On-the-fly Generation of Dictionary Articles for the DWDS Website

Alexander Geyken, Frank Wiegand, Kay-Michael Würzner

Berlin-Brandenburg Academy of Sciences and Humanities, Jägerstraße 22/23, 10117 Berlin, Germany
E-mail: {geyken|wiegand|wuerzner}@bbaw.de

Abstract

We present a method for generating on-the-fly dictionary articles for the DWDS website (https://www.dwds.de). The DWDS website contains electronic versions of large legacy dictionaries as well as very large corpora. On-the-fly articles are a fallback solution for user queries that cannot be matched with dictionary headwords or one of its inflected forms on the website. They depend on an automatic morphological analyser that segments complex words into parts that formally match existing dictionary headwords in a reliable way. On-the-fly articles are a useful mechanism for increasing the number of headwords with minimal manual effort. They are particularly useful for compounding languages like German. The generation method described in this article is fully integrated into the DWDS website.

Keywords: automatic creation of dictionary content; compound recognition; German morphology

1. Introduction

A major challenge for (monolingual) online dictionaries is to guarantee exhaustive vocabulary coverage, a goal that is time consuming, labour intensive and therefore generally considered as impossible to achieve. This is even more true for languages such as German, a language well known for its very large and theoretically even unlimited number of compounds. Therefore additional methods have to be developed to provide users with lexical information for as many words as possible with minimal manual intervention.

In this article we show how “out-of-headword-range” user queries, i.e. queries that cannot be directly matched to headwords in the dictionary, are dealt with in the Digital Dictionary of German language (DWDS), a comprehensive lexical information system of contemporary German. The problem of “out-of-headword” queries is a major practical problem for the DWDS system since there are numerous morphologically complex words (compounds and derived forms) in German that are not lexicographically described in neither of the largest monolingual dictionaries of New High German, including Duden (1999), Wahrig (Wahrig-Burfeind, 2011) and DWDS (Klein & Geyken, 2010; Geyken, 2015). These “hand-crafted” dictionaries have a size of between 150,000 and 200,000 headwords whereas the number of German words occurring in corpora is estimated as being well above five million (Klein, 2013). Even though many of those words may not require a full description from a lexicographer’s point of view, they are nevertheless targeted by regular user queries and therefore need to be handled by the lexical information system.

We propose a solution to this kind of user query by providing—wherever possible—dynamically generated dictionary articles on the DWDS platform with automatic methods. These articles generated “on the fly” are presented in the same way as dictionary articles compiled by lexicographers. Nevertheless, both automatic and hand-crafted articles are labeled as such. Thus, the dictionary user is provided with lexicographic information for many of those compounds that are not contained in the hand-crafted dictionaries.

The remainder of the article is organized as follows: in the next section the DWDS lexical information platform is presented. Section 3 briefly describes the quality management of DWDS platform that is used to identify missing entries as well as incomplete or false information of existing entries. Section 4 briefly introduces mechanisms of morphological productivity, shows how automatic morphological analysers deal with the problem of segmenting complex words and applies these methods on the problem addressed in this article, namely to relate “out-of-headword” compounds to headwords in the DWDS dictionary. Automatic morphological analysis is at the basis of the generation of “on-the-fly” dictionary articles. Its different components are presented in Section 5. Morphological analysis is just one mechanism to deal with “out-of-headword” queries. Section 6 shows how the automatic morphological analysis is combined with other fallback mechanisms dealing with queries that are commonly used to deal with “out-of-headword” user queries. The method presented here is fully integrated into the DWDS platform. In Section 7 some results together with an evaluation on the basis of DWDS user queries are presented. The article ends with a short conclusion (Section 8).

In this paper the following terminology is adopted. The term “headword” is used to denote the lemma string of a dictionary entry. The term “dictionary entry” refers to the lexicographic description of a headword that consists of a form and a sense description. The term “dictionary article” is used for aggregated information, including the dictionary entry as well as information from automatically extracted information from corpora or from external lexicographic resources.

