# Wiktionary

## The Metalexicographic and the Natural Language Processing Perspective

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

**Dissertation**

zur Erlangung des akademischen Grades
Doktor-Ingenieur

vorgelegt von

**Christian M. Meyer, M. Sc.**
geboren in Mannheim

Tag der Einreichung: 29. August 2013
Tag der Disputation: 28. Oktober 2013

Referenten: Prof. Dr. phil. Iryna Gurevych, Darmstadt
Prof. Dr. phil. Nicoletta Calzolari, Pisa
Prof. Dr.-Ing. Gerhard Weikum, Saarbrücken

Darmstadt 2013
D17

# Abstract

Dictionaries are the main reference works for our understanding of language. They are used by humans and likewise by computational methods. So far, the compilation of dictionaries has almost exclusively been the profession of expert lexicographers. The ease of collaboration on the Web and the rising initiatives of collecting open-licensed knowledge, such as in Wikipedia, caused a new type of dictionary that is voluntarily created by large communities of Web users. This collaborative construction approach presents a new paradigm for lexicography that poses new research questions to dictionary research on the one hand and provides a very valuable knowledge source for natural language processing applications on the other hand. The subject of our research is Wiktionary, which is currently the largest collaboratively constructed dictionary project.

In the first part of this thesis, we study Wiktionary from the metalexicographic perspective. Metalexicography is the scientific study of lexicography including the analysis and criticism of dictionaries and lexicographic processes. To this end, we discuss three contributions related to this area of research: (i) We first provide a detailed analysis of Wiktionary and its various language editions and dictionary structures. (ii) We then analyze the collaborative construction process of Wiktionary. Our results show that the traditional phases of the lexicographic process do not apply well to Wiktionary, which is why we propose a novel process description that is based on the frequent and continual revision and discussion of the dictionary articles and the lexicographic instructions. (iii) We perform a large-scale quantitative comparison of Wiktionary and a number of other dictionaries regarding the covered languages, lexical entries, word senses, pragmatic labels, lexical relations, and translations. We conclude the metalexicographic perspective by finding that the collaborative Wiktionary is not an appropriate replacement for expert-built dictionaries due to its inconsistencies, quality flaws, one-fits-all-approach, and strong dependence on expert-built dictionaries. However, Wiktionary's rapid and continual growth, its high coverage of languages, newly coined words, domain-specific vocabulary and non-standard language varieties, as well as the kind of evidence based on the authors' intuition provide promising opportunities for both lexicography and natural language

processing. In particular, we find that Wiktionary and expert-built wordnets and thesauri contain largely complementary entries.

In the second part of the thesis, we study Wiktionary from the natural language processing perspective with the aim of making available its linguistic knowledge for computational applications. Such applications require vast amounts of structured data with high quality. Expert-built resources have been found to suffer from insufficient coverage and high construction and maintenance cost, whereas fully automatic extraction from corpora or the Web often yields resources of limited quality. Collaboratively built encyclopedias present a viable solution, but do not cover well linguistically oriented knowledge as it is found in dictionaries. That is why we propose extracting linguistic knowledge from Wiktionary, which we achieve by the following three main contributions: (i) We propose the novel multilingual ontology OntoWiktionary that is created by extracting and harmonizing the weakly structured dictionary articles in Wiktionary. A particular challenge in this process is the ambiguity of semantic relations and translations, which we resolve by automatic word sense disambiguation methods. (ii) We automatically align Wiktionary with WordNet 3.0 at the word sense level. The largely complementary information from the two dictionaries yields an aligned resource with higher coverage and an enriched representation of word senses. (iii) We represent Wiktionary according to the ISO standard *Lexical Markup Framework*, which we adapt to the peculiarities of collaborative dictionaries. This standardized representation is of great importance for fostering the interoperability of resources and hence the dissemination of Wiktionary-based research. To this end, our work presents a foundational step towards the large-scale integrated resource UBY, which facilitates a unified access to a number of standardized dictionaries by means of a shared web interface for human users and an application programming interface for natural language processing applications. A user can, in particular, switch between and combine information from Wiktionary and other dictionaries without completely changing the software.

Our final resource and the accompanying datasets and software are publicly available and can be employed for multiple different natural language processing applications. It particularly fills the gap between the small expert-built wordnets and the large amount of encyclopedic knowledge from Wikipedia. We provide a survey of previous works utilizing Wiktionary, and we exemplify the usefulness of our work in two case studies on measuring verb similarity and detecting cross-lingual marketing blunders, which make use of our Wiktionary-based resource and the results of our metalexicographic study. We conclude the thesis by emphasizing the usefulness of collaborative dictionaries when being combined with expert-built resources, which bears much unused potential.

# Zusammenfassung

Wörterbücher bilden die wichtigste Referenz für unser Sprachverständnis. Sie werden von menschlichen Nutzern und von Computerprogrammen gleichermaßen eingesetzt. Bisher wurden Wörterbücher nahezu ausschließlich von professionellen Lexikographen verfasst. Neue Möglichkeiten der Interaktion im Web und die zunehmende Bestrebung frei-zugängliches Wissen zu dokumentieren, wie etwa in Wikipedia, schaffen einen neuartigen Typ von Wörterbuch, welcher von großen Nutzergemeinden freiwillig erstellt wird. Das gemeinschaftlich-kollaborative Vorgehen etabliert ein neues lexikographisches Paradigma, das einerseits zu völlig neuen Forschungsfragen im Bereich der Wörterbuchforschung führt und andererseits eine wertvolle Wissensquelle für sprachtechnologische Anwendungen darstellt. Wiktionary ist das derzeit größte gemeinschaftlich erstellte Wörterbuch und Gegenstand unserer Forschung.

Im ersten Teil der vorliegenden Arbeit untersuchen wir Wiktionary aus der metalexikographischen Perspektive. Metalexikographie bezeichnet die wissenschaftliche Beschäftigung mit der Lexikographie, Wörterbüchern und lexikographischen Prozessen. Wir diskutieren drei Forschungsbeiträge aus diesem Bereich: (i) Wir geben eine detaillierte Beschreibung von Wiktionary und den damit verbundenen vielfältigen Sprachversionen und Wörterbuchstrukturen. (ii) Wir analysieren die gemeinschaftliche Vorgehensweise der Wiktionary-Autoren. Unsere Ergebnisse zeigen, dass sich die bekannten Phasen des lexikographischen Prozesses nur mäßig auf Wiktionary anwenden lassen. Daher schlagen wir eine neue Prozessbeschreibung vor, die auf der häufigen und fortwährenden Überarbeitung und Diskussion der Wörterbuchartikel und der lexikographischen Instruktionen basiert. (iii) Wir vergleichen Wiktionary mit anderen Wörterbüchern hinsichtlich der enthaltenen Sprachen, Lemmazeichen, Bedeutungen, pragmatischen Markierungen, lexikalischen Relationen und Übersetzungen. Für die metalexikographische Perspektive kommen wir zu dem Schluss, dass Wiktionary kein adäquater Ersatz für professionell erstellte Wörterbücher ist, da Inkonsistenzen und qualitative Mängel überwiegen, weder Wörterbuchfunktionen noch Benutzerbezug festgelegt sind und professionelle Wörterbücher häufig zur Verifikation der Wörterbuchangaben dienen. Das rasante und stetige Wachstum, die große Zahl verschiedener Sprachen, Neologismen, domänenspezifisches Vokabular und nicht standardisierter Varietäten, sowie die Einbeziehung der Intuition und

subjektiven Meinungen von vielen Autoren zeigen dagegen vielversprechende Möglichkeiten für Lexikographie und Sprachtechnologie auf. Insbesondere erweist sich Wiktionary als größtenteils komplementär zu professionell erstellten Wortnetzen und Thesauri.

Im zweiten Teil der Arbeit adressieren wir die sprachtechnologische Perspektive, um das kodierte linguistische Wissen für softwaregestützte Anwendungen nutzbar zu machen. Solche Anwendungen benötigen enorme Mengen strukturierter Daten mit hoher Qualität. Während professionell erstellte Ressourcen häufig zu klein oder zu teuer in ihrer Erstellung und Unterhaltung sind, mangelt es bei vollautomatisch erzeugten Ressourcen häufig an der Qualität der extrahierten Angaben. Gemeinschaftlich erstellte Enzyklopädien werden vielfach als Lösung vorgeschlagen, enthalten jedoch kaum linguistisch orientiertes Wissen. Wir adressieren daher die Extraktion und Aufbereitung linguistischer Angaben aus Wiktionary und diskutieren dazu drei Forschungsbeiträge: (i) Wir erzeugen die neue mehrsprachige Ontologie OntoWiktionary basierend auf den in Wiktionary kodierten schwach strukturierten Angaben. Eine wesentliche Herausforderung liegt dabei in der Auflösung von Mehrdeutigen in semantischen Relationen und Übersetzungen, die wir mit Hilfe automatischer Methoden zur Lesartendisambiguierung erreichen. (ii) Wir alignieren Wiktionary und WordNet 3.0 auf der Ebene der Wortbedeutungen, was zu einer integrierten Ressource mit höherer Abdeckung und reichhaltigeren Angaben führt. (iii) Wir modellieren Wiktionary anhand des ISO-Standards *Lexical Markup Framework* und beschreiben das dazu nötige Lexikonmodell im Hinblick auf gemeinschaftlich erstellte Wörterbücher. Dies ist von großer Wichtigkeit, um die Interoperabilität zwischen verschiedenartigen Ressourcen zu gewährleisten. Unsere Arbeit ist damit ein fundamentaler Schritt zur umfangreichen integrierten Ressource UBY, welche einen einheitlichen Zugriff auf eine große Zahl standardisierter Wörterbücher erlaubt – sowohl über eine Online-Plattform für menschliche Nutzer als auch über eine Programmierschnittstelle für sprachtechnologische Anwendungen. Entwickler können insbesondere zwischen verschiedenen Wörterbüchern wechseln und deren Angaben kombinieren ohne die Anwendung von Grund auf neu zu konzipieren.

Unsere finale Ressource und die dazugehörigen Datensätze und Software-Tools sind frei verfügbar und können für vielfältige sprachtechnologische Anwendungen eingesetzt werden. Damit schließen wir speziell die Lücke zwischen den oftmals kleinen professionell erstellten Wortnetzen und den großen Mengen enzyklopädischer Angaben aus Wikipedia. In einer Bestandsaufnahme charakterisieren wir frühere Arbeiten zur sprachtechnologischen Nutzung von Wiktionary, bevor wir die Zweckmäßigkeit unserer Arbeit anhand zweier Fallbeispiele zur Messung von Verb-Ähnlichkeiten und zur Identifikation von sprachenübergreifenden Werbepannen aufzeigen. Dabei bauen wir auf den Erkenntnissen unserer metalexikographischen Analyse und unseren Wiktionary-basierten Ressourcen auf. Im abschließenden Fazit stellen wir den Nutzwert gemeinschaftlich erstellter Wörterbücher heraus, wenn diese insbesondere mit professionell erstellten Ressourcen kombiniert werden – eine Forschungsrichtung, die noch sehr viel ungenutztes Potential birgt.

# Acknowledgements

Writing this dissertation would not have been possible without the help of many people. I would like to thank Prof. Dr. Iryna Gurevych for affording me the opportunity to conduct this research and for giving excellent feedback throughout the past years. I am grateful to my referees Prof. Dr. Nicoletta Calzolari and Prof. Dr. Gerhard Weikum for finding the time to evaluate my thesis. This work was funded by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant № I/82806, which I gratefully acknowledge.

Moreover, I want to thank the co-authors of my scientific publications, Dr. Andrea Abel, Dr. Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Tri Duc Nghiem, Christian Wirth, and the Open Linguistics Working Group, for the countless brainstorming and paper planning sessions, as well as the multiple anonymous reviewers for adding the final polish. I am truly indebted to my student assistants Yevgen Chebotar and Christian Kirschner, who contributed an impressive share of tools and foundational work related to this thesis. In addition to that, I am grateful to Daniel Bär, Sergii Cherkavskyi, Richard Eckart de Castilho and the DKPro developers, Verena Henrich, Dr. Andrew A. Krizhanovsky, Bastian Laur, Christof Müller, Elisabeth Niemann, Lizhen Qu, Richard Steuer, Dr. Torsten Zesch, the students of the *Lexical-Semantic Methods for Language Understanding* course held in the summer term 2010, and the Wiktionary community for providing and contributing to the software, resources, and datasets created and used in the scope of this thesis.

I would like to give special thanks to the UBY team for the particularly successful collaboration, Michael Matuschek for extensive proofreading, Elisabeth Niemann for many fruitful discussions, and all my colleagues at the Ubiquitous Knowledge Processing Lab for offering a sympathetic ear and a helping hand at any time. The researchers around the DFG network "Internetlexikografie", in particular Dr. Annette Klosa, Prof. Dr. Robert Lew, Michael Mann, Prof. Dr. Stefan Schierholz, and Prof. Dr. Angelika Storrer, deserve special recognition for shaping my view of metalexicography.

Last, but not least, I would like to express my deepest gratitude to my family and friends who inspired and supported me through this work. Without the help and love of my family, this work would not have been possible.

# Contents

Chapter 1

# Introduction

*Dictionaries* play a major role in our society for thousands of years. The earliest dictionaries date back to about 2300 BCE. They served the purpose of reading and interpreting religious texts in India, China, and the Middle East. Besides religion, Hausmann (1989) considers poetry and fiction, education, public administration and politics, economy, and linguistics as the "driving forces" for compiling dictionaries. Today, dictionaries are used for understanding texts (*reception*), writing in a clear, comprehensible way (*production*), and for *translating* between different languages. We consult dictionaries when learning languages, preparing contracts and legal texts, authoring books and articles, communicating during a vacation or business trip, mediating in international conflicts, or studying ancient texts – to name just a few situations of our daily lives. In addition to fulfilling the information needs by individuals, dictionaries serve a collective need for recording and documenting language through generations and cultures.

*Lexicography* is the practice of dictionary making. Over the decades, new technologies and theories as well as experiences with previous dictionary projects have caused fundamental changes in the way dictionaries are compiled. We differentiate five major paradigms in the development of lexicography over the last hundred years:[1]

(1) editorial lexicography,
(2) corpus-based lexicography,
(3) electronic lexicography,
(4) computational lexicography, and
(5) collaborative lexicography.

[1] In a wider sense, lexicography encompasses also the compilation of encyclopedias, encyclopedic dictionaries, or other reference works. Unless otherwise indicated, we focus on language lexicography and thus the compilation of (language) dictionaries. We have chosen the five lexicographic paradigms to highlight the major developments in lexicography related to this thesis. It is not our intent to provide a comprehensive description on the history of lexicography, which has been the subject of Hartmann (1986) and Considine (2008), for example.

The discussed paradigms are not temporally distinct from each other. They are rather closely entwined and lead to multiple orthogonal dimensions of dictionary types, which we describe below and summarize in figure 1.1.

Most early dictionaries have been compiled by individual authors. While the pioneering work by Jean Nicot, Samuel Johnson, Noah Webster, the Brothers Grimm, Nicola Zingarelli, and many others on the first comprehensive dictionaries paved the way for modern lexicography, the effort has often been underestimated in those projects. The *Deutsches Wörterbuch* (1854–1961), for example, was initially planned for ten years, but it took already 25 years for its preparation phase and for the compilation of the articles between *"A"* and *"Frucht"*, which is the last article edited by the Brothers Grimm (Kirkness, 2011). As a consequence of this, the lexicographic work has been shared among a number of authors that constitute the *editors* of a dictionary. We call this change in the lexicographic practice *editorial lexicography*. Dictionaries constituted by editors – or *editorial dictionaries* for short – raise an increasing need for planning and organizing the work. This is necessary for ensuring consistent descriptions throughout the dictionary and for minimizing the communicative overhead among the editors. Many major dictionaries in modern times are constituted by editors, including the *Oxford English Dictionary* (1989) and the various *Merriam-Webster's* and *Duden* dictionaries.

Increasing demands on the quality and consistency of a dictionary have led to the rise of *corpus-based lexicography*. Instead of relying on their own intuition and feeling for language, the lexicographers rather strive for providing evidence for a word's meaning and usage (cf. Hanks, 1990; Atkins and Rundell, 2008). To achieve that, lexicographic corpora consisting of books, newspapers, charters, tape recordings, etc. have been created and systematically excerpted. There are, for instance, over four million index cards that document the lexicographic facts for formulating the dictionary articles of the *Deutsches Wörterbuch* (1854–1961).[2] Like in many other disciplines, the advent of computer technology has caused fundamental innovations in lexicography. Computers facilitate the creation of large electronic text corpora and thus a broader and more comprehensive access to language than possible before. Between the 1980s and 1990s, many dictionaries started to make use of electronic corpora – prominent examples are the *Collins COBUILD English Language Dictionary* (1987) and the *New Oxford Dictionary of English* (1998).

The emerging computerization and, in particular, the invention of the World Wide Web has established a third lexicographic paradigm: *electronic lexicography* (Granger and Paquot, 2012). Dictionaries are electronically compiled using *dictionary writing systems* that support the lexicographers with organizing their work, editing the articles, and preparing the final dictionary for publication (Abel, 2012). Database technologies and interoperable data formats allow for cross-media publication (e.g., releasing a general and a learner's dictionary both electronically and as printed books based on the same data model, cf. Alexa et al., 2002). A steadily increasing number of electronic dictionaries are published online. In such *online dictionaries*,

[2] `http://dwb.bbaw.de/arbeitsstelle/archiv` (24 May 2013)

| | |
|---|---|
| **editorial lexicography** | *dictionary types by author:*<br>– dictionaries by individual authors<br>– editorial dictionaries |
| **corpus-based lexicography** | *dictionary types by evidence:*<br>– dictionaries based on intuition and introspection<br>– dictionaries based on corpus evidence<br>  – dictionaries based on electronic corpora |
| **electronic lexicography** | *dictionary types by medium:*<br>– printed dictionaries<br>– electronic dictionaries<br>  – online dictionaries |
| **computational lexicography** | *dictionary types by purpose:*<br>– human-oriented dictionaries<br>– machine-oriented dictionaries<br>  – machine-readable dictionaries<br>  – computational dictionaries<br>  – wordnets<br>  – ontologies |
| **collaborative lexicography** | *dictionary types by user contribution:*<br>– expert-built dictionaries<br>– collaborative dictionaries |

Figure 1.1: Different dimensions of dictionary types and the corresponding lexicographic paradigms

there are practically no space restrictions which have been a pertinent problem of printed dictionaries. In addition to that, online user interfaces can be more flexible than paper-based dictionaries, for example, by using dynamic elements, complex visualizations, hyperlinks, advanced search options, or integrated language tools.

The availability of electronic dictionaries is also the cornerstone of *computational lexicography*, which denotes the use of dictionaries for natural language processing tasks that rely on vast amounts of linguistic knowledge. However, dictionaries are in the first place compiled for human users (i.e., *human-oriented dictionaries*) rather than for machines, which require different means of accessing the encoded information. Early computational approaches employed *machine-readable dictionaries*. That is, computationally accessible versions of existing dictionaries originally intended for human users (Litkowski, 2006), such as *Webster's Seventh New Collegiate Dictionary* (1965) or the *Longman Dictionary of Contemporary English* (1978). Apart from that, many *computational dictionaries* have been compiled – i.e., dictionaries that are designed and intended for computational use. A groundbreaking example is the Princeton

*WordNet* (1985 f.), which has been used in hundreds of applications and revised and extended for over 20 years now.

Other computational dictionaries include the multilingual *EuroWordNet* (1999) linking eight wordnets from different languages, the *FrameNet* (1997 f.) lexicon of semantic frames, and *VerbNet* (2000 f.), which encodes syntactic verb classes. More recently, *ontologies* such as *Cyc* (1995 f.), *DBpedia* (2007 f.), and *YAGO* (2007 f.) are being proposed, which allow for accessing the encoded knowledge in a standardized way and for inferring new facts based on what is explicitly defined. As of today, ontologies are employed in numerous natural language processing systems and represent the backbone of the *Semantic Web* (Berners-Lee et al., 2001). Their fine-grained structure particularly fosters the detailed modeling of domain-specific knowledge and terminology.

## 1.1 Collaborative Lexicography and Research Goals

In this thesis, we discuss a new paradigm of lexicography that has emerged in the last ten years: *Collaborative lexicography* denotes the compilation of *collaborative dictionaries*, which are based on the contributions of voluntary authors. Instead of relying on a small group of lexicographers and professional editors (i.e., the "wisdom of experts"), collaborative dictionaries are backed by the collective intelligence and the subjective opinions of many authors – often described as the "wisdom of crowds" (Surowiecki, 2005). The collaborative approach has its origin in the development of *social media technologies* of the so-called *Web 2.0*, such as blogs, wikis, and social networks, which have caused a transition from academic and professionally edited content to *user-generated content*. Closely connected are the sociological and cultural changes towards an information and knowledge society as well as multiple initiatives for collecting open-licensed knowledge (e.g., the Free Software Foundation, the Wikimedia Foundation, or the Open Knowledge Foundation).[3] One of the most popular collaboratively created works is the free online encyclopedia *Wikipedia* (2001 f.). Wikipedia has been found to be of competitive quality compared to expert-built encyclopedias (Giles, 2005; Casebourne et al., 2012) and it proved highly useful in numerous natural language processing applications (Medelyan et al., 2009).

While much work has been dedicated to Wikipedia, our knowledge of collaborative (language) dictionaries is still limited and presents a major research gap. The study of this new lexicographic paradigm is, however, highly relevant, given the fact that collaborative encyclopedias are about to replace expert-compiled ones and that user-generated content is increasingly dominating the Web. This is why the focus of our research is *Wiktionary* (2002 f.), which is currently the largest collaborative dictionary available. The present thesis is interdisciplinary in nature at the intersection of lexicography and computer science. More specifi-

[3] `http://www.fsf.org`; `http://www.wikimedia.org`; `http://okfn.org` (3 June 2013)

cally, we research Wiktionary from a metalexicographic perspective by studying the impact of collaborative lexicography on the prevailing lexicographic theories and processes, and we research Wiktionary from the perspective of natural language processing by harvesting its linguistic knowledge and making it usable for language processing systems. The goal of our work is to gain new insights into collaborative lexicography on the one hand and to assess the potential of using collaboratively constructed linguistic knowledge in the context of natural language processing on the other hand. The two fields of research are closely connected, since harvesting the linguistic knowledge from Wiktionary is not meaningful without a comprehensive analysis of its structure and organization, whereas researching the collaborative lexicographic paradigm requires qualitative and quantitative data analysis through natural language processing systems. In addition to that, Granger (2012, p. 2) notes that the line between human-oriented and computational dictionaries is "progressively narrowing" due to the increasing integration of computational tools and resources into human-oriented dictionaries and, vice-versa, the integration of vast amounts of linguistic knowledge into natural language processing applications.

In the following sections, we give a detailed overview of our contributions in the area of metalexicography and natural language processing. We then describe our previous publications and introduce the terminology and typographic conventions used in the remaining thesis.

## 1.2 The Metalexicographic Perspective

*Metalexicography* is the science of studying lexicography. That is, the research of lexicographic processes and practices as well as the analysis of lexicographic reference works. The field of metalexicography that is concerned with language dictionaries is also called *dictionary research*. From the perspective of metalexicography, our aim is to study the following research question:

**Research Question A**: *In which way is collaborative lexicography different from previous lexicographic paradigms, and what are the implications for lexicographers, dictionaries, and dictionary users?*

We discuss three metalexicographic contributions towards answering this research question. Our first contribution is describing Wiktionary, the subject of our study. In chapter 2, we provide a general overview of Wiktionary describing its macrostructure, mediostructure, access paths, microstructure, community, and outside matter. This description is the basis for all further contributions discussed in the thesis. Unlike previous works, we not only describe the information items (i.e., headwords, sense definitions, etymologies, etc.), but also take note of the outside matter of the dictionary including its guidelines, usability, and the participating users. We summarize this first contribution as:

**Contribution 1**: *We provide a comprehensive description of Wiktionary.*

Based on this detailed description, we get in a position to analyze the collaborative construction approach of Wiktionary. In chapter 3, we assess the dictionary conception and innovative features of Wiktionary, before we study the collaboration and coordination of its web community based on the revision history of the dictionary articles. The quality of collaborative works is among the most controversial topics, which is why we analyze the quality assurance measures and the sources of lexicographic evidence used in Wiktionary. As a result of our study, we propose a novel description of the lexicographic process of collaborative dictionaries. This contribution is summarized as:

**Contribution 2**: *We analyze the collaborative construction process of Wiktionary.*

Previous works on Wiktionary predominantly relied on qualitative analyses based on a few sample articles. We go beyond these works by carrying out a large-scale quantitative analysis of the English, German, and Russian Wiktionary editions, which we describe in chapter 4. Our goal is to identify well-covered topics and systematic gaps. We compare our quantitative results with other publicly available dictionaries to evaluate the potential of Wiktionary for lexicography and natural language processing. Our analysis addresses the available language editions and the encoded lexical entries, word senses, pragmatic labels, lexical relations, and translations. The contribution can be summarized as:

**Contribution 3**: *We perform a large-scale quantitative analysis of Wiktionary in comparison to other dictionaries.*

Based on the findings reported in the first part of the thesis, we conclude the metalexicographic study and lead over to the natural language processing perspective.

## 1.3 The Natural Language Processing Perspective

*Natural language processing* denotes the research of computational applications involving human language, such as machine translation, information retrieval, or question answering, including their algorithms, resources, and evaluation. From the natural language processing perspective, we aim at studying our second research question:

**Research Question B**: *How can we harvest linguistic knowledge from Wiktionary that can be effectively employed in natural language processing systems?*

We discuss four contributions related to natural language processing towards answering this research question.

Since Wiktionary is intended to be consulted by humans, we first need to extract the encoded knowledge by means of text mining methods. In chapter 5, we survey multiple software

tools for this purpose. A particular challenge is that the extracted information is not sense-disambiguated *per se* and thus yields close semantic relationships between actually unrelated words. This is why we propose and evaluate a method for disambiguating this information automatically. Using the disambiguated knowledge, we construct the multilingual ontology *OntoWiktionary*, which is based on a notion of lexicalized concepts inferred from the sense-disambiguated synonyms. Our work goes beyond the mere scraping of information from Wiktionary discussed in most previous works. It is rather targeted at providing a more expressive structure and at inferring implicitly encoded knowledge. The experiments carried out in the remaining thesis confirm that this approach yields better results than using the raw Wiktionary data. We summarize this contribution as:

**Contribution 4**: *We extract linguistic knowledge from Wiktionary and construct the multilingual, sense-disambiguated lexical ontology OntoWiktionary.*

Our large-scale quantitative comparison showed that Wiktionary is – to some extent – complimentary to other computational dictionaries. In accordance with de Melo and Weikum (2009), we expect an increased coverage and accuracy for natural language processing systems when relying on the combined evidence found in multiple sources. In chapter 6, we discuss how Wiktionary can be aligned with other computational dictionaries in order to obtain an increased coverage and an enriched representation of word senses. For the first time, we describe and evaluate a method for obtaining automatic word sense alignments for Wiktionary. We use the example of aligning Wiktionary and WordNet, but also point out multiple subsequent works on aligning Wiktionary with other resources. We evaluate our work using a newly compiled dataset for this task, for which we provide a detailed data analysis. This is necessary, because little work has been done on obtaining reliable and reproducible evaluation datasets for word sense alignments. This fifth contribution can be summarized as follows:

**Contribution 5**: *We align Wiktionary and WordNet at the level of word senses.*

Providing word sense alignments is only one step towards fully interoperable language resources. In chapter 7, we represent Wiktionary based on the international ISO standard *Lexical Markup Framework* (LMF) in order to overcome differences in the organization, terminology, coverage, information types, data formats, and access paths among multiple resources. This makes it possible to integrate our multilingual ontology OntoWiktionary into the large-scale unified resource *UBY* (2012). We particularly discuss how our data model and software tools deal with noise and inconsistencies induced by the weakly structured knowledge encoded in Wiktionary. By means of the standardized representation, we get in a position to switch easily between multiple computational dictionaries and to combine their information without further transformation. We summarize this contribution as:

**Contribution 6**: *We create a standardized representation of Wiktionary based on the Lexical Markup Framework.*

Our final standardized ontology with translations into over 1,000 languages is publicly available for the research community and can be applied to multiple natural language processing tasks.[4] In chapter 8, we provide a detailed survey of applications making use of Wiktionary data. We then exemplify the usefulness of our derived resources in two use cases: First, we employ OntoWiktionary for computing monolingual and cross-lingual verb similarity. This task benefits from the sense-disambiguated knowledge provided by our resource. Second, we discuss the use of Wiktionary in a translation context by identifying potential cross-lingual marketing blunders. Our solution directly builds on the insights from the quantitative data analysis and the standardized representation of knowledge. We summarize this contribution as:

**Contribution 7:** *We employ Wiktionary in natural language processing tasks.*

We conclude the thesis by observing that the collaborative construction approach of Wiktionary is an important new paradigm in the area of metalexicography and that its linguistic data can be effectively employed for natural language processing applications. The main findings of our work are summarized in chapter 9.

Unless otherwise indicated, all statistics and experimental results refer to Wiktionary data from February 25, 2013 (English Wiktionary), February 20, 2013 (German Wiktionary), and February 17, 2013 (Russian Wiktionary). We describe our software, lexical resources, and annotated datasets in appendix A and make them freely available from our homepage:

`http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/`
`http://www.christian-meyer.org/research/publications/dissertation/data/`

## 1.4 Publication Record

Some parts of this thesis have been previously published at peer-reviewed conferences and in internationally recognized edited books and journals in the fields of lexicography, computational linguistics, and computer science.

We describe the metalexicographic perspective of Wiktionary (chapters 2, 3, and 4) in the recently appeared handbook *Electronic Lexicography* published by the Oxford University Press (Meyer and Gurevych, 2012a). In two focused contributions that we have published as part of the OPAL series of the Institut für Deutsche Sprache in Mannheim, Germany (Meyer and Gurevych, 2013) and at the *Second Web Science Conference* in Raleigh, NC, USA (Meyer and Gurevych, 2010a), we research the lexicographic process of Wiktionary and how its community collaborates on encoding the dictionary articles. We have also presented our work as part of the academic network on internet lexicography[5] and at the symposium *Ihr Beitrag bitte!* –

[4] `http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/`
[5] `http://www.internetlexikografie.de` (30 April 2013)

*Der Nutzerbeitrag im Wörterbuchprozess* held at the annual congress of the Gesellschaft für Angewandte Linguistik in Erlangen, Germany.

With regard to the natural language processing perspective, we describe the ontology construction process of OntoWiktionary (chapter 5) as part of the edited collection *Semi-Automatic Ontology Development: Processes and Resources* published by IGI Global (Meyer and Gurevych, 2012b). We particularly focus on the task of disambiguating semantic relations and translations, for which we have published a pilot study at the *11th International Conference on Computational Linguistics and Intelligent Text Processing* (CICLing) in Iași, Romania and a state-of-the-art solution at the *24th International Conference on Computational Linguistics* (COLING) held in Mumbai, India (Meyer and Gurevych, 2010b, 2012c). Moreover, we have presented our work on aligning Wiktionary and WordNet (chapter 6) at the *5th International Joint Conference on Natural Language Processing* (IJCNLP) in Chiang Mai, Thailand (Meyer and Gurevych, 2011).

The standardized representation of Wiktionary (chapter 7) is part of a shared effort towards the large-scale integrated resource *UBY* (2012). We have actively contributed to the description of the lexicon model UBY-LMF (Eckle-Kohler et al., 2012), its instantiation UBY (Gurevych et al., 2012a), and the web interface for accessing the standardized data (Gurevych et al., 2012b), which have been published at leading computational linguistics conferences. Our Wiktionary-specific work on the UBY-LMF model is described in relation to the standardization of OmegaWiki, which we jointly published in the *Translation: Computation, Corpora, Cognition* journal (Matuschek et al., 2013). Recently, we contributed to the international handbook on the *Lexical Markup Framework* published by Wiley-ISTE (Eckle-Kohler et al., 2013).

In addition to that, we have presented our UBY-related work at the *23rd Meeting of Computational Linguistics in the Netherlands* (Gurevych et al., 2013), and we contributed to the dissemination of standardized resources in the context of the Semantic Web in a joint publication by the Open Linguistics Working Group (Chiarcos et al., 2012). Together with Andrea Abel, we have published an article on user contributions to online dictionaries at the *3rd Biennial Conference on Electronic Lexicography* (eLex) held in Tallinn, Estonia, in which we take up the subject of user contributions to Wiktionary in a broader context (Abel and Meyer, 2013).

## 1.5 Terminology and Typographic Conventions

When considering the interdisciplinary nature of this thesis, we face the challenge of contrasting vocabularies used in the fields of metalexicography, linguistics, natural language processing, and computer science. One example is the term *lexicon*, which has been used as a synonym for *dictionary*, to specifically denote the *mental lexicon* of a person, and for referring to *lexical resources* used merely by computer programs rather than by humans. Unless otherwise indicated, we use the terms defined by the *Dictionary of Lexicography and Dictionary Research* (Wiegand et al., 2010) and by the ISO standard *Lexical Markup Framework* (ISO 24613,

2008) throughout the thesis. Though we give a brief definition of these terms, it goes beyond the scope of this thesis to provide detailed comments on their meaning and how they relate to each other, which is why we refer the reader to the two original publications. Some terms are taken from the works by Atkins and Rundell (2008) and Guarino et al. (2009) as well as some other sources referenced in the text.

We adopt the common lexicographic practice of separating the bibliography by *scientific literature* and *dictionaries and resources*. For the latter, we use italics for providing citations in the running text, for example, *Wiktionary* (2002 f.). Electronically published expandable dictionaries (i.e., continually updated resources) are indicated by an "f." next to the year of their first publication. In addition to that, we use the following typographic conventions:

- newly introduced terms and example lemmas are typed in italics (e.g., *puppy*),
- synsets are enclosed by curly brackets (e.g., {*puppy, juvenile dog*}),
- concepts are enclosed by angle quotes and typed in small caps (e.g., ›YOUNG DOG‹),
- relations are written as pairs in parentheses (e.g., (*puppy, dog*)),
- classes of the LMF standard are printed in a monospace font starting with an upper case letter (e.g., `LexicalEntry`), and
- LMF data categories are printed in a monospace font starting with a lower case letter (e.g., `partOfSpeech`).

# Part I

# The Metalexicographic Perspective

Chapter 2

# Wiktionary

This chapter aims at describing the different kinds of knowledge encoded in Wiktionary and how the dictionary is organized. We first provide a general overview of Wiktionary and its historical development (section 2.1). Then, we introduce the different language editions (section 2.2), their macrostructure (section 2.3), mediostructure (section 2.4), access paths (section 2.5), microstructure (section 2.6), community, and outside matter (section 2.7).

## 2.1 Overview

Wiktionary is a collaboratively created, multilingual online dictionary. The project has been established in December 2002 as a "companion volume" to *Wikipedia* (2001 f.). It originated from a long discussion within the Wikipedia community concerning the exclusion of linguistic knowledge from its encyclopedic articles.[6] Daniel Alston, Brion Vibber, and Tim Starling were important initiators of this development.[7] The name "Wiktionary" was chosen as a portmanteau of "wiki" and "dictionary". A *wiki* is a web-based application allowing simple editing of hyperlinked web pages in a collaborative manner (see Leuf and Cunningham, 2001) and thus refers to the *dictionary writing system* and the compilation process. By 2004, Wiktionary had gradually turned into an independent project. Starting with the French and Polish Wiktionaries, a separate language edition had been created for all 143 active Wikipedia editions by May 2004 and Wiktionary moved to its current URL `http://www.wiktionary.org`. Since then, Wiktionary has rapidly grown and attracted an increasing number of users. By the end of 2006, seven Wiktionaries exceeded 100,000 articles and, by the time of writing, over 15 million dictionary articles in over 170 languages have been created by the Wiktionary community.

[6] See `http://en.wikipedia.org/w/index.php?oldid=294531` (1 November 2001), `http://meta.wikimedia.org/w/index.php?oldid=403` (22 November 2001)

[7] See `http://meta.wikimedia.org/w/index.php?oldid=3149` (25 November 2002), `http://en.wikipedia.org/w/index.php?oldid=378432551` (11 August 2010), `http://meta.wikimedia.org/w/index.php?oldid=1759149` (11 December 2009)

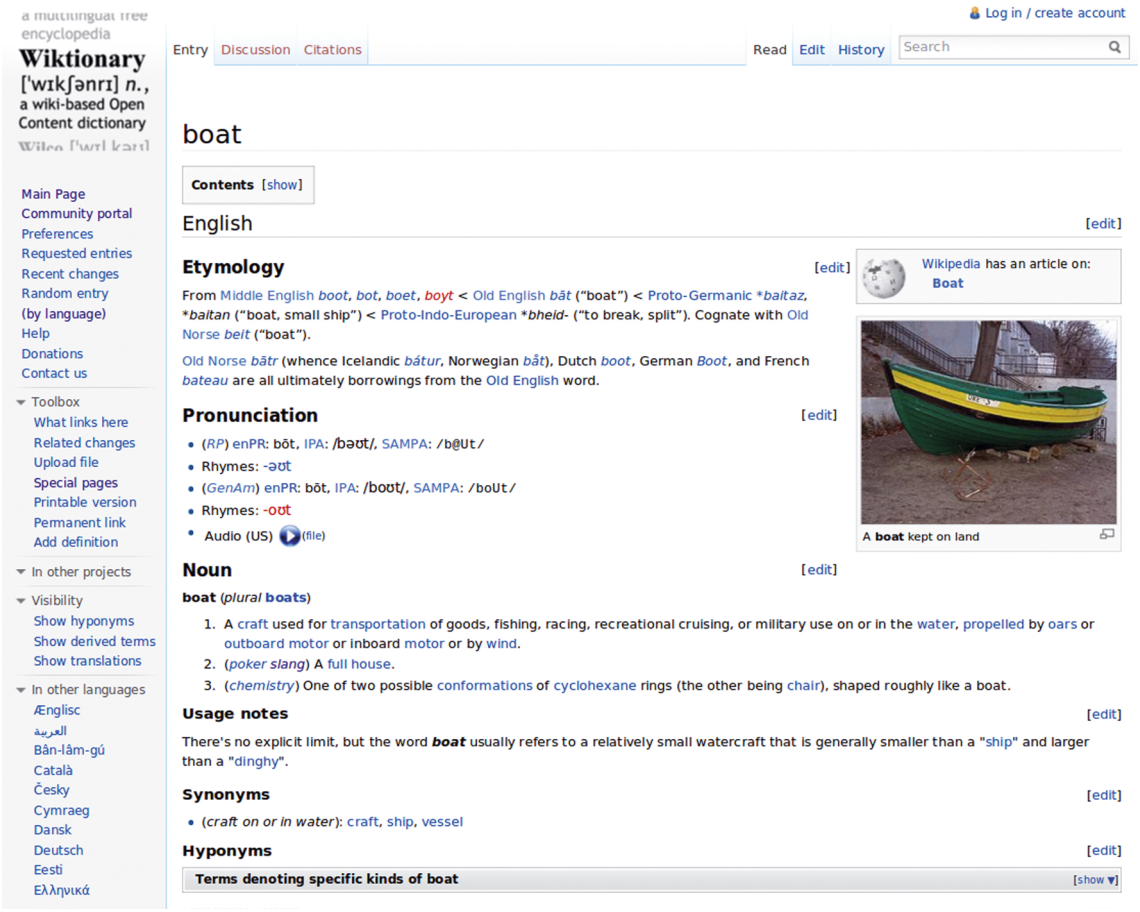



Figure 2.1: The article *boat* in the English Wiktionary (`http://en.wiktionary.org/wiki/boat`; 28 February 2013)

## 2.2 Dictionary Structure

**Language Editions.** Strictly speaking, Wiktionary is not a single dictionary, but a set of dictionaries, which we call *language editions*. Each language edition is associated with a certain *native language*. The native language is being used for the lexicographic descriptions of an edition and for its user interface. There are, for example, separate English, German, and Russian Wiktionary language editions. Figure 2.1 shows the dictionary article *boat* from the English Wiktionary. Each language edition can be accessed through the main Wiktionary web page and through its corresponding subdomain referring to the ISO 639-1 (2002) language code, for example, `http://ru.wiktionary.org` for the Russian edition.

