A database of German definitory contexts from selected web sources

Adrien Barbaresi, Lothar Lemnitzer, Alexander Geyken
Berlin-Brandenburg Academy of Sciences
Jägerstr. 22/23 – 10117 Berlin – Germany
{barbaresi,lemnitzer,geyken}@bbaw.de

Abstract
We introduce work on the detection of definitory contexts designed to speed up two lexicographical tasks: searching for the exact meaning(s) of terms and providing usable input for paraphrasing. Our database is built from a specialized web corpus using a robust pattern-based extraction method. The corresponding interface displays information for a large range of lexical units. The contributions of this article are threefold: we describe both acquisition and extraction, provide a qualitative assessment of the method, and present an interface to access the data.

Keywords: Definition extraction, definitory contexts, computational lexicography, resource acquisition

1. Introduction
German is well-known for making extensive use of compounds (Schlücker, 2012), as composition is a productive process of German word formation. In the context of lexical resources and dictionaries, the number of potential words to define is very large if not infinite. For that matter, corpus data is crucial in order to perform lexicographic work. Furthermore, automatic definition extraction can be used to complement existing entries or help lexicographers creating new ones. This article introduces both a corpus and a database built to support lexicographers as well as interested users in sorting out the meaning(s) of lexemes. The acquisition draws on selected web sources containing a large number of lexical descriptions, which are expected to contain potential definitory contexts, from which handwritten definitions will be derived. We describe an experimental setting and front-end for pattern-based definition extraction from such resources.

We first describe the context and motivations of our work. In section 2 we summarize related work and place our contribution into this context. In section 3 we present our acquisition and extraction method. Section 4 describes the database, an interface to it as well as the results of a qualitative evaluation. In section 5 we conclude with the insights gained from our investigation.

1.1. Context and motivation
The Digital Dictionary of the German Language (Digitales Wörterbuch der deutschen Sprache, DWDS)1 is a long term project of the Berlin-Brandenburg Academy of Sciences (BBAW). The goal of the DWDS project (Klein and Geyken, 2010) is to create a large-scale aggregated word information system based on legacy dictionaries, large corpora, word statistics and automated methods to provide additional types of linguistic information as well as to speed up the process of updating and amending the existing lexical resources.

The DWDS web platform provides access to this information and is among the most frequently used academic dictionary platforms. The dictionary component of the DWDS is based on high-coverage and detailed dictionaries of contemporary German, including the Wörterbuch der deutschen Gegenwartssprache and Duden Großwörterbuch der deutschen Sprache. Thus, the DWDS dictionary component with more than 170,000 entries provides a good coverage of the core vocabulary of German.

Having covered this core vocabulary, the current work focuses on the description of lexical units whose meaning reaches deeply into highly specialized domains. In order to revise the entries of existing legacy dictionaries and to work on the full lexicographic description of headwords that are not yet recorded, a team of 6 lexicographers has been employed for more than 4 years, with another 6 years to come. We would like to present to our users detailed information concerning the form and grammar as well as the meaning of as many headwords as possible. For instance, it has been shown in a corpus-based quantitative study that in a German newspaper corpus of one billion running words, a number of 4-5 million distinct base forms, i.e. lexical units in our sense, could be detected (Klein, 2013). In a more recent investigation using our morphological analyzer and lemmatizer and a corpus of ca. 5 billion running words, we have collected a total of 16.3 million base forms, most of which appear rather infrequently in the corpora.

However, providing hand-crafted definitions for such a large number of headwords is far beyond the scope of our project. That is why any corpus-based support is highly valuable, with two main areas of interest: on one hand the identification of senses and usage of specialized lexical units and on the other hand the elaboration of appropriate definitions (Geyken et al., 2017a).

In this context, it is expected that a specialized corpus and database with definitory contexts from selected web sources is highly relevant. Since we are reaching out into the periphery of the German vocabulary, i.e. targeting lexical units of specialized use in many domains, the research that is necessary for a correct description of the word senses is costly in time and effort. Typically, the lexicographers are not experts in such domains. That is why any support based on a corpus of reliable internet resources can be helpful. Secondly, it would also be an asset for the users of our dictionary if we can provide contexts for a large share of these words that help the users to sort out for themselves the meaning, hyperonyms, translation equivalents of such

1https://www.dwds.de/
lexical units. Geyken et al. (2017b) present a study that is based on user queries of the DWDS platform and show that 17% of the queries did not match a headword in our dictionary and would therefore not return any lexical information. A database of definitory contexts on top of our high-quality lexical descriptions would indeed be of great help and improve on coverage with our lexical information system.