2. The DWDS platform

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 process of updating and amending the existing lexical resources (Geyken, 2014). The platform integrates an automatic collocation extractor and a good example finder (Didakowski & Geyken, 2014). Furthermore, the DWDS draws on large corpora with a size of 12.5 billion running words (as of May 2017) that cover the period between 1600 and now. The DWDS website with all the data and functions described in the article can be consulted under https://www.dwds.de/. The dictionary component of the DWDS draws mainly on two legacy dictionaries: the Dictionary of the German Contemporary Language (Klappenbach & Steinitz, 1964–1977), a synchronic dictionary of 4,800 pages in six volumes 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 (Scholze-Stubenrecht, 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 dictionary plattform. Since 2013, a team of six lexicographers edits new articles and revises the existing entries. The 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.

3. Quality management within the DWDS platform

The revision process of the legacy dictionaries requires a check of all entries for their correctness and up-to-dateness on all 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. To this end we use *MantisBT*,[1] an open source, web-based issue tracker that is easy to install and requires only little time for users to familiarize with the system. Users of the issue management system can report either missing entries or inconsistencies on any type of lexicographic information, including spelling, morphology, sense, collocation, phraseology. 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 platform. As of 22nd May 2017 more than 18,500 issues have been submitted by a group of 30 people, the majority of them are employees of the BBAW. According to the summary page of the *MantisBT* the top three issues are: *missing entry* (11,500), *missing/wrong meaning* (4,850), and *grammar or word formation errors* (870).

It is important to note here that only those words are submitted to the issue management system as "missing entries" where major additional and manual lexicographic description is deemed necessary. However, as stated in the introduction, due to the very large number and the high productivity of (new) German compounds it is not possible to manually compile full lexicographic entries for all compounds. Therefore automatic methods are used to generate basic dictionary entries (cf. Section 4) that form one component of the aggregated dictionary article that is used for the DWDS platform (cf. Section 5).

4. Automatic morphological segmentation as a building block for dynamic dictionary articles

The idea of this section is to use automatic morphological analyses in order to split complex words which are not in the dictionary into less complex components for which dictionary entries exist. More precisely, we are looking for the least complex decomposition that corresponds best to the word formation of the complex word. In the remainder of this section, we briefly mention linguistic aspects of German word formation (4.1), we summarize the relevant aspects of automatic morphological analysers for German (4.2), and we present a method to map complex words to the appropriate headwords in the DWDS dictionary.

4.1 German word formation

The term *word formation* subsumes operations to create novel (complex) words[2] based on existing linguistic units (i.e. words and affixes). Together with *lexical borrowings* and *semantic shifts* it is one of the means to cover the need for "new" words. Word formation operations are usually distinguished in terms of their operands: The combination of two words is called *compounding* while the combination of a word and an affix is called *derivation*.[3]

The German language is not only known for its rich productivity of compounding. It has also some very productive affixes that can be used to form new compounds. Example (1)

[1] MantisBT: https://www.mantisbt.org/

[2] Note that this is the principal difference to *inflection* which does not result in novel words.

[3] *Conversion*, i.e., the covert changing a word's category may be treated as a special case of derivation involving an invisible affix.

below illustrates this combinatorial process. The noun *Vollstreckbarkeit* (engl. ‘enforceability’) is derived from the verb *strecken* by subsequently adding the verbal prefix *voll-*, the suffix *-bar*, and the suffix *-keit*:

$$\Big(((\mathit{voll}_{\mathtt{P}}\ (\mathit{streck}_{\mathtt{V}}))_{\mathtt{V}}\ \mathit{bar}_{\mathtt{S}})_{\mathtt{A}}\ \mathit{keit}_{\mathtt{S}}\Big)_{\mathtt{NN}} \tag{1}$$