A language edition is, however, not limited to describing words of its native language. It is rather a multilingual dictionary in itself. There is, for example, a dictionary article on the Russian word *лодка* (English: boat) both within the English and the Russian edition (the latter is shown in figure 2.2). The rationale behind this is to provide lexicographic descriptions in

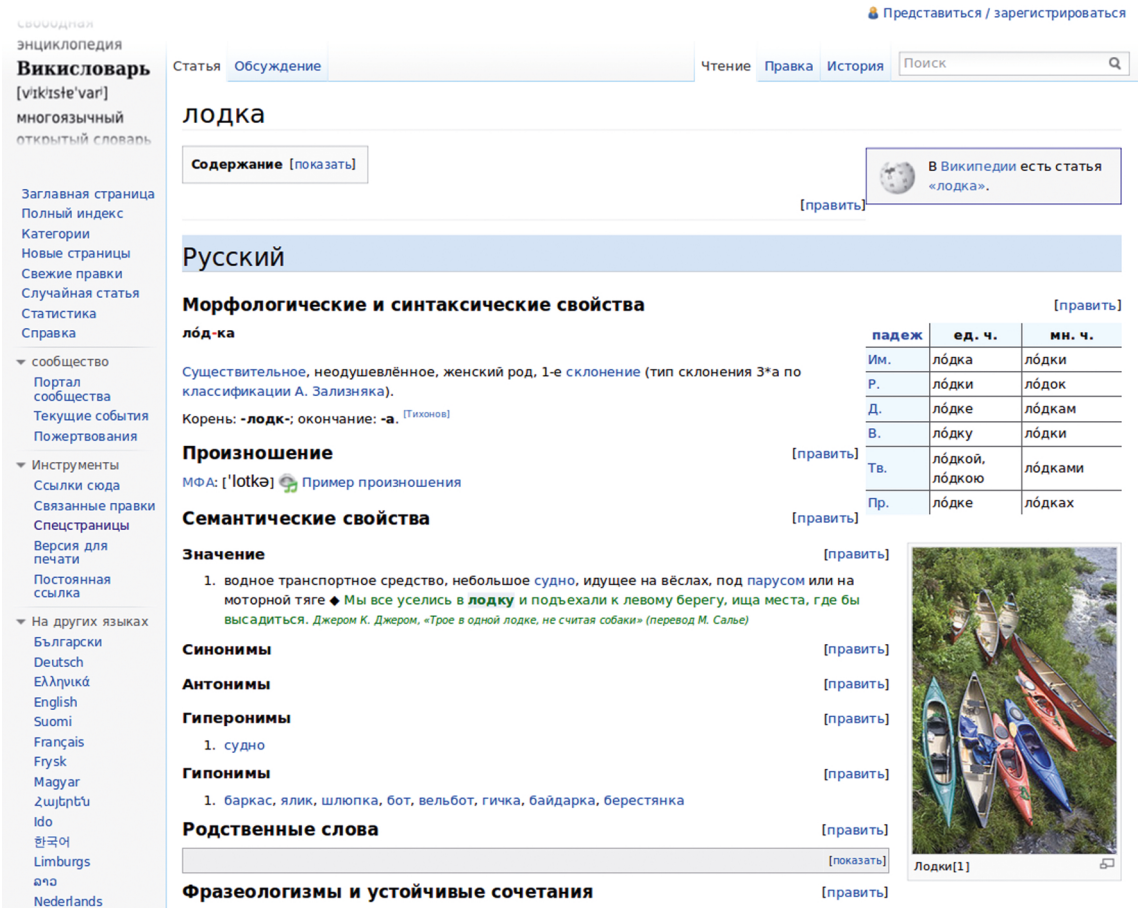



Figure 2.2: The article *лодка* in the Russian Wiktionary (`http://ru.wiktionary.org/wiki/лодка`; 28 February 2013)

different languages: the Russian Wiktionary uses Russian (like in a monolingual dictionary), whereas the article *лодка* in the English Wiktionary is written in English (corresponding to the practice of a bilingual dictionary).[8] This makes Wiktionary useful for both native speakers and language learners. Consider for instance the definition "водное транспортное средство, небольшое судно, идущее на вёслах, под парусом или на моторной тяге" (English: "a water-based means of transport, a small vessel powered by oars, sails, or a motor") for *лодка*. Although *лодка* is a basic vocabulary word, a learner needs to have a certain level of Russian in order to understand this definition. When looking up the word in the language edition of his or her native tongue, say English, the learner finds that *лодка* means "(nautical) boat, dinghy, gig, yawl". The language of the user interface also plays an important role here, since a menu

[8]Note that this distinction has not always been clear in previous work. Fuertes-Olivera (2009), for instance, uses the term "Spanish Wiktionary" to refer to the Spanish entries within the English Wiktionary edition. This has tended to exaggerate the claim that Wiktionary is dominated by the English language. The Spanish language edition has, however, not been considered in their study.

| Page type | English Wiktionary | German Wiktionary | Russian Wiktionary |
|---|---|---|---|
| *Dictionary articles* | | | |
| Article pages | 3,285,810 | 236,584 | 471,852 |
| Supplements | 7,115 | 8,260 | — |
| Appendix pages | 6,834 | — | 488 |
| Thesaurus pages | 1,292 | 105 | — |
| Rhyming dictionary | 8,708 | 244 | 28 |
| Picture dictionary | 25 | — | — |
| Sign language dictionary | 204 | — | — |
| *Lexicographic metatexts* | | | |
| Index pages | 3,729 | 490 | 1,314 |
| Categories | 91,092 | 4,012 | 107,543 |
| Concordance pages | 194 | — | 3 |
| Redirections | 19,896 | 513 | 105,648 |
| Help pages | 413 | 1,068 | 157 |
| Instruction pages | 3,425 | 1,038 | 1,115 |
| Discussion pages | 55,135 | 13,298 | 9,752 |
| User pages | 19,751 | 4,051 | 1,721 |
| *Templates, files, modules* | 47,465 | 3,824 | 22,117 |

Table 2.1: Number of wiki pages for the most important page types

item labeled with *Полный индекс* (English: full index) might not be easily comprehensible for a language learner of Russian. Using the index of the learner's native language edition to browse the Russian entries is much more convenient.

**Wiki pages.** The building blocks of each Wiktionary language edition are *wiki pages*. A wiki page is a document consisting of a formatted text body and a unique title describing the contents of the page. Each wiki page is associated with a so-called *namespace* that denotes the type of the page. Table 2.1 shows the number of wiki pages per page type encoded in the English, German, and Russian Wiktionary editions.

The wiki pages can be divided into *dictionary articles* and *lexicographic metatexts*. The former contain the actual lexicographic descriptions and thus represent the "heart" of the dictionary. We describe the different types of dictionary articles below. The lexicographic metatexts constitute the *outside matter* of the dictionary (i.e., the front and back matter in printed works), including guidelines on how to use the dictionary, lists of abbreviations and irregular verb conjugations, a grammar, etc. In Wiktionary, they are also used to establish the macrostructures (see section 2.3) and to organize the collaborative work (see section 2.7).

**Dictionary articles.** The majority of wiki pages are *article pages*. They contain lexicographic descriptions for a certain *headword* (or *lemma*), similar to *general language dictionaries* such as the *American Heritage Dictionary of the English Language* (2000). The figures 2.1 and 2.2 show two example articles. The title of an article page refers to the headword that is being described. This can be single words (such as the noun *boat* and the verb *sleep*), multiword expressions, and phrases (e.g., the proverb *Rome wasn't built in a day*). The title is case-sensitive and can distinguish diacritic variations. The words *cafe*, *café*, and *Café* are thus described on different article pages. The lexicographic descriptions obey a certain *microstructure* defining the order and type of information found in the articles. We describe this microstructure in section 2.6.

*Supplementary articles* are sometimes created for providing additional information on a certain headword, which does not fit well on the main article page. In the English Wiktionary, the community makes use of this option to document citations of early or illustrative usage of a headword. There is, for example, a citation page for the headword *fractal* referring to several scientific articles containing this word. Supplementary articles are also used to encode extensive inflection tables within the German Wiktionary. While the article page on the verb *fliegen* (English: (to) fly) only lists the most important verb forms, the dictionary user can access a supplementary article showing the conjugated verb forms in each combination of number, person, tense, mood, and voice. There are 430 word forms for *fliegen* encoded in this supplementary article, including, for instance, the first person plural pluperfect *"wir waren geflogen"* and the second person singular simple past forms *"du flögest"* and *"du flögst"*, which are alternative spellings.

**Specialized articles.** In addition to the article pages describing a certain headword (following the practice of a *general dictionary*), there are dictionary articles usually found in *specialized dictionaries* (e.g., rhyming dictionaries or thesauri). *Appendix pages* contain further background information on certain topics, like an overview of the *days of the week* in different cultures. They are used to constitute a *phrase book*, which describes phrases used in certain situations, such as greeting people, managing emergencies, or explaining health problems. The situations described are mostly targeted at traveling. The phrasebook page *communication* encodes, for instance, phrases like "I don't speak English", "how do you say … in English", and "please speak more slowly", which are further described on article pages.

An emerging sub project of the English Wiktionary is *Wikisaurus*, which aims at compiling a wiki-based online *thesaurus*. The goal of Wikisaurus is helping users in finding *related terms* and exploring *semantic fields*. Instead of the word form, the title of a Wikisaurus page denotes a subject that the related words belong to (e.g., *watercraft*). The text body of a Wikisaurus page provides a list of synonymous words and words with an opposite (*antonymy*), broader (*hypernymy*), or more specific meaning (*hyponymy*). In addition to that, words standing in a part–whole (*holonymy* and *meronymy*) or unspecified relation may be listed. The Wikisaurus page for *watercraft* encodes, for example, the synonym *vessel*, the hyponyms *canoe* and *speed-*

*boat*, the hypernym *vehicle* and the related words *captain* and *helm*, for which no specific type of relation is provided. The related words refer to their corresponding article pages and to other thesaurus pages by means of hyperlinks. So far, an active Wikisaurus only exists for the English Wiktionary edition. It encodes about 1,200 pages and is still very sparse. A similar thesaurus project within the German Wiktionary has not gained much interest, since many semantic relations are already encoded as parts of the article pages (see also section 4.7).[9]

Wiktionary's *rhyming dictionary* describes phonetic suffixes such as *-iːðɪŋ*. Each of the corresponding wiki pages encodes hyperlinks to the article pages whose headword has this phonetic suffix (e.g., *breathing*, *seething*, *teething*) and to similar rhyming suffixes (e.g., *ing*-forms of words ending with *-iːð*). The English Wiktionary additionally provides a *picture dictionary*, which organizes the articles graphically. The article *Solar System* displays, for example, an image in which a reader can click on the sun, the planets, a comet, an asteroid belt, etc. to refer to the corresponding article pages. The *sign language dictionary* covers the *American Sign Language* and makes vast use of explanatory images or videos. It can, however, be seen from table 2.1 that these specialized parts of Wiktionary are still of small size.

## 2.3 Macrostructure

A *macrostructure* denotes the ordering of the dictionary articles. The *index pages* in Wiktionary contain ordered lists of headwords linking to their corresponding article pages. Most printed dictionaries are ordered alphabetically. This enables a reader to also find headwords whose spelling she or he is uncertain about. Wiktionary's *full index* displays all headwords of the dictionary sorted by alphabet and thus yields such an *alphabetical macrostructure*. In addition to the full alphabetical index, there are indices grouping the articles by language, part of speech, and topic. These index pages yield a *systematic macrostructure*. That is, the articles are ordered by syntactic or semantic properties.

Grouping the article pages by language, part of speech, or topic is achieved by tagging them with *categories*. The article *boat* is, for instance, tagged with the categories *English nouns*, *English verbs*, *Chemistry*, *Middle English derivations*, *1000 English basic words*, and 13 more. Each category can again be tagged with multiple categories yielding a *taxonomy*. The category *English verbs* has, for instance, the (parent) categories *English parts of speech* and *Verbs by language*, while *en:Watercraft* is subsumed by *Watercraft*. The root of the taxonomy branches to *All languages*, *All parts of speech*, and *All topics*. A reader can browse the entire taxonomy of categories by using hyperlinks to broader and more specific categories as well as an alphabetical list of dictionary articles tagged with this category.

*Concordance pages* list words used in a certain book or document, such as Umberto Eco's *"Foucault's Pendulum"* or the *U. S. Declaration of Independence*. The listed words are hyper-

[9] `http://de.wiktionary.org/w/index.php?oldid=1072170` (21 September 2009)

linked with their corresponding article pages and sometimes associated with short explanations and the page numbers on which the word occurs in the book. While this procedure provides the means for corpus-based lexicography (i.e., compiling the dictionary articles based on systematically excerpted texts; see section 3.4), only a few concordance pages exist for the English Wiktionary and none for the German and the Russian Wiktionary edition.

## 2.4 Mediostructure

The *mediostructure* of a dictionary is defined by *cross-references* between different lexicographic texts – in particular between dictionary articles. The article *elevator* of the *Oxford Student's Dictionary of Current English* (1988) encodes, for example, a cross-reference to the synonymous word *lift* instead of providing a separate definition. The cross-references connect (or *mediate* between) different lexicographic structures and facilitate a non-linear organization – as opposed to the linear ordering induced by macrostructures.

Many electronic dictionaries employ *hypertext* for displaying the articles. That is, a text in which some words or phrases are marked as *hyperlinks* (Conklin, 1987). Similar to a cross-reference in a printed dictionary, a hyperlink is referring to a certain target article. A reader can, however, "follow" the hyperlink by clicking, touching, or using keystrokes in order to display its target text. This spares the effort of finding the target article in a, say, alphabetically ordered dictionary. Wiktionary is based on the wiki technology, which inherently makes use of hypertext and permits adding hyperlinks at practically any part of an article. We can distinguish four types of hyperlinks in Wiktionary: internal, external, and interwiki links, as well as redirections.

**Internal links**. *Internal links* are hyperlinks to other wiki pages. They are used to refer to articles containing additional information, as for the Lemmaelevator–*lift* example introduced above, where there is no dedicated sense definition for *elevator*, but merely an internal link to the descriptions of the article *lift*. In Wiktionary, this type of internal link is often employed for inflected word forms, which refer to the respective canonical form. The article *mice* encodes, for example, only form-related information (such as the pronunciation and grammatical properties), but asks the reader to refer to its canonical form *mouse* for information on the word's meaning and usage. In addition to that, internal links refer to related articles, such as synonyms, opposites, broader terms, etc., as well as to words that appear in the definition text and might require further explanation. The computation-related word sense of *mouse* is, for example, defined as "an input device that is moved over a pad or other flat surface to produce a corresponding movement of a pointer on a graphical display" (underlined words represent internal links). A reader who is unaware of the meaning of, say, *pad* in this context can follow these hyperlinks to open the corresponding dictionary article. Unlike electronic dictionaries that automatically link every word in the definition text, such as the *Macmillan English Dic-*

*tionary Online* (2009 f.), Wiktionary requires the internal links to be manually defined by the community. This allows the authors to decide which expressions they consider relevant for being hyperlinked in the current context – a process commonly called *wikification*.

Based on the typology by Wiegand (2003), we can distinguish (among others) between lemma-oriented, entry-oriented, sense-oriented, and metatext-oriented internal links. Most internal links in Wiktionary are *lemma-oriented* – they refer to an entire article page with the specified word form. Though it is possible to link to specific *lexical entries* (i.e., a certain language or part of speech, see section 2.6), such *entry-oriented internal links* are only rarely used. *Sense-oriented internal links* refer to a specific word sense of the target article. This would be useful for pointing a reader directly to the twenty-first word sense of *pad* (which describes the computer-related meaning used in the definition of *mouse*), rather than to the entire article. In other dictionaries, such as the *Oxford Dictionaries Online* (2010 f.), such links are usually marked by index numbers referring to the target word sense. As of today, sense-oriented internal links are not used in Wiktionary. We will discuss in chapter 5 that this presents a major challenge when using Wiktionary in natural language processing applications. The absence of sense-oriented internal links has two reasons: First, human readers do not require sense-oriented internal links in most cases, because they can easily identify the referred word sense by reading the corresponding sense definitions. The second reason is that the collaborative construction process of Wiktionary is focusing on editing individual articles without the necessity of checking articles connected via hyperlinks. This would cause broken or invalid hyperlinks if a word sense of the target article changes – for instance, if it is split into two more fine-grained descriptions. Missing sense-oriented internal links is not a specific issue of Wiktionary, but also applies to many other dictionaries, such as the *Cambridge Dictionaries Online* (1999 f.) or *Wordnik* (2009 f.). Finally, *metatext-oriented internal links* refer to a lexicographic metatext of Wiktionary, such as the instruction or help pages (see section 2.7).

**External links**. *External links* are hyperlinks to web pages outside of Wiktionary. They are mostly used for providing additional information on a dictionary article or giving evidence about a word's meaning and usage. External links typically point to news articles, digital books, or other online dictionaries. We will analyze them in detail in section 3.4.

**Interwiki links**. *Interwiki links* are hyperlinks between different Wiktionary language editions. We can distinguish between title-based and translation-based interwiki links. The former are displayed in the lower left corner as part of the user interface. They link to an article page with the same title that is encoded in another language edition. The article page *boat* of the English Wiktionary edition encodes, for example, a *title-based interwiki link* to the article page *boat* in the Russian Wiktionary. Note that this is different to *Wikipedia*, where such interwiki links are used to refer to translations of an article. Instead of the Russian translation of *boat*, the reader finds a Russian description of the English word *boat* when following this

interwiki link (cf. section 2.2). The reason is that there can be only a single title-based interwiki link per language, which would impose encoding multiple translations for a language (such as the Italian words *barca* and *battello* for *boat*). Translations are rather expressed by *translation-based interwiki links*, which are encoded as a part of the article microstructure. The English Wiktionary's article *boat* refers, for example, to the article *лодка* in the Russian Wiktionary and to the articles *barca* and *battello* in the Italian Wiktionary by means of translation-based interwiki links. Both title-based and translation-based interwiki links are *lemma-oriented* as described above for internal links. Translations are therefore not sense-disambiguated, which we will further discuss in chapter 5.

**Redirections**. *Redirections* are wiki pages solely consisting of an internal link that automatically navigates a reader to the target text of the hyperlink (i.e., the reader does not need to follow the hyperlink manually). Many redirections are used for instruction pages (see section 2.7) to become able to access them quickly. The wiki page *Wiktionary:RP* redirects, for instance, to *Wiktionary:Request Pages*. Other applications of redirections include typographic variations or capitalizations. The article *you've* (Unicode character U+2019) redirects, for example, to the article *you've* (Unicode character U+0027) without typographic apostrophe. Another example is *pdf* which automatically redirects to *PDF*.

It is interesting to note that redirections are usually not used as cross-references between synonyms, as it is the case in Wikipedia. They are neither used for inflected word forms (such as *forgot*) nor common misspellings (like *lier* instead of *liar*), which would be obvious applications for redirections, too. In Wiktionary, such cases rather yield separate dictionary articles referring to the synonyms, canonic forms, or correct spellings by internal links. The rationale behind this is to provide additional descriptions (e.g., explaining grammatical properties like "simple past of forget" for *forgot*) and to facilitate entries sharing the same word form. The adjective *parked* could, for example, not be described in case of a redirection of the simple past verb form to the lemma *park*. Likewise, the Swedish noun *faster* (English: paternal aunt), could not be described if there is a redirection for the comparative of the English adjective *fast*. An exception is Cyrillic Wiktionary editions, in which redirections are sometimes used for encoding inflected word forms. The Russian plural form *красные* redirects, for example, to *красный* (English: red). Note, however, that this is only possible if there are no entries on the same word form (in any language).

## 2.5 Access Paths

Bergenholtz and Gouws (2010, p. 103) argue that it is of "critical importance [...] that the target users of a specific dictionary gain unimpeded access to the data they need". They stress the requirement of user-friendly and quick-to-use *access paths* (or *access processes*). The most obvious access path in printed dictionaries is using the alphabetical macrostructure. Having

a word in mind, a user browses through the pages of the dictionary and finds the required information at the corresponding position in the alphabet. Other dictionaries (in particular specialized ones) are organized by semantic fields (thesaurus), language (multilingual dictionaries), or similar properties, which can be accessed using a subject index, a table of contents, etc. (see Wiegand, 1989). Besides browsing the list of articles, electronic dictionaries make it possible to search articles by headword, full article text, or specific article constituents, and to refer from one article to another by means of hyperlinks Storrer and Freese (1996) discuss dictionaries with access paths based on *browsing*, following hyperlinks (i.e., *hypertext-based*), using a retrieval engine (i.e., *search-based*) and combinations thereof. Sérasset (1993) distinguishes *direct access* (i.e., providing a certain headword) from *indirect access* via hyperlinks or lexical-semantic networks. Wiktionary provides access paths from all of these groups:

(1) *Direct access*: Each article page can be directly accessed by its URL. In addition to jumping quickly to an article, this makes it possible to bookmark or link to an article. The URL `http://en.wiktionary.org/wiki/boat` refers, for example, to the English Wiktionary's article *boat*.

(2) *Indirect access based on browsing*: The index, category, appendix, and thesaurus pages organize the dictionary articles by alphabet, language, part of speech, semantic field, etc. (see section 2.3), which enables a browsing-based access to the articles. Using the picture and the rhyming dictionary, a user can also browse the articles visually and by phonetic suffix. While the index pages provide a *semasiological perspective* on the dictionary (i.e., a user has a word in mind and wants to find out what it means), the thesaurus yields an *onomasiological perspective* (i.e., a user has a certain concept or idea in mind and explores the words expressing this concept).

(3) *Indirect access based on retrieval*: Wiktionary articles can be accessed using internet search engines. In addition to that, an internal search engine may be used, which has the advantage that the search is limited to the contents of Wiktionary. This search engine directly refers to an article page if the specified search term matches exactly with the title of a wiki page. Otherwise, a list of search results is displayed, which orders the retrieved articles by relevance. The search is based on the entire article text and not case-sensitive. The search terms are not automatically lemmatized, since inflected word forms are usually encoded as separate articles. The asterisk symbol '*' can be used to perform left- or right-truncated searches (i.e., searching for all headwords beginning or ending with the specified search term). More complex search constraints, such as restricting the search by part of speech, are not possible.

(4) *Indirect access based on hypertext*: As discussed in section 2.4, dictionary articles are connected via internal links. A reader can thus access related articles by following these hyperlinks.

## 2.6 Microstructure

An article page consists of multiple *article constituents*, which we define as composites of one or more information items formatted in a certain way. An *information item* is a lexicographic description of a property of the language (or, in a broader sense, of the *subject matter of the dictionary*), for example, the part of speech "noun" or the phonetic transcription "bəʊt" of *boat*. Each information item is of one particular *information type* (e.g., part of speech, phonetic transcription). The article constituent describing the pronunciation of *boat* shown in figure 2.1 encodes, for example, nine information items: six of them are of the type phonetic transcription (describing the British and American pronunciation in three different phonetic alphabets), two are rhyming suffixes, and one is an audio file of a speaker pronouncing the word. The format of an article constituent is defined by typographic variations (such as indentation, font size, weight, or style) and by explanatory texts and symbols. Most constituents are, for instance, marked by a headline, which denotes the information type(s) being addressed by this constituent. The constituent describing the pronunciation begins, for example, with the headline "Pronunciation". These headlines are used to organize the constituents in a hierarchy. The choice of article constituents, their format and position within the article text defines the *microstructure* of the article pages.[10]

Table 2.2 shows a schematic overview of the microstructure of article pages in the English, German, and Russian Wiktionary editions. As noted in Wiktionary's guidelines, the structure is not a "set of rigid rules", because the authors are free to "experiment with deviations".[11] The constituents describing the etymology and the pronunciation are, for instance, often encoded in reversed order in the English Wiktionary, and not every article contains all constituents. It can be seen from the table that there are differences between the language editions. This includes differences in

- the selection of the constituents: the German Wiktionary is, for instance, the only one with a separate constituent for dialectal expressions,
- the order of the constituents: a description of the etymology is encoded near the top (English Wiktionary), in the middle (German Wiktionary) or near the bottom (Russian Wiktionary) of an article,
- the hierarchy of the constituents: the English and Russian editions combine sense definition and example sentences per word sense, while the German Wiktionary encodes two separate constituents for sense definitions and usage examples,

[10] Our definitions in this section follow pragmatic reasons. Wiegand (2006) and Wiegand et al. (2010) differentiate, for example, between *items*, *item texts*, and *non-typographical microstructural indicators* as parts of the article constituents. They also introduce many other terms relevant for describing the dictionary articles, which are, however, not necessary for the understanding of this thesis and therefore not further distinguished. Our definition of information type is particularly targeted at the following chapters and serves as a bridge between the metalexicographic theories and the natural language processing terminology.

[11] `http://en.wiktionary.org/w/index.php?oldid=19130446` (26 December 2012)

– the format of the constituents: the English Wiktionary, for example, displays translation equivalents in separate tables for each word sense; the German Wiktionary uses only a single table and associates the individual translations with a certain word sense.

Some of the differences are due to culture- or language-specific peculiarities, for example, when using a different script (like Chinese) or for encoding references between male and female job titles, such as *Mechanikerin* and *Mechaniker* (English: mechanic) in the German Wiktionary, which are not distinguished in English. Other differences evolved over time within the language-specific Wiktionary communities. In the remaining section, we will describe the most important article constituents.

**Lexical entry and language**. At the uppermost level, an article page is divided into multiple *lexical entries* (or entries for short), which are characterized by the language, the etymology, and the part of speech of the headword they describe. The article page *boat* encodes, for example, six lexical entries: The English noun and verb, the Finnish and West Frisian nouns, and the Latin and Malay verbs. The lexical entries are alphabetically ordered by language name, which is defined by their first article constituent. An exception are lexical entries for the edition's native language, which are always described first. Those entries are usually the most detailed ones and they are expected to be looked up most frequently. Information that cannot be associated with a certain language, such as the letters of an alphabet, internationally used abbreviations (e.g., chemical symbols or the ISO language codes), and the scientific names of the biological taxonomy are encoded in a separate entry entitled "Translingual".

**Etymology**. The *etymology* describes the origin of a word (e.g., "from Middle English boot, bot, boet, boyd, from Old English bāt [...]" for the English *boat*). The Wiktionary community uses the etymology constituent to distinguish *homonyms* (i.e., words with the same written form, but different origin). The English noun *bass* distinguishes, for instance, two homonymous meanings originating from the Latin *bassus* for its musical meaning and from the Proto-Indo-European **bhors-* for its biological meaning. It is surprising to find such a distinction, because homonymy "is gradually being abandoned as an organizing principle in many types of dictionary" (Atkins and Rundell, 2008, p. 281). The reason is that it might confuse a reader when looking up a word without knowing its etymology (see also Moon, 1987). It has indeed been discussed for a long time, how the dictionary articles should be tailored to fit the article pages.[12] An early idea was to create a separate article page for each word sense. This suggestion was, however, abandoned in 2003 because the different senses could not be easily compared. The same applies to describing homonyms on separate article pages, which led to the currently used distinction of homonymous entries on a single article page. By 2006, it has been proposed to list all word senses without distinguishing homonyms, which was,

[12]See `http://en.wiktionary.org/wiki/Wiktionary_talk:Entry_layout_explained/` and the corresponding archive pages for a full discussion on this topic.

| English Wiktionary | German Wiktionary | Russian Wiktionary |
|---|---|---|
| Language | Language | Language |
| ▷ Alternative forms | ▷ Part of speech | ▷ Entry index |
| ▷ Etymology | ▶ Inflection table | ▶ Inflection table |
| ▶ Pronunciation | ▶ Illustration* | ▶ Illustration* |
| ▶ Part of speech | ▶ Hyphenation | ▶ Morphology |
| ▶ Illustration* | ▶ Pronunciation | ▶ Pronunciation |
| ▷ Inflection | ▶ Definitions* | ▶ Semantics |
| ▷ Definitions* | ▶ Abbreviations | ▷ Definitions* |
| ▷ Examples* | ▶ Etymology* | ▷ Examples* |
| ▷ Quotations* | ▶ Synonyms* | ▷ Synonyms* |
| ▶ Usage notes | ▶ Antonyms* | ▷ Antonyms* |
| ▶ Synonyms* | ▶ Hypernyms* | ▷ Hypernyms* |
| ▶ Antonyms* | ▶ Hyponyms* | ▷ Hyponyms* |
| ▶ Hypernyms* | ▶ Examples* | ▷ Meronyms* |
| ▶ Hyponyms* | ▶ Phrases* | ▷ Derived terms |
| ▶ Coordinate terms* | ▶ Collocations* | ▷ Collocations |
| ▶ Derived terms* | ▶ Derived terms* | ▷ Phrases |
| ▶ Related terms* | ▶ Translations* | ▶ Related terms* |
| ▶ Translations* | ▶ Dialects* | ▶ Anagrams |
| ▶ Descendants* | ▶ References* | ▶ Etymology |
| ▶ References | ▷ Similar words | ▶ Translations* |
| ▶ See also | | ▶ References |
| ▶ External links | | ▶ See also |
| ▷ Anagrams | | |

Table 2.2: Schema of the article microstructure found in the English, German, and Russian Wiktionary editions. Indention indicates a hierarchical relation between two constituents. Constituents marked with an asterisk (∗) can be associated with a sense marker.

however, rejected by the community, as etymologies are seen to play an important role in the lexicographic descriptions.

**Phonetic information**. The *pronunciation* of a word is often described using multiple *phonetic transcriptions*. They are represented using the *International Phonetic Alphabet* (IPA), the *Speech Assessment Methods Phonetic Alphabet* (SAMPA), and the notation of the *American Heritage Dictionary of the English Language* (1969). Different geographical variants can be distinguished using labels such as "Received Pronunciation" (Standard English of England), "General American", "Standard German", "Swiss German", etc. In addition to a transcription, *audio files* are frequently added in order to help the dictionary users with the correct pronunciation.

The pronunciation constituent may also contain a rhyming suffix, which refers to the rhyming dictionary discussed in section 2.2.

**Morphological information**. Information on the *conjugation* and *declension* of a headword is either encoded as a table listing the inflected word forms for each combination of person, number, tense, aspect, etc. (see figure 2.2) or as textual items – for example, "third-person singular simple present does, present participle doing, simple past did, past participle done" (underlined words indicate internal links) for the English verb *(to) do*. The formation of *compound words* is often described within the etymology constituent, for example, "Determinativkompositum aus den Substantiven Boot und Steg mit dem Fugenelement -s" for the German *Bootssteg* (English: landing stage; "determinative compound of the nouns *Boot* and *Steg* and the linking element -s"). A reader can access the members of the compound using the lemma-oriented internal links to the corresponding article pages. In addition to that, there are metatext-oriented internal links explaining determinative compounds and linking elements as part of Wiktionary's help pages (see section 2.7). Similarly, the *derivation* of a word is described within the etymology constituent by providing the basis and the affixes (e.g., "red + -ish" for the English *reddish*).

**Grammatical information**. Each lexical entry is marked with a *part of speech*. For single words, common categories such as "noun", "verb", "adjective", etc. are used and multiword expressions are described as "compound", "proverb", "idiom", etc. The part of speech tags used in Wiktionary are further analyzed in section 4.4. *Mass nouns* and *count nouns* are identified using the labels "countable" and "uncountable". Likewise, the *transitivity* of verbs is expressed with labels such as "transitive" and "intransitive". Additions like "may take two objects" are sometimes describing a verb's *valency*. Wiktionary does not encode deep lexical-syntactic knowledge, such as *subcategorization frames*.

**Semantic information**. Similar to other dictionaries, Wiktionary distinguishes between multiple *word senses* for describing the *meaning* of a word. For each word sense, the Wiktionary authors provide a *sense definition* (i.e., a *paraphrase of its meaning*). In addition to that, pragmatic labels, usage notes, semantic relations, and illustrations may be encoded. The third word sense of *boat* is, for instance, described as "(chemistry) One of two possible conformers of cyclohexane rings (the other being chair), shaped roughly like a boat" (underlined words are again internal links). The label "chemistry" is a *pragmatic label* describing the technical domain this word sense is used in. Particularities of using a certain words can be described in a separate article constituent "usage notes". For *boat*, the usage notes explain the preference of using *boat* for small vessels: "There's no explicit limit, but the word boat usually refers to a relatively small watercraft that is generally smaller than a 'ship' and larger than a 'dinghy"'.

A number of article constituents can be associated with a certain word sense (marked by a '∗' in table 2.2). The article constituents encoding *semantic relations* consist, for instance, of a list of lemma-oriented internal links to related words for a given word sense. They are titled with the *relation type* (e.g., synonymy, antonymy, hypernymy, hyponymy). Other sense-specific article constituents describe the context a word sense is usually used in, including *example sentences*, *quotations*, *phrases*, and *collocations*. We analyze the encoded semantic information in detail within chapter 4.

**Cross-lingual information**. The translation constituent contains *equivalents* of a word in other languages. Due to the multilingual nature of Wiktionary discussed in section 2.2, each equivalent can refer to the translated word within the same Wiktionary language edition and to the edition of the target language. For the German translation *Boot* of the noun *boat* within the English Wiktionary, there is, for instance, an internal link to *Boot* in the English Wiktionary and a translation-based interwiki-link to *Boot* in the German Wiktionary. For each word sense and target language, multiple translations can be encoded.

**References**. *References* are sometimes provided as evidence for the described information items. They should ensure the correctness of the descriptions made and raise their credibility. The article *boat* contains, for instance, the reference "Weisenberg, Michael (2000): The Official Dictionary of Poker. MGI/Mike Caro University. ISBN 978-1880069523", which is used verify its poker-related word sense. Besides references to published books and articles, there are often references to online sources. In section 3.4, we will analyze Wiktionary's references in detail.

## 2.7 Community and Outside Matter

**Outside matter**. The outside matter of a dictionary consists of lexicographic metatexts, such as usage guidelines on how to search for articles and how to understand the lexicographic descriptions. In Wiktionary, there are separate namespaces for help pages and instruction pages. *Help pages* contain explanations on the lexicographic descriptions including the part of speech tags, the definition of a certain type of semantic relation, and how translations are encoded. In addition to that, they provide technical information on the use of the Wiktionary user interface addressing, for example, the search options.

Most of the help pages are, however, targeted at authors rather than readers. They document the way a new article is created and how existing articles can be modified. This also applies to the explanations of lexicographic descriptions: the help page on translations focuses, for example, on inserting and modifying the encoded translations over describing the information items and their format. The usage guidelines of the dictionary therefore coincide

with the lexicographic instructions.[13] These *instruction pages* describe the goals of Wiktionary, its organization plan and dictionary conception, as well as the lexicographic instructions for new and experienced authors. The majority of the instructions address the article style sheet (i.e., the definition of the article microstructure). The set of instruction pages is also called the *community portal* of Wiktionary, which we analyze in more detail within section 3.3.

**Wiktionary users**. The *Wiktionary community* (i.e., the users of Wiktionary) can roughly be divided into *authors* and *readers*. Like for most dictionaries, it is not exactly clear, how many readers Wiktionary has. There have been about 262 million page views during January 2013 over all language editions (the majority of 96 million in the English edition), but naturally there are users looking up multiple articles or ending up at Wiktionary by accident without actually reading a single article.[14]

While only a few authors contribute to editorial dictionaries, there are many authors contributing to Wiktionary. Any reader can easily become an author, because the modification and creation of articles is not limited to a predefined group of people. The idea of Wiktionary is rather to foster the collaboration of many authors. We can distinguish four types of authors:

(1) *Administrators* constitute the smallest group of authors. They have the permission to delete pages, change user rights, inhibit the modification of certain articles, and block users from making further contributions. Administrators must be nominated and elected by the Wiktionary community. There are 98 administrators in the English Wiktionary, 26 in the German Wiktionary, and 8 in the Russian Wiktionary.
(2) *Registered users* are authors, who have created a personal account. This allows them to sign modifications and comments with their name and, for example, make use of the *watchlist* for keeping track of certain articles. There are currently 1,029,698 registered users for the English Wiktionary, 72,079 for the German, and 95,751 for the Russian edition.[15] However, only a fraction of them is actively contributing to the project: 7,025 (English Wiktionary), 1,424 (German Wiktionary), and 618 (Russian Wiktionary) of them performed at least ten edits.
(3) A third type of author is *unregistered users*. Unregistered users perform about 5 % of the modifications to the dictionary. Their edits are anonymous and solely distinguishable by their *IP address*, which can be shared by many web users. Hence, it is not possible to determine the exact number of distinct unregistered users.
(4) *Bots* are computer programs that automatically crawl through the article pages and perform changes according to the patterns and rules specified by their developers. Although

[13] This is why the numbers reported in table 2.1 can only serve as an approximation. They are based on the namespaces "Help" and "MediaWiki" (help pages) and "Wiktionary" (instruction pages).

