaries. In: Proceedings EURALEX, Oslo, p. 362-366.

WDG [1961-1977]: Wörterbuch der deutschen Gegenwartssprache in 6 volumes (4,800 pages). Akademie-Verlag : Berlin