In addition to such iterated derivation operations, German, in contrast to e.g. English or French, knows “non-spaced” compounding: compounds are realized as a continuous sequence of characters optionally agglutinated with non-empty linking elements such as *-s* or *-er*; the subparts may very well be complex words again:

$$\Big(((\mathit{voll}_{\mathtt{P}}\ \mathit{streck}_{\mathtt{V}})_{\mathtt{V}}\ \mathit{bar}_{\mathtt{S}})_{\mathtt{A}}\ \mathit{keit}_{\mathtt{S}}\Big)_{\mathtt{NN}}\ s_{\mathtt{Link}}\ ((\mathit{er}_{\mathtt{P}}\ \mathit{kl\ddot{a}r}_{\mathtt{V}})_{\mathtt{V}}\ \mathit{ung}_{\mathtt{S}})_{\mathtt{NN}} \tag{2}$$

The sequence of operations leading to a complex word is called its *derivational history.* A fundamental problem of (word-based) morphological analysis is ambiguity; often, multiple analyses for a single word are available. Lemnitzer & Würzner (2015) distinguish four types of ambiguities:

segmentation ambiguities A complex word may be split into several morpheme sequences: `Musik<NN>Erleben<+NN>` (‘musical experience’) vs. `Musiker<NN>Leben<+NN>` (‘a musician’s life’).

categorial ambiguities A word belongs to more than one category: *weiß* (adj. ‘white’ vs. verb ‘[I] know’).

lexical ambiguities Multiple lexemes are realized with the same word: *Bank* as financial institution and as seating-accommodation.

morpho-syntactic ambiguities Multiple forms of the same morphological paradigm have an identical realization: *übe* (‘practice’) as first person singular indicative active as well as imperative singular.

Complex morphological processes must therefore be employed to generate one or more plausible segmentations of a complex word, and eventually, to link these segments to existing dictionary entries. This is discussed in more detail in the next section.

4.2 Automatic morphological analysis

The overall goal of the morphological analysis of a (possibly) complex word form is its decomposition into smaller segments consisting of a combination of affixes and stems together with symbols marking segment separators. It can thus be understood as the identification of operations and operands which led to formation of that complex word.

Finite-state morphology is a technique to implement the analysis of productive word formation processes using a set of *rational rules* (cf. Lawson, 2003) over a finite alphabet. It is a very popular model in computational morphology and has been applied to a large number of languages (cf. Beesley & Karttunen, 2003). Rational rules can be efficiently represented and applied using (weighted) finite-state transducers. There are several finite-state morphologies available for German, most notably GERTWOL (Haapalainen & Majorin, 1995), TAGH (Geyken & Hanneforth, 2006) and SMOR (Schmid, 2004). While GERTWOL is not freely available for large-scale testing and application, TAGH and SMOR have a comparable coverage of German word formation. SMOR allows for a segmentation into atomic morphemes whereas TAGH regroups morphemes to larger units. Since we need a 1:1 mapping of automatically analysed morphemes onto headwords of the DWDS dictionary, SMOR is more flexible and therefore better suited for the task at hand. Figure 1, as an example, shows the output of SMOR for the German compound *Kürzungen* (‘shortages’, ‘cuts’):

```
> Kürzungen
Kür<NN>:<>Z:zunge<+NN>:<><Fem>:<><>:n<Nom>:<><Pl>:<>
Kür<NN>:<>Z:zunge<+NN>:<><Fem>:<><>:n<Gen>:<><Pl>:<>
Kür<NN>:<>Z:zunge<+NN>:<><Fem>:<><>:n<Dat>:<><Pl>:<>
Kür<NN>:<>Z:zunge<+NN>:<><Fem>:<><>:n<Acc>:<><Pl>:<>
k:Kürze:<>n:<><V>:<>ung<SUFF>:<><+NN>:<><Fem>:<><>:e<>:n<Nom>:<><Pl>:<>
k:Kürze:<>n:<><V>:<>ung<SUFF>:<><+NN>:<><Fem>:<><>:e<>:n<Gen>:<><Pl>:<>
k:Kürze:<>n:<><V>:<>ung<SUFF>:<><+NN>:<><Fem>:<><>:e<>:n<Dat>:<><Pl>:<>
k:Kürze:<>n:<><V>:<>ung<SUFF>:<><+NN>:<><Fem>:<><>:e<>:n<Acc>:<><Pl>:<>
```