[14] `http://stats.wikimedia.org/wiktionary/EN/TablesPageViewsMonthly.htm` (4 February 2013)

[15] `http://en.wiktionary.org/wiki/Special:Statistics`, `http://de.wiktionary.org/wiki/Special:Statistics`, `http://ru.wiktionary.org/wiki/Special:Statistics` (28 March 2013)

bots are not authors in the narrower sense, they are programmed and maintained by the Wiktionary authors. Bots have different responsibilities, including automatic data imports, reformatting sections, finding interwiki links, etc. Currently, there are 49 bots in the English Wiktionary, 39 bots in the German Wiktionary, and 30 bots in the Russian Wiktionary. We analyze their impact on the lexicographic process in section 3.5.

*User pages* are wiki pages that can be created by any registered Wiktionary user. They serve as a space for a personal profile and for organizing one's own contributions. They also provide access to the list of all changes by a certain user and a personal discussion page (see below).

**Means of collaboration**. The collaborative creation process is based on the individual contributions of the authors and on their discussion. This is why *discussion pages* (or *talk pages*) are available for any other wiki page, including article, instruction, and user pages. They are used for discussing the lexicographic descriptions, collecting ideas, expressing criticism, asking questions, and organizing the work. Besides mere discussion, a consensus can also be reached by voting, for example, on deleting an article or on a proposed change to the lexicographic instructions.

Every edit operation within Wiktionary is recorded and archived in a *revision history*. In this way, a previous version of an article can be returned to at any time in order to inspect how the article has evolved, or which user made a certain change. Wiktionary authors often use the revision history to revert vandalism (i.e., changes that introduced spam or deleted important parts of an article). Together with the discussion pages, the revision history lets us study the lexicographic process of Wiktionary, since all decisions on an article become transparent and trackable. This kind of information is usually either undocumented or kept private for editorial dictionaries. The revision history and the discussion pages also constitute the basis for our analysis of the collaborative construction approach described in chapter 3.

## 2.8 Chapter Summary

In this chapter, we introduced Wiktionary and its main structures. We found that Wiktionary is divided into multiple language editions, each of them describing words from multiple languages. We have described different access paths the dictionary users can take to fulfill their information needs and also commented on the article microstructure consisting of a hierarchy of article constituents and information items. Finally, we explained the different types of users participating in Wiktionary, and we took a look at the outside matter of the dictionary. While this chapter is focused on *describing* Wiktionary, we analyze several of the points raised here in the following chapters of the thesis.

Chapter 3

# Collaborative Lexicography in Wiktionary

This chapter aims at analyzing the collaborative construction process of Wiktionary. We give an overview of our motivation and research goals for this metalexicographic study in section 3.1 and discuss related work in section 3.2. Then, we examine the dictionary conception and the innovative features of Wiktionary (section 3.3). In a more detailed analysis of the English and German Wiktionaries, we study the impact of lexicographic evidence and quality assurance measures (section 3.4), as well as the coordination and cooperation of the Wiktionary authors (section 3.5). We conclude the chapter by proposing a novel description of the collaborative lexicographic process (section 3.6) and summarizing our findings (section 3.7).

## 3.1 Motivation

Since the mid 1990s, an increasing number of collaborative dictionaries become available on the Web and their collaborative compilation process has established as a new paradigm of lexicography. The success of collaboratively built resources is backed by the phenomenon of collective intelligence – often referred to as the "wisdom of crowds" (Surowiecki, 2005) – and thus the consolidation of the subjective opinions of many authors. However, there has not been much research on collaborative lexicography until now. Wiegand et al. (2010, p. 138) note that "the lexicographical processes as one can observe them in the formation of collaboratively produced online dictionaries such as the 'Wiktionary' cannot be adequately described by the traditional classification of phases; the steps they use have only been researched cursorily up to now".

The research questions on how the authors coordinate the lexicographic work and which impact collaborative lexicography has on editorial dictionaries are highly relevant for the compilation of new dictionaries and for future research in the area of lexicography. Since the

collaboratively compiled *Wikipedia* (2001 f.) is about to replace expert-built reference works like the *Encyclopædia Britannica* (1985), we can even more radically ask if collaborative dictionaries will replace expert-compiled ones. Recent news articles reporting the end of the print version of the *Macmillan English Dictionary*, the fundamental changes at the German publisher *Duden*, the unclear future of the *Deutsches Wörterbuch*, and the winding up of the *Brockhaus* encyclopedia mark a turning point in lexicography and emphasize the importance of our research questions in the context of collaborative lexicography.[16] We discuss the following contributions towards analyzing the collaborative construction process of Wiktionary:

**Contribution 2.1**: *We classify Wiktionary based on existing dictionary typologies and critically assess its innovative features (section 3.3).*

**Contribution 2.2**: *We discuss the quality of the lexicographic descriptions in Wiktionary, the quality assurance measures taken, and the sources used to provide evidence (section 3.4).*

**Contribution 2.3**: *We analyze the cooperation and coordination of the voluntary authors based on the revision history and discussion pages of Wiktionary (section 3.5).*

**Contribution 2.4**: *We propose a novel description of the lexicographic process in collaborative dictionaries (section 3.6).*

## 3.2 Related Work

**Collaborative lexicography**. Storrer and Freese (1996) and Carr (1997) have been among the first to describe collaborative lexicography. Carr (1997, p. 214) defines the notion of *bottom-up lexicography* in which dictionaries are "evolving upward from readers" – as opposed to *top-down lexicography* "from editors, through publishers, to readers". He also raises the question of quality in collaborative works and discusses an example dictionary of low quality and one of high quality. Storrer (1998, 2013) discusses different forms of user participation including the contribution of error corrections and headword suggestions, as well as entire dictionary articles by domain specialists or laypeople. Further works, for example, by de Schryver and Prinsloo (2000) and Køhler Simonsen (2005), strengthen the need for involving the users in the lexicographic process, but do not take account of user contributions to collaborative dictionaries. Instead, they focus on different types of feedback and supplementary material that the users can submit to editors.

Mann (2010) examines 88 online dictionaries for various criteria, such as access paths and usability. He finds that 23 % allow for creating new articles and 14 % for modifying articles. In addition to that, he investigates the risk of diminishing the motivation of the participating

[16] http://www.macmillandictionaryblog.com/bye-print-dictionary (5 November 2012), http://www.sz-online.de/nachrichten/vom-internet-ueberrollt-der-niedergang-des-duden-verlages-1631057.html (21 October 2012), http://www.welt.de/print/die_welt/kultur/article115384288/Das-boese-Ende-des-Grimm.html (18 April 2013), http://www.sueddeutsche.de/kultur/brockhaus-am-ende-wissen-das-nie-am-rechten-ort-ist-1.1695116-2 (12 June 2013)

users if their modifications are controlled by editors. As opposed to *community-driven* online dictionaries such as Wiktionary, Lew (2011) defines *collaborative-institutional dictionaries*, which are collaboratively built but maintained by commercial publishers, for example, the *Macmillan Open Dictionary* (2009 f.) or the *Merriam-Webster Open Dictionary* (2005 f.).

**Collaborative dictionaries.** More detailed analyses focus on specific collaborative projects. A particularly large strand of research is dedicated to the processes of the online encyclopedia *Wikipedia* (2001 f.). This includes the study of collaborative writing (Emigh and Herring, 2005), information quality (Wilkinson and Huberman, 2007; Stvilia et al., 2008), social structure (Stegbauer, 2009) and conflicts (Viégas et al., 2004; Kittur et al., 2007). Ferschke et al. (2013) give a comprehensive survey on the collaborative writing process in Wikipedia. While these works address the collaborative compilation of encyclopedic articles, it remains to be seen how well this applies to collaborative dictionaries.

Naber (2005) introduces *OpenThesaurus* (2004 f.), a collaboratively built synonym dictionary. He studies the number of edits and finds that voting on recently added entries did not work well, because not enough users contributed to the polls. This is why OpenThesaurus relies on editorial control of the additions and modifications contributed by the users. Melchior (2012) calls such works *semi-collaborative dictionaries*. In a detailed analysis of the *LEO Deutsch–Italienisch* (2008 f.) dictionary, Melchior categorizes the different types of users contributing to this project. The study shows that there are conflicting opinions regarding the content of the dictionary and its presentation. Some users argue, for example, in favor of adding newly coined terms even though they might be used only for a very short period of time. This conflicts with other users complaining about confusing and overfull search results.

Penta (2011) finds that collaborative dictionaries are advantageous in providing synchronic information, since they can be modified at practically any time. Instead, he observes more accurate and more comprehensive diachronic information in editorial dictionaries.

**Wiktionary.** Descy (2006, p. 4) introduces Wiktionary as a "neat", easy-to-use dictionary and gives a brief overview of the wiki project. Abel (2006) studies the article style sheet and reviews some example entries on basic vocabulary words. She mentions the citation of sources and raises the question of quality by exemplifying a lax definition of collocations found in the German Wiktionary. In her conclusion, Abel argues in favor of editorially controlled collaborative dictionaries, but notes – in accordance to Storrer (1998) – that it is too early to assess the prospects of collaborative lexicography.

Meijssen (2009) analyzes the multilingual *OmegaWiki* (2006 f.), an offshoot of Wiktionary, and compares the structures used by this dictionary with those of Wiktionary. The initiators of OmegaWiki aim at creating a collaborative dictionary with a more rigidly structured view on the dictionary data in order to overcome Wiktionary's language-specific differences outlined in section 2.6. The more detailed comparison of OmegaWiki and Wiktionary by Matuschek

and Gurevych (2011) with a particular focus on multilingualism shows that the definition of a fixed (OmegaWiki) or more expressive microstructure (Wiktionary) has a large impact on the size, usage, and extensibility of a collaborative dictionary. They propose a combination between these two approaches.

Fuertes-Olivera (2009) critically assesses the usefulness of Wiktionary for Spanish users, who aim at understanding and translating English texts on business-related topics. Hanks (2012) analyzes some selected Wiktionary articles and comments on the low quality of the lexicographic descriptions and the absence of corpus evidence. Rundell (2012) sees most potential in the cooperation of lexicographers, translators, and subject-specialists. As a result of increasing collaboration, he predicts the possible demise of considering the dictionary as an "authority" (ibid., p. 81). In section 3.4, we will discuss those three works in detail.

Our work goes beyond the discussed previous works by assessing the collaborative lexicographic process and the sources of lexicographic evidence used in Wiktionary. For the first time, we analyze the revision history of Wiktionary and shed light on the coordination and cooperation of the Wiktionary authors.

## 3.3 Dictionary Conception

**Dictionary typology.** Having described the structure of Wiktionary in the previous chapter, we will now assess how well Wiktionary can be described with existing dictionary typologies and in which ways its conception is innovative. The starting page of the English Wiktionary mentions:

> Welcome to the English-language Wiktionary, a collaborative project to produce a free-content multilingual dictionary. It aims to describe all words of all languages using definitions and descriptions in English.[17]

Thus, Wiktionary is a *general language dictionary*, whose *subject matter* addresses all individual languages. It is a *collaboratively compiled* rather than an *editorial dictionary*. Based on the notation of Wiegand et al. (2010), we can further classify Wiktionary as a *polyinformative* (describing multiple language-related properties), *polyselective* (not restricted to a certain group of headwords), and *polyaccessive dictionary* (providing multiple access paths). In addition to the *semasiologic access* (i.e., from word to meaning) using the index pages or the internal search engine, the thesaurus pages provide *onomasiologic access* to the lexicographic descriptions (i.e., from meaning to word). The dictionary articles address mainly present-day language. Besides the *standard variety* of language, we find dialectal and technical terms, as well as articles on archaic expressions. To verify the encoded descriptions, the authors are asked to provide *evidence* in form of references and citations:

[17] `https://en.wiktionary.org/w/index.php?oldid=19402071` (23 January 2013)

> A term should be included if it's likely that someone would run across it and want to know what it means. This in turn leads to the somewhat more formal guideline of including a term if it is attested [...] verified through
> - clearly widespread use,
> - use in a well-known work, or
> - use in permanently recorded media, conveying meaning, in at least three independent instances spanning at least a year[18]

Wiktionary is an *expandable dictionary*. That is, its dictionary articles are not finalized, but can be updated at any time. Furthermore, Wiktionary can be described as an *electronic dictionary* that is being published online and as a *multimedia dictionary*, since its dictionary articles make use of text, sound, and images.

In all these dimensions of existing dictionary typologies, Wiktionary can be clearly categorized. The intended user relationship and the dictionary functions are, however, not specified. Defining the type of *user relationship* is considered crucial for the conception of modern dictionaries, as it allows to compile the dictionary according to the specific needs of its users (e.g., using simple language for the definitions in a learner dictionary). We discuss the different types of users that come into question in section 4.8. Regarding the *dictionary function*, Wiktionary provides the means for *reception* (understanding text), *production* (formulating texts), *active translation* (translating into another language), and *passive translation* (translating from another language), because of its broad variety of structures and information types. The dictionary functions are, however, not explicitly mentioned.

**Innovative features**. In his influential article "Lexicographers' Dreams in the Electronic-Dictionary Age", de Schryver (2003) scrutinizes 118 visions about future dictionaries uttered in the scientific literature. We discuss which of these visions are being achieved in Wiktionary. Like many other electronic dictionaries, Wiktionary benefits from the absence of size restrictions and aims at including much additional information that would have been omitted from most printed dictionaries (dreams #1–7). Automatically compiled information (e.g., concordances) is, however, not included in the dictionary (#2). A unique feature of Wiktionary is its large variety of language editions which merge information that is traditionally found in monolingual and bilingual dictionaries (#8–9). In addition to that, the collaborative construction approach provides the ability that changes in the language are quickly included in the dictionary (#12).

Wiktionary makes vast use of multimedia in the form of illustrative images (#13) and audio files on how to pronounce a word including different geographical variants (#20). As Hanks (2012) notes, Wiktionary does usually not include video clips or audio files for describing the meaning of a word, for instance, the sound of a trumpeting elephant or a video of its typical behavior (#14–15, #21). The use of multimedia is up to the standards of other online dictionaries, such as *Duden online* (2011 f.), but does not go beyond them.

[18] `https://en.wiktionary.org/w/index.php?oldid=19569878` (15 February 2013)

A frequently discussed topic in the context of electronic dictionaries is the connection of the dictionary with text corpora and language-processing software (#31–50). The rationale behind this is to provide additional example sentences, more evidence on the lexicographic descriptions, and tools for assisting the user in making better use of the dictionary. Wiktionary does not provide links to a certain background corpus and although there are multiple external applications making use of Wiktionary data (see chapter 8), none of them has been directly integrated into Wiktionary. Hence, it does not go beyond state-of-the-art projects such as *ordnet.dk* (2004 f.) and the *Louvain EAP Dictionary* (2010 f.) in this respect.

In the context of accessibility, Wiktionary provides several access paths for browsing and searching the articles. Using the categories, thesaurus, index, and appendix pages, Wiktionary provides both a semasiological and an onomasiological view on its data and effectively liberates its users from the alphabetical order (#66–67, #72). The dictionary articles are densely interconnected by means of hyperlinks, including external links to the online encyclopedia *Wikipedia* (2001 f.) and other web pages (#79–82, #87–90). The hyperlinks are, however, not sense-oriented (#91). While the use of hypertext is up to the standards of online dictionaries, the search engine included in Wiktionary does not provide much innovative features. It is, for example, not possible to restrict the search by language or part of speech, and one cannot search for specific article constituents (e.g., searching words with a certain morphological word formation), which is possible in *OWID* (2008 f.), for example (cf. Müller-Spitzer, 2011). The lexicographic descriptions of a Wiktionary article are static in the sense that every user gets to see the same information in the same format and order. Apart from a separate user interface for mobile phones, little innovation is made in terms of customization and user-adaptation (#97–112). An exception is the recently published feature of the English Wiktionary to restrict the displayed translations to the languages selected by a user.

Finally, de Schryver summarizes lexicographer's dreams related to dictionary compilers, researchers and publishers. Though this is only partly applicable to Wiktionary, as it is not an editorial dictionary, the ease of modifying the dictionary articles and the availability of discussion pages provides direct feedback (#113). Providing separate discussion pages on each dictionary article is particularly innovative, as none of the large editorial dictionaries makes use of a similar concept. Closely related is the availability of the entire edit history of the dictionary articles. Although none of the dreams discussed by de Schryver addresses the access to an article's edit history and archived versions, we identified this as an important future topic of electronic lexicography, for which the wiki technology provides a viable solution. In section 3.5, we therefore discuss the edit history in more detail.

**Development of the lexicographic instructions**. In editorial dictionaries, the development and planning of the dictionary conception is usually part of a preparatory phase, which takes place before starting the actual work on the dictionary articles. The practical decisions and guidelines on the dictionary are documented in a *lexicographical instruction book*, which serves

as a manual for the lexicographers in their daily work. The collaboratively compiled Wiktionary encodes such guidelines as part of its outside matter in form of help and instruction pages (see section 2.7). As opposed to editorial dictionaries where the lexicographic instructions are fixed (with few exceptions), Wiktionary allows changing them at any time.

Using the example of the article microstructure, we study the development of Wiktionary's lexicographic instructions. In 2002, the page *Wiktionary:Entry layout explained* (i.e., the most important instruction page on the microstructure) encoded a simple template for new articles consisting of separate constituents for hyphenation, pronunciation, language, part of speech, sense definition, etymology, thesaurus, and translations.[19] About a year later, the instruction page has largely changed. Besides being a mere template for new entries as before, the page commented on how pronunciations and polysemous words should be encoded.[20] By 2006, the lexicographic instructions have been further fleshed out and there were extensive descriptions on each article constituent.[21] The page also included new article constituents which had not been used before (such as inflected word forms, usage notes, semantic relations, references). It turns out that the latest version of the lexicographic instructions from 2012 provides, by and large, the same information as the version from 2006.[22] Besides minor changes on the layout and the wording, the additions address more detailed descriptions on providing expanded forms of abbreviations and the definition of pragmatic labels. We hence observe that there is also a preparatory phase in Wiktionary, during which the lexicographic instructions are constituted. After this phase, the authors seem to content themselves with minor changes.

We can also confirm this observation for the German Wiktionary. The article microstructure is defined by the article style sheets there, which serve as a basis for all new articles. The first version of the style sheet from 2004 did not distinguish peculiarities of entries on different languages or parts of speech.[23] It defined the article constituents language, part of speech, inflected forms, pronunciation, etymology, meanings, synonyms, antonyms, broader terms, narrower terms, examples, sayings, references, derived terms, and translations (in this order). Each article constituent has been briefly described and exemplified with information on the headword *Kraftfahrzeug* (English: motor vehicle), for example, "Automobil {Bedeutungsgleiche oder bedeutungsverwandte Wörter}" (English: automobile {words with the same or related meaning}) for the synonym constituent.

About a year later, the style sheet has substantially changed and separate style sheets have been created for different languages and parts of speech.[24] New article constituents have been added, namely hyphenations, abbreviations, characteristic word combinations, dialectal expressions, and similar words. Translations and derived terms have been moved above the

[19] `http://en.wiktionary.org/w/index.php?oldid=202` (13 December 2002)
[20] `http://en.wiktionary.org/w/index.php?oldid=34118` (12 December 2003)
[21] `http://en.wiktionary.org/w/index.php?oldid=1912829` (28 December 2006)
[22] `http://en.wiktionary.org/w/index.php?oldid=19130446` (26 December 2012)
[23] `http://de.wiktionary.org/w/index.php?oldid=13719` (5 May 2004)
[24] `http://de.wiktionary.org/w/index.php?oldid=99282` (29 August 2005)

references. In addition to that, the formatting of the constituents has been revised. Translations are, for example, displayed in a table with two columns rather than a single-column itemization.

When comparing the noun style sheet from 2005 with a current version from 2012,[25] we again observe only minor differences: some article constituents have been retitled with alternative captions (e.g., from *Silbentrennung* to *Worttrennung* for the constituent on hyphenations), broader and narrower terms should now be associated with a specific word sense, and a number of default entries have been added to the references (see section 3.4). Thus, we observe also a preparation phase for the German Wiktionary. Later changes to the lexicographic instructions are possible, but mostly address smaller issues.

One explanation for this observation is that changes in the lexicographic instructions yield a need to adapt the existing dictionary articles to the new decisions in a collaborative effort. The more articles reside in the dictionary, the longer it takes to revise them, which might be a disincentive to performing larger revisions of the instructions. Another reason is that authors who have been active for a long time are likely to oppose against new proposals, because they are used to and maybe even proposed the existing instructions. Stegbauer (2009) discusses similar problems related to the power and leadership of administrators and standing contributors in the context of *Wikipedia* (2001 f.).

## 3.4 Lexicographic Evidence and Quality

**Quality.** The quality of the lexicographic descriptions in collaborative dictionaries is among the most controversial topics being discussed. Lepore (2006, p. 87) has a critical opinion of the information quality in Wiktionary:

> "Be your own lexicographer!" might be Wiktionary's motto. Who needs experts? Why pay good money for a dictionary written by lexicographers when we can cobble one together ourselves?

Wiktionary is often criticized for providing unspecific or too general definitions. Fuertes-Olivera (2009, p. 123) points out that the definition "the purchase of one company by another; a merger without the formation of a new company" of the noun *takeover* does not differentiate well between the general *purchase* of a company and the specialized concepts of a *takeover* and a *merger*. Other issues are spelling errors in the dictionary articles, such as the use of "bootle [sic!] feeding" in the article *bottle feed*.

Hanks (2012) observes many old-fashioned descriptions in Wiktionary, which mostly stem from importing information from copyright-expired dictionaries. His analysis of some selected Wiktionary articles shows that the sense definitions lack explanatory power and that example

[25] `http://de.wiktionary.org/w/index.php?oldid=2172258` (9 April 2012)

sentences illustrate extreme cases of using a word rather than its normal usage. He ascribes this to the lack of corpus evidence.

Rundell (2012, p. 81) notes the "randomness of what contributors produce" (i.e., providing articles at different degrees of completion), but points out that such a project could otherwise never "get off the ground". Compare, for instance, the Wiktionary articles *epizootic* and *misfortune*.[26] The former is excessively elaborated giving many details and referring to multiple sources, while the information provided by the latter is rather modest, both in terms of the number of sources and information types. The idea of Wiktionary is to complete its articles in a collaborative effort, which is why inconsistent articles are explicitly permitted.

Nevertheless, we found some cases in our analysis for which Wiktionary provides relevant information. This particularly holds for newly coined terms, regional usages, or technical terminology, which we discuss in chapter 4 (cf. Meyer and Gurevych, 2010a, 2012a). Penta (2011) reports a similar observation when comparing Wiktionary, the *Urban Dictionary* (1999 f.), and the *Oxford English Dictionary Online* (2002 f.): He finds the collaborative dictionaries to have a good account of the contemporary meanings of a word in particular in slang-related contexts.

The Wiktionary community is aware of quality problems arising from its openness. The German Wiktionary, for instance, explicitly informs its readers:

> Aufgrund der anonymen und freiwilligen Mitarbeit kann die Richtigkeit nie garantiert werden. Das kann sie im Übrigen in gedruckten Werken aber auch niemals. Eine leichte Skepsis den Inhalten gegenüber kann grundsätzlich nicht schaden […]. Da zahlreiche Benutzer die letzten Änderungen im Auge behalten und das Wiktionary nach Fehlern durchforsten, sollten sich Fehler in Grenzen halten.[27]
>
> [English: Because of the anonymous and voluntary participation, correctness can never be guaranteed. But this, by the way, also holds for printed works. A slight skepticism about the contents generally does no harm […]. Since numerous users keep track of the latest changes and comb through the Wiktionary for flaws, errors should be kept within a limit.]

Studies on *Wikipedia* (2001 f.) indeed showed that the collaborative creation has the ability to be neck and neck with expert-built works: Giles (2005) found on average four inaccuracies in Wikipedia articles compared to about three in the *Encyclopædia Britannica Online* (1994 f.). Casebourne et al. (2012) report similar results in a larger comparative study of Wikipedia and a number of online encyclopedias.

It remains an open question if collaborative dictionaries will yield similar results in the future. But as of today, the discussed works show that inaccuracies predominate in the comparative studies of Wiktionary and expert-built dictionaries. Despite the competences of the

[26] `https://en.wiktionary.org/w/index.php?oldid=16546340` (27 March 2012), `https://en.wiktionary.org/w/index.php?oldid=19342437` (13 January 2013)

[27] `http://de.wiktionary.org/w/index.php?oldid=1747323` ( 29 March 2011)

contributing authors, we consider two major influencing factors on the quality of the lexicographic descriptions: the quality assurance measures that are taken by the community and the sources that are provided to verify the correctness of the descriptions.

**Quality assurance**. The Wiktionary community makes use of two major quality assurance measures: requests and flagged revisions. *Requests* are added to an article page by authors observing a quality issue – but who are not capable or willing to resolve it immediately. The request then shows up for subsequent authors as a colored banner naming the observed flaw. Requests can address the verification of the lexicographic descriptions (e.g., providing additional sources), the extension of an article (e.g., describing the etymology), the cleanup of articles requiring substantial revision (in terms of content or formatting), and the deletion of an article in case it contains nonsense, its descriptions could not be verified, or it appears to be a copyright infringement.

The German, Icelandic, Polish, Russian, and Ukrainian Wiktionary editions provide the *flagged revisions* feature to mark articles that reached a certain level.[28] Depending on the definition of the flagged revisions, this can indicate (1) that the article is free of vandalism or (2) that its descriptions have been checked for correctness. *Vandalism* is defined as a "deliberate disruption of Wiktionary"[29] in the sense that unrelated, provoking, or nonsensical insertions, modifications, or deletions are being performed (e.g., replacing the entire article text with a swear word). So far, only the former definition of flagged revisions is applied. This is why it is planned to distinguish between a *sighted flag* (definition 1) and a *quality flag* (definition 2) in the future. A flagged revision can only be created by authors with a corresponding user permission. Authors receive this permission automatically or can ask for it if they have a personal account, made at least 200 edits, and have been active within the last two months.

**References and sources**. The lexicographic descriptions of most editorial dictionaries are based on a *dictionary basis*. That is, a set of all sources that the lexicographers utilize to formulate the dictionary articles. The set of *primary sources* constitutes the *lexicographic corpus* containing, for instance, fiction, technical literature, scientific articles, newspapers, etc. The corpus is often balanced by different text types, time of publication, or author. The *secondary* and *tertiary sources* consist of other dictionaries and reference works (such as grammars). When planning an editorial dictionary, the dictionary basis is usually defined as a fixed catalog of sources. In contrast to that, no fixed catalog exists in Wiktionary. Although there is a list of about 110 frequently used sources in the English Wiktionary and about 750 sources in the German Wiktionary, the authors have the possibility to include any source without adding it to one of these lists.[30]

[28] `http://meta.wikimedia.org/w/index.php?oldid=5270006` (23 February 2013)

[29] `http://en.wiktionary.org/w/index.php?oldid=18662560` (28 October 2012)

[30] `https://en.wiktionary.org/w/index.php?oldid=16999309` (23 June 2012), `http://de.wiktionary.org/w/index.php?oldid=2716574` (31 October 2012)

The sources of the lexicographic descriptions are encoded by means of *references* naming the bibliographic details of the source. We distinguish between article-related and item-related references. *Article-related references* address the dictionary article as a whole. They are used to provide evidence for the existence and correctness of the described word and its meanings. *Item-related references* are directly associated with a certain information item and serve as a source for quotations, example sentences, or etymological information.

In a quantitative analysis of the 3.2 million articles of the English Wiktionary, we found that only 3 % of them encode article-related references and that item-related references are even used in only 0.2 % of the articles. This is fundamentally different in the German Wiktionary, where 49 % of the about 245,000 articles encode at least one article-related reference and 14 % at least one item-related reference. On average, we find 2.27 references per article there – as opposed to 0.05 in the English Wiktionary. In total, there are about 153,000 (English Wiktionary) and 556,000 (German Wiktionary) references, which point to 15,000 (English Wiktionary) and 23,000 (German Wiktionary) unique sources.[31] We thus observe different cultures regarding the use of references in Wiktionary: They play a major role in the German Wiktionary where lexicographic descriptions are likely to be deleted if no references are provided. In the English Wiktionary, they are sparse and we find the majority of articles not providing any references. In addition to that, the bulk of the references from the English Wiktionary are encoded for entries on the Chinese (20 %) and the Armenian language (6 %). The most often referenced dictionary is the *Han yu da ci dian* (漢語大詞典, 2007) there.

Table 3.1 shows the most frequently referenced sources in entries on the native language of a Wiktionary edition. Dictionaries with expired copyright, such as *Webster's Revised Unabridged Dictionary* (1913) and the *The Century Dictionary and Cyclopedia* (1911), serve as evidence for most English entries. This is in line with Hanks' (2012) finding of many old-fashioned sense definitions within the English Wiktionary. The German Wiktionary particularly refers to online dictionaries describing contemporary German, like *canoonet* (2000 f.) and *Duden online* (2011 f.), whereas the German authors also refer to multiple printed dictionaries and grammars. It is not surprising to find the first six German sources among the most frequently used ones, since they are suggested as default entries by the article style sheet. It remains unclear whether these six sources have actually been consulted by the authors and used for formulating the sense definitions. The remaining sources are, however, explicitly added by the authors. Since it is common practice to provide explicit page numbers or a definite URL to the referenced information, this indicates that the sources have at least been looked up by the authors.

[31] The references are very heterogeneously formatted in Wiktionary, which is why the number of unique sources should be considered as an upper bound. Duplicates occur, for instance, if online resources have a different URL, but point to the same content, different editions of a printed work are used, and individual news articles are treated as separate sources instead of grouping them by newspaper.

| № | English sources | References |
|---|---|---|
| 1) | *Webster's Revised Unabridged Dictionary* (1913) | 4,088 |
| 2) | *The Century Dictionary and Cyclopedia* (1911) | 2,862 |
| 3) | *Mineralogy Database* (1997 f.) | 2,645 |
| 4) | *Wikipedia* (2001 f.) | 2,456 |
| 5) | *Oxford English Dictionary* (multiple editions) | 1,633 |
| 6) | *OneLook Dictionary Search* (1996 f.) | 1,532 |
| 7) | *Dictionary.com* (1995 f.) | 373 |
| 8) | *Online Etymology Dictionary* (2001 f.) | 246 |
| 9) | *The American Heritage Dictionary of the English Language* (2000) | 146 |
| 10) | *Webster's Seventh New Collegiate Dictionary* (1965) | 122 |
| 11) | *American Dictionary of the English Language* (1828) | 120 |
| 12) | *Random House Webster's Unabridged Dictionary* (1997) | 108 |
| 13) | *Taber's Encyclopedic Medical Dictionary* (1993) | 88 |
| 14) | *The New Geordie Dictionary* (1987) | 84 |
| 15) | *Canadian Oxford Dictionary* (2004) | 76 |

| № | German sources | References |
|---|---|---|
| 1) | *Wikipedia* (2001 f.) | 48,732 |
| 2) | *Wortschatz-Lexikon* (1998 f.) | 47,907 |
| 3) | *Digitales Wörterbuch der deutschen Sprache* (2004 f.) | 45,123 |
| 4) | *canoonet* (2000 f.) | 43,864 |
| 5) | *TheFreeDictionary.com* (2003 f.) | 18,936 |
| 6) | *Deutsches Wörterbuch* (1854–1961) | 14,021 |
| 7) | *Duden online* (2011 f.) | 13,653 |
| 8) | *Duden: Deutsches Universalwörterbuch* (multiple editions) | 2,859 |
| 9) | *Duden: Die deutsche Rechtschreibung* (multiple editions) | 2,089 |
| 10) | *Duden: Das große Fremdwörterbuch* (multiple editions) | 1,249 |
| 11) | *wissen.de* (2000 f.) | 1,006 |
| 12) | *Metzler-Lexikon Sprache* (2005) | 1,004 |
| 13) | *OWID* (2008 f.) | 930 |
| 14) | *Lexikon der Sprachwissenschaft* (2002) | 743 |
| 15) | *Der Neue Herder* (1949) | 688 |
| 16) | *Brockhaus* (1996–1999) | 654 |
| 17) | *Goethe-Wörterbuch* (1966 f.) | 563 |
| 18) | *Redensarten-Index* (2001 f.) | 541 |
| 19) | *Neues Lexikon der Vornamen* (1993) | 533 |
| 20) | *Großes Abkürzungsbuch* (1980) | 531 |

Table 3.1: Most frequently referenced sources in native language entries of the English (top) and the German (bottom) Wiktionary edition

It is notable that all frequently used sources are secondary and tertiary sources. We have already noted the absence of systematic corpus evidence above (i.e., the lack of primary sources). The few primary sources referenced by the authors point mainly to online texts such as news articles, digital books, and to scientific literature (including juristic, medical, or economic articles). This means that the lexicographic descriptions in Wiktionary are either predominantly based on other dictionaries or that other kind of evidence is used to formulate them, which is not referenced. For the former, another fundamental question is whether the lexicographic descriptions have been copied from the secondary sources without change or newly worked out with awareness of them.

The large Wiktionary editions have automatically imported dictionary articles from copyright-expired dictionaries. The English Wiktionary, for instance, imported many articles from *Webster's Revised Unabridged Dictionary* (1913). These articles are associated with a special category, such that they can be revised over time by the community. In these cases the information is initially copied from the secondary sources. The rare usage of references in the English Wiktionary, however, makes it hard to study if further sources have then been consulted during the revision of these articles.

Figure 3.1 shows the article *Betreuungsgeld* (English: money given to stay-at-home parents) from the German Wiktionary. The article contains three article-related references as evidence for the existence of this term and one item-related reference pointing to a news article, from which the example sentence has been taken. The specified references cover a detailed encyclopedic description (*Wikipedia*), multiple corpus examples and collocations (*Digitales Wörterbuch der deutschen Sprache*), as well as corpus frequencies, morphological and grammatical information (*Wortschatz-Lexikon*). None of the referenced sources contain information on the etymology and pronunciation of the word. In addition to that, we could not find the sense definition encoded in Wiktionary in any of the referenced sources nor in other frequently used, but not referenced sources, such as *Duden online* (2011 f.). This suggests that the encoded information has not been copied, but newly worked out based on the referenced sources and the language intuition of the authors. There are, however, also cases which are far from being clearly determinable. The article *schneeweiß* encodes, for instance, the sense definition "weiß wie Schnee" (English: snow white; as white as snow). Since *Duden online* encodes "weiß wie (frisch gefallener) Schnee" (English: as white as (newly fallen) snow), it is not possible to make clear assumptions if the sense definition has been copied (and maybe slightly varied) or worked out on the basis of intuition and introspection.

In summary, we find that the lexicographic descriptions in Wiktionary are mainly based on secondary and tertiary sources. A question that we cannot conclusively clarify is whether the Wiktionary authors tend to copy information from such secondary and tertiary sources or newly formulate their descriptions with awareness of the sources and based on their own intuition and language feeling. Storrer (2012) raises the question whether collaborative dictionaries supersede scientific and professional lexicography. She suggests that – due to the

# Betreuungsgeld

## Betreuungsgeld (Deutsch) [Bearbeiten]

### Substantiv, *n* [Bearbeiten]

**Worttrennung:**
Be·treu·ungs·geld, Plural: Be·treu·ungs·gel·der

**Aussprache:**
IPA: [bəˈtʁɔɪ̯ʊŋsˌgɛlt], Plural: [bəˈtʁɔɪ̯ʊŋsˌgɛldɐ]
Hörbeispiele: —, Plural: —

**Bedeutungen:**
[1] finanzielle, staatliche Unterstützung für Eltern, die ihr Kind zu Hause betreuen, anstatt es in eine Krippe zu schicken

| Kasus | Singular | Plural |
|---|---|---|
| Nominativ | **das Betreuungsgeld** | **die Betreuungsgelder** |
| Genitiv | des Betreuungsgelds<br>des Betreuungsgeldes | der Betreuungsgelder |
| Dativ | dem Betreuungsgeld | den Betreuungsgeldern |
| Akkusativ | das Betreuungsgeld | die Betreuungsgelder |

**Herkunft:**
Determinativkompositum aus den Substantiven *Betreuung* und *Geld* mit dem Fugenelement *-s*

**Beispiele:**
[1] „Merkel ist indes fest entschlossen, den seit Monaten schwelenden Streit über das *Betreuungsgeld* jetzt zu beenden: Noch in diesem Monat soll ein Gesetz auf den Weg gebracht werden, dass Mütter und Väter, die ihre Kleinstkinder nicht in eine Krippe geben, mit zunächst 100 Euro im Monat unterstützt werden."[1]

**Übersetzungen** [Bearbeiten]

Einklappen

- Englisch: [1] [[|]] → en
- Französisch: [1] [[|]] → fr
- Italienisch: [1] [[|]] → it
- Spanisch: [1] [[|]] → es

**? Referenzen und weiterführende Informationen:**
[1] Wikipedia-Artikel „Betreuungsgeld"
[*] Digitales Wörterbuch der deutschen Sprache „Betreuungsgeld"
[1] Uni Leipzig: *Wortschatz-Lexikon* „Betreuungsgeld"

*Quellen:*

1. ↑ "*Das ist ein K.-o.-Schlag*". In: *Welt Online*. 15. Mai 2012, ISSN 0173-8437 (URL, abgerufen am 22. Mai 2012).

Figure 3.1: The article *Betreuungsgeld* in the German Wiktionary (`http://de.wiktionary.org/w/index.php?oldid=2690921`; 28 September 2012)

lack of corpus evidence in Wiktionary – there is still a demand for empirical and lexicologically sound descriptions, and thus the need for expert lexicographers. Our analysis confirms that there is no "original" scientific lexicographic work in Wiktionary, as the descriptions rely largely on secondary sources. However, we consider the intuition of the authors as an important kind of evidence, which is fundamentally different from corpus evidence. We therefore analyze the coordination of and the discussions among of the authors in the next section in order to research the degree of intuition contributed by the authors.