In short, the corpus and database of contexts will support lexicographers in crafting definitions and help users of our information systems to glean useful semantic information even for headwords which fall beyond the scope of lexicographic work planned for the DWDS.

2. Related work

Automatic definition extraction has already been applied in a range of different contexts and tasks in computational linguistics (Navigli and Velardi, 2010), and the extraction of definitory contexts from free texts is currently a trending topic in lexicography and neighboring disciplines. Method and corpora are two key issues. The method of choice has been the drafting of lexico-syntactic patterns or features (Hearst, 1992) that more or less precisely describe the surface forms of prototypical definitory contexts. Empirical approaches such as the use of conditional random fields in order to label beginning and content of a definition as well as non-related material (Anke, 2013) describe common patterns such as the use of the verb “to be”. While most of the work has focused on English (Borg, 2009; Zhang et al., 2014), there is also work to be found for Slavic languages (Przepiorkowski et al., 2007) and Dutch (Westerhout, 2010). Work on German includes the search for technical terms with a series of structural patterns with verbal predicates (Storrer and Wellinghoff, 2006) as well as methods for proper definition extraction based on hand-crafted rules or patterns, evaluated in numerical terms (Cramer, 2011) or with regard to their statistical relevance (Schumann, 2013). We chose to first follow the work of Cramer (2011) and employ some of her patterns, namely the ones she classifies as most efficient. While a qualitative evaluation of the patterns as well as a classification in terms of efficiency is included, her study does not present any accountable results in empirical, quantitative terms. In this sense, our article can be seen as a replication study on our data. We use the categories introduced by Cramer (2011) and compare them with a baseline consisting of more loosely defined patterns.

The sources used in related work mostly include small domain-specific datasets, e.g. instructional texts compiled for teaching purposes (Borg, 2009); large encyclopedic resources such as Wikipedia (Kovár et al., 2016); or general-purpose web corpora (Navigli et al., 2010); unfiltered noisy data such as the CommonCrawl (Seitner et al., 2016); or small sets of webpages (Schumann, 2014). Our approach is centered on a larger but still specialized web corpus of glossaries and similar lexical resources. While Kovár et al. (2016) for example restrict themselves to a version of the Wikipedia that is part of the Sketch Engine, we aim at a richer and more diverse collection of web sources. Having focused on specialized resources, the extraction in our case is slightly easier than on general-purpose texts, work reported e.g. by Zhang et al. (2014), and slightly more complex than in work targeting Wikipedia. The main challenge therefore is to find relevant information in loosely and diversely structured data, namely the headword (definendum) and the defining context (definitens).

3. Acquisition, extraction and exploitation

3.1. Definitions and definitory contexts

In a strict sense that is applied e.g. in language philology and terminology, a definition supplies information that is sufficient to explicate either the content (or intension) of a concept or the set of individuals that form the extension of the concept. The classical logical form of such a definition is the formulation of genus proximum (the hyponym) and differentiae specificae (exactly those features of the concept that makes it distinguishable from all other concepts which are represented by co-hyponyms). For example, in Princeton WordNet the lexical unit broom is defined as “a cleaning implement for sweeping”. We do not employ this strict sense of definition in our work.

In the wider sense that is commonly employed in lexicography, a definition or, more precisely, a meaning paraphrase, is provided as part of the explication of the meaning of a particular sense of a lexical unit. Such a paraphrase is typically more elaborate than a definition stricto sensu, it is derived from examples of the usage of a word (and not necessarily based on an abstract concept) and in many dictionaries it is accompanied by a set of crafted or manually selected usage examples that illustrate the rules and conventions of the usage of the lexical unit. It is therefore not restricted to conceptual information but should also provide typical and relevant world knowledge that is related with the lexical unit. It might be worth mentioning in an extended meaning paraphrase that brooms are typically made of stiff fibers and typically have a long handle. We will, in the following, use the terms definition and definitory context in that wider sense.