Figure 1: SMOR analyses for *Kürzungen*

Notation	**Meaning**
<NN>	morpheme category: normal noun
<V>	morpheme category: verb
<A>	morpheme category: adjective
<+*x*>	denotes the category of the word (part of speech)
<SUFF>	suffix
<Fem>	feminine gender
<Nom>, <Gen>, <Dat>, <Acc>	grammatical case
<Pl>	plural
<>	empty string (epsilon)
x : *y*	mapping from lemma to word-form level

Table 1: SMOR syntax[4]

A number of strategies have been proposed to deal with the aforementioned ambiguity phenomena, usually employing the context of a word’s occurrence. In our use-case, i.e., the analysis of dictionary queries, context is not available. We therefore make use of a simple heuristic which goes back to Volk (1999) in order to reduce the number of analyses. Each word formation operation is assigned a specific cost (e.g., 2.5 for suffixation and 5 for compounding). From the two possible analyses for *Kürzungen* (i.e., `Kür<NN>Zunge<+NN>n`

[4] Note that the analysis contains the lemma as well as the word-form level. Differences between the two are denoted by the colon symbol. Symbols only present on the lemma level are mapped onto the empty string. For details, the reader is referred to Schmid (2004).

vs. `kürzen<V>ung<SUFF><+NN>en`), the latter is ‘cheaper’ and thus considered to be more likely. In addition, we increase the total cost of a segmentation by the edit-distance between the lemmas associated with the segmentation and the input word. Favoring orthographically closer analyses helps for example resolving ambiguities introduced by the optional dative suffix *-e* in cases like *Hängebuche* (‘hanging book’ or ‘weeping beech’) with analyses `hängen<V>Buch<+NN>e` and `hängen<V>Buche<+NN>`.

4.3 Mapping morphological analysis to dictionary entries

After performing the morphological analysis of the queried word and the ranking of the resulting analyses according to the weighting sketched above, only the best (i.e., cheapest) analyses are considered as candidates for linkage. Instead of simply linking to the entries of the identified (atomic) morphemes, we try to be as specific as possible by linking to the most complex available dictionary entries. This is done by constructing the set of all possible derivational histories leading from the morphemes to the complex word form for each remaining analysis. Derivational histories can be depicted as trees:

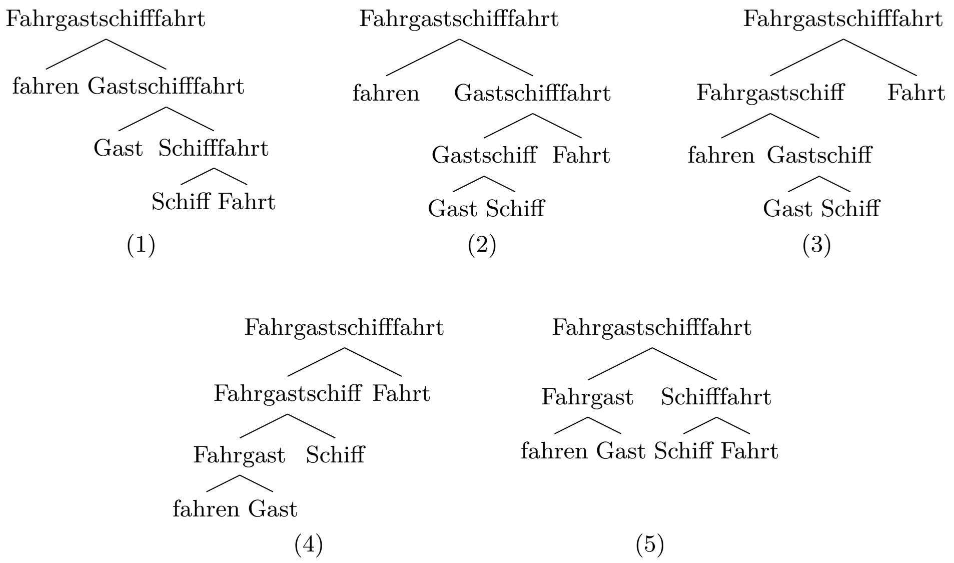
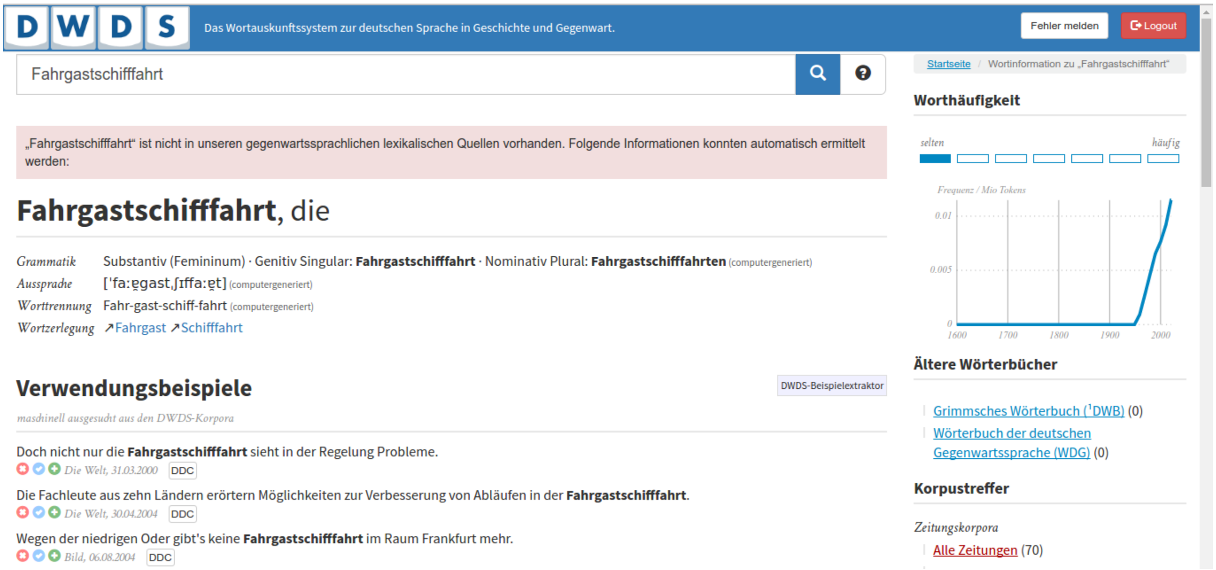


Figure 2: Derivational histories for the word *Fahrgastschifffahrt* depicted as trees

The components of each level in each tree are looked up in the dictionary. The least complex segmentation is used for the mapping, i.e., the highest in a tree where each segment matches a dictionary headword. For *Fahrgastschifffahrt* (‘passenger shipping’) the selected segmentation is on level one of tree number five since both *Fahrgast* (‘passenger’) and *Schifffahrt* (‘shipping’) are listed in the DWDS dictionary.

5. Components of dynamically generated dictionary articles

Dynamically generated dictionary articles consist of all components of the DWDS system which can be generated automatically for a given word from various resources, including information about its form (spelling, grammar, word formation), word frequency, thesaurus information (synonyms, antonyms, hyponyms, and hyperonyms) re-



„Fahrgastschifffahrt“ ist nicht in unseren gegenwartssprachlichen lexikalischen Quellen vorhanden. Folgende Informationen konnten automatisch ermittelt werden:

Fahrgastschifffahrt, die

Grammatik Substantiv (Femininum) · Genitiv Singular: **Fahrgastschifffahrt** · Nominativ Plural: **Fahrgastschifffahrten** (computergeneriert)
Aussprache ['faːɐ̯gastˌʃɪffaːɐ̯t] (computergeneriert)
Worttrennung Fahr-gast-schiff-fahrt (computergeneriert)
Wortzerlegung ↗Fahrgast ↗Schifffahrt

Verwendungsbeispiele

maschinell ausgesucht aus den DWDS-Korpora

Doch nicht nur die **Fahrgastschifffahrt** sieht in der Regelung Probleme.
Die Welt, 31.03.2000 DDC

Die Fachleute aus zehn Ländern erörtern Möglichkeiten zur Verbesserung von Abläufen in der **Fahrgastschifffahrt**.
Die Welt, 30.04.2004 DDC