## 3.5 Coordination and Cooperation

**Revision history.** The wiki software used by Wiktionary saves each modification of an article in a *revision history*. This allows for studying the development of language – and *semantic shift* in particular – by comparing the dictionary articles from different dates. We find, for example, that the definition of *hand-held* has changed from "a computing device (e.g. organizer,

Internet-enabled cell phone) that is operated while held in the hands" in 2003 to "personal digital assistant or video game console that is small enough to be held in the hands" in 2012, which accounts for the use of *hand-held* for referring to video game consoles, too.[32]

In addition to that, the revision history facilitates linking to a specific, stable version of an article, which does not change over time.[33] Such a feature is of particular importance, since the World Wide Web is continually changing, which impedes properly citing articles from electronic dictionaries. Drude et al. (2012) and Schüller (2004) point out that much language-related data is highly endangered because of the usage of fragile formats, inappropriate storage technology, missing archive and versioning, etc. The revision history of Wiktionary makes a valuable contribution in this context by providing access to old article versions, which is not possible in major online dictionaries, such as the *Oxford Dictionaries Online* (2010 f.) or *Duden online* (2011 f.).

Based on Daxenberger and Gurevych (2012), we distinguish between edit operation and revision. An *edit operation* denotes a change to a certain article constituent, in which information is being

- *inserted* (e.g., providing a new phonetic transcription),
- *modified* (e.g., reformulating the sense definition),
- *deleted* (e.g., removing an example sentence),
- *formatted* (typographic changes, such as bold types or italics), or
- *reverted* (i.e., reset to the previous version due to vandalism).

The set of edit operations performed by an author at a certain point in time is called a *revision*. Thus, revising an article can embrace changes to multiple article constituents, for example, adding a new sense definition (insertion) while reformulating another one at the same time (modification). Wiktionary's *revision history* provides access to the set of all revisions of a dictionary article.

**Analysis of Wiktionary revisions.** In a quantitative analysis of the revision history, we find 17.2 million revisions in the English Wiktionary and 2.5 million revisions in the German Wiktionary. The number of revisions per article page follows a Zipf law (i.e., there are many articles with only one revision and a few articles with a large number of revisions). On average, an article has between five (English Wiktionary) and ten (German Wiktionary) revisions. While less than 5 % of the revisions are changes by unregistered users, it is interesting to find about 50 % of the revisions in the English Wiktionary and even 64 % in the German Wiktionary to be authored by a bot (i.e., an automatic computer program, see section 2.7). Since bots have usually very fine grained tasks, which do not add new information items, this should, however,

[32] `http://en.wiktionary.org/w/index.php?oldid=26785` (7 November 2003), `http://en.wiktionary.org/w/index.php?oldid=18254941` (22 September 2012)

[33] We use such links to specific Wiktionary revisions throughout the thesis.

be taken with a grain of salt. In fact, the human authors account for 61 % (English Wiktionary) and 87 % (German Wiktionary) of the total number of changed characters over all revisions. We focus on the 7.9 million (English Wiktionary) and 0.8 million (German Wiktionary) revisions by registered users during the remaining study.

Over 90 % of these revisions have been described by a *comment*, which is not part of the dictionary article, but a short text explaining the changes made by the author. This is intended for subsequent authors to track which changes have been made and why. The Wiktionary article *water* (2,014 revisions) is the most frequently revised article of the English Wiktionary. The majority of the frequently changed articles describe words from the basic vocabulary, for example, the nouns *dog* (1,012 revisions) and *cat* (855), the pronoun *you* (830), the verb *love* (825), and the German words *Wasser* (255), *Haus* (235), and *Wort* (195).[34] In addition to that, we identified many terms from colloquial speech and slang – in particular vulgar expressions.

On average, an article has between 1.6 (English Wiktionary) and 2.2 (German Wiktionary) distinct authors (which again follows Zipf's law). The list of articles with many authors is very similar to the list of articles with many revisions (i.e., the more often an article is being revised, the more authors participate). It is not surprising that older articles have in general a higher number of distinct authors, since they reside longer in the dictionary than recently created ones.

We call the author of the most revisions (but at least five) of a certain article its *main author*. Main authors perform on average between 34 % (English Wiktionary) and 29 % (German Wiktionary) of the revisions of an article. This ranges from articles written entirely by one author to articles in which the main author is not clearly distinguishable from other authors. From our analysis, we can identify different types of main authors. Consider, for instance, the authors #9200 and #720 of the German Wiktionary. Author #9200 is the main author of 15 articles, in which she or he has performed on average 66 % of the revisions. Of the 463 revisions that #9200 contributed in total, he or she performed 57 % of them as the main author of the article. This includes the articles *Privatrecht*, *Sozialrecht*, *Arbeitsrecht*, *Gleichheitsrecht* (over 85 % of the revisions), *Grundrecht* (44 %), and *Satzung* (22 %).[35] We observe that #9200 is obviously an author focusing on jurisprudence and law.

In contrast, author #720 is the main author of over 3,500 articles. But she or he performed only about 31 % of the revisions in those articles and there are many articles with only a few number of revisions. By taking a closer look at these articles, we observe that #720 does not have a thematic focus, but predominantly contributes translation equivalents. Hence, there are different types of authors contributing to Wiktionary and different responsibilities they take. Besides authors with a thematic focus and those focusing on particular article constituents, we can identify authors caring about foreign language entries (e.g., Russian words within the

[34]English: water; house; word

[35]English: private law; social law; labor law; legal equality; basic rights; constitution/by-law

English Wiktionary), authors revising existing information items or their formatting rather than contributing new information, and authors mostly reverting vandalism.

When taking a look at articles with a *dominant secondary author* (i.e., articles in which two authors have performed the majority of the revisions), we can study their cooperation and potential for conflict. Articles with a dominant secondary author are of particular interest, because two authors contributing an equally high number of revisions can either effectively work together or reject and overwrite the modifications of the respective other. The latter is known as an *edit war*, which has been frequently observed in *Wikipedia* (cf. Viégas et al., 2004). In a small qualitative study of 20 articles with a dominant secondary author taken from the German Wiktionary, we find that the two main authors worked either cooperatively or during different time periods. The article *backen* (English: (to) bake) has, for example, been edited by its main author mostly after 2010, while the secondary author revised the article predominantly before 2006. The two main authors of *Evolution* (English: evolution) worked during the same period of time, but did neither revert the changes of each other nor show any potential for conflict.

**Discussion pages**. There is a separate *discussion page* for every wiki page in Wiktionary, which can be used for planning, organizing, and evaluating the lexicographic work. The discussion pages belong to the metatexts of the dictionary: Instead of lexicographic descriptions, they consist of a number of *turns* (Ferschke et al., 2012). A turn is a statement added by a Wiktionary author to either start a new, yet undiscussed topic or reply to another turn added previously. It is good practice to sign a turn with one's own user name and the current date. The discussion page of the article *week* contains, for example:

> Either our sense is not flexible enough and should be re-worded more like that of month, or we need to add a sense to cover weeks of lengths other than seven days used in ancient or exotic calendars. For instance, the Aztecs used a week of five days. — Hippietrail 14:47, 14 October 2005 (UTC)
>
> What about a phrase like "I'll see you on Monday week"? Is that covered by the existing definitions? 109.154.75.4 10:10, 21 July 2010 (UTC)
>
> - It is now. SemperBlotto 10:15, 21 July 2010 (UTC)[36]

This discussion consists of three turns. The second turn provides an example sentence that is not covered by any of the encoded sense definitions. Just five minutes later, the author of the third turn added a new word sense "Seven days after (sometimes before) a specified date" and a corresponding example sentence to the article page, before she or he posted the reply to the discussion page. The first turn, however, has not been commented on by any author and the issue raised has not caused any changes to the dictionary article yet.

[36] `http://en.wiktionary.org/w/index.php?oldid=9493042` (21 July 2010)

Besides proposing extensions, the discussion pages serve also as a method to avoid or resolve conflicts. Consider, for example, the discussion page for the German word *Kreuzung* (English: crossroads):

> Entschuldigt meine heftige Wortwahl, aber ich weiß nicht, was hier für „Fachleute“ revertieren und sichten!!
>
> - ein Ort, wo sich 2 Straßen treffen, nennt man „Straßenknick“ oder „Straßenecke“, im einfachsten Fall einfach nur Straße, wenn eine gerade verlaufende Straße von der a-Straße zur b-Straße wird.
> - ein Ort, wo sich 3 Straßen treffen, wird Straßengabel oder auch Abzweigung genannt
> - ein Ort, wo sich 4 oder mehr Straßen treffen, wird Kreuzung genannt. Das ist sinngleich mit der Aussage: ein Ort, wo sich 2 (oder mehr) Straßen kreuzen oder ein Ort, wo eine Straße eine zweite Straße quert.
>
> [...] Ich möchte jetzt hier keinen Editwar anzetteln, weshalb ich die stattgefundene Revertierung meiner Änderungen nicht erneut ändere und bitte, jemand mit Sachkunde und Verständnis möge sich der Angelegenheit annehmen.[37]
>
> [English: Please excuse my tough words, but I don't know, which kind of "experts" are reverting and reviewing this article!!
>
> - a place where 2 roads meet is called a "road bend" or "road corner", in the simplest case just road if a straight road becomes from a-road to b-road.
> - a place, where 3 roads meet is called a road fork or branch.
> - a place, where 4 or more roads meet is called a crossroads. This is synonymous to the statement: a place, where 2 (or more) roads are crossing or a place, where one road traverses a second road.
>
> [...] I do not want to start an edit war, which is why I have not made my previously reverted changes again and ask someone with expertise and understanding to take care of this matter.]

The author of this turn refers to a modification of the sense definition from "Stelle, wo sich zwei oder mehrere Straßen treffen" (English: place, where two or more roads meet) to "Stelle, wo sich vier oder mehr Straßen treffen" (English: place, where four or more roads meet), which has been reverted shortly after the change has been saved. The further turns on this discussion page reveal that there are indeed two interpretations of a crossroads, which differ in defining a road to end in the crossroads (four or more roads meet) or pass through the crossroads (two or more roads meet). Although the opinion of the initiator of this discussion cannot convince the other authors in the course of the discussion, such contributions show that the subjective opinions and language intuitions of the authors are being controversially discussed.

This opens up the possibility of including in the dictionary how language is used and understood by a language community. Following Køhler Simonsen (2005), we may denote this

[37] `http://de.wiktionary.org/w/index.php?oldid=1992616` (20 October 2011)

as *lexicographic democracy*. Thereby, Wiktionary goes even beyond Køhler Simonsen's definition, which leaves the final decision on a user contribution always to the editors of the dictionary. In Wiktionary, the community decides collaboratively on what remains in the dictionary articles. To this end, the sum of opinions of the Wiktionary authors has the potential to provide a different kind of evidence than found in a mere corpus-based dictionary.

## 3.6 The Collaborative Lexicographic Process

The term *lexicographic process* is often used in the metalexicographic literature to describe the steps and phases towards the compilation of a dictionary. Wiegand et al. (2010, p. 132) distinguish, for instance, the following five phases:

- the preparation phase,
- the data collection phase,
- the data editing phase,
- the data analysis phase, and
- the preparation of setting and printing phase.

We now discuss each of these phases and evaluate to what extent the existing model can be used to describe the collaborative process of Wiktionary. It should be noted that there are multiple other descriptions of the lexicographic process. While some of them might be more suitable for describing electronic dictionaries than the one by Wiegand et al. – for example, the works by Müller-Spitzer (2004) and Klosa (2013) – we are not aware of any work focusing particularly on the lexicographic process of collaborative dictionaries. We therefore use the five phases by Wiegand et al., as they are fairly generic and hence allow us to highlight the peculiarities of Wiktionary. As a result of this discussion, we propose a novel description of the lexicographic process in collaborative dictionaries.

**Phase 1: the preparation phase**. When compiling an editorial dictionary, the editors usually develop a *dictionary plan* in a separate preparation phase. That is, they decide on the dictionary type, plan the organization of work, and formulate the lexicographic instructions. Organizational matters like compiling a staff plan, financial plan, or time schedule are only applicable to Wiktionary to a certain extent. But regarding the lexicographic issues, Wiktionary similarly reveals a preparation phase, in which the dictionary conception and the lexicographic instructions are largely defined (see section 3.3).

As opposed to editorial dictionaries, Wiktionary allows for modifications of the lexicographic instructions at practically any time. Although we have seen that the community tends to perform rather minor changes to the lexicographic instructions after these have settled (i.e., after the preparation phase comes to an end), the lexicographic process description should

model this aspect. This is particularly important, because such changes raise a need for revising the existing articles with respect to the new instructions.

**Phase 2: the data collection phase.** The goal of the second phase is the creation of a *dictionary basis* (i.e., the selection of primary, secondary, and tertiary *sources*). We have discussed in section 3.4 that Wiktionary does not make use of a fixed set of sources. The authors are rather free to include any type of source while editing a dictionary article. There is thus no explicit data collection phase in Wiktionary, as such efforts merge with the subsequent phases.

**Phase 3: the data editing phase.** As part of the data editing phase, the lexicographers usually create a *provisional lexicographical database*. That is, a structured set of preprocessed data from the lexicographic corpus, which serves as a basis for formulating the dictionary articles (which is then part of phase 4). The rationale behind this is selecting the facts required for writing the lexicographic descriptions (i.e., the *inner selection*). Atkins and Rundell (2008, p. 322) note that the entries on a certain headword of such a database "should be at least two or three times bigger than the final dictionary entry". The process of constructing the provisional lexicographical database is traditionally called *systematic excerption*, which addresses checking the primary sources for illustrative usages of a word and documenting the corresponding excerpts on *index cards*. In modern dictionaries, this is usually replaced by processing electronic text corpora, from which the relevant excerpts can be selected. Wiegand et al. (2010, p. 135) note:

> Because dictionary articles are not normally written in such a way that, for example, if the lexicographer wants to add a lexicographical example to a lemma, he/she will select some document in a sub-database of the lexicographical source data that is regarded as a primary source, and look for a suitable citation text there.

This would not be sensible in an editorial dictionary, since the lexicographers had to perform the steps of the inner selection over and over again.

The absence of a dictionary basis and the predominant use of secondary and tertiary sources, which we observed in section 3.4, indicates that there is no separate data editing phase in Wiktionary. The large number of authors and frequent revision of the authors yields a decentralized and non-systematic collection of evidence. As opposed to the procedure described by Wiegand et al. above, the Wiktionary authors choose the sources (phase 2), select the facts (phase 3), and formulate the lexicographic descriptions (phase 4) in a single step.

**Phase 4: the data analysis phase.** The fourth phase is targeted at formulating the dictionary articles based on the *provisional lexicographical database*. The lexicographers working on an editorial dictionary usually formulate either entire dictionary articles or elaborate a certain set of information items (e.g., etymologies) for a number of articles from a predefined list of

headwords (e.g., all words starting with *"D"* or all adjectives). This is different in Wiktionary, where an author can begin a new dictionary article on any word. She or he is neither limited to adhere to a certain list of headwords nor required to formulate the entire dictionary article at once. The idea of the collaborative work is rather to rely on subsequent modifications and complete the article constituents step by step in a number of revisions. As we described in section 3.5, revising an article plays a major role in Wiktionary, which should be modeled by the lexicographic process.

In addition to that, we found that discussion pages are being used to organize the lexicographic work and discuss the lexicographic descriptions. Such discussions are made transparent in Wiktionary and provide a form of lexicographic metatext, which is either not documented or kept private in editorial dictionaries. Our analysis in section 3.5 shows that discussion pages can provide the subjective opinions and language intuitions of the Wiktionary community, which is a different kind of information than corpus data and may yield dictionary articles backed by the community's understanding of language.

**Phase 5: the preparation of setting and printing phase**. The final layout of the dictionary (including the articles and the lexicographic metatexts) is being defined during the last phase of the lexicographic process. The corresponding activities in Wiktionary differ only marginally from editorial online dictionaries: While a dictionary writing system is formatting the lexicographic descriptions for print or online publication in editorial dictionaries, this is done by the wiki software used by Wiktionary.

It is, however, notable that many modern dictionaries clearly separate the lexicographic descriptions from the way they are presented – with the intent of achieving cross-media publication (cf. Müller-Spitzer, 2004). The wiki pages in Wiktionary are encoded in a *wiki markup* language, which features the use of typographic variations, hyperlinks, etc. to ease the readability of encoded descriptions. The wiki markup representation in Wiktionary is very close to its presentation format in HTML (i.e., the interpretation of the wiki markup). This hinders cross-media publication on the one hand and the computational exploitation of the lexicographic descriptions on the other hand. In section 5.3, we discuss that this presents a major challenge to using Wiktionary in natural language processing applications. Apart from that, the formatted representation of the dictionary articles is updated directly after the article has been created or modified. There is hence no separate phase in Wiktionary for preparing the entire dictionary for publication at once.

**New process description**. Comparing the lexicographic process of Wiktionary and editorial dictionaries reveals three major differences:

(1) Subsequent changes to created dictionary articles are hardly covered by existing descriptions of the lexicographic process. The dictionary conception of editorial dictionaries usually defines fixed workflows for proof-reading and approving the dictionary

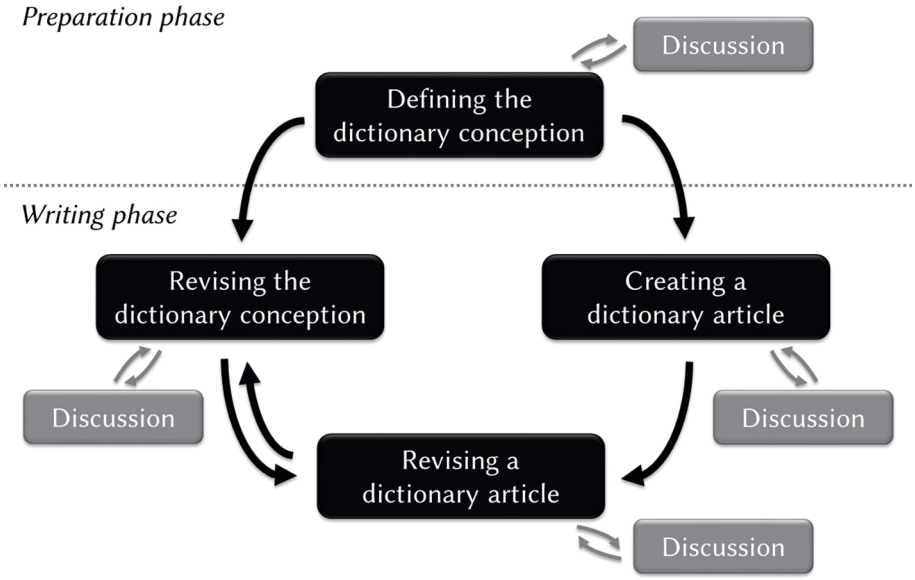



Figure 3.2: Schema of our description of the collaborative lexicographic process

articles. This is different in Wiktionary, where the dictionary articles are written iteratively yielding multiple revisions of an article. While changes to an article are to be minimized in editorial dictionaries for economic reasons, they are part and parcel of the lexicographic process in Wiktionary.

(2) The second difference is the unclear distinction of the phases 2–5 of the lexicographic process in Wiktionary. Authors collect and select facts directly while formulating the description texts and the modified article is being automatically prepared for publication. The increasing conflation of these phases has also been noted for other online dictionaries, for example by Klosa (2013).

(3) The lexicographic descriptions are usually discussed among the editors of a dictionary. In Wiktionary, such discussions are explicitly encoded in form of discussion pages. They reflect the subjective opinions and language intuitions of the authors, and thus provides a different kind of lexicographic evidence than found in corpora. To this end, discussions are more important in Wiktionary than in editorial dictionaries.

We propose a novel description of the lexicographic process in collaborative dictionaries based on these differences. Our process description consists of a *preparation phase* and a *writing phase* and a number of steps that are being performed during these phases. Figure 3.2 shows the individual steps and their relation with each other.

The preparation phase is – in accordance with editorial dictionaries – targeted towards planning the dictionary. The writing phase is interlaced with the preparation phase and starts by creating the first dictionary articles. Creating a new article implies collecting, editing, and analyzing the data. As opposed to the previous process descriptions, we consider the

revision of an article as a separate step, which can be performed multiple times after an article has been created. In addition to integrating new facts into the article, the revision step may also include proof-reading and verifying the information provided by the previous revision. The revision of a dictionary article might cause changes in the dictionary conception (for example, introducing a new article constituent) even though the preparation phase might be already concluded. This is why we modeled the revision of the dictionary conception as a separate step of the writing phase. These changes might yield inconsistencies in the existing dictionary articles, and hence again cause a revision of the dictionary articles by the authors. This is indicated by the bidirectional link between the revision of the dictionary articles and the dictionary conception. Each of the four steps is backed by a discussion step. The discussions are used to plan and evaluate the work, and – in the context of formulating the lexicographic descriptions – to provide evidence and exchange language intuitions for being integrated in the dictionary article.

## 3.7 Chapter Summary

In this chapter, we described the collaborative compilation process of Wiktionary by analyzing its dictionary conception, the quality of the lexicographic descriptions, and the cooperation and coordination of the authors. A particularly notable finding was the vast use of references to existing dictionaries and reference works in the German Wiktionary. While this suggests that no original lexicographic work is done in Wiktionary, we also found descriptions, which we could not trace back to a particular source. On the contrary, we observed that the discussion pages yield a kind of evidence that is based on the subjective opinions and language intuitions which is usually not found in corpora. We concluded the study by presenting a novel description of the lexicographic process in Wiktionary. The newly developed model accounts for the collaborative, revision-driven compilation of the articles, the indistinguishable steps of data collection, editing, and analysis, as well as the discussion processes to organize and reflect the lexicographic work.

Chapter 4

# Dictionary Comparison

This chapter aims at comparing Wiktionary to multiple machine-oriented dictionaries using quantitative data analysis methods. After describing our motivation and related work in section 4.1 and section 4.2, we analyze the availability and the development of the Wiktionary language editions in section 4.3. Then, we present the results of our quantitative comparison at the level of lexical entries (section 4.4), word senses (section 4.5), pragmatic labels (section 4.6), and relations (section 4.7). Our analysis encompasses eight dictionaries in English, German, and Russian. Based on our findings in this and the previous chapter, we assess the potential of using Wiktionary both as a human-oriented dictionary and as a machine-oriented dictionary (section 4.8). Thereby, we conclude our metalexicographic study and lead over to the natural language processing perspective.

## 4.1 Motivation

In the previous chapters, we have described Wiktionary in isolation. Now, we turn towards assessing its potential in relation to other publicly available dictionaries. Our goal is to identify well-covered topics and systematic gaps of Wiktionary. This kind of insight is crucial for researching collaborative dictionaries and their implications on lexicography, and it is at the same time necessary to effectively make use of Wiktionary in natural language processing applications.

The comparison of dictionaries is usually a part of *dictionary reviews* and the practice of *dictionary criticism*. Chan and Taylor (2001) provide a general overview of this topic. Most dictionary comparisons are based on the qualitative analysis of a few example entries. While this enables interesting insights into a dictionary in general, it is debatable to which extent these findings apply to all articles in the entire dictionary. This is usually not considered a severe issue for editorial dictionaries, since their dictionary articles are complete and designed

to be consistent throughout the entire dictionary. In collaborative dictionaries, the level of detail can, however, differ tremendously, which raises the need for quantitative comparisons.

To meet this requirement, we carry out a large-scale quantitative analysis of the English, German, and Russian Wiktionary editions in comparison to a variety of other dictionaries. We put a special focus on publicly available machine-oriented dictionaries, since their data can be electronically accessed and automatically processed, which is a prerequisite for our quantitative analysis. For the English language, we compare Wiktionary with *WordNet 3.0* (2006), which is being developed by psycholinguists at Princeton University since 1985 (see Fellbaum, 1998), and the electronic version of *Roget's International Thesaurus* (1911) created by Jarmasz and Szpakowicz (2003). For German, we analyze the German wordnet *GermaNet 6.0* (2011) developed at the University of Tübingen (Kunze and Lemnitzer, 2002) and the semi-collaborative synonym dictionary *OpenThesaurus* (2004 f.),[38] which has been introduced by Naber (2005). For Russian, we compare Wiktionary with the *Russian WordNet 3.0* (2008), an equivalent to the English *WordNet* initiated by Гельфейнбейн et al. (2003). Our analysis addresses the following contributions:

**Contribution 3.1:** *We analyze Wiktionary's coverage of language editions and the development of the dictionary over time (section 4.3).*

**Contribution 3.2:** *We compare the set of dictionaries at the lexical level and assess their lexical overlap (section 4.4).*

**Contribution 3.3:** *We compare the set of dictionaries at the level of word senses and evaluate the distribution of pragmatic labels (section 4.5 and section 4.6).*

**Contribution 3.4:** *We compare the set of dictionaries at the level of lexical relations and translations (section 4.7).*

## 4.2 Related Work

**Comparing dictionaries**. There is a long tradition of reviewing and comparing dictionaries. Osselton (1989), Chan and Taylor (2001), and Nielsen (2009) give recommendations for reviewing dictionaries and provide a good overview of existing reviews. While dictionaries are often reviewed in isolation, Atkins (1991) reports, for example, a qualitative comparison of the *Collins English Dictionary* (1986), *Webster's New World Dictionary* (1988), the *Oxford Advanced Learner's Dictionary* (1989), the *Longman Dictionary of Contemporary English* (1978), and *Collins COBUILD English Language Dictionary* (1987) by analyzing the sense definitions of the headwords *admire*, *acknowledge*, *admit*, *safety*, *danger*, and *reel*.

Most comparisons are based on the qualitative analysis of a few words. In his comparison of the *Oxford Advanced Learner's Dictionary of Current English* (1995), the *Longman Dictionary*

[38] We use a full database dump of OpenThesaurus from 12 March 2013.

*of Contemporary English* (1995), the *Collins COBUILD English Dictionary* (1995), and the *Cambridge International Dictionary of English* (1995), Herbst (1996) makes also use of quantitative methods to assess the lexical coverage, which he finds between 66–89 % (based on a list of 70 orthographic word forms). Litkowski (1999) compares the degree of polysemy and the overlap of lexical entries for 18 English words taken from *HECTOR* (1993), *WordNet 1.6* (1998), *Webster's Third New International Dictionary* (1961), the *Oxford Advanced Learner's Dictionary*, the *American Heritage Dictionary*, and *Dorr's Lexical Knowledge Base* (Olsen et al., 1998).[39] He identifies differences in the coverage for the chosen set of verbs and discusses the benefits of combining dictionaries. In the natural language processing community, large-scale quantitative comparisons have been predominantly used for comparing wordnets. Burgun and Bodenreider (2001) compare, for example, the *Unified Medical Language System* (2001 f.) and *WordNet 1.6* (1998) and find large differences in their lexical coverage.

**Comparing Wiktionary**. Correspondingly, most previous works on Wiktionary rely on the qualitative analysis of a few sample articles. Fuertes-Olivera (2009) discusses a number of manually selected Wiktionary articles on business topics, while Hanks (2012) examines the articles *admit*, *dog*, and *elephant* in great detail (see section 3.4 for a discussion of their findings). As opposed to these works, we carry out a large-scale quantitative analysis.

Garoufi et al. (2008) compare the topology of the relational structure of Wiktionary, *GermaNet 5.0* (2006), and *Wikipedia* (2001 f.). They find similar properties for the individual dictionaries and conclude that all graphs are scale-free, small-world networks. Zesch et al. (2008a) analyze the number of lexical entries encoded in the English and the German Wiktionary edition and their parts of speech. Similarly, Navarro et al. (2009) compare the number of article pages and native lexical entries of the French, English, German, Polish, and Chinese Wiktionary editions with each other, which reveals major differences in the coverage of the individual language editions. However, neither of the two works relate their results to other dictionaries. Matuschek and Gurevych (2011) compare Wiktionary and *OmegaWiki* (2006 f.) at the level of lexical entries and find a much higher lexical coverage of Wiktionary.

Krizhanovsky and Lin (2009), Krizhanovsky (2010), and Крижановский (2011) are the most similar works to ours as they perform a quantitative analysis of the English and Russian Wiktionary in comparison to the Princeton *WordNet 3.0* (2006). They find only minor differences in the average degree of polysemy between the editorial and the collaborative dictionaries. We extend those works by incorporating more dictionaries into the quantitative comparison and, for the first time, we take a closer look at Wiktionary's pragmatic labels, lexical relations, and translations in comparison to other dictionaries.

[39] The exact edition of the *Oxford Advanced Learner's Dictionary* and the *American Heritage Dictionary* is not mentioned in the original paper.

| Edition | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | Growth |
|---|---|---|---|---|---|---|---|---|---|---|---|
| English | 30 k | 48 k | 104 k | 309 k | 610 k | 1.1 M | 1.5 M | 2.2 M | 2.1 M | 2.5 M | +1181 |
| French | 31 | 2 k | 97 k | 220 k | 666 k | 1.1 M | 1.6 M | 1.9 M | 2.1 M | 2.3 M | +396 |
| Malagasy | — | — | 496 | 893 | 2 k | 2 k | 4 k | 229 k | 941 k | 2.0 M | +2039 |
| Chinese | 31 | 418 | 10 k | 92 k | 115 k | 117 k | 263 k | 962 k | 1.2 M | 1.2 M | +14 |
| Lithuanian | — | — | 63 | 481 | 18 k | 95 k | 409 k | 555 k | 595 k | 610 k | +194 |
| Russian | — | 762 | 2 k | 102 k | 131 k | 186 k | 232 k | 268 k | 304 k | 352 k | +135 |
| Korean | — | 306 | 2 k | 12 k | 34 k | 56 k | 87 k | 177 k | 317 k | 351 k | +80 |
| Polish | 28 | 27 k | 37 k | 50 k | 80 k | 108 k | 144 k | 202 k | 267 k | 308 k | +101 |
| Greek | — | 83 | 426 | 77 k | 142 k | 119 k | 143 k | 158 k | 187 k | 303 k | +323 |
| Turkish | — | 647 | 1 k | 68 k | 182 k | 252 k | 266 k | 270 k | 280 k | 295 k | +39 |
| Tamil | — | 26 | 1 k | 6 k | 6 k | 102 k | 102 k | 192 k | 239 k | 276 k | +104 |
| Swedish | — | 2 k | 7 k | 21 k | 50 k | 87 k | 97 k | 121 k | 147 k | 144 k | +250 |
| Kannada | — | — | 36 | 40 | 138 | 222 | 37 k | 85 k | 175 k | 226 k | +68 |
| Vietnamese | — | 34 | 509 | 209 k | 225 k | 228 k | 228 k | 229 k | 229 k | 230 k | +13 |
| German | — | 3 k | 16 k | 44 k | 68 k | 87 k | 102 k | 128 k | 201 k | 225 k | +70 |
| Finnish | — | 442 | 18 k | 42 k | 72 k | 103 k | 136 k | 164 k | 180 k | 217 k | +86 |
| Ido | — | — | 30 k | 97 k | 124 k | 144 k | 164 k | 177 k | 190 k | 211 k | +54 |

Table 4.1: Number of article pages encoded by the largest Wiktionary editions (as counted by December each year; k = thousand; M = million). The rightmost column 'growth' shows the average number of new articles per month in 2012.

## 4.3 Coverage of Language Editions

**Dictionary size and growth**. There are currently 170 Wiktionary language editions of which 138 are active (i.e., encoding more than ten article pages and having been edited at least ten times during the last month). Table 4.1 shows the seventeen language editions exceeding 200,000 article pages and how their number of articles developed over time.[40] While the French and English Wiktionary editions had been neck and neck at being the largest edition, the English Wiktionary has recently outpaced the other editions. The Wiktionary editions are generally growing, but their speed differs markedly. The Malagasy edition, for instance, grew by over one million articles between 2011 and 2012. This is usually an indicator for automatically importing articles from existing dictionaries – a technique that is often used as a primer for creating new dictionary articles (see section 3.4). In the same period of time, the Vietnamese Wiktionary only increased by about 1,000 articles, and the Swedish edition even decreased by about 3,000. A decreasing number of articles is the result of a consolidation process, for exam-

[40]Note that the statistics discussed in this section take only active editions and only those articles with at least ten edits into account. The numbers are based on `http://stats.wikimedia.org/wiktionary/EN/TablesArticlesTotal.htm` and `http://meta.wikimedia.org/wiki/Wiktionary` (1 February 2013).

ple, after a change of the lexicographic instructions or when removing automatically imported articles which are no longer required.

In comparison to editorial dictionaries, we find a particular high coverage of language editions in Wiktionary. Even the worldwide effort by the *Global WordNet Association* towards providing wordnets in every language resulted in only 71 resources in 55 languages so far – compared to 138 active Wiktionary editions.[41] The same holds true for the growth of the dictionaries: The recently published *WordNet 3.1* (2011) contains, for example, only 132 synsets more than its previous version *WordNet 3.0* (2006). In an experiment, Hanks (2012) created a new article page on the lemma *rogue elephant* by providing a definition for its figurative meaning. Within minutes, the article had been modified by another Wiktionary author, who formatted the page using the article style sheet and added the (literal) meaning of the animal rogue elephant. This shows that the collaborative compilation approach has the ability to grow very rapidly due to the division of lexicographic work.

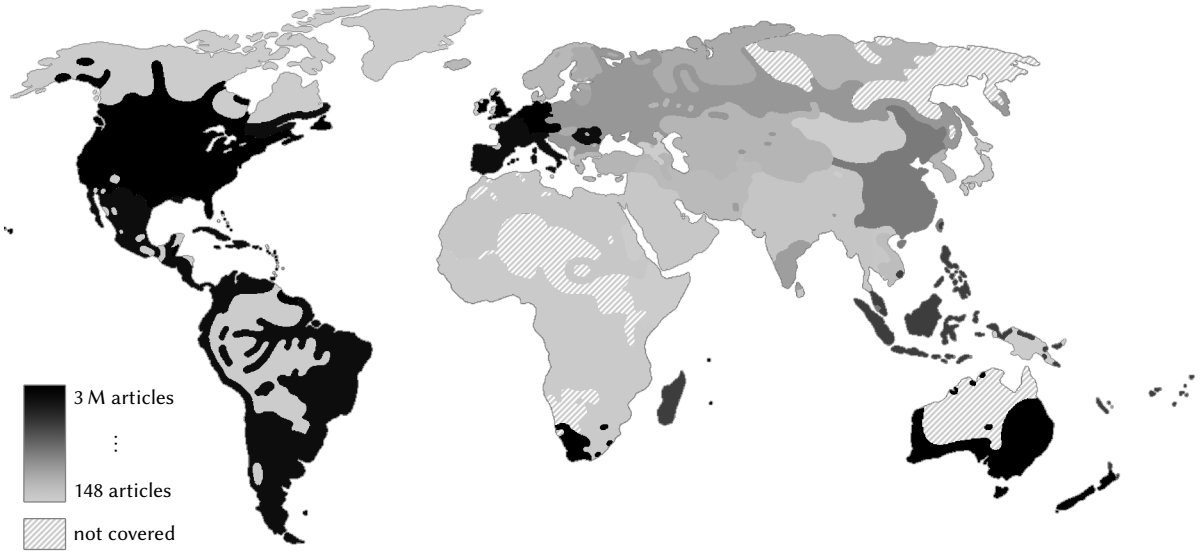
**Coverage of language families.** An active Wiktionary language edition exists for 67 % of the 185 languages defined by ISO 639-1 (2002). In order to clarify whether Wiktionary covers the full variety of languages in the world or is solely dominated by certain countries, continents, or cultures, we study the *language families* and their geographical distribution for which an active Wiktionary edition exists. Our classification is based on Ruhlen (1987) and Lewis (2009).[42] For the linguistically diverse regions in the Americas, Australia, New Guinea, and the Russian Far East, we utilize "American Indian", "Australian", "Papuan", and "Paleo-Siberian" as groups of multiple language families and *language isolates* occurring in these regions. In addition to that, we mark the Wiktionary editions on the constructed languages Esperanto, Ido, Interlingua, Interlingue, Lojban, and Volapük, the dead languages Latin and Old English, and the Simple Wiktionary as "Other". The *Simple Wiktionary* is a language edition that uses only a controlled vocabulary of English for its lexicographic descriptions. The idea of this edition is to provide easier descriptions for "people who do not speak English well".[43]

Figure 4.1 shows the number of language editions and the number of article pages per language family as well as their geographical distribution. Although each continent is covered by at least one Wiktionary language edition, there are large differences in the corresponding dictionary size. Europe, Australia, and the Americas are well-covered because of the large proportion of major languages from the Indo-European family spoken there. The Indo-European languages constitute the largest language family with over 2.7 billion speakers (Lewis, 2009). A Wiktionary edition exists for each of its main branches and about half of all active Wik-

[41] `http://www.globalwordnet.org/gwa/wordnet_table.html` (28 March 2013)

[42] We are aware that some language families are subject to discussion (e.g., the classification of the Korean language), and that a clear allocation to certain geographic regions is very fuzzy and debatable. Nevertheless, we do not aim at a full ethnological study but at gaining insights into the type of languages for which a Wiktionary language edition exists.

[43] `http://simple.wiktionary.org/w/index.php?oldid=262811` (23 January 2013)




| Language family | Editions | Articles |
|---|---|---|
| Indo-European | | |
| *Albanian* | 1 | 7,400 |
| *Armenian* | 1 | 7,300 |
| *Baltic* | 2 | 615,800 |
| *Celtic* | 6 | 46,010 |
| *West Germanic* | 8 | 3,058,900 |
| *Greek* | 1 | 303,000 |
| *Indo-Aryan* | 13 | 128,597 |
| *Iranian* | 4 | 257,973 |
| *Nordic* | 6 | 312,231 |
| *Romance* | 14 | 2,858,264 |
| *Slavic* | 15 | 819,400 |
| Afro-Asiatic | | |
| *Berber* | 0 | 0 |
| *Chadic* | 0 | 0 |
| *Cushitic* | 2 | 472 |
| *Semitic* | 4 | 60,694 |
| Altaic | | |
| *Mongolian* | 1 | 511 |
| *Tungusic* | 0 | 0 |
| *Turkic* | 8 | 331,019 |
| American Indian (areal) | 4 | 7,941 |
| Australian (areal) | 0 | 0 |
| Austro-Asiatic | 2 | 233,300 |
| Austronesian | 10 | 2,141,494 |
| Basque | 1 | 37,000 |
| Caucasian | 1 | 5,600 |
| Dravidian | 4 | 673,000 |
| Eskimo-Aleut | 2 | 1,452 |
| Japonic | 1 | 87,000 |
| Khoisan | 0 | 0 |
| Korean | 1 | 351,000 |
| Niger-Congo | 7 | 19,328 |
| Nilo-Saharan | 0 | 0 |
| Paleo-Siberian (areal) | 0 | 0 |
| Papuan (areal) | 1 | 148 |
| Sino-Tibetan | | |
| *Chinese* | 2 | 1,217,000 |
| *Tibeto-Burman* | 1 | 120,000 |
| Tai-Kadai | 2 | 83,000 |
| Uralic | 3 | 500,000 |
| Other | 9 | 290,545 |

Figure 4.1: Number of language editions and article pages per language family (bottom) and their geographical distribution (top). Darker colors indicate a higher number of article pages, lighter colors a lower number. The image has been modified by the author based on the "Human Language Families Map" by Wikipedia user Industrius, available under the Creative Commons License (CC BY 3.0) from `http://en.wikipedia.org/wiki/File:Human_Language_Families_Map.PNG` (2 October 2012).

tionary editions describe Indo-European languages. Languages spoken in Asia and Oceania are also well backed by corresponding Wiktionary editions, including languages spoken in China, India (both from the Indo-Aryan and the Dravidian families), and Korea.