3.2. Examples

Above we argued that for our lexicographical work we are particular interested in definitory contexts for lexical units of highly specialized domains. We will illustrate this point with some examples from our database. The examples consist of the headword, an English equivalent of it (in brackets), the definition proper, a pointer to the source of the definition and some further comments.

Auseinandersetzungsbilanz (dissolution balance)

Eine Auseinandersetzungsbilanz – auch Abschichtungs- bilanz genannt – ist die Bilanz einer Personengesellschaft, die als Grundlage für die Auszahlung eines oder mehrrer Gesellschafter dienen soll. Das Ergebnis dieser Bilanz ist das Auseinanderset-
zungsguthaben. (Source: Anlegerlexikon\textsuperscript{5}). This is a highly specialized and non-transparent term that form the domain of company / corporation law. A synonymous term is given as well as the factually related term Auseinandersetzungsguthaben (credit balance).

Barett, Birett (Biretta)

Barett – auch Birett – ist die Kopfbedeckung von Akademikern und Geistlichen, ausgezeichnet durch die vier- oder fünf Eckige Form. (Source: Das Heiligenlexikon\textsuperscript{6}).

This is a simplex from the domain of church and religion. The term Barett itself is ambiguous. In addition, the reader learns something about the form of this kind of headdress.

Direktionsrecht (employer’s executive prerogatives)

Unter Direktionsrecht – auch Weisungsrecht genannt – wird das Recht des Arbeitgebers verstanden, die Leistungspflichten des Arbeitnehmers einseitig näher auszugestalten. (Source: Lexikon Recht\textsuperscript{7}).

This is a non-transparent compound from the domain of labour relations. Both parts of the compound are highly ambiguous. In addition, a synonymous term, Weisungsrecht is given.

Pelletheizung (pellet stove)

Pelletheizung ist eine Holzheizung, in deren Heizkessel zu Stäbchen geformte Holzabfälle – sogenannte Holzspellets – verbrannt werden. (Source: Baulexikon\textsuperscript{8}).

This is a very recent term from the domain of heating engineering. This definition follows closely the pattern of genus proximum and differentiae specificae.

3.3. Acquisition

We chose to build a specialized web corpus, that is a collection of web documents targeting web pages which are defined in advance\textsuperscript{9}. After identifying and manually selecting a series of relevant websites. We make the hypothesis that there are webpages in which it is probable to find definitions for highly specialized domains, because some feature a higher density of specialized vocabulary than others and some are explicitly characterized as explanatory. Thus, in order to find potential sources, we sift through large lists of URLs collected during web corpus construction for DWDS corpora and look for expressions such as “lexicon” or “glossary”. Heuristic guesses on the URLs determine the probable home page of the given website. The retrieved URLs are then manually reviewed with respect to their potential: out of more than 4200 candidates, a list of 285 websites forms the basis of the present experiment. They include highly specialized lexical domains such as apiculture, astronomy, chemistry, electronics, finance, fishing, metallurgy, politics, religion, and wine-making.

The corpus has been acquired by using focused crawling techniques (Olston and Najork, 2010). This strategy involves finding all pages located at levels deeper than the starting page, which means here that in the best case all definitive material is downloaded and stored. There is also a certain amount of noise: there might be pages with a poor yield in terms of definitions, e.g. impressum or unrelated content. This is a difficulty common to most web corpora, which require filtering operations. Because of the diversity of the websites to crawl, it is not possible to define a precise retrieval strategy. Furthermore, not all pre-selected domains are suitable for crawling, as they are deeply ramified and feature a large number of sub-pages. As a result, the corpus of downloaded pages consists of 268 different websites with a total of 501,308 web documents and about 29 Gb of data.

Additionally, we take already existing, generic web corpora into account in order to manually assess if this specially acquired corpus effectively provides more definitive contexts than a broad search in larger and more diverse corpora.