Wegen der niedrigen Oder gibt's keine **Fahrgastschifffahrt** im Raum Frankfurt mehr.
Bild, 06.08.2004 DDC

Worthäufigkeit

Ältere Wörterbücher

Grimmsches Wörterbuch (¹DWB) (0)
Wörterbuch der deutschen Gegenwartssprache (WDG) (0)

Korpustreffer

Zeitungskorpora
Alle Zeitungen (70)

Figure 3: Generated article for *Fahrgastschifffahrt* on the DWDS website

trieved from the OpenThesaurus[5] dataset as well as automatically selected usage examples from DWDS corpora by using the DWDS-Beispielextraktor[6] and collocations by the DWDS-Wortprofil.[7] The extraction of usage examples and collocations are described in more detail elsewhere (cf. Section 2). Therefore this section focuses on the description of frequency and form information.

Using the DDC search engine[8] indices, we can provide information about the word frequencies within the DWDS corpora. Using a level meter, a value between one and seven on a logarithmic scale shows how often the requested lemma occurs within the corpus texts.[9] Since all corpus documents are marked with reliable metadata (including its date of publication), a graph of the distribution of the word frequencies from 1600 until today can be computed. This graph is shown on the website below the frequency meter.[10] The graph image is linked to an extended version of our corpus search plotting tool. In addition, hyperlinks to occurrences of the keyword in the public searchable corpora are provided as well.

The form part of the dynamically generated article consists of several parts: Information about the word's pronunciation[11] and hyphenation is provided by the gramophone webservice.[12] The grammatical information (i.e. inflection and the Part-of-Speech tag, more precisely the mapping of an STTS tag to the principal word classes of the dictionary such as nouns, verb, adjective and adverb) is obtained via the SMOR analysis. If applicable, morphological segmentation is displayed and all components are linked to their respective dictionary articles.

[5] OpenThesaurus: https://www.openthesaurus.de/
[6] DWDS-Beispielextraktor: https://www.dwds.de/d/beispielextraktor
[7] DWDS-Wortprofil: https://www.dwds.de/d/ressources#wortprofil
[8] DDC (DWDS/Dialing Concordance), the search engine used in the DWDS project: https://www.dwds.de/d/suche
[9] https://www.dwds.de/d/api#frequency
[10] DWDS-Wortverlaufskurve: https://www.dwds.de/d/plot
[11] Only for users with a DWDS user account.
[12] http://kaskade.dwds.de/~kmw/gramophone.py (Würzner & Jurish, 2015).

6. Combination with other fallback mechanisms

In Section 4 it was shown how user queries corresponding to "out-of-headword" compounds can be correctly mapped to headwords in the DWDS dictionary. However, this is only one way to handle query strings that do not directly match dictionary entries. In the current implementation of the DWDS platform the following fallback mechanisms take effect:

1. If the query string can be morphologically analysed via SMOR then
   (a) if the query string corresponds to an inflected form of a dictionary headword, the user query is redirected to the dictionary article of that headword.
   (b) else if SMOR provides a valid segmentation into two or more morphologically valid segments and if all components of the word are itself valid dictionary entries in the DWDS system, an aggregated dictionary article is generated "on-the-fly".
2. If the morphological analysis fails, a "Did you mean?" function is triggered. It aims to refer the user to orthographically close (defined in terms of edit distance) dictionary entries.
3. If the "Did you mean?" function fails, i.e. no close dictionary headword can be identified, the user is referred to a corpus search and corpus concordances for the query string are offered.

7. Results and evaluation

The method for generating on-the-fly articles presented here is fully integrated into the DWDS platform. The results in Table 2 are based on an evaluation of the user queries for a period of one month from 23rd April to 23rd May 2017. The logfile for that period contains a total of 190,554 unique lexical queries (types), i.e. only those queries that consist of "bare words" without special characters. Among those queries, 17% (i.e. 33,134) do not have a direct match with a dictionary headword of the DWDS dictionary.