With the exception of Malagasy and South African English, only a few Wiktionary editions exist for languages spoken in Africa. There are no active editions for languages from the Nilo-Saharan family (including, for example, the Maasai language), the Berber family (e.g., Tarifit), or the Khoisan family. The latter are known for their click sounds which might have impeded the creation of a corresponding Wiktionary edition due to the complicated pronunciation rules and script of the words (e.g., *ǂqʻáa ká*; English: mud). Nevertheless, we find seven editions on Niger-Congo languages, four editions on Semitic languages, and two editions on Cushitic languages, including Swahili, Amharic, and Oromo. Prinsloo (2010, p. 183) notes that also the professional lexicography of African languages is in a "development phase". In this respect, a collaborative dictionary can be of interest for obtaining lexicographic descriptions.

Besides many African languages, Paleo-Siberian and Tungusic languages spoken in East Russia, Siberia, and the Manchuria do not account for any Wiktionary edition. Many of such languages are endangered because of their small number of speakers, which might explain the lack of respective language editions. The languages of Native Americans and Australian aborigines as well as the hundreds of languages in the New Guinea region are also hardly covered by any language edition. The few exceptions are Guarani (from the Tupian language family), Quechua (Quechuan), Nahuatl (Uto-Aztecan), Cherokee (Iroquoian), Aymara (Aymaran), and the Creole language Tok Pisin. The lack of written knowledge about indigenous languages as well as the missing technical infrastructure in cultures that live close to nature might be the reasons for the absence of respective Wiktionary editions.

Although the vast majority of Wiktionary articles concern the most widespread languages, also the less developed Wiktionary editions make a valuable contribution, since we get in a position to obtain more lexicographic resources for smaller languages. Wiktionary is continually growing, which lets us expect the number of Wiktionaries for minor languages to reach a considerable size in the future. In this context, Wiktionary can provide an important, easy-to-use platform for linguists who study endangered or yet unclassified languages and want to share their research.

**Case Study: Greenlandic.** Greenlandic with its main dialect Kalaallisut is a language of the Eskimo-Aleut family. Since 2009, it is the official language of Greenland spoken by about 50,000 speakers. The language is considered vulnerable by the UNESCO *Atlas of the World's Languages in Danger* (Moseley, 2010), which means that its use is "restricted to certain domains (e.g., home)".[44] There are only a few resources for Greenlandic. One example is the *Greenlandic English Dictionary* (2007 f.) with about 20,000 articles. It is being published online by the Greenland Language Secretariat Oqaasileriffik and the Education Sciences department

[44] `http://www.unesco.org/culture/languages-atlas/en/atlasmap/language-id-687.html` (1 February 2013)

Inerisaavik of the University of Greenland as a revised version of the *Dictionary of the West Greenland Eskimo language* (1927) and the *Grønlandske Ordbog: Grønlandsk–Dansk* (1926). Oqaasileriffik is aware of the "scarcity of dictionary material from Greenlandic into English".[45]

The Greenlandic Wiktionary edition currently contains about 1,200 articles. Although there are many gaps in this Wiktionary edition, we could find information, which is not covered by the *Greenlandic English Dictionary* (2007 f.), for example, the article on the noun *paarlaaffik* (English: ski lift). The multilingual nature of Wiktionary presents a great opportunity for such resource-poor languages, since the lexicographic documentation is not limited to professional lexicographers, whose work usually relies on funding. Wiktionary has the potential to grow rapidly, since it is open to both native speakers and language learners having different native tongues. There is, to the best of our knowledge, no dedicated Greenlandic–Italian dictionary.[46] From consulting Wiktionary, we can, however, learn that *paarlaaffik* translates to the Italian word *seggiovia*. In addition to that, other Wiktionary language editions encode translations into Greenlandic. There are currently, for example, 668 Greenlandic translations encoded in the English Wiktionary, 26 in the German Wiktionary, and 225 in the Russian Wiktionary. Presumably, those projects could benefit much if professional lexicographers and lay authors combine their efforts to systematically close the gaps of the dictionary.

## 4.4 Coverage of Lexical Entries

By taking a closer look at the eight dictionaries introduced in section 4.1, we now compare their lexicographic descriptions at the lexical level. We distinguish between lexical items and headwords in our analysis. As a *lexical item*, we define any unique word form that appears as a target of the lexicographic descriptions (see also Atkins and Rundell, 2008). This can be single words (such as *plant*) or *multiword expressions* (e.g., *freedom of speech*). In Wiktionary, the lexical items correspond to the titles of the article pages. A *headword* is a lexical item that is additionally characterized by its language, part of speech, and etymology. The headword is usually the first information item of each *lexical entry* (which, in turn, refers to the sum of information items described for the headword). Recall the distinction between article page and lexical entry from section 2.6: There is a single article page on *boat*, which encodes six lexical entries in five languages. Correspondingly, we find six headwords for the one lexical item *boat* (i.e., the English, Finnish, and West Frisian nouns and the English, Latin, and Malay verbs). For the two homonymous lexical entries on *plant*, we also count two separate headwords. Table 4.2 shows the number of lexical items and headwords found in our eight dictionaries.

**Lexical items**. The first step towards our quantitative study is to establish the conditions for a fair comparison. This is why we decompose the *synsets* (i.e., sets of synonymous word

[45] `http://www.oqaasileriffik.gl/en/resources/greenlandicenglishdictionary` (15 February 2013)

[46] Leaving aside dictionaries that are based on Wiktionary data, such as `http://www.glosbe.com`.

senses) found in the wordnets into lexical items, headwords, and word senses. We focus on the lexical items on the native language of each Wiktionary edition, as the other dictionaries are monolingual. In addition to that, we filter out Latin terms from the biological taxonomy predominantly found in the wordnets and the inflected word forms encoded as separate lexical items in Wiktionary. This yields the number of *comparable lexical items* shown in table 4.2.

The English Wiktionary exceeds the size of the Princeton WordNet by about two times and that of the Roget's thesaurus by more than five times. The difference is not as big for the Russian Wiktionary, which exceeds the Russian WordNet by about 36,000 lexical items. As opposed to that, the German Wiktionary is slightly smaller than both GermaNet and OpenThesaurus. One reason for this might be the differences in the use of references, which we observed in section 3.4: because references are to be provided for each article in the German Wiktionary, this seems to prevent some authors from contributing. Besides comparing the absolute sizes of the dictionaries, we also analyze which types of word forms are predominantly found in one of the dictionaries by assessing the coverage of basic vocabulary and neologisms.

**Basic vocabulary**. The basic vocabulary of a language is known to change very slowly and should be well-represented in a general dictionary. Table 4.3 shows the proportion of covered lexical items from several basic vocabulary word lists. We use the *Swadesh lists* (Dyen et al., 1992) for English, German, and Russian; Ogden's (1938) *Basic English word list*, West's (1953) *General Service List* (GSL), and Nation's (2006) BNC 1–4 lists based on the *British National Corpus* (1991–1994) for English; the *GUT1 Wortschatz*[47] 100 and 500 for German; and Штейнфельдт's (1963) list of common terms in modern Russian (Steinfeldt).

Each Wiktionary edition covers the basic vocabulary very well. The English Wiktionary seems to be the most thorough, as it is the only dictionary that covers the full Swadesh list and over 99 % of the other word lists. While the other dictionaries also have a good coverage of the English basic vocabulary, their coverage is much lower for German and Russian. Wiktionary can help closing this gap, as it retains a high coverage of over 97 %.

**Neologisms**. We also assess the coverage of newly coined words using a list of 555 English neologisms[48] from 1997 to 2008 provided by the Birmingham City University, a list of 36,220 German neologisms[49] taken from the Wortwarte project for the years 2000–2010, and 7,482 Russian neologisms[50] provided by the Russian Academy of Sciences. Note that, due to the different size of the neologism lists and the different language characteristics, the numbers are not comparable across the three languages. This particularly holds for the very low coverage

[47] `http://www.gut1.de/download/download.html` (24 May 2012)
[48] `http://rdues.bcu.ac.uk/neologisms.shtml` (24 May 2012)
[49] `http://www.wortwarte.de/` (24 May 2012)
[50] `http://dict.ruslang.ru/gram.php?act=search&orderby=word` (24 May 2012)

| | English dictionaries | | | German dictionaries | | | Russian dictionaries | |
|---|---|---|---|---|---|---|---|---|
| | Wiktionary | WordNet | Thesaurus | Wiktionary | GermaNet | Thesaurus | Wiktionary | WordNet |
| *Lexical items* | 3,296,943 | 148,730 | 59,391 | 245,100 | 89,816 | 76,325 | 576,136 | 130,062 |
| Foreign languages | 2,802,289 | 7,080 | 22 | 128,565 | 20 | 20 | 414,505 | 5,025 |
| Inflected forms | 163,024 | 0 | 0 | 49,795 | 0 | 0 | 378 | 0 |
| Comparable | 331,630 | 141,650 | 59,369 | 66,740 | 89,796 | 76,305 | 161,253 | 125,037 |
| *Headwords* | 364,663 | 149,502 | 62,797 | 69,631 | 89,832 | 78,502 | 163,456 | 126,224 |
| Noun | 201,461 | 97,534 | 29,854 | 39,227 | 72,763 | 40,940 | 66,312 | 97,224 |
| Verb | 32,111 | 11,531 | 15,150 | 5,119 | 8,812 | 11,991 | 21,410 | 8,995 |
| Adjective | 82,054 | 21,156 | 12,739 | 6,759 | 8,257 | 16,595 | 27,419 | 16,048 |
| Adverb | 14,013 | 4,481 | 3,017 | 867 | 0 | 1,562 | 5,815 | 3,885 |
| Named entity | 19,715 | 14,213 | 0 | 5,668 | 0 | 7,027 | 16,840 | 72 |
| Phraseme | 2,128 | 0 | 2,037 | 2,196 | 0 | 4 | 1,043 | 0 |
| Interjection | 1,871 | 0 | 0 | 200 | 0 | 1 | 156 | 0 |
| Abbreviation | 7,174 | 0 | 0 | 3,505 | 0 | 0 | 299 | 0 |
| Affix | 1,820 | 0 | 0 | 539 | 0 | 22 | 241 | 0 |
| Pronoun | 408 | 0 | 0 | 129 | 0 | 77 | 93 | 0 |
| Determiner | 469 | 587 | 0 | 117 | 0 | 31 | 160 | 0 |
| Adposition | 484 | 0 | 0 | 149 | 0 | 124 | 143 | 0 |
| Conjunction | 227 | 0 | 0 | 88 | 0 | 76 | 49 | 0 |
| Other tags | 728 | 0 | 0 | 5,068 | 0 | 52 | 23,476 | 0 |

Table 4.2: Comparison of lexical items and headwords

| English dictionaries | Wiktionary | WordNet | Thesaurus | List size |
|---|---|---|---|---|
| Swadesh list | 205 (100.0 %) | 190 (92.7 %) | 197 (96.1 %) | 205 |
| Nation's BNC 1–4 | 3,997 (99.9 %) | 3,909 (97.7 %) | 3,620 (90.5 %) | 3,999 |
| West's GSL | 2,282 (99.9 %) | 2,221 (97.2 %) | 2,204 (96.5 %) | 2,284 |
| Ogden's Basic English | 845 (99.4 %) | 824 (96.9 %) | 829 (97.5 %) | 850 |
| Neologisms | 78 (14.1 %) | 4 (0.7 %) | 1 (0.2 %) | 555 |

| German dictionaries | Wiktionary | GermaNet | Thesaurus | List size |
|---|---|---|---|---|
| Swadesh list | 214 (98.6 %) | 189 (87.1 %) | 206 (94.9 %) | 217 |
| GUT1 Wortschatz 100 | 98 (99.0 %) | 76 (76.8 %) | 91 (91.9 %) | 99 |
| GUT1 Wortschatz 500 | 500 (99.6 %) | 363 (72.3 %) | 430 (85.7 %) | 502 |
| Neologisms: | 214 (0.6 %) | 705 (1.9 %) | 249 (0.7 %) | 36,220 |

| Russian dictionaries | Wiktionary | WordNet | | List size |
|---|---|---|---|---|
| Swadesh list | 224 (97.0 %) | 195 (84.4 %) | | 231 |
| Steinfeldt | 2,451 (97.8 %) | 1,701 (67.9 %) | | 2,506 |
| Neologisms | 3,822 (51.1 %) | 353 (4.7 %) | | 7,482 |

Table 4.3: Coverage of basic vocabulary words and neologisms

of German neologisms, which stems from many *hapax legomena* included in Wortwarte that one would usually not expect to be found in a dictionary.

Both the English and the Russian Wiktionary editions encode significantly more neologisms than their respective machine-oriented dictionaries. This can be explained by the collaborative construction approach of Wiktionary, which allows updating the dictionary at any time, without being restricted to certain release cycles as it is the case for most editorial dictionaries. In contrast, the German Wiktionary encodes only 0.6 % of the neologisms; 491 fewer than GermaNet. This can again be explained by the expectation of providing references in the German Wiktionary. Neologisms are often not tolerated unless they are widely used and described in other reference works as well. The commonly used internet slang word *lol* is, for instance, still marked as "sprachwissenschaftlich *nicht* erfasst" (English: not linguistically covered), which highlights the strong dependency of the German Wiktionary on secondary sources.[51]

**Parts of speech**. Table 4.2 shows the number of headwords encoded in each dictionary and their part of speech distribution. As described above, we separate out headwords that are

[51] `http://de.wiktionary.org/w/index.php?oldid=2847684` (14 February 201)

not directly comparable (i.e., foreign language entries and inflected word forms). The English Wiktionary is again the largest dictionary. It encodes more than twice as many nouns, verbs, adjectives, and adverbs as WordNet and Roget's thesaurus. The German Wiktionary is the smallest dictionary in comparison to GermaNet and OpenThesaurus. Verbs seem to be particularly under-represented. The Russian Wiktionary encodes more verbs, adjectives, and adverbs than the Russian WordNet but, in turn, contains a lower number of nouns.

In total, we found 69 different part of speech tags within the three Wiktionary editions. Since many tags are very fine-grained, we grouped them into the fourteen general categories shown in the table. The Wiktionary community uses, for instance, three different tags for abbreviations: initialisms (pronounced letter by letter, e.g., *CD* for *Compact Disc*), *acronyms* (pronounced like a regular word, e.g., *ROM* for *read only memory*), and abbreviations terminated by a full stop (such as *Apr.* for *April*). A similar distinction is made for pronouns (demonstrative, reflexive, and possessive pronouns), particles (comparative, intensifying, and answering particles), affixes (prefixes and suffixes), and *phrasemes*. The latter are tagged as proverbs (e.g., *love is blind*), idioms (e.g., *in the same boat*), or collocations (like *strong tea*). Wiktionary encodes a high number of phrasemes. This is particularly useful in combination with the corresponding translations into other languages, since idioms and proverbs are usually hard to translate. The high number of named entities in the English Wiktionary is also notable. In comparison to the English WordNet, we predominantly find given names (e.g., *Alice* or *Nadine*), and toponyms (e.g., *Berlin* or *Ohio*) in Wiktionary, as well as named entities from the non-US culture (such as the Arabic broadcaster *Al Jazeera* or the Swiss canton *Aargau*. Interestingly, phrasal verbs (like *turn off*), as well as compounds and multiword expressions (like *toothpaste* or *academic discipline*) do not receive a special tag.

**Overlap of lexical entries**. To examine whether the dictionaries largely overlap or contain complementary information, we aligned the lexical entries that share the same headword. We do not perform any linguistic preprocessing such as stemming or case-folding in order to account for different spelling variants and grammatical forms. Figure 4.2 shows a Venn diagram of the number of lexical entries shared by each pair of dictionaries. We find that the total overlap of the dictionaries is very small. For the English language, only 8 % of the lexical entries in Wiktionary, 19 % of the entries in WordNet, and 47 % of the entries in Roget's thesaurus occur as headwords within the respective other dictionaries. The highest number of lexical entries is shared by Wiktionary and WordNet. In comparison to the total number of lexical entries, it is, however, still quite low.

This is a surprising result, since one would expect two general language dictionaries to encode, by and large, a similar list of headwords. We particularly find named entities (e.g., *Grammy*), multiword expressions (e.g., *grain of salt*), and domain-specific terms to be encoded in only one of the dictionaries. In Wiktionary, we predominantly observe terms from information sciences (e.g., *sound card*), natural sciences (e.g., *benzoyl*), and sports (e.g., *libero*),

as well as informal (e.g., *ear candy*), dialectal (e.g., from the Geordie dialect), and archaic terms (e.g., *abaculus*). In WordNet, we find terms from the biological or medical domain (e.g., the *napa* plant, or the *axial muscle*), named entities (e.g., *Fourth Council of Constantinople*, *Horatio Walpole*) and numerous headwords covering shades of color (such as *reddish-pink*). While nouns represent the main group of headwords found in only one dictionary, we also observe verbs (e.g., *relaunch* in Wiktionary, *louden* in WordNet) and adjectives (*superfluid* in Wiktionary, *ventilated* in WordNet) among them. The overlap between the lexical entries is similarly small for the Russian dictionaries, and – although slightly higher – also the German dictionaries reveal large differences in their coverage.

As a consequence, we consider the combination of multiple dictionaries a viable option in order to increase the lexical coverage of a dictionary. We will further discuss this topic in chapter 6.

## 4.5 Coverage of Word Senses

Each lexical entry can distinguish multiple *word senses* describing a certain meaning of a word. The English noun *boat* shown in figure 2.1 encodes, for instance, three word senses (i.e., the vehicle, the poker term, and the cyclohexane conformation). Table 4.4 compares the total number of word senses found in the eight dictionaries. The English Wiktionary encodes the most word senses: more than twice the number of WordNet and over four times as many as Roget's thesaurus. The German dictionaries do not differ much in this respect: Wiktionary, GermaNet, and OpenThesaurus are neck and neck. The Russian Wiktionary encodes a lower number of word senses than the Russian WordNet.

**Degree of polysemy**. Comparing the absolute number of word senses only allows us to draw limited conclusions, since lexicographers can choose different sense granularities for their sense descriptions. Thus, a higher number of word senses does not necessarily imply a higher coverage of meanings. This is why we compare the number of word senses per lexical entry (i.e., the *degree of polysemy* of a lexical entry).

Table 4.4 shows the number of stubs, monosemous, and polysemous entries in our eight dictionaries as well as the average and maximum number of word senses per entry. Editorial dictionaries do not contain lexical entries without any word senses. This is different in Wiktionary, where users may encode entries without providing descriptions for each article constituent. An author can, for instance, describe the pronunciation of a word, but leave the formulation of sense definitions to other authors. We call such lexical entries *stubs*. The low number of stubs in the English Wiktionary indicates that it is in a stable state and contains definitions for the vast majority of the encoded entries. This is different for the Russian Wiktionary, which lacks word sense definitions for 56 % of its entries and hence requires much work in order to close its gaps.

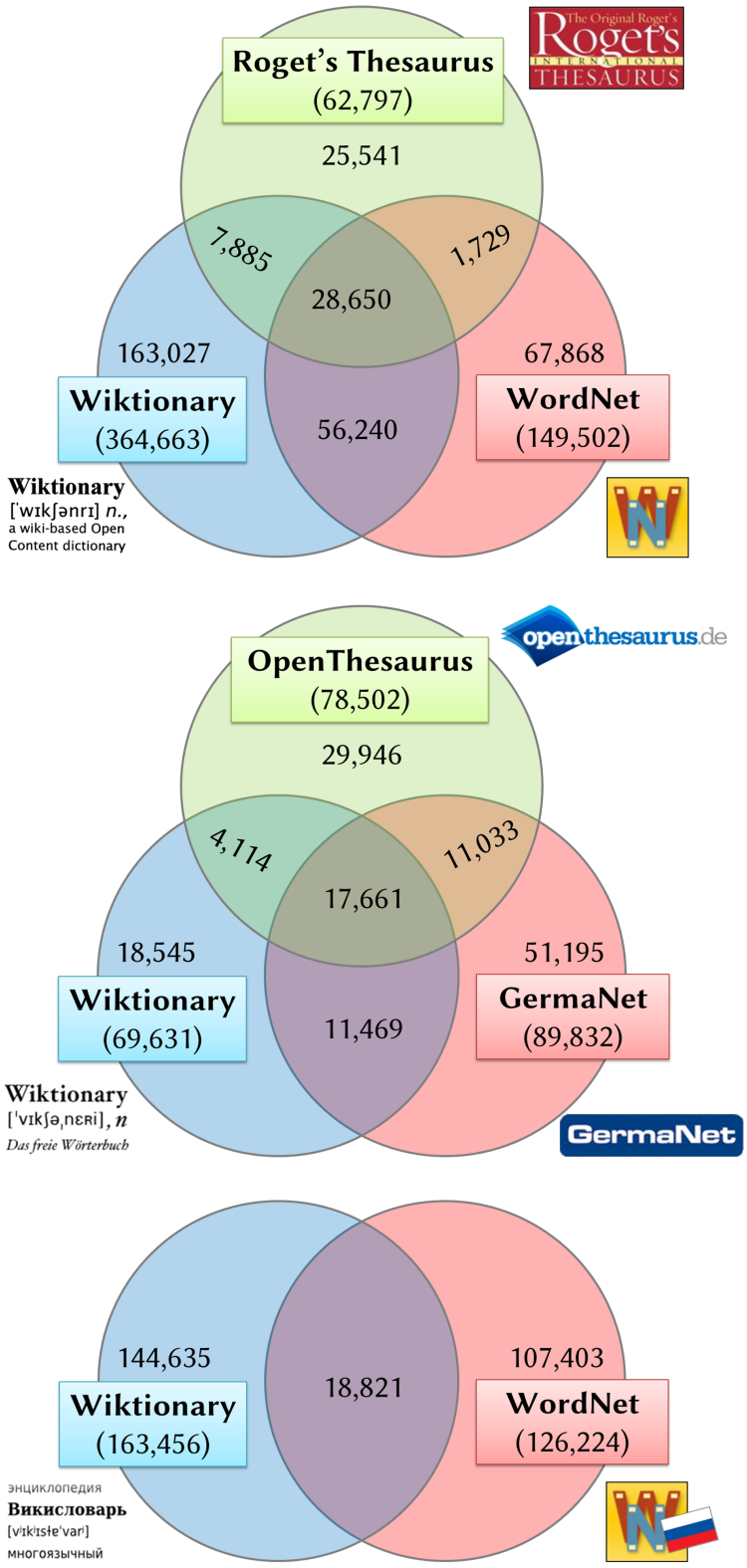



Figure 4.2: Overlap of lexical entries between the English (top), German (middle), and Russian (bottom) dictionaries. The total number of lexical entries is shown in parentheses.

| | English dictionaries | | | German dictionaries | | | Russian dictionaries | |
|---|---|---|---|---|---|---|---|---|
| | Wiktionary | WordNet | Thesaurus | Wiktionary | GermaNet | Thesaurus | Wiktionary | WordNet |
| Word Senses | 470,691 | 206,978 | 98,464 | 99,263 | 102,999 | 96,054 | 130,629 | 182,448 |
| Stubs | 78 | 0 | 0 | 1,881 | 0 | 0 | 74,008 | 0 |
| Monosemous entries | 304,826 | 123,187 | 44,299 | 50,453 | 80,813 | 67,532 | 67,964 | 105,385 |
| *Nouns* | 185,954 | 96,570 | 21,258 | 35,019 | 67,013 | 42,572 | 35,204 | 84,569 |
| *Verbs* | 22,964 | 6,280 | 10,193 | 2,426 | 6,323 | 9,518 | 4,156 | 5,145 |
| *Adjectives* | 71,132 | 16,589 | 8,644 | 4,808 | 7,477 | 15,231 | 4,378 | 12,428 |
| *Other tags* | 24,776 | 3,748 | 4,204 | 8,200 | 0 | 211 | 24,226 | 3,243 |
| Polysemous entries | 59,759 | 26,315 | 18,498 | 17,297 | 9,019 | 10,970 | 21,484 | 20,839 |
| *Nouns* | 35,180 | 15,384 | 8,596 | 9,739 | 5,750 | 5,395 | 10,224 | 12,688 |
| *Verbs* | 9,143 | 5,251 | 4,957 | 2,688 | 2,489 | 2,473 | 6,326 | 3,850 |
| *Adjectives* | 10,921 | 4,947 | 4,095 | 1,955 | 780 | 3,042 | 3,089 | 3,659 |
| *Other tags* | 4,515 | 733 | 850 | 2,915 | 0 | 60 | 1,845 | 642 |
| Avg. senses per entry | 1.29 | 1.34 | 1.57 | 1.46 | 1.15 | 1.22 | 1.46 | 1.41 |
| *Nouns* | 1.28 | 1.24 | 1.56 | 1.37 | 1.11 | 1.17 | 1.38 | 1.28 |
| *Verbs* | 1.59 | 2.17 | 1.67 | 2.03 | 1.52 | 1.32 | 2.45 | 2.35 |
| *Adjectives* | 1.20 | 1.39 | 1.61 | 1.45 | 1.11 | 1.28 | 1.72 | 1.64 |
| *Other tags* | 1.14 | 1.25 | 1.24 | 1.51 | — | 1.41 | 1.57 | 1.33 |
| Max. senses per entry | 66 | 59 | 18 | 52 | 26 | 14 | 30 | 54 |
| *Nouns* | 55 | 33 | 18 | 37 | 15 | 14 | 19 | 54 |
| *Verbs* | 58 | 59 | 16 | 35 | 26 | 12 | 30 | 49 |
| *Adjectives* | 22 | 27 | 16 | 15 | 10 | 10 | 14 | 48 |
| *Other tags* | 66 | 13 | 7 | 52 | — | 12 | 25 | 22 |

Table 4.4: Comparison of word senses and degree of polysemy

Between 80 % and 90 % of the entries in the machine-oriented dictionaries are *monosemous* (i.e., they encode only one word sense). In the English Wiktionary, we find the number of monosemous entries in the same range (81 %). The German Wiktionary, however, contains only 68 % monosemous entries and hence encodes a higher number of *polysemous entries* (i.e., entries encoding more than one sense). A possible explanation for this discrepancy is that the Wiktionary community is more likely to create articles for polysemous words, since they can cause confusion when understanding a text and are thus felt to be more important to describe. This also applies to the Russian Wiktionary, in which 72 % of the entries are monosemous.

The average number of encoded word senses is similar in all the dictionaries, ranging between 1.15 and 1.57. The largest difference can be observed for verbs. The English WordNet is known to be very fine-grained (cf. Palmer et al., 2007) and therefore generally encodes a higher number of verb senses than Wiktionary and Roget's thesaurus. When comparing the German Wiktionary and GermaNet, we find, however, more word senses per Wiktionary entry. This is in line with our finding that more polysemous words are encoded there.

**Polysemic difference**. The English Wiktionary has exactly one verb with 58 word senses, while WordNet has exactly one with 59 word senses. This seems to show strong similarity. The verb in question is, however, *(to) break* in WordNet and *(to) go* in Wiktionary. To accommodate this issue in our analysis, we define the *polysemic difference*

$$\Delta_{d_1,d_2}(\ell) = |senses_{d_1}(\ell) - senses_{d_2}(\ell)|$$

as the difference in the number of word senses of a lexical entry $\ell$ encoded in the dictionaries $d_1$ and $d_2$. The verb *(to) break* from the example above has 27 word senses in Wiktionary and hence yields a polysemic difference of $|59 - 27| = 32$.

In the English Wiktionary, 61 % of the entries shared with WordNet and 42 % of the entries shared with Roget's thesaurus have a polysemic difference of zero (i.e., they encode the same number of word senses). This is even higher for the German Wiktionary: 66 % of the entries shared with GermaNet and 53 % of the entries shared with OpenThesaurus have $\Delta(\ell) = 0$. Over 90 % of the English and German Wiktionary entries have a polysemic difference of not more than two (i.e., $\Delta(\ell) \leq 2$) when compared to the wordnets and thesauri. This indicates that the sense granularity of the lexical entries shared by these dictionaries does not differ considerably.

**Overlap of word senses**. Although the adjective *buggy* has two word senses in WordNet and three word senses in Wiktionary (i.e., $\Delta(buggy) = 1$), this does not necessarily imply that two of the described meanings are identical. In fact, we find only the meaning "infested with bugs" in both dictionaries. Wiktionary additionally encodes "resembling an insect" and "containing programming errors", while WordNet defines "informal or slang terms [sic!] for mentally irregular". In order to gain a clearer insight into the coverage of word senses, we

need to align the word senses of the dictionaries with each other and quantify the number of shared word senses – similar to our study concerning the overlap of lexical entries reported in the previous section. Aligning the dictionaries at the level of lexical entries can be achieved using a simple headword matching (i.e., finding the adjective *buggy* in both dictionaries). An alignment at the level of word senses is, however, a very complex task, which requires the identification of highly similar sense definitions. We address this issue in chapter 6.

**Sense ordering**. Fellbaum (1998) notes that the word senses in WordNet are ordered according to the occurrence frequencies in *SemCor* (1993 f.). This promotes the most frequently used word sense to the first position, which is a common strategy in practical lexicography (see Atkins and Rundell, 2008). However, using a corpus such as *SemCor* to obtain these occurrence frequencies might not yield very realistic data because sense-tagged corpora are usually very small and often limited to certain text types (e.g., newspaper articles). Although there is no specific guideline for the sense ordering in Wiktionary, we observed that the first entry is often the most frequently used one. For the noun *tattoo*, the first word sense in Wiktionary is "an image made in the skin with ink and a needle", but "a drumbeat or bugle call that signals the military to return to their quarters" in WordNet. Intuitively, the Wiktionary word sense is the more frequently used one nowadays. We confirmed this intuition using the *British National Corpus* (1991–1994), where we found 42 % of 180 sentences containing *tattoo* referring to the meaning of a mark in the skin (as opposed to 28 % in the military sense). Hence, the sum of subjective opinions on the usage of word senses that coins Wiktionary's sense ordering can alleviate the limitations and sparseness of sense-tagged corpora and provide a viable resource for ordering word senses by usage.

## 4.6 Coverage of Pragmatic Labels

Following the practice of other dictionaries, Wiktionary encodes *pragmatic labels* to mark non-standard varieties of language, for example, poetic expressions, Yorkshire dialect, or technical terms used in archeology.[52] Pragmatic labels are specified before the definition text of a word sense and usually enclosed in parentheses, typed in italics, or separated by a colon.

About 34 % of the word senses in the English Wiktionary, 28 % in the Russian Wiktionary, and even 43 % in the German Wiktionary are marked by at least one pragmatic label. This indicates that Wiktionary provides many lexicographic descriptions on non-standard language varieties (word senses from the standard variety usually remain unmarked). Although there are lists of commonly used pragmatic labels, the Wiktionary authors are free to introduce

[52]There is a large variety of other terms for pragmatic labels used in the literature, including *semantic labels* (ISO 24613, 2008), *linguistic labels* (Atkins and Rundell, 2008), *lexicographic labels* (Beyer, 2011), *usage labels*, and *diasystematic labelling* (Hartmann and James, 1998).

additional labels, which yields a high number of different labels: 1,132 in the English Wiktionary, 1,308 in the German Wiktionary, and 179 in the Russian Wiktionary. The labels differ in their granularity (e.g., *calculus* and *mathematics*) and often use highly related expressions (e.g., *programming*, *programming language*, and *software engineering*). We can distinguish the following main types of pragmatic labels:

- *Domain labels* mark the subject field or technical domain of a word sense (the *diatechnical variety* of language), for example, *chemistry*, *calculus*, *sports*.
- *Geographical labels* mark regional or dialectal expressions (the *diatopic variety*), for example, *Yorkshire*, *Scottish*, *Ireland*.
- *Sociolectal labels* mark jargon used by a certain culture, social group, or social class (the *diastratic variety*), for example, *army slang*, *argot*, *computer jargon*, *working class*, *children's language*.
- *Register* and *style labels* mark word senses used in certain communicative situations (the *diaphasic variety*). The former refer to a particular social setting including *formal*, *familiar*, and *slang*. The latter indicate a stylistic variation, such as *poetic*, *humuorous*, as well as *literal* and *figurative* meanings.
- *Temporal labels* mark the period of time a word sense is used in (the *diachronic variety*), for example, *archaic*, *19th century*, *nonce word*.
- *Evaluative labels* mark offensive terms and word senses with a certain connotation (the *diaevaluative variety*), for example, *approving*, *rude*, *derogatory*.
- *Normative labels* are prescriptive markings used for expressions that deviate from cultural standards (the *dianormative variety*), for example, *incorrect*, *non-standard*, *hypercorrect*.
- *Frequency labels* indicate how often a term is used (the *diafrequential variety*), for example, *rare*, *less common*, *often*.
- *Syntax labels* specify grammatical properties of a word sense. They do not belong to the pragmatic labels in the narrower sense, but they are encoded at the same position of the article. Examples are *intranstive*, *ergative*, *countable*.

We manually group the pragmatic labels used in the English, German, and Russian Wiktionary editions according to these types.[53] Table 4.5 shows the number of labels and the number of marked word senses for each label type. The majority of the pragmatic labels are domain labels, which we describe in more detail below. Among the other types of labels, we notice a large number of dialectal and slang expressions. The latter also represent the scope of another large collaborative dictionary, the *Urban Dictionary* (1999 f.), and thus represent a variety of language that can particularly benefit from collaborative approaches.

[53] We only consider labels used at least three times, and we count a label twice if it belongs to two label types, such as *Australian slang*, which denotes both a geographical and a register label.

| Label type | English Wiktionary Labels | English Wiktionary Word senses | German Wiktionary Labels | German Wiktionary Word senses | Russian Wiktionary Labels | Russian Wiktionary Word senses |
|---|---|---|---|---|---|---|
| Domain | 794 | 99,460 (45 %) | 839 | 42,480 (57 %) | 119 | 15,148 (42 %) |
| Geographical | 144 | 23,479 (11 %) | 130 | 3,779 (5 %) | 7 | 344 (< 1 %) |
| Sociolectal | 15 | 528 (< 1 %) | 34 | 784 (1 %) | 1 | 26 (< 1 %) |
| Register | 10 | 14,837 (7 %) | 24 | 6,385 (9 %) | 10 | 2,741 (8 %) |
| Style | 35 | 12,922 (6 %) | 67 | 4,043 (5 %) | 14 | 10,950 (30 %) |
| Evaluative | 20 | 3,789 (2 %) | 40 | 1,664 (2 %) | 10 | 1,024 (3 %) |
| Normative | 6 | 1,096 (< 1 %) | 9 | 35 (< 1 %) | 0 | 0 – |
| Temporal | 29 | 26,617 (12 %) | 30 | 3,862 (5 %) | 3 | 3,539 (10 %) |
| Frequency | 10 | 6,367 (3 %) | 9 | 351 (< 1 %) | 1 | 125 (< 1 %) |
| Syntax | 69 | 30,594 (14 %) | 126 | 11,463 (15 %) | 14 | 2,308 (6 %) |
| Total | 1,132 | 219,689 (100 %) | 1,308 | 74,846 (100 %) | 179 | 36,205 (100 %) |

Table 4.5: Comparison of pragmatic labels in the English, German, and Russian Wiktionary editions

A broad coverage of domains, sociolects and dialects can be explained by the community-based approach, as each contributor has a certain field of expertise yielding a broad diversity of the encoded word senses. This has also been confirmed by other researchers: Rundell (2012, p. 80) notes that "you can be an expert on *homeopathy*, *permafrost* or the *nitrogen cycle*, but not on *decide*, *limitation*, or *dull*". This suggests that collaborative lexicography is very valuable for technical terminology and expressions colloquially used in everyday speech – terms that are also rarely found in corpora. But in terms of general language, the collaborative dictionaries seem to largely rely on editorial dictionaries, as observed by the vast use of secondary sources in section 3.4.

**Domain labels**. In table 4.6, we compare Wiktionary's domain labels with *WordNet Domains 3.2* (2007) introduced by Bentivogli et al. (2004). *WordNet Domains* marks 128,669 word senses (62 %) of *WordNet 2.0* (2003) with 157 different domain labels. We group similar labels into the 26 general categories reported in the table. The labels *cycling* and *weightlifting* are, for example, assigned the category *sports*. For labels that do not fit into one of these categories, we use residual categories labeled "other" (e.g., *numismatics*).