3.4. Extraction

The extraction process also has to be generic, because of the large number of lexica it would be too cumbersome to craft a targeted extraction algorithm for each webpage. However, the pages do not offer structural patterns which could allow for a reliable boilerplate and metadata extraction (Barbaresi, 2016). These are often loosely structured and mostly provide information in the form of tables, lists or simple paragraphs. On website level, the information can be divided in two different ways: either a web page features a series of entries, for example for each letter of the alphabet, or they provide a single page for each lexical entry. We used the vocabulary coverage (known words and to-do list) of the DWDS project in order to provide lexemes to look for. Then a series of syntactic cues had to be defined. We chose a pattern-based method in order to extract the contexts in a robust manner. Most patterns are derived from the work of Cramer (2011), which is also the occasion to apply them systematically and to review them in terms of adequacy and efficiency. Our patterns first match a lexeme of our list and then look for cues left and right of the definiendum. We adopt the categories defined by Cramer (“strict”, “less strict”, “opportunistic”) and evaluate them with respect to their helpfulness for definition writing. Additionally, heuristic criteria are used to determine which definitions may be directly transferable or not. For example, a “strict” pattern is a constrained syntactic structure which is expected to be a strong case for a definitive context. The patterns by Cramer (2011) are described in human language, we translate them into regular expressions, which seem powerful enough to capture the desired context. The example below integrates a number of potential variants for the pattern “under a X1, one understands X2”, where X1 is a definiendum and X2 a catchall

\textsuperscript{5}www.anleger-beteiligungen.de/htm/de/html/Info_Center-Glossart74a.html
\textsuperscript{6}www.heiligenlexikon.de/Glossar/Priester-Ordensgewaender.html
\textsuperscript{7}www.musterkanzlei.info/1041317/portal/lexikon/-recht/d/direktionsrecht
\textsuperscript{8}www.das-baulexikon.de/lexikon/Pelletheizung.htm
\textsuperscript{9}In the original: sehr genau, genau and mäßig genau.
expression expected to contain a definition: /[Uu]нтер
(?::eine[mr])? $definiendum$ versteht(?::en) (?::man|wir)
$catchall/

A “less strict” pattern is a structure which features less con-
straints and whose output in less certain, for example “X1 is
used for X2”: /$definiendum$ verwendet man für $catchall/

This pattern also raises issues concerning the definitory
context, since the extracted sentences may not be strictly
of lexicographic nature but rather entail practical advice.
We believe that such patterns are still valuable, since they
help determining what the lexeme is about.

An “opportunistic” pattern is a loosely constrained struc-
ture which may be a cue for a definitory context but which
is also expected to be subject to noise. The most basic struc-
ture in our patterns is accordingly “a X1 is a X2”, which can
be translated into a pattern such as /[Ee]ine? $definiendum$

is eine? $catchall/

4. Database

We acquired web data corresponding to specialized lex-
ical resources and extracted definitory contexts which were
loaded into a database. In a first run, we could identify
191,951 contexts. 14,460 text snippets are supposed to
be relevant contexts according to the extraction patterns of
Cramer. In addition, we established a baseline by apply-
ning some looser, opportunistic patterns, that yielded another
177,491 contexts. 14,460 text snippets are supposed to
be returned for at least two of the senses

This pattern also raises issues concerning the definitory
context, since the extracted sentences may not be strictly
of lexicographic nature but rather entail practical advice.
We believe that such patterns are still valuable, since they
help determining what the lexeme is about.

An “opportunistic” pattern is a loosely constrained struc-
ture which may be a cue for a definitory context but which
is also expected to be subject to noise. The most basic struc-
ture in our patterns is accordingly “a X1 is a X2”, which can
be translated into a pattern such as /[Ee]ine? $definiendum$

is eine? $catchall/

4. Database

We acquired web data corresponding to specialized lex-
ical resources and extracted definitory contexts which were
loaded into a database. In a first run, we could identify
191,951 contexts. 14,460 text snippets are supposed to
be relevant contexts according to the extraction patterns of
Cramer. In addition, we established a baseline by apply-
ning some looser, opportunistic patterns, that yielded another
177,491 hits. In the following subsection, we will describe
an interface to this database and a lexicographic evaluation
of the data.

4.1. Interface

We made the data available through a simple interface
where users can search for a particular term and get a
weighted list of resulting definitory contexts along with
metadata such as origin and pattern type.

Figure 1 shows a prototype version of the interface with a
wildcard search displaying results for $Trennscheibe$, where
useful contexts can be returned for at least two of the senses
of the word, and the adjective $trennscharf$, with several con-
texts to choose from. Figure 2 demonstrates the display of
a single definition with an evaluation menu. In that case,
the definition for $Freihandelsabkommen$ (free-trade agree-
ment), is a typical example of specialized vocabulary for
which automatic definition extraction can be of great help.