A quick evaluation of the 100 most frequent of these queries led to the classification in Table 2, which shows that 35% of these "out-of-headword" queries correspond to inflected forms of existing dictionary entries and for another 20%, an on-the-fly article can be dynamically generated. For another 28% it was possible to identify candidates via a "Did you mean?" function. Only for 17% of the "out-of-headword" queries the user had to be redirected to a corpus query.

Fallback method	**% of total**	**% correct**
1. Inflected input, redirected to lemma entry	35%	91%
2. On-the-fly dictionary article generated	20%	95%
3. Suggestions "Did you mean?"	28%	68%
4. Redirection to corpus search	17%	n/a

Table 2: Proportion of processed user queries with no direct match for a DWDS dictionary headword

The correctness of this classification is displayed in the last column of Table 2. It shows that more than 91% of the entries were lemmatised correctly and for even 95% a correct dictionary article was generated on-the-fly.

Since the main topic of this paper is on the dynamic generation of dictionary articles, we will focus on a discussion of the second fallback method. Figure 4 lists various examples and how they are dealt with by our approach. A main observation is that the ambiguity problem of automatic morphological analysers is solved remarkably well in our case. This is due to the fact that wrong segmentations can be eliminated in general because at least one of their segments does not have a match with a dictionary headword. This is illustrated by the Examples (1)–(3). Ambiguities due to linking elements can often be solved with the least weight method of the morphological analysis (cf. Section 4.2) as in Examples (4) and (5). Much more difficult is the mapping to the correct word category. Example (6) is a case where the mapping of the morphological analyser works correctly whereas in Example (7) it is incorrect.

(1) *Angsthasenpolitik* ('politics of cowardice'): correct segmentation is found: `Angsthase<>:n<NN>P:politik`, but not `Angst<NN>H:hase<>:n<NN>P:politik<+NN>`.
(2) *Autobahnmeisterei* ('highway maintenance area'): correct segmentation is `Autobahn<NN>M:meisterei<+NN>` (*Meisterei* is `Meister<N>ei<SUFF><+NN>`) and not `Meister<NN>E:ei<+NN>`.
(3) *Krötenlaubfrosch* ('tree frog'): correct segmentation is `Kröte<>:n<NN>L:laubfrosch<+NN>`, but not `Kröte<>:n<NN>L:laub<NN>F:frosch<+NN>` or `Kröte<>:n<NN>L:laub<NN>F:frosch<+NN>`.
(4) *Reiseabschnitt* ('travel segment'): correct segmentation is `Reise<NN>A:abschnitt<+NN>`, not `Reis<>:e<NN>A:abschnitt<+NN>`.
(5) *Arbeitsamtsbericht* ('job center report'): correct segmentation is `Arbeitsamt<NN>B:bericht<+NN>`, and not `Arbeit<>:s<NN>A:amt<NN>B:bericht<+NN>` or even `Arbeit<NN>S:samt<>:s<NN>B:bericht<+NN>`.
(6) *Treibschnee* ('drift snow'): correct expansion is `t:Treibe:<>n:<><V>S:schnee<+NN>`.
(7) *Grillfest* ('barbecue party'): automatic analysis `G:grill<NN>fest<+A>`, whereas the correct segmentation `g:Grille:<>n:<><V>F:fest<+NN>` is not found.
(8) *Arbeitsstellenleiter* ('work place leader', masc., or 'work place ladder', fem.).
(9) *Ballbesitzfußball* ('football game with possession of the ball'): wrong plural.
(10) *Schweinsteiger* (a family name which should not be segmented).

Figure 4: Various examples for correct or incorrect compound segmentations

There are also cases where the ambiguity is undecidable. An example for this case is the homography of *Leiter* ('ladder', fem. vs. 'leader', masc.) as in Example (8). In this case two dictionary entries are generated and the decision about the correctness is left to the user. Another problem for the automatic morphological analyser is the correct generation of inflected forms. For example, in the case of ambiguities between count nouns and non-count nouns, the system has to decide if a plural is possible (count noun sense) or not (non-count noun sense), see Example (9). Finally, the ambiguity between a proper noun and a common noun is a difficulty for our method. This is generally true for all family names that can be segmented into two or more common nouns like in Example (10).