About a quarter of the domain labels in WordNet Domains are from biology, because WordNet covers a large share of the biological taxonomy of plants and animals. The Wiktionaries have a stronger focus on the other natural sciences – most prominently on chemistry with 14,910 word senses in the English Wiktionary. Well-represented domains are also physics (3–5 %), computer science (3–7 %), maths (4–5 %), engineering (4–8 %), medicine (9–11 %), and sports (3–6 %). Clearly under-represented are the humanities and social sciences, which are better covered within WordNet. While linguistics and engineering seem to be predominantly

| Domains | English Wiktionary | German Wiktionary | Russian Wiktionary | WordNet Domains |
|---|---|---|---|---|
| *Humanities* | | | | |
| Architecture | 848 (< 1 %) | 1,429 (3 %) | 183 (1 %) | 4,233 (2 %) |
| Art | 3,388 (3 %) | 2,193 (5 %) | 508 (3 %) | 8,300 (5 %) |
| History | 640 (< 1 %) | 388 (< 1 %) | 1,319 (9 %) | 2,746 (2 %) |
| Linguistics | 2,502 (3 %) | 6,003 (14 %) | 976 (6 %) | 2,913 (2 %) |
| Literature | 641 (< 1 %) | 334 (< 1 %) | 187 (1 %) | 2,671 (2 %) |
| Philosophy | 692 (< 1 %) | 304 (< 1 %) | 225 (1 %) | 712 (< 1 %) |
| Other | 340 (< 1 %) | 90 (< 1 %) | 18 (< 1 %) | 226 (< 1 %) |
| *Social sciences* | | | | |
| Communication | 1,040 (1 %) | 404 (< 1 %) | 147 (< 1 %) | 2,186 (1 %) |
| Economics | 2,383 (2 %) | 1,133 (3 %) | 491 (3 %) | 5,437 (3 %) |
| Law | 3,056 (3 %) | 1,117 (3 %) | 300 (2 %) | 2,723 (2 %) |
| Pedagogy | 412 (< 1 %) | 121 (< 1 %) | 53 (< 1 %) | 1,634 (< 1 %) |
| Politics | 709 (< 1 %) | 564 (1 %) | 251 (2 %) | 2,812 (2 %) |
| Psychology | 483 (< 1 %) | 293 (< 1 %) | 144 (< 1 %) | 3,876 (2 %) |
| Sociology | 317 (< 1 %) | 394 (< 1 %) | 94 (< 1 %) | 4,402 (3 %) |
| Other | 1,002 (1 %) | 360 (< 1 %) | 218 (1 %) | 3,470 (2 %) |
| *Natural sciences* | | | | |
| Biology | 15,315 (15 %) | 6,439 (15 %) | 1,744 (12 %) | 60,940 (36 %) |
| Chemistry | 14,910 (15 %) | 1,768 (4 %) | 579 (4 %) | 8,005 (5 %) |
| Geology | 6,634 (7 %) | 1,477 (3 %) | 964 (6 %) | 10,942 (6 %) |
| Physics | 5,434 (5 %) | 1,421 (3 %) | 652 (4 %) | 4,647 (3 %) |
| Other | 0 – | 25 (< 1 %) | 0 – | 100 (< 1 %) |
| *Structural sciences* | | | | |
| Computer science | 6,855 (7 %) | 1,173 (3 %) | 427 (3 %) | 983 (< 1 %) |
| Math | 5,159 (5 %) | 1,533 (4 %) | 584 (4 %) | 1,495 (< 1 %) |
| *Miscellaneous* | | | | |
| Agriculture | 189 (< 1 %) | 1,281 (3 %) | 215 (1 %) | 585 (< 1 %) |
| Engineering | 4,226 (4 %) | 2,123 (5 %) | 1,220 (8 %) | 4,622 (3 %) |
| Health & nutrition | 296 (< 1 %) | 864 (2 %) | 254 (2 %) | 5,342 (3 %) |
| Medicine | 11,131 (11 %) | 3,883 (9 %) | 1,605 (11 %) | 13,134 (8 %) |
| Military | 1,627 (2 %) | 932 (2 %) | 455 (3 %) | 2,898 (2 %) |
| Religion | 1,996 (2 %) | 1,456 (3 %) | 744 (5 %) | 3,223 (2 %) |
| Sport | 5,509 (6 %) | 2,258 (5 %) | 502 (3 %) | 3,235 (2 %) |
| Other | 1,726 (2 %) | 720 (2 %) | 89 (< 1 %) | 3,139 (2 %) |

Table 4.6: Distribution of domain labels in the English, German, and Russian Wiktionary editions in comparison to WordNet Domains

encoded by the German (14 % and 5 %) and the Russian Wiktionary communities (6 % and 8 %), these domains are more rare in the English Wiktionary (3 % and 4 %). The different focus of the Wiktionary language editions and WordNet Domains can help closing domain-specific vocabulary gaps of a dictionary. We discuss an approach towards this goal in chapter 6.

The different coverage of technical domains additionally provides insight into the composition of the Wiktionary community: our results suggest a higher number of contributors from natural and structural sciences as well as authors working in a technical profession. Moreover, the large proportion of sport-related vocabulary indicates that the voluntary authors focus on topics of their leisure time.

## 4.7 Coverage of Relations

A number of article constituents in Wiktionary encode *relations* in form of hyperlinks connecting two related dictionary articles or article constituents. We use the term *source* for referring to the article (constituent) encoding the relation, and the term *target* to refer to the article (constituent) the relation is pointing to. Following our notation from section 2.4, we can distinguish between *internal relations*, which are realized as internal links (such as synonyms), and *translations* realized as interwiki links spanning two Wiktionary language editions.

**Internal relations**. Internal relations can be subdivided into different *relation types*, which describe the relationship between the source and the target of a relation. *Semantic relations* are defined between two word senses, for which they express a paradigmatic relationship.[54] The semantic relations induce a *semantic network*, in which related word senses are connected by means of a path of semantic relations. This is useful for many natural language processing tasks, such as computing the degree of semantic relatedness by measuring the proximity of two word senses in the semantic network (cf. Budanitsky and Hirst, 2006). We discuss this in more detail in the next chapter, when we construct an ontology from Wiktionary. As opposed to that, *form-based relations* are defined between two lexical entries. They express a morphological or syntagmatic relationship. We distinguish the following relation types:

- *Synonymy* denotes a semantic relation in which the source and the target have the same meaning (e.g., *free* and *gratis*).
- *Antonymy* denotes a semantic relation in which the source and the target have opposite meanings (e.g., *even* and *odd*).
- *Hypernymy* denotes a semantic relation in which the target has more general (broader) meaning than the source (e.g., from *thumb* to *finger*).

[54]The semantic relations in Wiktionary are encoded as lemma-oriented internal links and are hence defined between a word sense and a lexical item. We discuss in section 5.4 how they can be turned into relations between two word senses. We omit this distinction here, since it does not affect the results of our study.

- *Hyponymy* denotes a semantic relation in which the target has more specific (narrower) meaning than the source (e.g., from *finger* to *thumb*).
- *Related term* denotes a semantic relation in which the source and the target have a related meaning, but the type of relationship is not further specified (e.g., between *autodidact* and *self-educated*).
- *Collocation* denotes a syntagmatic relation in which the source and the target occur frequently together, but have a transparent, non-idiomatic meaning (e.g., between *strong* and *tea*).
- *Derived term* denotes a morphological relation in which the target is a morphological derivative of the source (e.g., from *drive* to *driver*).

In addition to that, we use "other types" to refer to several less-frequent types of internal relations, such as siblings (co-hyponymy), part-whole relations (meronymy, holonymy), and etymologically related words.

Table 4.7 shows the number of internal relations encoded by our eight dictionaries. The first and most obvious observation derived from these numbers is that the English Wiktionary contains only a fraction of the relations encoded in WordNet. This means that, although the English Wiktionary shows a good coverage of lexical items and word senses, its induced semantic network is sparse. We find many articles in Wiktionary that do not provide any internal relations yet (e.g., the article *lengthy*), while others are well elaborated in this respect (e.g., *free*). Krizhanovsky (2010) notes that the article style sheet of the Russian Wiktionary fosters the encoding of relations as it provides empty article constituents for them, which can be easily filled by the authors. Another difference is that the English Wiktionary uses linguistic terminology (i.e., hypernyms, antonymy, etc.), whereas the German Wiktionary relies on descriptions, such as broader terms or opposites. This might be more intuitively understandable by lay authors.

We also observe many unidirectional links. As opposed to *bidirectional links*, a *unidirectional link* does not have a counterpart pointing backwards from the target to the source. There is, for instance, a synonymy relation pointing from *free* to *unhindered*, but no symmetric relation (*unhindered, free*) – although this would be a reasonable addition. In section 5.5, we will discuss how the semantic network induced by Wiktionary can be enriched based on this finding.

Though we find the German Wiktionary to be the smallest of the three language editions, it encodes the highest number of internal relations. It exceeds the expert-built GermaNet in terms of synonyms and antonyms. With regard to the very high number of antonyms, we observe that Wiktionary has a rather loose definition for them, since we find cases of co-hyponymy such as the relation between *März* and *November* (English: March; November) as well as cases of loosely related words having contrary meanings, for example, between *subjektiv* and *parteilos* (English: subjective; politically independent). The German Wiktionary is the only language edition explicitly encoding collocations by means of internal relations,

| | English dictionaries | | | German dictionaries | | | Russian dictionaries | |
|---|---|---|---|---|---|---|---|---|
| | Wiktionary | WordNet | Thesaurus | Wiktionary | GermaNet | Thesaurus | Wiktionary | WordNet |
| *Internal relations* | 162,840 | 1,411,522 | 529,374 | 774,642 | 546,544 | 935,674 | 134,424 | 567,129 |
| Synonymy | 59,025 | 315,984 | 332,462 | 189,304 | 78,154 | 605,088 | 59,324 | 95,723 |
| Antonymy | 8,187 | 7,979 | 0 | 81,924 | 3,353 | 407 | 18,165 | 14,861 |
| Hypernymy | 1,091 | 329,402 | 98,456 | 111,601 | 212,507 | 140,175 | 38,692 | 159,161 |
| Hyponymy | 2,498 | 329,402 | 98,456 | 145,010 | 212,066 | 0 | 15,738 | 128,613 |
| Related term | 64,084 | 0 | 0 | 42,835 | 11,269 | 190,004 | 0 | 0 |
| Collocation | 0 | 0 | 0 | 120,685 | 0 | 0 | 0 | 0 |
| Derived term | 24,472 | 74,716 | 0 | 83,283 | 0 | 0 | 0 | 0 |
| Other types | 3,483 | 354,039 | 0 | 0 | 29,195 | 0 | 2,505 | 168,771 |
| *Translations* | 1,230,539 | 207,378 | – | 949,203 | – | – | 314,113 | – |
| Arabic | 15,361 | – | – | 6,022 | – | – | 1,275 | – |
| Chinese | 55,395 | – | – | 12,785 | – | – | 1,136 | – |
| Czech | 27,645 | 12,824 | – | 25,103 | – | – | 6,107 | – |
| Dutch | 45,552 | 53,448 | – | 30,270 | – | – | 4,659 | – |
| English | – | – | – | 123,413 | – | – | 48,801 | – |
| Estonian | 7,881 | 9,004 | – | 3,229 | – | – | 2,703 | – |
| Finnish | 83,725 | – | – | 13,511 | – | – | 3,849 | – |
| French | 59,293 | 22,730 | – | 88,484 | – | – | 14,937 | – |
| German | 65,088 | 16,347 | – | – | – | – | 23,677 | – |
| Italian | 38,381 | 71,789 | – | 55,307 | – | – | 12,546 | – |
| Japanese | 37,865 | – | – | 23,651 | – | – | 2,289 | – |
| Polish | 26,985 | – | – | 43,146 | – | – | 5,128 | – |
| Portuguese | 31,802 | – | – | 23,990 | – | – | 5,387 | – |
| Russian | 75,639 | – | – | 40,955 | – | – | – | – |
| Spanish | 53,481 | 21,236 | – | 69,938 | – | – | 15,993 | – |
| Swedish | 32,914 | – | – | 126,829 | – | – | 3,348 | – |
| *Target languages* | 2,008 | 7 | – | 351 | – | – | 388 | – |

Table 4.7: Comparison of internal relations and translations

for example between *fällen* and *Urteil* (English: return a verdict). While collocations can be of great help for language learners, this also allows for mining and using them in natural language processing applications.

**Translations**. *Translations* indicate equivalent word senses in a certain *target language*. For each word sense, there can be multiple translations into different languages, for example, the German translations *Boot* and *Schiff* and the Russian translation *лодка* of the English noun *boat*. In Wiktionary, translations are encoded as translation-based interwiki links (see section 2.4).

Table 4.7 shows the number of translations found within the English, German, and Russian Wiktionary editions for some major languages and in total. Since none of the other dictionaries contain translations, we compare Wiktionary with *EuroWordNet* (1999), a multilingual wordnet introduced by Vossen (1998). EuroWordNet is based on *WordNet 1.5* (1995) and provides translation equivalents for seven languages (we show the corresponding numbers in the English "WordNet" column of table 4.7).[55] Wiktionary exceeds the number of translations in EuroWordNet by far for Czech, French, German, and Spanish, but encodes fewer translations for Dutch, Estonian, and Italian. The comparison with EuroWordNet shows that Wiktionary can help closing the gaps of expert-built multilingual wordnets without being an adequate replacement in any case. A key advantage is, however, that Wiktionary provides translations into far more languages than covered by any multilingual wordnet. We have already noted in section 4.3 that Wiktionary provides a large diversity of language editions. Now, we find that this also holds for the translations encoded within the individual language editions.

The number of languages seems to be extremely high, especially for the English Wiktionary. Thereby, it should be noted that there are only a few translations for many of the encoded languages. For Abua (a language spoken in southern Nigeria), there is, for example, only a single translation yet (*water* translates to *ə̀mùum*). Besides the main languages of a country, also dialects (e.g., Geordie or Swabian) and ancient languages (like Ancient Egyptian) are encoded in the form of translations.

## 4.8 Discussion and Further Perspectives

Our first research question addresses the implications of collaborative dictionaries such as Wiktionary on lexicography, dictionaries, and dictionary users. In the course of this and the previous chapter, we found that the collaborative construction approach of Wiktionary yields promising results, as Wiktionary is available in many languages and continually growing. The vast number of authors allows the encoding of changes to the language more quickly than in editorial dictionaries. In addition to that, the different backgrounds of the authors yield a broad

[55] The numbers are taken from `http://www.illc.uva.nl/EuroWordNet/finalresults-ewn.html` (5 April 2013)

coverage of technical domains and language varieties. However, we also discussed different cultures of providing sources and multiple quality issues, which mostly stem from automatic imports of copyright-expired dictionaries, the lack of corpus evidence, and providing general rather than specific descriptions. Apart from that, we observed a strong dependence on existing dictionaries and lexicographic reference works, which serve as references for the lexicographic descriptions.

In section 3.3, we raised the question of Wiktionary's user relationship, as the target audience of the dictionary is not clearly defined. Based on our study, we discuss below which users come into question and how they can benefit from or are hampered by Wiktionary:

(1) *Laypeople* can quickly look up the definition of an unknown word in Wiktionary. They have the possibility to ask questions on its usage or share their own subjective opinions and language intuitions on the respective discussion page. The quality issues observed for the lexicographic descriptions can, however, cause an inaccurate answer to the user's language need. Given the enormous amount of information available on the World Wide Web, the competence of telling apart reliable and unreliable information is hence more required than ever and becomes highly relevant for dictionaries – or as Rundell (2012, p. 81) puts it: "The old idea of the dictionary as an 'authority' on language may not survive".

(2) *Language learners* benefit from the densely interlinked multilingual organization (see section 2.2), the good coverage of basic vocabulary, and the use of graphics to illustrate word senses. But for them, inaccurate descriptions are especially challenging, since they cannot rely on the language intuition of a native speaker.

(3) *Professional translators* can exploit the vast number of translations of proverbs, idioms, interjections, and domain-specific vocabulary, which are often absent from other dictionaries and therefore hard to find. This particularly holds for resource-poor languages and language pairs not covered by other dictionaries (such as Greenlandic–Italian). Inaccurate descriptions might not be a too severe problem for them if Wiktionary is used in addition to corpora, language tools, and other reference works, which provide further evidence.

(4) *Journalists* can take advantage of Wiktionary's up-to-dateness regarding neologisms and newly coined word senses. For their (language-related) information needs on emerging news topics, time plays a critical role. The quick reaction time of a web community can provide a promising solution for meeting this information need.

(5) *Social scientists* have the opportunity of studying the collaboration of web communities, how they coordinate the work, and how conflicts are being resolved. This particularly includes cultural peculiarities across languages, which becomes possible with the multilingual organization of Wiktionary.

(6) *Linguists* benefit from Wiktionary's revision history and discussion pages, with which they can investigate the evolution of language and semantic shift. The broad coverage

of languages allows for studying resource-poor and endangered languages, whereas the easy-to-use wiki interface provides the opportunity for the collaboration of linguists and speakers.

(7) *Lexicographers* can gain totally new insights about the users of their dictionaries. This includes questions of what is important to include in a dictionary and what is comprehensible to the reader. The sum of subjective opinions and language intuitions of the different authors yields a different kind of evidence that is usually not found in corpora and might represent a valuable addition to the dictionary.

(8) *Computational linguists* can effectively make use of Wiktionary data in natural language processing applications, as it shows a number of improvements over computational dictionaries, such as *WordNet*. This will be the subject of the remaining chapters of the thesis.

The broad range of the potential target audience indicates that many users can benefit from Wiktionary. However, neither the specific needs of a certain group of users nor the different situations of consulting the dictionary are being addressed by Wiktionary. Bergenholtz and Tarp (2003, p. 172) note:

> Consequently, all theoretical and practical considerations must be based upon a determination of these needs, i.e. what is needed to solve the set of specific problems that pop up for a specific group of users with specific characteristics in specific user situations.

The attempt of a one-fits-all approach displaying the same information to every user without explicitly defining the target audience limits the usefulness of Wiktionary as a "utility product" (ibid., p. 172).

In section 3.1, we formulated the radical thesis whether collaborative dictionaries will or are about to replace expert-built ones. To date, we can reject this claim and conclude that collaborative and expert-built dictionaries exist side by side. As noted previously by Rundell (2012) and others, there is much potential in the cooperation between professional lexicographers and the large number of contributors in collaborative dictionaries. Lexicographers provide descriptions based on corpus evidence, whereas the vast number of collaborative authors contribute information based on their language intuition particularly addressing newly coined words and non-standard varieties of language as well as translation equivalents in a broad range of languages.

At least in the context of commercial dictionaries, a great challenge is, however, the conflicting goals of the two groups: Wiktionary strives for providing freely available knowledge, which is a major reason for its large number of authors and its rapid growth. The lexicographers (or, more precisely, the publishers) are bound to economic success of their work and employ restrictive licenses to their products. This might be one reason why collaborative-institutional projects such as the *Macmillan Open Dictionary* (2009 f.) – although providing

valuable information for the lexicographers – hardly pick up speed in comparison to Wiktionary. Future research should concentrate on bridging this gap in order to make better use of the collaborative work.

## 4.9 Chapter Summary

In this chapter, we compared Wiktionary to a number of machine-oriented dictionaries by means of quantitative data analysis methods. We found a broad diversity of languages in Wiktionary – both in terms of available language editions and translations. Using the example of Greenlandic, we showed that resource-poor languages can particularly benefit from Wiktionary. The collaborative construction process causes the dictionary to grow rapidly and continually without being limited to fixed release cycles. In comparison to other dictionaries, Wiktionary covers lexical entries on any part of speech and shows good coverage of basic vocabulary and neologisms, whereby we found large differences between the English and the German Wiktionary editions, which we ascribe to the different practices of providing references. A surprising result was the very low overlap of headwords encoded by the eight dictionaries. We observed that Wiktionary has a strong focus on technical terms and non-standard varieties of language. By analyzing Wiktionary's pragmatic labels in comparison to WordNet Domains, we predominantly found vocabulary from structural and natural sciences as well as sports in Wiktionary, whereas WordNet showed a stronger focus on humanities and social sciences. We concluded our metalexicographic study by assessing which types of users can benefit from Wiktionary and leading over to the natural language processing part of the thesis.

# Part II

# The Natural Language Processing Perspective

Chapter 5

# From Dictionary to Ontology

This chapter aims at creating a multilingual, sense-disambiguated ontology from information harvested from Wiktionary. We introduce our motivation in section 5.1 and discuss related work in section 5.2. In section 5.3, we survey multiple text-mining-based software tools for extracting Wiktionary's information items and describe our extensions to them. We then propose and evaluate our methods for disambiguating the extracted semantic relations and translations (section 5.4) and for composing consistent sets of synonymous word senses, which results in our final ontology *OntoWiktionary* (section 5.5). We conclude the metalexicographic perspective in section 5.6 and summarize the chapter in section 5.7.

## 5.1 Motivation

Traditional natural language processing approaches are knowledge-based using wordnets or ontologies as a source of background knowledge. To date, these approaches are getting more and more replaced by statistical models, although it has been found that knowledge resources have the ability to substantially contribute to the performance of a system (Oepen et al., 2007; Herbert et al., 2011). This is particularly true for the emerging research on learning structured embeddings of such knowledge resources (Bordes et al., 2011), which has the potential to replace purely statistical methods that dominated our field of research over the last years. One reason for the knowledge-based approaches being rarely employed is the challenging construction process of resources with a high coverage, a broad diversity of information types, and a certain level of quality.

Early approaches of compiling such knowledge resources utilize *machine-readable* versions of editorial dictionaries (Amsler, 1982), such as the *Longman Dictionary of Contemporary English* (1978). This poses the challenge of extracting information from the weakly structured, often highly compressed dictionary articles. The extraction of lexical relations is a particular problem, as this kind of information is usually not explicitly encoded in human-oriented

dictionaries (Calzolari, 1984; Wang and Hirst, 2012). Another problem is the use of restrictive licenses for most commercial projects, which often impedes integrating them in language processing applications.

*Computational dictionaries* are specially crafted for being used by machines rather than humans and are thus rigorously structured. The development of the Princeton *WordNet* (1985 f.) has proven to be a success and caused the compilation of wordnets for many languages other than English. Wordnets have the advantage of providing high quality information, since they are manually created by experts. However, they suffer from their small coverage, which particularly holds for multilingual wordnets. There are, for instance, only about 16,000 translations from English to German in *EuroWordNet* (1999). Expert-built *ontologies*, such as *Cyc* (1995 f.) and its extensions (Lenat, 1995; Reed and Lenat, 2002), show essentially the same limitations, and most of them are limited to a narrow domain. An additional challenge of all expert-built resources (ontologies, machine-readable, and computational dictionaries) is keeping them up to date, because the release of a revised version – if done at all – is the result of a costly and time-consuming manual compilation process.

Fully automatic methods for acquiring linguistic knowledge from large corpora or the World Wide Web yield resources with a huge coverage. Although such systems have recently shown impressive progress in their precision, they still cannot reach the quality of human judgments. The most prominent approaches in this line of research rely on the redundant nature of a large number of documents, usually acquired from the Web, and try to infer semantic knowledge from them, even though only a small fraction of the input data contains evidence for a certain information type. Banko et al. (2007) call this approach *open information extraction* (OIE), which is being used by their implementation *TextRunner*. Besides quality issues in the resulting resources, the utilized corpora are often biased towards certain topics, styles, registers, or text types, which prevents the creation of general-purpose dictionaries.

Social media technologies and the vast amount of freely available user-generated content have led to a "renaissance of knowledge-rich approaches" (Hovy et al., 2013, p. 3). The collaborative creation of wordnets and ontologies, as it is addressed in the *Open Mind Common Sense* (1999–2012) project has the ability to acquire rapidly growing knowledge resources (Singh, 2002). The wiki technology has been found to be especially promising in this context. The most prominent representative *Wikipedia* (2001 f.) provides a huge amount of data at a considerable quality and has been a topic of research in over 100 institutions worldwide (Medelyan et al., 2009, p. 748). But most of the articles in Wikipedia are about nouns and named entities and there are no information items covering pronunciations, inflected word forms, pragmatic labels, etc. and no descriptions on verbs, adjectives, and the like. Therefore, natural language processing applications that depend on linguistic knowledge can only make limited use of Wikipedia, for example, sentiment analysis (which particularly requires adjectives), lemmatization (requiring word forms), speech synthesis (requiring pronunciation information), or measuring verb similarity (which we discuss in section 8.3).

| Dictionary type | Knowledge | Coverage | Extraction | Releases | Quality |
|---|---|---|---|---|---|
| machine-readable dictionary | linguistic | high | hard | seldom | very high |
| computational dictionary | linguistic | small | easy | seldom | very high |
| OIE-based dictionary | linguistic | huge | easy | often | low |
| Wikipedia | encyclopedic | high | medium | often | high |
| Wiktionary | linguistic | high | medium | often | high |

Table 5.1: Properties of different dictionary types with regard to natural language processing applications (OIE = open information extraction)

Wiktionary is targeted at linguistic rather than encyclopedic knowledge. We have described in the previous chapters that it covers a broad variety of information types, parts of speech, and languages. It is continually growing and its descriptions are manually contributed by human users asserting a considerable degree of quality – especially in comparison to automatically acquired dictionaries. Our idea is that Wiktionary fills the gap between the small expert-built wordnets and the large amount of encyclopedic knowledge from Wikipedia. Table 5.1 summarizes the advantages and limitations of the different dictionary types in terms of the kind of knowledge they encode, their coverage, how their knowledge can be automatically extracted, the time needed to update them, and the quality of their lexicographic descriptions.

**OntoWiktionary.** In the scope of this chapter, we construct a new multilingual lexical ontology *OntoWiktionary* from the lexicographic descriptions in Wiktionary. Following Guarino et al. (2009), we define an *ontology* as a knowledge resource that is able to model a certain universe (not necessarily the real world). The building blocks of an ontology are *concepts* (or *classes*, *categories*) and *relations* (or *predicates*, *properties*). The former is a conceptualization of a phenomenon observed in the universe being described. An example is the idea of a ›DOG‹. The backbone of an ontology are *subsumption relations* between concepts (i.e., a relationship that forms a hierarchy of concepts). The concept ›DOG‹ can, for example, be subsumed by a (super-)concept ›ANIMAL‹, which represents any type of animal, including dogs. In addition to subsumption relations, there can be other types of relations, for example, to express that the concept ›DOG‹ is similar to the concept of a ›GRAY WOLF‹.

A *lexical ontology* – a term proposed, for example, by Veale et al. (2004) – is an ontology in which the concepts represent a certain meaning, which can be referred to by multiple words or phrases of a language. We call these words or phrases the *lexicalizations* of a concept. The concept ›DOG‹ has, for instance, the lexicalizations *dog* and *hound*. In a *multilingual lexical ontology*, lexicalizations can refer to different languages (e.g., lexicalizing ›DOG‹ as *Hund* in German and as *собака* in Russian). The necessity of associating lexicalizations and corresponding linguistic information types with ontologies has been previously raised by Buitelaar et al. (2009) in the context of the *LexInfo* project.

We construct our multilingual lexical ontology *OntoWiktionary* in multiple steps: We first extract and disambiguate the individual information items from Wiktionary. Then, we use the extracted word senses and the disambiguated synonymy relations to conceptualize Wiktionary and infer a large number of semantic relations based on this ontological structure. We make OntoWiktionary publicly available to foster further research on constructing knowledge resources and to improve knowledge-rich applications that employ OntoWiktionary as a source of background knowledge. Our contributions can be summarized as:

**Contribution 4.1**: *We extend a text mining software tool in order to extract the information items encoded in the English, German, and Russian Wiktionary editions (section 5.3).*

**Contribution 4.2**: *We propose and evaluate a method for disambiguating semantic relations and translations in Wiktionary (section 5.4).*

**Contribution 4.3**: *We propose and evaluate a method for composing synsets from disambiguated Wiktionary relations (section 5.5).*

**Contribution 4.4**: *We construct the multilingual lexical ontology OntoWiktionary from Wiktionary's synsets, semantic relations, and translations (section 5.5 and section 5.6).*

## 5.2 Related Work

An obvious choice for deriving lexical ontologies is using wordnets, such as the English *WordNet* (Fellbaum, 1998) or the German *GermaNet* (Kunze and Lemnitzer, 2002). Besides monolingual wordnets, a number of multilingual wordnets have been created: *EuroWordNet* (1999) described by Vossen (1998) covers Czech, Dutch, English, Estonian, French, German, Italian, and Spanish. *MultiWordNet* (2002 f.) introduced by Pianta et al. (2002) encodes English and Italian descriptions and has later been extended to Hebrew, Latin, Portuguese, Romanian, and Spanish. Tufiş et al. (2004) describe *BalkaNet* (2004), which has been developed for Bulgarian, Czech, Greek, Romanian, Serbian, and Turkish. Gangemi et al. (2003) propose an approach for adding additional structure to wordnets. They create *OntoWordNet* (2003) from *WordNet 1.6* (1998). As outlined above, the professionally crafted wordnets suffer from their small coverage and high development cost, which is why we utilize Wiktionary for constructing our ontology.

Singh (2002) presents the collaborative ontology *Open Mind Common Sense* (1999–2012). This project asks voluntary web users to add machine-readable common-sense knowledge that can directly be used for creating ontologies. This is achieved by selecting a predefined relationship and submitting two lexicalizations that are to be connected by this relationship. One can, for instance, define that *shoes* "are made of" *leather*, which directly describes an ontological relationship. A problem is, though, that *Open Mind Common Sense* only models lexicalizations without organizing them in a conceptual structure. There are thus different relations for the synonymous words *pullover* and *sweater*. An additional problem is ambiguity,

which prevents, for example, distinguishing between the relations for the animal-related and the music-related meaning of *bass*.

Wikipedia-based ontologies are another strand of research. The most influential works in this area are *DBpedia* (2007 f.), *YAGO* (2007 f.), and *WikiNet* (2010 f.), which construct a multilingual ontology from information extracted from *Wikipedia* (2001 f.). Bizer et al. (2009b) describe different approaches for harvesting information from Wikipedia infoboxes, its link structure, redirects, categories, and many more. They model the acquired information as RDF triples yielding the large-scale ontology DBpedia. The goal of this project is providing freely available data that can be interlinked with other resources in the so-called *Linked Data cloud* (Bizer et al., 2009a). Likewise, YAGO introduced by Suchanek et al. (2007) utilizes a number of extraction components to harvest knowledge from *Wikipedia* (2001 f.) and *WordNet 2.1* (2005). Particular emphasis is put on deriving ontological relations from categories. For the concept ›ALBERT EINSTEIN‹, Suchanek et al. extract, for example, a "bornInYear" relation to the concept ›1879‹, because the article *Albert Einstein* is associated with the category *1879 births*. With WikiNet, Nastase et al. (2010) provide a multilingual resource from Wikipedia based on similar approaches. They focus on mining relations from the Wikipedia article texts, and they compile a consistent interlingual index spanning multiple Wikipedia language editions.

Wikipedia-based ontologies typically use the categories of an article to infer a subsumption hierarchy. Although this yields a densely connected network of concepts, Ponzetto and Strube (2007, p. 1440) point out that the categories "do not form a taxonomy with a fully-fledged subsumption hierarchy" and hence yield a noisy resource. The DBpedia concept ›IRON (APPLIANCE)‹ is, for instance, not only a subsumption of ›HOME APPLIANCE‹, but also of ›LAUNDRY‹.[56] This is not a generalization of ›IRON‹, but represents the domain the concept is used in. Another problem is related to the lexicalizations of the concepts. In order to reduce redundancy, each concept is encoded only once within Wikipedia – usually within the article of its most common lexicalization. The concept ›IRON‹ in the sense of the chemical element is, for example, described within the article *Iron*. Alternative lexicalizations can be extracted from the redirections pointing to this article. There are redirections from *Fe*, *Ferryl*, and *Element 26* to *Iron*, which provide sensible lexicalizations. However, there are also redirections for spelling errors (e.g., *Iorn*) and loosely related expressions (e.g., *Iron rope*, *Iron compounds*) that should not serve as lexicalizations for ›IRON‹. Of the fifteen redirections to the article *Iron*, only six represent valid lexicalizations of this concept. Suchanek et al. (2007) addressed this issue when constructing YAGO by combining information from Wikipedia with the taxonomy of *WordNet 2.1* (2005). Navigli and Ponzetto (2010) follow a similar approach for *BabelNet* (2012 f.) by aligning Wikipedia and *WordNet 3.0* (2006) at the word sense level.

Although this yields large knowledge resources with a high quality, the information extracted from Wikipedia is almost entirely about nouns and there are hence no lexicalizations for verbs, adjectives, and the like. Our work provides a viable option towards closing this

[56] The examples from DBpedia discussed here have been checked on 27 July 2011.

gap, as it makes use of the linguistic knowledge of Wiktionary covering any part of speech. A closely related research effort is the *Universal WordNet* (2010 f.) introduced by de Melo and Weikum (2009), which is based on the combined evidence found in existing wordnets, parallel corpora, and machine-readable dictionaries. It covers over 800,000 terms from 200 languages and is publicly available. De Melo and Weikum also incorporate data from Wiktionary, but merely use the encoded translations, which are not fully disambiguated (see section 5.4). Since Universal WordNet is limited to the semantic network of *WordNet* (2006), it does not make use of semantic relations and lexicalizations from Wiktionary, which is the subject of our work.

## 5.3 Extracting Knowledge

Wiktionary is intended to fulfill linguistic information needs of humans. This is why the community has put a focus on providing a graphical user interface optimized for human perception rather than for automatic data processing. For our purpose, this raises the need for obtaining the Wiktionary data and for developing a *text mining software*, which is capable of processing the article constituents and harvesting the individual information items encoded therein.

**Obtaining Wiktionary data.** One way of obtaining Wiktionary data is crawling article pages from the web by means of the interface that is being used by the human readers of Wiktionary. This requires processing the formatted HTML pages, which are being generated by the wiki software (e.g., the representation shown in figure 2.1). The individual information items in this HTML representation are, however, only weakly structured and their type is often solely characterized by their position in the text or their typographic properties.

Before being transformed into the HTML format, the dictionary articles are written and stored in a *wiki markup* language, which we introduced in section 3.6. Though being closely connected, this representation in wiki markup is often more expressive than the corresponding HTML format – in particular because of the vast use of templates. *Templates* are reusable patterns that are defined in a central place and then *invoked* by multiple articles. Invoking a template means that its (unique) name is added to the article text and enclosed by two curly brackets. Upon transforming the article to HTML, the invoked template is substituted with the template's text. A *request* for adding a missing etymology (see section 3.4) is, for example, added by invoking the *rfe* template (i.e., using the wiki markup `{{rfe}}` in the article text). This yields a box "This entry lacks etymological information [...]" and automatically associates the article with the category *Requests for etymology* to allow searching for such entries easily.

Templates may be further parameterized with different user inputs. The *sense* template is, for instance, invoked as `{{sense|<marker>}}`, where `<marker>` can be replaced by what is called a *sense marker*. That is, an indicator for associating an information item (e.g., an example sentence) with a certain word sense. The sense marker can be a running number (e.g., "1." or "[3a]") or a short text representing the meaning of the word sense (like "water craft" for the

```
=====Synonyms=====
* {{sense|private vehicle that moves independently}} [[auto]],
[[motorcar]], [[vehicle]]; [[automobile]] {{qualifier|US}},
[[motor]] {{qualifier|British colloquial}}, [[carriage]]
{{qualifier|obsolete}}
* {{sense|non-powered part of a train}} [[railcar]], [[wagon]]
* {{sense|unit of quantity}} [[carload]], [[wagonload]]
* {{sense|passenger-carrying light rail unit}} [[carriage]]
* {{sense|part of an airship}} [[gondola]], [[basket]]
{{qualifier|balloons only}}
```

(a)

**Synonyms**

- (*private vehicle that moves independently*): auto, motorcar, vehicle; automobile (*US*), motor (*British colloquial*), carriage (*obsolete*)
- (*non-powered part of a train*): railcar, wagon
- (*unit of quantity*): carload, wagonload
- (*passenger-carrying light rail unit*): carriage
- (*part of an airship*): gondola, basket (*balloons only*)

(b)

Figure 5.1: Wiki markup (a) and HTML representation (b) of the synonymy constituent of *car*

first word sense of *boat*). The use of the *sense* template is exemplified in figure 5.1 showing (a) the wiki markup and (b) the corresponding HTML format of the synonymy constituent of the article *car*. While the *rfe* template primarily serves as an abbreviation for a frequently used element of an article, the template for sense markers has obviously a different motivation, because it requires more characters to invoke this template than to add its substitution directly to the article text. The rationale behind this is to encourage a consistent encoding of the articles and obtain more control over the structure and format of its constituents. They allow the community to quickly modify the formatting of all articles at once, for example, if it would be decided to use square brackets for the sense markers instead of round ones or using a different headline for an article constituent (as it has been done for "Worttrennung" in the German Wiktionary, see section 3.3).

A text mining software that is capable of processing the wiki markup of an article can achieve more detailed and accurate results, since the templates are more expressive than their substitution. The wiki markup of Wiktionary is publicly available in the form of database dumps.[57] These dumps are encoded in a simple XML format. Although they are intended for hosting mirror sites, storing backups, and providing alternative user interfaces, they yield an ideal starting point for extracting knowledge from Wiktionary.

**Text mining software.** Implementing a text mining software for extracting the information items from Wiktionary is very challenging because of the large differences in the article microstructure between the Wiktionary language editions and the deviations from the style sheet (cf. section 2.6). In addition to that, the continual development of Wiktionary demands regular updates to the software in order to cope with modifications in the representation of the article constituents. It is not surprising that there is currently no software being capable of processing all Wiktionary language editions. The individual projects rather focus on either a small number of language editions or on certain information types. Table 5.2 shows a selection of available software tools and which language editions they cover. The table lists also if the software is able to extract the majority of information types rather than focusing on a small

[57] http://dumps.wikimedia.org

| Project/URL | Languages | ANY | OSS | UPD |
|---|---|---|---|---|
| Texai [2007]<br>http://sourceforge.net/projects/texai | en | ✓ | ✘ | ✘ |
| JWKTL (Zesch et al., 2008a)<br>http://www.ukp.tu-darmstadt.de/software/jwktl | en, de | ✓ | ✓ | ✓ |
| Lexvo (de Melo and Weikum, 2008)<br>http://www.lexvo.org/ | en, de, fr, es, ca, el, pt, sv | ✘ | ✘ | ✓ |
| Wikokit (Krizhanovsky and Lin, 2009)<br>http://code.google.com/p/wikokit | en, ru | ✓ | ✓ | ✓ |
| PanDictionary (Mausam et al., 2010)<br>http://ai.cs.washington.edu/projects/panlingual-translation | en, fr | ✘ | ✘ | ✘ |
| WISIGOTH (Sajous et al., 2010)<br>http://redac.univ-tlse2.fr/wisigoth | en, fr | ✓ | ✘ | ✘ |
| Zawilinski (Kurmas, 2010)<br>http://www.cis.gvsu.edu/~kurmasz | en | ✘ | ✘ | ✘ |
| NULEX (McFate and Forbus, 2011)<br>http://www.qrg.northwestern.edu/resources/nulex.html | en | ✘ | ✘ | ✘ |
| wiktionaryConvertor (Mörth et al., 2011)<br>http://corpus3.aac.ac.at/showcase/index.php/tools/wiktionaryconvertor | de | ✓ | ✘ | ✘ |
| Dbnary (Sérasset, 2012)<br>http://dbnary.forge.imag.fr | en, de, fr, pt | ✓ | ✓ | ✓ |
| WiktionaryToXML [2012]<br>https://sites.google.com/site/korhonenjoel | en | ✘ | ✘ | ✓ |
| lemon parser (McCrae et al., 2012)<br>http://monnetproject.deri.ie/lemonsource | en, de, fr, es, nl, jp | ✓ | ✘ | ✓ |
| DBpedia Wiktionary (Hellmann et al., 2013)<br>http://dbpedia.org/Wiktionary | en, de, fr, ru | ✓ | ✓ | ✓ |

Table 5.2: Selection of text mining software for extracting knowledge from Wiktionary, the language editions they cover, whether they can extract the majority of information types (ANY), if they are available as open source software (OSS), and if they have been updated recently (UPD)

subset of them (column ANY), whether the source code of the software is available (OSS), and whether it has been recently updated (UPD).