4.2. Results

In order to perform a qualitative evaluation, we selected a
random sample of 1000 definitory contexts containing both
targeted and baseline contexts. A trained lexicographer as-
essed the quality of the data as follows:

0 Not a definition at all (634; including missed targets,
truncated definitory contexts and doublets);

1 Provides helpful information but is not suitable for dis-
play (276);

2 Appropriate for unfiltered presentation to a user of the
dictionary website, with some minor flaws or errors
(86);

3 Can be directly integrated (14).

All in all, more than a third of the data proved to be help-
ful (classes 1 to 3), which considering the specialization
degree of the lexemes is already worthwhile. In order to re-
fine the extraction process, we performed an error analysis
and listed the most common characteristics that disquali-
fy a text snippet as a good definitory context. We found three
main reasons: the $definiendum$ could often not be detected
correctly; there are many doublets in the data; or the text of
the $definiens$ is not complete due to markup issues or the
definition is split into several sentences.

We assume that our extraction process has to be refined.
Contrarily to the expectations raised by the patterns, regu-
lar expressions alone do not always perform well in prac-
tice on structured data, most notably on tables. The overall
quality of the data can be improved with a more in-depth
analysis of the structural properties of the resources found
online, which involves answering questions relative to the
structure of a $definiendum$. Removing doublets would also
be beneficial, this issue is directly linked to the extraction
which may have to be made less tolerant.

Incomplete contexts could be addressed by a surface anal-
ysis of the syntax, although the content is fragmentary on a
few webpages or do not lead to full sentences due to HTML
extraction; 100 contexts could be properly extracted and
proved to be appropriate for presentation on the website
(classes 2 and 3).

The opportunistic patterns performed just as well as the
“strict” patterns so that we can afford to rely on loose con-
straints for the extraction. This finding can be explained by
the quality of the input data: the URLs of the homepages
have been screened manually before download, and the re-
trieved data seem to confirm that we nearly exclusively deal
with explanatory contexts. Thus, we can confirm that the
selection of our corpus is optimized for the task at hand
since it contains a large number of definitions by its con-
struction principle. The questions we raised are rather of
lexicographic nature, concerning the detection we do not
need to discriminate patterns based on their efficiency $a
priori$ and can afford to perform an opportunistic extraction.

Finally, we started looking for definitions in other web cor-
pora to provide a comparison. We used free texts as input
which have no relation to lexicography whatsoever. The ex-
traction results seem to be far worse than in the lexical re-
sources. This confirmed our intuition that it makes sense to
acquire a dedicated corpus of lexical resources of all kinds
and build a database of definitory contexts based on such a
corpus.
| Begriff       | Definition                                                                 |
|--------------|---------------------------------------------------------------------------|
| Trennscheibe | alle 14 Tage für insgesamt eine halbe Stunde besucht werden, Kontakte sind nur mit **Trennscheibe** erlaubt. Auch die Anwälte können mit ihren Mandanten nur mit **Trennscheibe** sprechen, die Vereidigungsprotokoll wird kontrolliert. Das einzige was sich haben die Betroffenen haben zu schuldem kommen lassen ist, dass sie |
5. Conclusions

We introduced work on the detection of definitory contexts which is meant to support lexicographers in writing definitions for a dictionary of contemporary German. We described a specialized web corpus built for this task as well as robust pattern-based extraction processes on bare HTML documents. The resulting database entails about 191,000 candidates taken from about half a million webpages. The corresponding interface displays information for a large range of lexical units which can then be assessed by lexicographers.

Altogether, the definitory contexts in the database can be really helpful in alleviating two key tasks of the lexicographical workflow, firstly because it speeds up the search for the exact meaning(s) of terms – in the present case highly specialized terms – and secondly because it provides usable input for the task of formulating an appropriate meaning paraphrase.

Our method yielded some useful results but can benefit from some improvements, mostly in structural analysis and selection. Given the type of web texts we use, a qualitative review of the extraction patterns does not seem to be very relevant, less constraints lead to more potential definitions, so that it can be said in our case that “looser is better”. Beyond an opportunistic setting, future challenges reside in finding the right balance between generic approaches and in-depth analysis.