8. Conclusion

We have presented a method to generate on-the-fly articles for the DWDS platform as a fallback solution for user queries that cannot be directly matched with dictionary headwords of the DWDS system. The strategy of generating dynamic dictionary articles on the fly is closely related to the activities in the issue management system: cases of wrong or insufficient articles generated “on the fly” can be reported to the system and eventually a full lexicographic dictionary entry can be compiled manually.

An evaluation of the logfiles of the DWDS platform for a one month period shows that approximately one out of six queries corresponds to a “out-of-headword” query. For 20% of those queries a DWDS dictionary article can be successfully generated on-the-fly. Thus the method presented in this article proves to be useful to augment the number of—actually used—headwords of the DWDS dictionary system.

9. References

Beesley, K.R. & Karttunen, L. (2003). *Finite State Morphology.* Stanford, CA: CSLI.

Didakowski, J. & Geyken, A. (2014). From DWDS corpora to a German word profile–methodological problems and solutions. *OPAL – Online publizierte Arbeiten zur Linguistik*, 2/2014, pp. 39–47.

Geyken, A. (2014). Methoden bei der Wörterbuchplanung in Zeiten der Internetlexikographie. *Lexicographica*, 30(1), pp. 77–111.

Geyken, A. (2015). Recent developments in German lexicography. In *Kernerman Dictionary News*, volume 23. pp. 16–19.

Geyken, A. & Hanneforth, T. (2006). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In *Finite State Methods and Natural Language Processing. 5th International Workshop, FSMNLP 2005, Helsinki, Finland, September 1-2, 2005. Revised Papers*, volume 4002. Springer, pp. 55–66.

Geyken, A. & Lemnitzer, L. (2012). Using Google Books Unigrams to Improve the Update of Large Monolingual Reference Dictionaries. In *Proceedings EURALEX 2012.* Oslo, pp. 362–366.

Haapalainen, M. & Majorin, A. (1995). GERTWOL und morphologische Disambiguierung für das Deutsche. In *Proceedings of the 10th Nordic Conference of Computational Linguistics.* University of Helsinki, Department of General Linguistics.

Klappenbach, R. & Steinitz, W. (eds.) (1964–1977). *Wörterbuch der deutschen Gegenwartssprache (WDG).* Berlin: Akademie-Verlag.

Klein, W. (2013). Von Reichtum und Armut des deutschen Wortschatzes. In *Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache.* Berlin/Boston: De Gruyter Mouton, pp. 15–56.

Klein, W. & Geyken, A. (2010). Das 'Digitale Wörterbuch der Deutschen Sprache DWDS‘. In *Lexicographica*, volume 26. pp. 79–96.

Lawson, M.V. (2003). *Finite Automata.* CRC Press.

Lemnitzer, L. & Würzner, K.M. (2015). Das Wort in der Sprachtechnologie. In U. Haß & P. Storjohann (eds.) *Handbuch Wort und Wortschatz.* De Gruyter, pp. 297–319.

Schmid, H.e.a. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In *Proceedings of LREC.*

Scholze-Stubenrecht, W. (ed.) (1999). *Duden – Das große Wörterbuch der deutschen Sprache in 10 Bänden.* Mannheim: Bibliographisches Institut, 3. edition.

Volk, M. (1999). Choosing the right lemma when analysing German nouns. In *Multilinguale Corpora: Codierung, Strukturierung, Analyse. 11. Jahrestagung der GLDV.* Frankfurt a. M., p. 304–310.

Wahrig-Burfeind, R. (2011). *Wahrig, Deutsches Wörterbuch. Mit einem Lexikon der Sprachlehre.* Gütersloh/München: wissenmedia in der inmedia ONE] GmbH.

Würzner, K.M. & Jurish, B. (2015). A hybrid approach to grapheme-phoneme conversion. In *Proceedings of the 12th International Workshop on Finite State Methods and Natural Language Processing.* URL http://www.aclweb.org/anthology/W/W15/W15-4811.pdf.