To our knowledge, *Texai* by Stephen Reed is the first system that has been used to extract information from Wiktionary. This project has, however, not been updated since its introduction in 2007 and neither source code nor binary files are currently available. The *Java-based Wiktionary Library* (JWKTL) introduced by Zesch et al. (2008a) has been the first publicly available software tool for researchers. It is capable of processing the English and the German Wiktionary edition and is regularly updated. Similarly, the *Wiki tool kit* (Wikokit) by Krizhanovsky and Lin (2009) can treat the English and the Russian Wiktionary edition and is continually adapted to changes of the corresponding Wiktionary language editions. Both

JWKTL and Wikokit represent the extracted data in a Wiktionary-specific database format with the intent of providing a detailed, preferably lossless, modeling of the information items encoded in Wiktionary. A related project for the English and the French Wiktionary edition by Sajous et al. (2010) is called *WIktionarieS Improvement by Graphs-Oriented meTHods* (WISIGOTH). It has, however, not been updated since 2010 and no source code is provided.

A number of software tools have recently been released, which aim at providing standardized data (e.g., RDF) for a larger collection of language editions: De Melo and Weikum (2008) extract a number of Wiktionary information types in eight languages for their RDF web service *Lexvo*, Mörth et al. (2011) create a standardized representation of the German Wiktionary, Sérasset (2012) processes four Wiktionary language editions for constructing the standardized *Dbnary*, and similarly, McCrae et al. (2012) incorporate six language editions for their *lemon* ontology. Hellmann et al. (2013) argue that their *DBpedia*-based Wiktionary RDF extraction can be quickly adapted to new language editions by configuring declarative patterns, which can be written also by non-programmers. Such patterns are currently available for processing four Wiktionary language editions. In section 7.2, we discuss the standardized data models produced by these systems.

Finally, there are a few software tools that extract Wiktionary data for a specific purpose. *Zawilinski* (Kurmas, 2010) extracts inflected word forms of Polish words from the English Wiktionary, and McFate and Forbus (2011) obtain English morphological properties in the context of the *NULEX* (2011) resource. De Melo and Weikum (2009) and Mausam et al. (2010) collect a large number of translations from multiple Wiktionary language editions. The *WiktionaryToXML* software by Joel Korhonen is designed to convert the English Wiktionary into e-book formats such as ePUB. These software tools do not allow accessing all information items encoded in Wiktionary.

**Extensions**. For the research described in this thesis, we utilize JWKTL and extend this software by a novel adapter to Wikokit, such that we become able to extract information from the English, German, and Russian Wiktionary editions. In addition to that, we enable the extraction of new information types yet missing from the software (such as inflected word forms and references), and we regularly contribute updates for being able to handle modifications of the article style sheets.

## 5.4 Disambiguation of Information Items

Much Wiktionary-related work has been carried out without taking semantic information into account. Navarro et al. (2009, p. 22) note, for instance, that they flattened all word senses, because the way they appear is "unpredictable". We argue, however, that it is important to distinguish the encoded word senses in order to facilitate a meaningful use of Wiktionary in natural language processing applications, which we further evaluate in chapter 8. The word

senses of a lexical entry are encoded as an enumeration providing the pragmatic labels and the sense definitions. Additional semantic information items such as example sentences, semantic relations, or translations are specified for a certain word sense by means of *sense markers*, which we described in the previous section. The JWKTL software is able to resolve these sense markers and thus to obtain sense-disambiguated information from Wiktionary.

The semantic relations (i.e., synonymy, hypernymy, hyponymy, etc.) and the translations are, however, encoded as *lemma-oriented internal links*. That is to say, the word sense of the target of a relation remains unspecified (see section 2.4). Consider, for example, the dictionary article on the English verb *hang* shown in figure 5.2. The eighth word sense of *hang* is defined as "to exhibit (an object)". It encodes synonymy relations pointing to *exhibit* and *show* and translations into German *ausstellen*, French *exposer*, Dutch *ophangen*, and a number of other languages. When taking a look at the article *exhibit*, we find that the synonym link can refer to the meaning of (1) displaying something (e.g., exhibiting a drawing) or (2) demonstrating a skill (e.g., exhibiting a talent in acting). For humans, it is easy to recognize that *hang* is synonymous to the former, but not to the latter. Natural language processing systems, however, cannot disambiguate such relations easily. The same applies to translations: The German *ausstellen* has a meaning of (1) exhibiting an object, (2) certificating a document, and (3) turning off smth., but only the first one describes the same meaning as the English *hang*.

In this section, we propose and evaluate an automatic method for disambiguating semantic relations and translations (called *relations* henceforth). Sense-disambiguated relations are a necessary precondition for many applications, such as computing semantic relatedness by measuring path lengths (see Budanitsky and Hirst, 2006; Matuschek and Gurevych, 2013): if undisambiguated relations were used, then *exhibit* and *loiter* would be highly related as they both have a relation to *hang*. We discuss this in more detail in section 8.3.

### 5.4.1 Previous Approaches

**Semantic relation disambiguation**. The task of disambiguating semantic relations has previously been addressed by Krovetz (1992), who analyzes the disambiguation of cross-references found in the articles of the *Longman Dictionary of Contemporary English* (1978). Krovetz proposes measuring the word overlap between two sense definitions and analyzing their morphological properties. His *sense linking* approach is not quantitatively evaluated.

In the context of ontology learning, Pantel and Pennacchiotti (2008) discuss a *relation anchoring* method, which has essentially the same goal as our task of disambiguating relations: They extract a large number of ontological relations from the World Wide Web using their *Espresso system*. Both the source and the target of these relations are words that need to be "ontologized" (i.e., disambiguated), whereby all word senses from *WordNet 2.0* (2003) serve as candidates for the relation's source and target word senses. The candidates are disambiguated

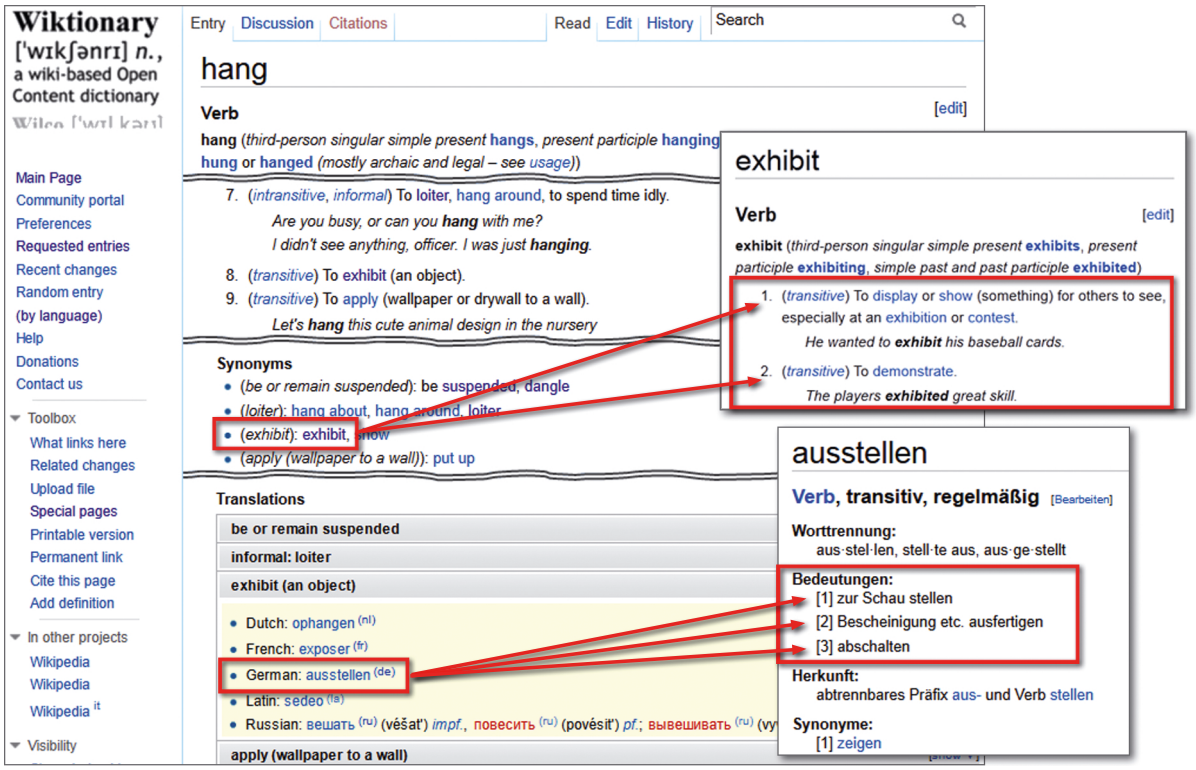



Figure 5.2: Wiktionary's semantic relations and translations are not sense-disambiguated: The synonym *(to) exhibit* of the English Wiktionary entry *(to) hang* and its German translation *ausstellen* have multiple possible target word senses.

using measures based on distributional similarity. They report an $F_1$ score between .36 and .53 on manually annotated gold standards specific for their task.

In this setting, the disambiguation of relations is a fairly complex task, since both the source and the target lexical entries need to be disambiguated. Consider, for instance, the hyponymy relation (*boat, canoe*). If there are three word senses for *boat* and two word senses for *canoe*, all six possible combinations have to be compared by the disambiguation method. In our setting, the source is already disambiguated by the Wiktionary authors. This reduces the computational complexity by one degree of freedom and, at the same time, allows for a higher quality due to the manual disambiguation. For the example (*boat, canoe*), only two pairs of word senses remain to be processed.

The disambiguation of all words in a sense definition (i.e. *gloss disambiguation*), which has been proposed for the *eXtendend WordNet* project (Harabagiu et al., 1999; Mihalcea and Moldovan, 2001), is very similar to the disambiguation of semantic relations. This allows us to make use of many of the features introduced by Moldovan and Novischi (2004) for this task. However, we use explicitly defined semantic relations rather than sense definitions as our disambiguation subjects. Moreover, we cannot utilize any WordNet-specific features, but need to adapt our method to Wiktionary. In addition to that, we extend our task to a cross-lingual

setting by disambiguating translations. Moldovan and Novischi report a precision of .76 for their method.

Recently, Flati and Navigli (2012) proposed a graph-based method to gloss disambiguation, which is able to outperform previous approaches. While this method could generally be adapted to disambiguating the relations in Wiktionary, we have observed in section 4.7 that the semantic network spanned by Wiktionary's relations is rather sparse. This hinders finding the cycles and quasi-cycles required by the method.

**Translation disambiguation.** The disambiguation of translations has been studied in the context of bilingual dictionaries and corpora (Kikui, 1999; Tsunakawa and Kaji, 2010). Mausam et al. (2010) discovered new translations in Wiktionary using a graph-based inference algorithm for Wiktionary translations. Although this also involves a disambiguation of translations, their work is not directly comparable to ours, since they do not strictly use the word senses encoded in Wiktionary but define them based on the translations shared across multiple languages. Since we aim at exploiting all information items encoded on Wiktionary's article pages, we use the word senses explicitly specified in Wiktionary for our disambiguation algorithm. To our knowledge, no previous work addressed the task of disambiguating the relations in Wiktionary that would be fully comparable to ours.

### 5.4.2 Relation Disambiguation Method

In this section, we describe our method for automatically disambiguating semantic relations and translations in Wiktionary. We first define a number of features and then combine them in a rule-based approach and in a machine-learning-based classifier.

Let $t_j \in t$ be one of multiple possible target word senses for a relation (either a semantic relation or a translation) $r = (s_i, t)$. We define the following features based on our analysis of 200 randomly sampled Wiktionary relations (referred to as *development data*).

**Definition overlap.** A widely used method for word sense disambiguation is based on counting the word overlap between sense definitions (Lesk, 1986). Let definition($s_i$) and definition($t_j$) be the lemmatized and stop-word-filtered sense definitions of $s_i$ and $t_j$. Their overlap is the number of shared words:

$$f_{\text{Lesk}} := \left|\text{definition}(s_i) \cap \text{definition}(t_j)\right|.$$

We additionally define $f_{\text{ExtLesk}}$ by employing the extension by Banerjee and Pedersen (2003). That is, we assign squared scores to consecutive sequences of words. If both definitions contain, for example, "large carnivorous animal", we assign a score of $3^2 = 9$.

**Source lemma**. A special case of overlapping definitions is that the lemma of the source word sense is contained in the definition of the target word sense:

$$f_{\text{src}} := \text{lemma}(s_i) \in \text{definition}(t_j).$$

This happens frequently, since a definition usually contains synonymous words or follows the *genus-differentia* pattern – i.e., providing a more specialized term (the *genus*) and the properties that distinguish the word from its co-hyponyms (the *differentia*). Consider, for instance, two word senses for *peck*: (1) "[...] a dry measure of eight quarts" and (2) "a great deal; a large or excessive quantity". The second one happens to be the correct disambiguation for the synonymy relation between *deal* and *peck* – and it contains the source lemma *deal*.

**Pragmatic labels**. Many word senses are domain-specific, such as the use of *host* as a certain kind of server in computer science, or they are marked as belonging to a certain variety of language. In section 4.6, we have seen that Wiktionary encodes many types of pragmatic labels including domain, register, and style labels. An example is the sense definition "(UK, pejorative) A working-class youth [...]" of *chav*, which is marked by a *geographical label* and an *evaluative label*. Relations usually connect two word senses used in the same language variety (e.g., being used in the same technical domain), which is why, we add a feature

$$f_{\text{lbl}} := \left|\text{label}(s_i) \cap \text{label}(t_j)\right|$$

counting the number of pragmatic labels shared by $s_i$ and $t_j$. Instead of the raw labels, we use the label groups described in section 4.6 (e.g., grouping "zoology" and "ornithology" as "biology").

**Inverse relation**. Consider a relation between two polysemous words, such as the antonymy relation between $fall_i$ and *increase*. If there is a word sense $j$ of *increase* for which an inverse antonymy relation ($increase_j$, *fall*) is encoded, then it is very likely that $j$ is the correct disambiguation for ($fall_i$, *increase*). Let relations($t_j$) be the set of relations of $t_j$. We define

$$f_{\text{inv}} := \text{lemma}(s_i) \in \text{relations}(t_j)$$

as the feature checking for inverse relations.

**Relation overlap**. The idea of inverse relations can be further extended by finding relations to other words shared by both the source and the target sense. A relation (*sweater, cloth*) can, for instance, be disambiguated by finding that one of their word senses shares a relation to

*pullover* (a synonym of *sweater* and a hyponym of *cloth*). We define

$$f_{\text{rel}} := \frac{\left|\text{relations}(s_i) \cap \text{relations}(t_j)\right|}{\left|\text{relations}(s_i) \cup \text{relations}(t_j)\right|},$$

which is similar to the link-based similarity measure proposed by Milne and Witten (2008), who use hyperlinks from Wikipedia.

**Commonness and monosemy**. The word senses of a dictionary are often ordered according to their usage frequencies in a corpus or the intuitions of the lexicographers. This has led to a very strong baseline for word sense disambiguation by always choosing the first sense. The same applies to the disambiguation of relations when choosing the first target sense in any case. Therefore, we introduce a feature $f_{\text{idx}} := j$ that is set to the index of the target sense $t_j$.

Finally, we add a feature $f_{\text{mono}}$ that is true if, and only if, the target word has only one word sense (i.e., if it is *monosemous*). In these cases, it is most likely (though not guaranteed) that this sense is the correct disambiguation. An example is the synonymy relation between *eggplant* and the monosemous word *brinjal*.

**Cross-lingual features**. Most of the features described above are also applicable in a multilingual setting when using translations instead of semantic relations. In order to also use the features based on sense definitions, we automatically translate them using the *Bing translation*[58] service. This opens up interesting research opportunities, since the definition of either the source or the target sense can be translated:

$$f_{\text{Lesk,SL}} := \left|\text{definition}(s_i) \cap \text{definition}(\text{translate}(t_j))\right|,$$

$$f_{\text{Lesk,TL}} := \left|\text{definition}(\text{translate}(s_i)) \cap \text{definition}(t_j)\right|.$$

There can even be a combined feature:

$$f_{\text{Lesk,SL\&TL}} := \tfrac{1}{2}\left(f_{\text{Lesk,SL}} + f_{\text{Lesk,TL}}\right).$$

Regarding the pragmatic label feature $f_{\text{lbl}}$, we manually map English and German labels that represent the same meaning (e.g., *biology* and *Biologie*). This yields a list of 75 label groups covering 2,969 distinct pragmatic labels.

**Constraints**. In addition to the features introduced above, we can apply a threshold to convert a numeric feature into a boolean one. The notation $f_{\text{Lesk}\geq k}$ defines, for instance, a feature that is true if the sense definitions share at least $k$ words. We use the notation $\hat{f}$ when only

[58] http://www.microsofttranslator.com/

the target word sense with the highest feature value is used. The feature $\hat{f}_{\text{Lesk}\geq\text{k}}$ is thus true if, and only if, $f_{\text{Lesk}}$ is higher than $k$ and the maximum $f_{\text{Lesk}}$ of all possible target word senses of $t$.

**Disambiguation method**. Let again $r = (s_i, t)$ be a relation and $t_j \in t$ a possible target word sense. We define a *relation disambiguation method* as a function

$$D\colon\ (r, t_j) \mapsto \{0, 1\}$$

returning 1 if $t_j$ is a correct disambiguation for $r$ and 0 otherwise.

Let $F_{r,t_j}$ be a set of features for the relation $r$ and the possible target word sense $t_j$. A basic disambiguation method $D_f(r, t_j) = f$ uses only a single boolean feature $f \in F_{r,t_j}$. Thereby we can model a most frequent sense baseline

$$\text{MFS} = f_{\text{idx}=1} = \begin{cases} 1 & \text{if } f_{\text{idx}} = 1 \\ 0 & \text{otherwise} \end{cases}$$

always using the first target word sense. One way of combining features is to concatenate them using a backoff strategy. That is, a method

$$D_{f_1 \circ f_2} = f_1 \circ f_2 = \begin{cases} f_1 & \text{if } f_1 \in F_{r,t_j} \\ f_2 & \text{otherwise} \end{cases}$$

relying on feature $f_1$ (if present) and $f_2$ otherwise. For example, $D_{f_{\text{inv}} \circ f_{\text{idx}=1}}$ disambiguates those relations that have an inverse relation using $f_{\text{inv}}$. The remaining relations are disambiguated using a most frequent sense approach.

This definition allows us to formulate our rule-based disambiguation method

$$\text{WKTWSD} = f_{\text{mono}} \circ f_{\text{lbl}\geq 1} \circ f_{\text{rel}\geq 0.5} \circ f_{\text{src}} \circ f_{\text{inv}} \circ \hat{f}_{\text{ExtLesk}\geq 2} \circ f_{\text{idx}=1}$$

that concatenates all features introduced above. For the disambiguation of translations, we use $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL\&TL}}$ instead of $\hat{f}_{\text{ExtLesk}\geq 2}$. The ordering and the thresholds have been chosen based on our analysis of the development data.

As a comparison, we train a number of machine learning classifiers using the same set of features. Below, we discuss the results for a Naïve Bayes classifier (BAYES) and a J48 decision tree (a C4.5 clone), although we try other classifiers as well, which yield similar results in general. The training has been done in a 5-fold cross validation using the *Weka* toolkit for machine learning (Hall et al., 2009). Note that we did not optimize the configuration of the algorithms in order to avoid overfitting to the datasets.

| Method | $A_{O,1}$ | $A_{O,2}$ | $\kappa_1$ | $\kappa_2$ |
|---|---|---|---|---|
| MFS | .78 | .79 | .45 | .50 |
| TEXTSIM | .79 | .82 | .48 | .57 |
| WKTWSD | **.84** | **.85** | **.59** | **.65** |
| HUMAN | .89 | .89 | .73 | .73 |

Table 5.3: Results of our pilot study for disambiguating German semantic relations

### 5.4.3 Evaluation

The evaluation of our approach is two-fold. First, we conduct a pilot study comparing our disambiguation method with a method based on text similarity. Then, we create four gold standard datasets and evaluate the performance of our method in comparison to machine-learning classifiers.[59]

**Pilot study**. For the disambiguation of German semantic relations, we utilize a text similarity measure based on *Explicit Semantic Analysis* (ESA) using Wikipedia (Gabrilovich and Markovitch, 2007). To this end, we represent each token of the lemmatized and stop-word-filtered sense definition as a so-called *concept vector* consisting of the *term frequency* $*$ *inverse document frequency (tf.idf)* weights of the token in relation to the underlying semantic space (Wikipedia articles in our case). The two concept vectors of the source word sense and the possible target word sense can then be compared using the *cosine similarity* metric (i.e., the cosine of the angle between the two concept vectors) and the target word sense with the highest similarity score is chosen. In Meyer and Gurevych (2010b, 2012b), we provide more details on this method.

Table 5.3 shows the performance of this text similarity measure (TEXTSIM) in comparison to our WKTWSD method and the most frequent sense baseline (MFS). In order to quantify the results, we measure the agreement between the algorithmic result and each of the two raters. We use the observed agreement $A_O$ and Cohen's $\kappa$ statistic as defined by Artstein and Poesio (2008) for this purpose. The utilized dataset consists of 250 randomly chosen semantic relations of the German Wiktionary. The inter-rater agreement between the two human raters accounts for $A_O = .89$ and $\kappa = .73$ (HUMAN). Table 5.4 provides the descriptive statistics for this pilot study dataset. We observe that our WKTWSD method outperforms both TEXTSIM and the MFS baseline by a large margin. The improvement is statistically significant using McNemar's test ($p < .05$).

[59]The pilot study is based on Wiktionary data from June 2009 and the gold standard datasets are based on Wiktionary data from April 2011.

| | pilot study | en:en | de:de | en:de | de:en |
|---|---|---|---|---|---|
| Number of relations | 250 | 394 | 459 | 204 | 204 |
| Number of annotations | 920 | 1,117 | 1,119 | 614 | 656 |
| Number of raters | 2 | 2 | 2 | 3 | 3 |
| Balanced sampling | ✘ | ✓ | ✓ | ✓ | ✓ |
| Observed agreement $A_O$ | .89 | .91 | .92 | .89 | .90 |
| Kappa statistics $\kappa$ | .73 | .82 | .85 | .73 | .75 |
| Upper bound $F_1$ | – | .89 | .92 | .80 | .83 |

Table 5.4: Statistics on our evaluation datasets for disambiguating Wiktionary relations

| № | $s_i$ = *phenomenal* | $D$ | $t_j$ = *awesome* |
|---|---|---|---|
| (1) | (colloquial) Very remarkable; highly extraordinary; amazing. | 0 | Causing awe or terror; inspiring wonder or excitement. |
| (2) | (colloquial) Very remarkable; highly extraordinary; amazing. | 1 | (informal) Excellent, exciting, remarkable. |

Table 5.5: Annotated example from the (en:en) evaluation dataset

**Gold standard datasets.** To our knowledge, no gold standard datasets of disambiguated Wiktionary relations exist so far. This is why we create four new annotated datasets, which consist of English semantic relations (en:en), German semantic relations (de:de), English–German translations (en:de), and German–English translations (de:en). The relations are sampled according to their type, the part of speech, and the number of candidates (i.e., possible target word senses) in order to create a balanced dataset. Balancing out the datasets is very useful for avoiding datasets with a strong bias (e.g., on synonyms between nouns). None of the sampled relations occurs in our development data. Table 5.4 shows the number of sampled relations and possible target senses (i.e., the number of annotations required).

We have asked two human raters to annotate the monolingual datasets (en:en) and (de:de) and three raters to annotate the cross-lingual datasets (en:de) and (de:en). The raters should annotate each possible target word sense as being a correct ($D = 1$) or incorrect ($D = 0$) disambiguation for a given relation. Table 5.5 shows an example for the relation (*phonomenal, awesome*). It is permitted to rate all target senses of a relation as incorrect (e.g., if the correct target sense has not yet been encoded in Wiktionary) or to rate more than one target sense as correct (e.g., if the target senses are more fine-grained than the source sense). The raters are free to consult external sources, such as dictionaries, encyclopedias, etc., and in particular Wiktionary itself. They should, however, not contact each other to ensure independent judgments. The raters are native in German and speak English fluently. They have been trained using some example cases and an annotation guidebook explaining the task.

To assess the reliability of our datasets, we measure the inter-rater agreement. Table 5.4 shows the observed agreement $A_O$ and the kappa statistics $\kappa$ for each dataset. We report Cohen's $\kappa$ for the two rater case and Fleiss' $\kappa$ (multi-$\pi$) for the three rater case (cf. Artstein and Poesio, 2008). The raters agree on about 90 % of the cases. The $\kappa$ statistics of over .80 for the monolingual datasets suggests good reliability. The cross-lingual datasets have a slightly lower agreement. The disambiguation of translations hence seems to be more difficult for our raters. However, the $\kappa$ scores are well above .67 and therefore allow us to draw tentative conclusions (Artstein and Poesio, 2008, p. 576). While subtle distinctions of meaning are the main reason for disagreement between the raters, we did not find systematic disagreement stemming from misunderstandings or unclear instructions. To provide an upper bound for evaluating our methods, we also provide $F_1$ scores for the datasets as suggested by Hripcsak and Rothschild (2005).

Finally, we create gold standard datasets based on the majority vote of the raters. As a tie breaker for the monolingual datasets, an additional adjudicator has been asked for a final decision. All datasets including the sampling properties and the annotation guidebook are freely available from our homepage.

**Evaluation results.** Table 5.6 shows the performance of our disambiguation method on the four gold standard datasets. We have counted the number of correct decisions $TP + TN$ (true positives plus true negatives), the number of false positives $FP$ and false negatives $FN$, which we use to report accuracy $A$, precision $P$ (proportion of correctly disambiguated relations in the system result), recall $R$ (proportion of correctly disambiguated relations in the gold standard), and the $F_1$ score:

$$A = \frac{TP + TN}{TP + TN + FP + FN}, \quad P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F_1 = \frac{2 \cdot P \cdot R}{P + R}.$$

As a lower bound, we use a method making a random decision (RAND) and the most frequent sense baseline (MFS). The upper bound is human performance (HUMAN) estimated by the inter-rater agreement $A_O$ and the inter-rater $F_1$ score introduced above. Our WKTWSD method outperforms the baselines for each dataset with the exception of (de:en), whose precision is slightly lower than the precision of MFS. The improvement is statistically significant for the monolingual datasets ($p < .001$) and the (en:de) dataset ($p < .1$).

Besides the lower and upper boundaries, we compare WKTWSD with the machine learning classifiers BAYES and J48. Our WKTWSD method generally reaches a similar or even better performance than the machine learning classifiers. The main reason for this is the largely varying number of possible target word senses. While one relation might have only a single possible target sense, another one can have ten or even more. This tends to cause more false negatives in the machine learning methods and thus less relations that can be disambiguated. The finding is in line with previous work on gloss disambiguation: Moldovan and Novischi

| | en:en | | | | de:de | | | | en:de | | | | de:en | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | $A$ | $P$ | $R$ | $F_1$ | $A$ | $P$ | $R$ | $F_1$ | $A$ | $P$ | $R$ | $F_1$ | $A$ | $P$ | $R$ | $F_1$ |
| RAND | .74 | .65 | .65 | .65 | .70 | .69 | .66 | .67 | .72 | .52 | .60 | .55 | .67 | .44 | .46 | .45 |
| MFS | .81 | .75 | .74 | .74 | .79 | .78 | .76 | .77 | .79 | .62 | .72 | .67 | .79 | .64 | .66 | .65 |
| WKTWSD | .84 | .78 | **.80** | **.79** | .84 | .83 | **.83** | **.83** | .81 | .64 | **.75** | **.69** | .79 | .62 | **.71** | **.67** |
| BAYES | .85 | **.81** | .78 | **.79** | .84 | **.84** | .81 | .82 | .81 | .67 | .69 | .68 | .82 | .74 | .61 | **.67** |
| J48 | .83 | **.81** | .71 | .76 | .84 | .83 | .82 | **.83** | .79 | **.69** | .53 | .60 | .82 | **.82** | .53 | .64 |
| BEST | .85 | .79 | .80 | .80 | .85 | .84 | .83 | .84 | .80 | .63 | .75 | .69 | .81 | .67 | .73 | .70 |
| HUMAN | .91 | | | .89 | .92 | | | .92 | .89 | | | .80 | .90 | | | .83 |

Table 5.6: Performance of our disambiguation methods on the four evaluation datasets

(2004) note that compiling a sufficient set of training examples is not possible in many cases. Despite this, the machine learning methods achieve a higher precision. J48 even yields $P = .82$ for the (de:en) dataset. However, this always comes at the cost of a lower recall.

**Feature and error analysis**. Table 5.7 shows the precision $P$ and coverage $C$ (proportion of items covered by this feature) of using each feature $f \in F_{r,t_j}$ individually. With the exception of $f_{\text{idx=1}}$ (most frequent sense strategy), none of the features is able to disambiguate the whole dataset, but most of them achieve a very high precision on the covered items. It is not surprising that $f_{\text{mono}}$ performs extremely well ($P \in [.88, .96]$), since there is only one target word sense available for these cases. The feature $f_{\text{src}}$ performs well on the monolingual datasets ($P \in [.87, .97]$), but does not work at all on the cross-lingual task ($P \in [.38, .50]$). The reasons for this are ambiguities in the sense definitions that are often not resolved by the machine translation service. Parallel ambiguities such as *commission* and *Kommission*, which both mean either a group of people or a transaction fee of a broker, is a main source of errors here. Similar errors also occur for $f_{\text{inv}}$.

The word overlap feature $\hat{f}_{\text{ExtLesk}}$ generally yields a high precision. It is, in particular, higher than usually reported for word sense disambiguation tasks (cf. Navigli, 2009). The reason might be that we do not compare a sense definition with context words, but two definitions with each other and hence benefit from comparing texts that are specially crafted to characterize word senses. Interestingly, the imprecise translation of certain words noted for $f_{\text{src}}$ is less problematic for $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL\&TL}}$, as there are usually at least some correctly translated words in the sense definition. In our experiments, we found that $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL}}$ outperforms $\hat{f}_{\text{ExtLesk}\geq 2,\text{TL}}$, whereas $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL\&TL}}$ is only marginally better than $\hat{f}_{\text{ExtLesk}\geq 2,\text{SL}}$. The English Wiktionary is very sparse in encoding semantic relations (cf. section 4.7). The coverage of $f_{\text{rel}\geq 0.5}$ is therefore very low for all datasets involving English data.

Our final system output using WKTWSD concatenates the individual features and hence yields the same types of error. Since we manually defined the ordering of the features, we

| Feature | en:en | | de:de | | en:de | | de:en | |
|---|---|---|---|---|---|---|---|---|
| | *P* | *C* | *P* | *C* | *P* | *C* | *P* | *C* |
| $f_{\text{mono}}$ | .91 | 21 % | .94 | 22 % | .96 | 8 % | .88 | 8 % |
| $f_{\text{inv}}$ | .78 | 13 % | .89 | 31 % | .68 | 49 % | .67 | 41 % |
| $f_{\text{lbl}\geq 1}$ | .82 | 7 % | .90 | 5 % | .86 | 2 % | .60 | 4 % |
| $f_{\text{src}}$ | .87 | 10 % | .97 | 7 % | .50 | 20 % | .38 | 18 % |
| $f_{\text{rel}\geq 0.5}$ | .94 | 4 % | .90 | 14 % | .33 | 1 % | .75 | 1 % |
| $\hat{f}_{\text{ExtLesk}\geq 2}$ | .89 | 27 % | .99 | 12 % | .87 | 15 % | .93 | 17 % |
| $f_{\text{idx}=1}$ | .75 | 100 % | .78 | 100 % | .62 | 100 % | .64 | 100 % |

Table 5.7: Precision and coverage of each disambiguation feature

compare WKTWSD to a method BEST that concatenates the features in descending order of their precision on each dataset. The rationale behind this is that we make use of the best feature before moving to the next one. By comparing WKTWSD with BEST, we can measure the influence of our manually chosen ordering. Note, however, that BEST needs to be considered as an upper bound for WKTWSD rather than a separate method, because it made use of our analysis of the test data. The results are included in table 5.6. We observe that the order of the features plays only a minor role: WKTWSD and BEST are only slightly different although they concatenate the features in totally different ways. The largest difference accounts to .03 for the (de:en) dataset and is mostly due to the low performance of $f_{\text{src}}$.

**Summary.** We conclude that our approach is better suited for disambiguating Wiktionary relations than using off-the-shelf textual similarity measures. The features are effectively applied using a concatenation method. The training of machine learning classifiers could not improve these results in our experiments.

## 5.5 Constructing OntoWiktionary

In the previous sections, we have focused on the extraction and disambiguation of the information items encoded in Wiktionary. In order to build *OntoWiktionary*, we need to transform this knowledge into ontological structures. Pantel and Pennacchiotti (2008, p. 171 f.) call this process "ontologizing" the harvested knowledge. To this end, we determine concepts and lexicalizations based on the word senses and the sense-disambiguated relations, and we link them by means of ontological relationships.

### 5.5.1 Determining Concepts

OntoWiktionary consists of concepts and lexicalizations, such as the concept ›DOG‹ with the English lexicalizations *dog*, *hound*, *canine*. The lexicalizations of our ontology correspond to the word senses encoded in Wiktionary. In order to establish its concepts, we take advantage of the synonymy of the lexicalizations: *dog* is a synonym of *hound*, *hound* is a synonym of *canine*, and so forth. In wordnets, such sets of synonymous word senses are known as *synsets*, which may be used as the concepts of a lexical ontology (cf. Gangemi et al., 2003).

**Synset inference**. There are, however, no explicitly encoded synsets in Wiktionary. This is why we propose a method for inferring them based on the encoded synonymy relations. Mathematically, synsets are defined as the equivalence classes of a *partial equivalence relation*. That is to say, the synonymy relation defining the synsets fulfills two properties:

1. **(Symmetry)** $\forall s, t.\, (s, t) \Rightarrow (t, s)$.
2. **(Transitivity)** $\forall s, t, u.\, (s, t) \wedge (t, u) \Rightarrow (s, u)$.

The synonymy relations encoded in Wiktionary do not fulfill these two properties. We have rather observed in section 4.7 that they are encoded for each article individually. There is, for example, a synonymy relation from *boat* to *ship*, but not vice-versa, which violates the symmetry property.

We therefore compute the *transitive closure* of the encoded synonymy relation. This means we add all symmetric counterparts, and we infer all transitive relations by combining each possible pair of word senses sharing the same neighbor. Consider, for instance, the explicitly encoded synonymy relations (*island, oasis*), (*oasis, island*), and (*oasis, refuge*). We first add the missing symmetric counterpart (*refuge, oasis*) and then infer the transitive relations (*island, refuge*), and (*refuge, island*). This yields the synset {*island, oasis, refuge*} and accordingly the corresponding concept within OntoWiktionary.

**Data analysis**. Table 5.8 shows the number of synsets generated for the English, German, and Russian Wiktionary editions in comparison to the computational dictionaries introduced in chapter 4. We observe that the number of synsets for Wiktionary is considerably lower than that for the corresponding wordnets. Let $S = \{s_1, s_2, \ldots, s_n\}$ be a synset. We define the *synset size* $|S| = n$ (i.e., the number of synonymous word senses), the *number of explicit relations*

$$\mathrm{rel}(S) = |\{(s, t) \in d \mid s, t \in S\}|$$

denoting the number of explicitly defined synonym relations encoded in dictionary $d$, and the *synset cohesion*

$$\mathrm{coh}(S) = \frac{\mathrm{rel}(S)}{(|S| - 1) \cdot |S|},$$

which denotes the ratio of explicitly encoded synonymy relations to the number of possible synonymy relations. The synset {*island, oasis, refuge*} discussed above has, for example, the size $|S| = 3$, three explicitly encoded relations ($\text{rel}(S) = 3$), and the synset cohesion $\text{coh}(S) = {}^{3}/_{2 \cdot 3} = 50\,\%$. From the table, we find that the generated Wiktionary synsets have a relatively low synset cohesion. This corresponds to our previous observation that many potential edges are missing.

We find multiple examples in which a Wiktionary synset encodes complementary information to wordnets and that the generated synsets are generally of reasonable quality. The English WordNet synset {*stenographer, amanuensis, shorthand typist*} corresponds, for instance, to the Wiktionary synset {*stenographer, shorthand typist, stenographist, court reporter*} containing two additional lexicalizations. The organic compound {*kepone, chlordecone*} (used as an insecticide) is, for example, not encoded at all in WordNet. Larger synsets consist particularly of slang and jargon expressions, such as the interjection *dear me* with synonyms *heavens*, *oh my*, *great Scott*, *by George*, *good God*, and 44 other variants.

It is, however, notable that there exist one or two synsets with over 1,000 word senses. These are obviously invalid ones in which unrelated word senses are lumped together, for example, *distance*, *era*, and *leisure*. Besides errors of the relation disambiguation step described in the preceding section, a main reason for these huge synsets are loosely encoded synonymy relations connecting word senses that are only marginally similar. Consider the synonymy relations (*liberty, freedom*), (*freedom, exemption*), (*exemption, immunity*), which connect related word senses, but yield a gradual shift of meaning. We remove the synsets with over 1,000 word senses from OntoWiktionary.

**Evaluation**. To substantiate our observations, we randomly sample 100 English and 100 German synsets and ask two human raters to judge their quality.[60] We only consider synsets with at least three synonyms, because smaller synsets are not influenced by the problem of lumped word senses described above. Table 5.9 shows three examples illustrating the annotation task. The human judges receive the lemma and the sense definition of each synonym of the synset. Their task is to annotate the synset as "consistent", "lexically consistent", or "inconsistent".