6. Bibliographical References

Anke, L. E. (2013). Towards Definition Extraction Using Conditional Random Fields. In Proceedings of the Student Research Workshop associated with RANLP, pages 63–70.

Barbaresi, A. (2015). Ad hoc and general-purpose corpus construction from web sources. Ph.D. thesis, École Normale Supérieure de Lyon.

Barbaresi, A. (2016). Efficient construction of metadata-enhanced web corpora. In Paul Cook, et al., editors, Proceedings of the 10th Web as Corpus Workshop, pages 7–16. Association for Computational Linguistics.

Borg, C. (2009). Automatic definition extraction using evolutionary algorithms. Ph.D. thesis, University of Malta.

Cramer, I. (2011). Definitionen in Wörterbuch und Text. Ph.D. thesis, TU Dortmund.

Geyken, A., Barbaresi, A., Didakowski, J., Jurish, B., Wiegand, F., and Lemnitzer, L. (2017a). Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). Zeitschrift für germanistische Linguistik, 45(2):327–344.

Geyken, A., Wiegand, F., and Würzner, K.-M. (2017b). On-the-fly Generation of Dictionary Articles for the DWDS Website. In Proceedings of the eLex 2017 Conference, pages 560–570.

Heairst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of COLING, volume 2, pages 539–545. Association for Computational Linguistics.

Klein, W. and Geyken, A. (2010). Das digitale wörterbuch der deutschen sprache (dwd). In Lexicographica, pages 79–96. De Gruyter.

Klein, W. (2013). Von Reichtum und Armut des deutschen Wortschatzes. In Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache, pages 15–56. De Gruyter.

Kovář, V., Močiaríková, M., and Rychlý, P. (2016). Finding Definitions in Large Corpora with Sketch Engine. In Proceedings of LREC, pages 391–394. ELRA.

Navigli, R. and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1318–1327. Association for Computational Linguistics.

Navigli, R., Velardi, P., and Ruiz-Martínez, J. M. (2010). An Annotated Dataset for Extracting Definitions and Hypernyms from the Web. In Proceedings of LREC, pages 3716–3722. ELRA.

Olston, C. and Najork, M. (2010). Web Crawling. Foundations and Trends in Information Retrieval, 4(3):175–246.

Przepiórkowski, A., Degórski, L., Wójtowicz, B., Spousta, M., Kuboř, V., Simov, K., Osenova, P., and Lemnitzer, L. (2007). Towards the automatic extraction of definitions in Slavic. In Proceedings of the 6th Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies, pages 43–50. Association for Computational Linguistics.

Schlücker, B. (2012). Die deutsche Kompositionsfreudigkeit. Übersicht und Einführung. In Livio Gaeta et al., editors, Deutsche als kompositionsfreudige Sprache. Strukturelle Eigenschaften und systembezogene Aspekte, pages 1–25. de Gruyter.

Schumann, A.-K. (2014). Linguistische Analyse und korpusbasierte Extraktion deutscher und russischer wissenshaltiger Kontexte. Ph.D. thesis, Universität Wien.

Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H., and Ponzetto, S. P. (2016). A Large Database of Hypernymy Relations Extracted from the Web. In Proceedings of LREC, pages 360–367. ELRA.

Storrer, A. and Wellinghoff, S. (2006). Automated detection and annotation of term definitions in German text corpora. In Proceedings of LREC, pages 2373–2376. ELRA.

Westherout, E. (2010). Definition extraction for glossary creation: A study on extracting definitions for semi-automatic glossary creation in Dutch. Ph.D. thesis, University of Utrecht.

Wiegand, H. E. (1989). Die lexikographische Definition im allgemeinen einsprachigen Wörterbuch (The lexicographic definition in the general monolingual dictionary. In F.J. Hausmann, et al., editors, Wörterbücher. Dictionnaires, volume 5.1 of Handbücher zur Sprach- und Kommunikationswissenschaft, pages 530–588. De Gruyter.

Zhang, J., Wang, Y., and Yang, D. (2014). Automatic learning common definitional patterns from multi-domain Wikipedia pages. In Data Mining Workshop (ICDMW), 2014 IEEE International Conference on, pages 251–258. IEEE.