The two word senses in (1) describe the same meaning, namely a singer in the bass range. Although there are subtle differences, such as that *basso* is used especially in the context of an opera, we ask our raters to ignore such subtle differences and judge the corresponding synsets as "consistent". Example (2) contains the additional synonym *singer*. We consider such examples "inconsistent", because a *bass* is a certain kind of *singer*, which we would expect in a separate synset that is connected to {*bass, basso*} by means of a subsumption relation. Example (3) is a special case, since the two words *bass* and *basso* are synonyms which one would expect in the same synset. However, the sense definitions reveal that the word senses describe totally different meanings. This is usually the result of an erroneous disambiguation

[60] The annotation study is based on Wiktionary data from February 2011.

| | English dictionaries | | | German dictionaries | | | Russian dictionaries | |
|---|---|---|---|---|---|---|---|---|
| | Wiktionary | WordNet | Thesaurus | Wiktionary | GermaNet | Thesaurus | Wiktionary | WordNet |
| *Number of synsets* | | | | | | | | |
| $\|S\| = 2$ | 9,244 | 33,914 | 11,406 | 1,685 | 14,582 | 10,705 | 3,981 | 35,902 |
| $\|S\| = 3$ | 2,144 | 11,678 | 4,707 | 1,375 | 2,862 | 5,786 | 987 | 8,597 |
| $\|S\| = 4$ | 811 | 4,668 | 2,393 | 843 | 1,323 | 3,156 | 403 | 2,316 |
| $\|S\| = 5$ | 346 | 1,853 | 1,407 | 498 | 312 | 1,775 | 205 | 618 |
| $5 < \|S\| \leq 10$ | 461 | 1,579 | 2,346 | 1,052 | 313 | 2,743 | 336 | 392 |
| $10 < \|S\| \leq 100$ | 113 | 119 | 515 | 243 | 12 | 869 | 109 | 7 |
| $100 < \|S\| \leq 1000$ | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| $\|S\| > 1000$ | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 0 |
| *Average synset size* | 2.67 | 2.66 | 3.43 | 4.36 | 2.41 | 3.78 | 2.98 | 2.35 |
| *Synset cohesion* | | | | | | | | |
| $\|S\| = 2$ | 59 % | 100 % | 100 % | 69 % | 100 % | 100 % | 62 % | 100 % |
| $\|S\| = 3$ | 42 % | 100 % | 100 % | 42 % | 100 % | 100 % | 44 % | 100 % |
| $\|S\| = 4$ | 33 % | 100 % | 100 % | 34 % | 100 % | 100 % | 32 % | 100 % |
| $\|S\| = 5$ | 28 % | 100 % | 100 % | 27 % | 100 % | 100 % | 28 % | 100 % |
| $5 < \|S\| \leq 10$ | 20 % | 100 % | 100 % | 20 % | 100 % | 100 % | 20 % | 100 % |
| $10 < \|S\| \leq 100$ | 10 % | 100 % | 100 % | 13 % | 100 % | 100 % | 10 % | 100 % |
| $100 < \|S\| \leq 1000$ | – | – | – | – | – | 100 % | – | – |
| $\|S\| > 1000$ | 0 % | – | – | 0 % | – | – | 0 % | – |
| *Average synset cohesion* | 52 % | 100 % | 100 % | 42 % | 100 % | 100 % | 52 % | 100 % |

Table 5.8: Comparison of synsets

| № | Lemma | Sense definition |
|---|---|---|
| (1) | *bass* | A male singer who sings in the bass range. |
| | *basso* | A bass singer, especially in opera. |
| (2) | *bass* | A male singer who sings in the bass range. |
| | *basso* | A bass singer, especially in opera. |
| | *singer* | person who sings, is able to sing, or earns a living by singing. |
| (3) | *bass* | The perch; any of various marine and freshwater fish resembling the perch. |
| | *basso* | A bass singer, especially in opera. |

Table 5.9: Example synsets illustrating our annotation study

of relations. We separate out these cases by asking our raters to judge these examples as "lexically consistent". The rationale behind this is to get in a position to estimate the impact of an erroneous disambiguation on the inference of synsets and to assess the usefulness of our synsets when considered at the lexical level. The latter corresponds to the use of a synonymy dictionary, such as the *Moby Thesaurus* (1996), which provides sets of synonymous words without describing their meaning or distinguishing word senses.

Each of the two raters has previous experience in linguistic annotation studies. In addition to the task description, we encourage the raters to consult other knowledge resources including Wiktionary to grasp the meaning of the individual word senses. They should not contact each other to ensure independent ratings. Table 5.10 shows the number of annotations associated by the two raters to the three categories and the corresponding inter-rater agreement. The high overall observed agreement of over .87 and the chance-corrected Cohen's $\kappa$ of .71 and .79 indicate that the dataset is reliable (Artstein and Poesio, 2008). Note that $\kappa$ is known to yield lower values if the distribution of categories is skewed, which is the case for our dataset. Following Fleiss (1971), we calculate the observed agreement $A_{O,i}$ and the kappa statistics $\kappa_i$ separately for each category $i$. While we find almost perfect agreement for the "inconsistent" category of the English dataset, none of the categories happens to be an outlier.

The vast majority of the synsets (59–70 % in the English and 65–77 % in the German dataset) are judged as "consistent", which demonstrates the general validity of our approach. An even larger portion is considered at least "lexically consistent": over 80 % of the English dataset and over 90 % of the German dataset are either "consistent" or "lexically consistent". This suggests that further improvement should concentrate on the relation disambiguation step (see section 5.6).

### 5.5.2 Determining Lexicalizations

Since the concepts of OntoWiktionary are based on synsets, it is straightforward to define the lexicalizations of the concepts: Each synonymous word sense within the synset serves as one

| | Rater A | Rater B | $A_O$ | $A_{O,i}$ | $\kappa$ | $\kappa_i$ |
|---|---|---|---|---|---|---|
| *English dataset* | 100 | 100 | .89 | | .79 | |
| consistent | 70 | 59 | | .92 | | .76 |
| lexically consistent | 13 | 21 | | .77 | | .72 |
| inconsistent | 17 | 20 | | .92 | | .90 |
| *German dataset* | 100 | 100 | .87 | | .71 | |
| consistent | 77 | 65 | | .92 | | .71 |
| lexically consistent | 17 | 25 | | .76 | | .70 |
| inconsistent | 6 | 10 | | .75 | | .73 |

Table 5.10: Evaluation of our synset inference method

lexicalization for the corresponding concept. The concept {*island, oasis, refuge*} receives, for instance, the three lexicalizations *island*, *oasis*, *refuge*.

In addition to that, we incorporate the translations encoded in Wiktionary to derive a multilingual lexical ontology. For the example above, we add the Finnish lexicalization *suojapaikka* (English: refuge), the French *île* (English: island), the Latin *refugium* (English: refuge), the Russian *убежище* (English: asylum), and many others. This task is only meaningful if word senses are taken into account. Otherwise, we obtain erroneous translations addressing the sense of a fertile region in a desert (*oasis*) or the area of land surrounded by water (*island*) instead of the figurative meaning. The quality of the derived multilingual lexicalizations hence depends on the disambiguation of the extracted information, which we discussed in section 5.4.

### 5.5.3 Determining Ontological Relations

A major limitation of the English Wiktionary is its sparseness of semantic relations, which we observed in section 4.7. This is why we propose enriching the semantic network of Wiktionary, which we achieve by generating bidirectional links and by inferring new semantic relations based on the disambiguated translations. After that, we transfer the semantic relations to the level of concepts.

**Bidirectional links.** We noted above for the inference of synsets that Wiktionary's synonymy relation is not symmetric. In contrast, wordnets usually consider all semantic relations as *bidirectional links*. Let $r = (s, t)$ be a relation of type $\tau$. We call $r$ bidirectional if, and only if, a symmetric or inverse counterpart $r' = (t, s)$ of type $\tau'$ exists. Synonymy and antonymy yield *symmetric relations* (i.e., $\tau' = \tau$), while hyponymy and hypernymy as well as holonymy and meronymy yield pairs of *inverse relations*. For example, the English article *boat* encodes a synonymy relation (*boat, ship*) and a hyponymy relation (*boat, canoe*), for which we generate a

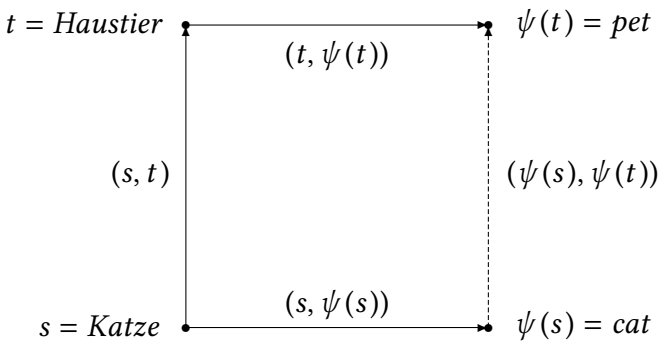



Figure 5.3: Cross-lingual inference of the semantic relation $(\psi(s), \psi(t))$

symmetric synonymy relation (*ship, boat*) and a hypernymy (i.e., inverse hyponymy) relation (*canoe, boat*). This approach has been previously discussed by Navarro et al. (2009), who also consider Wiktionary relations as bidirectional links. However, they do not take word senses into account and thus regard each relation only at the level of words. With the disambiguation of the relations discussed in section 5.4, we get in a position to go beyond that and consider the bidirectional links at the level of word senses. About three quarters of the relations in Wiktionary are *unidirectional links*. In total, we generate 44,739 symmetric and inverse counterparts for the English Wiktionary, 272,438 for the German Wiktionary, and 74,264 for the Russian Wiktionary.

**Cross-lingual relation inference.** Regarding the sparse semantic network of the English Wiktionary, we can especially benefit from the disambiguated translations by inferring semantic relations from other Wiktionary language editions. Let $(s, t) \in d_1$ be a disambiguated semantic relation in a Wiktionary language edition $d_1$ and let $(s, \psi(s))$ and $(t, \psi(t))$ be disambiguated translations of $s$ and $t$ into another language. Assuming a correct disambiguation of these three relations, we can infer a fourth relation $(\psi(s), \psi(t)) \in d_2$ in the foreign language edition $d_2$, since the meaning of $s$ and $t$ is preserved under the disambiguated translations. Figure 5.3 shows an example: For the German hypernymy relation (*Katze, Haustier*) and the corresponding translations (*Katze, cat*) and (*Haustier, pet*), we can infer the English hypernymy relation (*cat, pet*) that is currently not encoded in the English Wiktionary. Note that the inferred relation is also sense-disambiguated (i.e., both *cat* and *pet* refer to the animal-related word sense. Using this approach, we infer 614,573 additional semantic relations for the English Wiktionary.

**Synset-based relations.** To obtain ontological relationships, we transfer the semantic relations from the level of word senses to the level of concepts. Each word sense corresponds to exactly one synset (word senses without any synonyms represent a synset of size 1). We can thus unambiguously identify the concept of the source and the target word sense of a seman-

| OntoWiktionary | English | German | Russian |
|---|---|---|---|
| Concepts | 604,365 | 120,535 | 105,167 |
| Lexicalizations | 1,870,019 | 1,125,040 | 444,742 |
| *monolingual* | 639,480 | 175,837 | 130,629 |
| *multilingual* | 1,230,539 | 949,203 | 314,113 |
| Semantic relations | 822,152 | 1,116,544 | 208,688 |
| *explicitly encoded* | 162,840 | 774,642 | 134,424 |
| *inferred* | 659,312 | 341,902 | 74,264 |
| Ontological relations | 700,880 | 619,985 | 132,008 |

Table 5.11: Size of OntoWiktionary

tic relation. The hypernymy relation (*submarine, boat*) yields, for example, the subsumption relation (*{submarine, U-boat}, {boat, craft, ship}*), since *boat* is a lexicalization of the concept {*boat, craft, ship*}, whereas *submarine* is a lexicalization for {*submarine, U-boat*}. In addition to the subsumption relations "subsumes" and "subsumed by" which correspond to the semantic relations of type hypernymy and hyponymy, we transfer antonymy relations to "opposites" and loosely defined relations such as "see also" to ontological relationships of type "related" and include them into the final OntoWiktionary.

## 5.6 Discussion and Further Perspectives

**OntoWiktionary**. Table 5.11 shows the size of our lexical ontology OntoWiktionary. The English OntoWiktionary is about two times larger than *OpenCyc* (2012) with about 239,000 concepts, four times larger than *WordNet* (2006) with about 117,000 synsets, and ten times larger than *OntoWordNet* (2003) with about 60,000 concepts. The Wikipedia-based ontology *DBpedia* (2007 f.) contains about 3.7 million concepts, which are, however, mostly proper names, such as places, organizations, people, etc. Wiktionary focuses on any part of speech and, for example, also contains concepts for adverbs, such as {*forgivably, excusably, pardonably*}. From the German and the Russian Wiktionary editions, a considerably lower number of concepts can be generated, because these language editions are smaller than the English Wiktionary.

However, we observe a greater number of synonymy relations in those editions, which yields a higher number of monolingual lexicalizations provided for each concept: a concept has on average only 1.06 lexicalizations in the English OntoWiktionary, but 1.46 in the German and 1.24 in the Russian OntoWiktionary. The German OntoWiktionary contains the most multilingual lexicalizations (i.e., translations). Each concept has, on average, about 8 translations there compared to 2 in the English OntoWiktionary and 3 in the Russian OntoWiktionary.

Using our approach for determining ontological relations, we can effectively overcome the sparseness of the semantic network induced by the English Wiktionary. This is particularly due to the inference of semantic relations based on the disambiguated translations. We make OntoWiktionary publicly available from our homepage including a graphical user interface to browse its contents.

**Perspectives.** Our work proposes solutions for the three main challenges of compiling a lexical ontology from Wiktionary and motivates future research in these areas. First, sophisticated text mining software is required to extract the knowledge encoded in Wiktionary. A particular problem is processing any Wiktionary language edition. The emerging initiative *Wikidata*[61] is an important step towards this direction, as the goal of this project is providing a centralized repository of structured data that can be shared by different Wikimedia projects. This would allow obtaining consistent information across different Wiktionary language editions without noise. Translations and semantic relations could, for example, be stored in a wordnet-like fashion and referenced by the specific Wiktionary editions – similar to *OmegaWiki* (2006 f.). So far, the Wikidata project is, however, limited to *Wikipedia* (2001 f.), and the integration of other Wikimedia projects is planned for future releases.

Second, the disambiguation of translations has room for further improvement, as our algorithm relies on automatic translation of sense definitions. Using statistical machine translations for this introduces noise and is limited to certain language pairs. We see much potential in graph-based approaches. Flati and Navigli (2012) and Matuschek and Gurevych (2013) recently proposed promising algorithms in this direction based on cycles and shortest paths, respectively. Before such methods can be applied to Wiktionary, it is, however, necessary to overcome the sparseness of the semantic network, which is why we propose enriching the semantic network with inferred relations. An additional source for semantic relations is Wikisaurus. But by the time of writing, Wikisaurus is still in a development phase and limited to a few wiki pages (see section 2.2).

Third, the inference of synsets can be further improved, for example, by using clustering algorithms. Ideally, the result of a clustering approach corresponds to a consistent synset structure avoiding the overly large synsets that lump together marginally related word senses. The use of clustering methods, however, raises the question of finding an appropriate number of clusters. This is not obvious *per se*, but highly related to ongoing research efforts in the context of *word sense induction* (cf. Schütze, 1998).

Other future research options in the context of OntoWiktionary include the integration with other resources, providing its data in a standardized format, and employing it in natural language processing applications, which is the subject of the following chapters.

[61] `https://www.wikidata.org/` (28 May 2013)

## 5.7 Chapter Summary

In this chapter, we compiled a lexical ontology from the harvested Wiktionary knowledge. This is achieved by extracting knowledge from Wiktionary by means of a text mining software, disambiguating the extracted knowledge, and inferring an ontological structure from it. We identified the main challenges to be the disambiguation of semantic relations and translations as well as constructing synsets from the encoded synonymy relations, for which we propose and evaluate a solution. Our final ontology OntoWiktionary exceeds the size of expert-built wordnets, and it contains a large number of lexicalizations in many languages. Unlike Wikipedia-based ontologies, OntoWiktionary is not limited to nouns, but encodes concepts from any part of speech. Thus, OntoWiktionary fills the gap between the small expert-built wordnets and the large amount of encyclopedic knowledge from Wikipedia.

Chapter 6

# Resource Integration at the Word Sense Level

This chapter aims at integrating Wiktionary with other computational dictionaries at the level of word senses. We present our motivation in section 6.1 and discuss previous work on integrating resources in section 6.2. Then, we describe our method for automatically aligning the word senses of two computational dictionaries using the example of Wiktionary and WordNet (section 6.3). By carrying out an extensive annotation study, we get in a position to analyze how humans approach this task (section 6.4) and to evaluate the quality of our automatic alignment approach (section 6.5). In section 6.6, we discuss the resulting aligned resource, before we conclude the chapter in section 6.7.

## 6.1 Motivation

Though WordNet has been extensively used in knowledge-rich natural language processing systems, there is no best computational dictionary for all purposes. Jarmasz and Szpakowicz (2003) found, for example, better results for solving word choice problems when using *Roget's Thesaurus of English Words and Phrases* (1998) instead of *WordNet* (2002). In fact, we find a huge number of different computational dictionaries: The *ACL Special Interest Group on the Lexicon*[62] lists, for instance, over 50 different dictionary projects on their homepage, and the *LRE Map*[63] contains more than 3,900 resources of different type, which have been proposed as a source of background knowledge for natural language processing systems. In the realm of human-oriented dictionaries, there is even a larger diversity: the *OBELEX bibliography*[64] lists about 17,000 electronic dictionaries that could potentially be used as a machine-readable

[62] http://www.siglex.org (18 April 2013)
[63] http://www.resourcebook.eu (18 April 2013)
[64] http://www.owid.de/obelex/dict/en (18 April 2013)

dictionary. These resources typically differ in their coverage of words and word senses and in the information types they encode. This ranges from fundamental differences (like the distinction between linguistic and encyclopedic knowledge) to more specific ones, such as putting a focus on certain lemmas or including synonyms and translations into the dictionary articles. We have found in chapter 4 that the overlap of the eight dictionaries we analyzed is surprisingly low and that WordNet predominantly encodes word senses from social sciences and the humanities, whereas Wiktionary has a better coverage of technology-related domains and non-standard varieties of language. Using WordNet without further considerations thus limits the performance of a system, since each resource has its individual advantages.

This has caused increasing research in the area of resource integration. The main challenge in this process is the identification of identical word senses being differently represented by the dictionaries – a task that has become known as *word sense alignment*. It has been shown that sense-aligned resources yield synergies, which lead to better performance than using the resources individually. For instance, Shi and Mihalcea (2005) improve semantic parsing using the knowledge of an aligned resource of FrameNet, WordNet, and VerbNet, whereas Ponzetto and Navigli (2010) observe improvements for coarse-grained and domain-specific word sense disambiguation by using an alignment between WordNet and Wikipedia.

In another line of research, Wiktionary has been successfully applied in multiple natural language processing tasks, such as cross-lingual image retrieval (Etzioni et al., 2007), named entity recognition (Richman and Schone, 2008), or synonymy mining (Sajous et al., 2010). Zesch et al. (2008b) compare different semantic relatedness measures using either Wiktionary, Wikipedia, or WordNet and find the best results for Wiktionary (see also chapter 8). Its high coverage of languages, words, and word senses and its ability to grow continually due to the collaborative effort lets us expect an improved resource when combining Wiktionary with other computational dictionaries.

This is why we propose aligning Wiktionary and WordNet. The resulting alignment has two important advantages over using the dictionaries individually: First, an increased coverage and second, an enriched representation of word senses including semantic relations, translations, pronunciations, and many other information types. In the course of this chapter, we describe the following contributions:

**Contribution 5.1**: *We present and evaluate a method for automatically aligning Wiktionary and WordNet (section 6.3 and section 6.5).*

**Contribution 5.2**: *We create and analyze a new human-annotated evaluation dataset for aligning Wiktionary and WordNet (section 6.4).*

**Contribution 5.3**: *We analyze the characteristics of our aligned resource and how it can benefit natural language processing tasks (section 6.6).*

## 6.2 Related Work

During the last twenty years, there have been many works on aligning computational dictionaries at the level of word senses. Almost all alignment approaches for the English language include WordNet, which is the *de facto* standard resource in the field. Knight and Luk (1994) align WordNet with the *Longman Dictionary of Contemporary English* (1978) [LDOCE], Kwong (1998) similarly with LDOCE and *Roget's Thesaurus of English Words and Phrases* (1987), Litkowski (1999) with *HECTOR* (1993), Burgun and Bodenreider (2001) with the *Unified Medical Language System* (2001 f.) [UMLS], Reed and Lenat (2002) with *Cyc* (1995 f.), Shi and Mihalcea (2005) with *VerbNet* (2005) and *FrameNet* (2005), Navigli (2006) with the *Oxford Dictionary of English* (2003) [ODE], and Laparra and Rigau (2010) with *FrameNet* (2006).

The great potential of the collaboratively compiled *Wikipedia* (2001 f.) has motivated multiple works on aligning WordNet and Wikipedia. One line of research is the alignment of WordNet synsets and Wikipedia categories, which has been done based on the shared taxonomic structure (Toral et al., 2008), textual entailment and semantic relatedness methods (Toral et al., 2009), as well as graph algorithms (Ponzetto and Navigli, 2009). In addition to that, Wikipedia article pages have been aligned with WordNet synsets. The first work in this direction has been carried out by Ruiz-Casado et al. (2005) for the *Simple English Wikipedia* (2003 f.), which is a smaller version of the full Wikipedia using a restricted vocabulary. Most of the subsequent work focuses on the articles of the full English Wikipedia, which have been aligned to WordNet synsets based on: human judgments (Mihalcea, 2007), giving preference to WordNet's first sense (Suchanek et al., 2007), word overlap (de Melo and Weikum, 2010; Navigli and Ponzetto, 2010), and using semantic relatedness measures (Niemann and Gurevych, 2011). The resulting integrated resources *Universal WordNet* (2010 f.), *WordNet++* (2010), *BabelNet* (2012 f.), and *UBY* (2012) are among the most often used resources in natural language processing.

Table 6.1 shows an overview of the related work on aligning WordNet with different computational dictionaries. The table indicates whether only a subset or the entire dictionaries has been aligned and which type of method has been utilized. We classified them into methods: aligning the first sense [mfs], counting weighted or normalized word overlaps (including cosine similarity) [overlap], using syntactic patterns [syntax], considering the (graph) structure of the resource [structure], utilizing measures of semantic relatedness, such as semantic vectors or Personalized PageRank [relatedness], and aligning senses manually [manual].

Each word sense alignment approach has been evaluated on a separate, manually annotated dataset: De Melo and Weikum report a precision of $P = .85$, Navigli and Ponzetto observe $F_1 = .79$, and the alignment described by Niemann and Gurevych evaluates to $F_1 = .78$. It should be noted that these numbers are not comparable to each other, since they are based on different datasets and annotation schemes, which we describe in detail in section 6.4.

| Work | Method | Aligned with | Entire |
|---|---|---|---|
| Knight and Luk (1994) | overlap | LDOCE | ✓ |
| Kwong (1998) | overlap | LDOCE and Roget | ✘ |
| Litkowski (1999) | syntax | HECTOR | ✘ |
| Burgun and Bodenreider (2001) | overlap | UMLS | ✘ |
| Reed and Lenat (2002) | manual | Cyc | ✘ |
| Shi and Mihalcea (2005) | manual/structure | VerbNet and FrameNet | ✓ |
| Navigli (2006) | relatedness | ODE | ✓ |
| Laparra and Rigau (2010) | structure | FrameNet | ✓ |
| Toral et al. (2008) | structure | Wikipedia categories | ✓ |
| Toral et al. (2009) | relatedness | Wikipedia categories | ✓ |
| Ponzetto and Navigli (2009) | structure | Wikipedia categories | ✓ |
| Ruiz-Casado et al. (2005) | overlap | Simple Wikipedia articles | ✓ |
| Mihalcea (2007) | manual | Wikipedia articles | ✘ |
| Suchanek et al. (2007) | mfs | Wikipedia articles | ✓ |
| de Melo and Weikum (2010) | overlap | Wikipedia articles | ✓ |
| Navigli and Ponzetto (2010) | overlap | Wikipedia articles | ✓ |
| Niemann and Gurevych (2011) | relatedness | Wikipedia articles | ✓ |
| This work | relatedness | Wiktionary senses | ✓ |

Table 6.1: Previous work on aligning WordNet

To the best of our knowledge, our work is the first word sense alignment covering Wiktionary. In section 6.6, we discuss multiple future efforts based on our proposed alignment between Wiktionary and WordNet.

## 6.3 Aligning Wiktionary and WordNet

**Word sense alignment.** As part of our quantitative comparison of computational dictionaries in section 4.4, we have created an alignment between two dictionaries at the level of lexical entries (i.e., a *lexical alignment*). That is to say, we align two lexical entries if, and only if, they share the same lemma and part of speech. Obtaining a lexical alignment is exact and unambiguous, as it solely requires string matching.

In section 4.5, we found that analyzing the overlap of word senses requires an alignment of the dictionaries at the level of word senses. This is, however, a substantially more complex task due to the large differences in the representation of meaning. The encoded word senses typically vary in their number, granularity, domain-specificity, and conciseness of their definition. The WordNet synset {*plant, works, industrial plant*} defined as "buildings for carrying on industrial labor" describes, for instance, the same meaning as the fifth Wiktionary word sense

of *plant* defined as "a factory or other industrial or institutional building or facility". These two word senses should be *aligned* – i.e., they should be part of a word sense alignment between the two dictionaries. Another Wiktionary word sense of *plant* defined as "an organism that is not an animal [...]" denotes a different meaning and should therefore not be aligned to the WordNet synset.

Accordingly, a *word sense alignment* between two computational dictionaries $d_1$ and $d_2$ is a set of word sense pairs

$$\textsc{Align}(d_1, d_2) = \{(s_{d_1}, s_{d_2}) \mid \text{meaning}(s_{d_1}) = \text{meaning}(s_{d_2})\}$$

sharing the same meaning.[65] The goal of an automatic word sense alignment method is thus to compare two word senses $s_{d_1}$ and $s_{d_2}$ and decide whether they are *corresponding* (i.e., if they share the same meaning). To exemplify our work towards this goal, we compile an automatic word sense alignment between Wiktionary and *WordNet 3.0* (2006). The method itself can, however, be generalized to other resources, which we discuss in section 6.6

**Approach**. Most previous word sense alignments are based on a one-to-one alignment assumption. This means that each sense is aligned with exactly one sense in the other dictionary:

$$\forall s \in d_1.\ \forall u, v \in d_2.\quad (s, u) \in \textsc{Align}(d_1, d_2) \wedge (s, v) \in \textsc{Align}(d_1, d_2) \rightarrow u = v,$$

$$\forall s, t \in d_1.\ \forall u \in d_2.\quad (s, u) \in \textsc{Align}(d_1, d_2) \wedge (t, u) \in \textsc{Align}(d_1, d_2) \rightarrow s = t.$$

Niemann and Gurevych (2011, p. 206 f.) argue that there are word senses requiring none, one, or multiple aligned senses. This also holds for aligning Wiktionary and WordNet. The Wiktionary word sense "the people who decide on the verdict; the judiciary" for the noun *bench* can, for example, be aligned to the two WordNet synsets {*judiciary, bench*} "persons who administer justice" and {*Bench*} "the magistrate or judge or judges sitting in court in judicial capacity to compose the court collectively". Accordingly, the Wiktionary word sense "the bottom part of a sand casting mold" for the noun *drag* is not covered by any WordNet synset and should therefore not be aligned.

This is why we follow the alignment approach by Niemann and Gurevych, which includes the state-of-the-art word sense disambiguation method by Agirre and Soroa (2009). This method is known to outperform purely word overlap based measures. The method consists of the two steps candidate extraction and candidate alignment, which we describe below.

**Candidate extraction**. In the *candidate extraction* step, the algorithm iterates over all word senses from one computational dictionary and extracts suitable candidates from the other dic-

[65] Other terms previously used in the literature are *sense mapping* or *sense matching*. The notion of a word sense alignment is not to be mixed up with *word alignment* or *sentence alignment*, which are used for processing parallel texts (e.g., in machine translation). For brevity, we do not distinguish between sense and synset here.

tionary that *might* form a valid alignment. In our case, we iterate over all synsets in WordNet and extract all word senses from Wiktionary that are encoded for one of the synset's synonymous words. For example, we extract all nine Wiktionary word senses from the article *plant* and all four word senses from the article *works* as the candidates for the WordNet synset {*plant, works, industrial plant*}. Note that there is no corresponding article for the lemma *industrial plant* in Wiktionary.

**Candidate alignment**. In the *candidate alignment* step, each candidate is then scored with two similarity measures based on word overlap (cosine similarity) and semantic relatedness (Personalized PageRank).

The *cosine similarity* (COS) denotes the cosine of the angle between a vector representation of the two senses $s_1$ and $s_2$:

$$\mathrm{COS}(s_1, s_2) = \frac{\mathrm{BoW}(s_1) \cdot \mathrm{BoW}(s_2)}{||\mathrm{BoW}(s_1)||\ ||\mathrm{BoW}(s_2)||}$$

To represent a sense as a vector, we use a bag-of-words approach. That is, a vector BoW($s$) containing the term frequencies of all words in the description of $s$. The Wiktionary sense "a factory or other industrial or institutional building or facility" can, for example, be represented as BoW($s$) = {a(1), building(1), facility(1), factory(1), industrial(1), institutional(1), or(3), other(1)}. Note that there are different options for choosing the description of sense $s$: For WordNet, the definition of the synset can be used alone or in combination with its hyponyms and/or hypernyms. For Wiktionary, we can choose between sense definition, usage examples, and related words of the word sense. We will discuss the best configuration during our evaluation in the following section.

The measure based on the *Personalized PageRank* algorithm (PPR) estimates the semantic relatedness between two word senses $s_1$ and $s_2$ by representing them in a semantic vector space and comparing these semantic vectors $\mathbf{Pr}_{s_1}$ and $\mathbf{Pr}_{s_2}$ by computing

$$\mathrm{PPR}(s_1, s_2) = 1 - \sum_i \frac{\left(\mathbf{Pr}_{s_1,i} - \mathbf{Pr}_{s_2,i}\right)^2}{\mathbf{Pr}_{s_1,i} + \mathbf{Pr}_{s_2,i}},$$

which is a $\chi^2$ variant introduced by Niemann and Gurevych (2011). The main idea of choosing **Pr** is to use the Personalized PageRank algorithm for identifying those synsets that are central for describing the meaning of a word sense. The word sense "buildings for carrying on industrial labor" is, for instance, well-represented by the WordNet noun synsets {*plant, works, industrial plant*}, {*building complex, complex*}, and the adjective synset {*industrial*}. These synsets should have a high centrality (i.e., a high PageRank score), which is calculated as

$$\mathbf{Pr} = c\, M\, \mathbf{Pr} + (1 - c)\, \mathbf{v},$$

```
1  function Align(WordNet, Wiktionary)
2    alignment := ∅;
3    for each synset ∈ WordNet.getSynsets() do
4      // Step 1: Candidate extraction
5      candidates := ∅;
6      for each word ∈ synset.getSynonyms() do
7        candidates := candidates ∪ Wiktionary.getWordSenses(word);
8      // Step 2: Candidate alignment
9      for each candidate ∈ candidates do
10       sim_COS := COS(synset, candidate);
11       sim_PPR := PPR(synset, candidate);
12       if sim_COS ≥ τ_COS ∧ sim_PPR ≥ τ_PPR then
13         alignment := alignment ∪ (synset, candidate);
14   return alignment;
15 end.
```

Figure 6.1: Pseudo code of the automatic alignment algorithm

with the damping factor $c$ controlling the random walk, the transition matrix $M$ of the underlying semantic graph, and the probabilistic vector $\mathbf{v}$, whose $i^{\text{th}}$ component $\mathbf{v}_i$ denotes the probability of randomly jumping to node $i$ in the next iteration step. Unlike in the traditional PageRank algorithm, the components of the jump vector $\mathbf{v}$ are not uniformly distributed, but personalized to the sense $s$ by choosing $\mathbf{v}_i = 1/m$ if at least one synonymous word of synset $i$ occurs in the bag-of-words description BoW of sense $s$, and $\mathbf{v}_i = 0$ otherwise. The normalization factor $m$ is set to the total number of synsets that share a word with the sense description, which is required for obtaining a probabilistic vector. As part of our implementation for this method, we use the publicly available UKB software (Agirre and Soroa, 2009) for calculating the PageRank scores, and we utilize the *WordNet 3.0* (2006) graph augmented with the *Princeton WordNet Gloss Corpus* (2008) as the transition matrix $M$ of the underlying semantic graph. We set the damping factor $c$ to .85.

**Alignment decision.** Having calculated the similarity scores, we add the pair of the WordNet synset and the Wiktionary sense to our alignment if both similarity scores are above a certain threshold $\tau_{\text{COS}}$ and $\tau_{\text{PPR}}$. We learn these thresholds in a ten-fold cross validation on our annotated dataset that is explained in the following section. The optimal thresholds have been determined independently from each other using a simple binary split of the fold's items. The final thresholds are $\tau_{\text{COS}} = .13$ and $\tau_{\text{PPR}} = .49$. Figure 6.1 shows the described alignment algorithm in pseudo code.

## 6.4 Annotation Study

A common way to evaluate the quality of an automatic word sense alignment algorithm is comparing its results with human judgments from an *annotated dataset*. Creating a reliable annotated dataset turns, however, out to be a very challenging task. Datasets of previous studies are either very small or suffer from being annotated by only one human rater, which might yield biased results. In addition to that, we are not aware of any annotated datasets on aligning Wiktionary and WordNet, which is why we create a new annotated dataset. Our goal is to overcome the limitations of previous annotated datasets regarding size and reliability. To this end, we carry out a careful analysis of the agreement among the human raters. Our dataset is publicly available for future work on aligning WordNet and Wiktionary.

**Data sampling**. Niemann and Gurevych (2011) introduce a well-balanced dataset for evaluating an alignment of WordNet and Wikipedia. Their sampled WordNet synsets are uniformly distributed in the number of synonyms, distance to the root node, and unique beginners. Because the word senses of two dictionaries are very diverse in terms of number, granularity, subject field, and specificity, such a balanced dataset is important for being able to judge the quality of an alignment algorithm as unbiased as possible. Therefore, we reuse 320 synsets from their dataset as a primer for our evaluation dataset. For each synset, we extract all possible word senses from Wiktionary according to the candidate extraction step introduced in the previous section. This results in 2,423 sense pairs that are to be annotated by the human raters.[66] This dataset is substantially larger than reported for previous works: Ruiz-Casado et al. (2005) annotate 180 pairs, Fernando and Stevenson (2010) 200 pairs, Wolf and Gurevych (2010) 297 pairs, Ponzetto and Navigli (2010) 1,000 pairs, and Niemann and Gurevych (2011) annotate 1,815 pairs. With the exception of the last work, all datasets have been sampled randomly, which does not guarantee a well-balanced distribution over the entire dictionary.

**Annotation**. We create a spreadsheet listing each of the 2,423 sense pairs in a separate row containing the lemma, sense definitions, and example sentences (if available) of the sense pair. Table 6.2 shows an excerpt of this spreadsheet with some example annotations on the lemma *hedonism*.[67] We have asked ten human raters (marked as A–J henceforth) to annotate each sense pair as describing the same meaning (class 1, "should be aligned") or describing a different meaning (class 0, "should not be aligned"). For each word sense, there can be multiple aligned word senses of the same lemma. This is, for example, the case for a coarse-grained word sense from the first dictionary, which has been split into multiple, more specific word senses within the second dictionary. Since we also expect word senses that are solely found

[66] The annotation study is based on Wiktionary data from April 2010.

[67] To support the annotator's work, we marked different sense definitions with alternating background colors in the original annotation spreadsheet.

| № | Lemma | WordNet | Same Sense | Wiktionary |
|---|---|---|---|---|
| (1) | hedonism | the pursuit of pleasure as a matter of ethical principle | 0 | (ethics) The belief that pleasure or happiness is the highest good in life. [...] |
| (2) | hedonism | the pursuit of pleasure as a matter of ethical principle | 1 | A general devotion to the pursuit of pleasure. |
| (3) | hedonism | an ethical system that evaluates the pursuit of pleasure as the highest good | 1 | (ethics) The belief that pleasure or happiness is the highest good in life. [...] |
| (4) | hedonism | an ethical system that evaluates the pursuit of pleasure as the highest good | 0 | A general devotion to the pursuit of pleasure. |

Table 6.2: Excerpt of our annotated dataset for aligning Wiktionary and WordNet

in one of the dictionaries, we also permit the annotation of each word sense candidate with class 0. For unclear cases, we provide an optional field in each row, which may be used by the raters to take a short note or comment.

The raters are students in computer science, math, or linguistics. Two of them (I and J) have previous experience with annotation studies. We describe the annotation task in an annotation guidebook and train the raters with some example cases. The raters are allowed to consult both WordNet and Wiktionary for clarifying a certain word sense (e.g., by taking semantic relations into account). Other resources, such as dictionaries or encyclopedias, are also permitted. However, the raters should not discuss their decisions with each other in order to foster an independent judgment of each rater.

**Inter-rater agreement**. To ensure the reliability of our annotated dataset, we calculate the inter-rater agreement between the raters. All ten raters agreed on 1,987 sense pairs. Their average observed agreement is $A_O = .93$. In addition to that, we measure the chance-corrected inter-rater agreement using Fleiss' $\kappa = .70$ (Fleiss, 1971) and Krippendorff's $\alpha = .74$ (Krippendorff, 2004).[68]

While $\kappa$ measures the agreement at the level of the 2,423 annotation items, we apply $\alpha$ at the level of the 165 unique lemmas in our dataset. For each lemma, we create the set of the corresponding word sense pairs and employ the MASI distance function (Passonneau, 2006) for calculating the $\alpha$ score. For two sets of annotated sense pairs, MASI returns 0 if they are

[68] As explained in detail by Artstein and Poesio (2008), Fleiss' $\kappa$ is a generalization of Scott's $\pi$ rather than a generalization of the commonly used Cohen's $\kappa$ in the two-rater case. We report Fleiss' $\kappa$ here, since it is the most often used measure for more than two raters. Apart from that, the difference to Davies and Fleiss' generalization of Cohen's $\kappa$ is only marginal (in our case .0003).

| $\kappa$ | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| B | .72 | | | | | | | | | |
| C | .60 | .64 | | | | | | | | |
| D | .72 | .75 | .60 | | | | | | | |
| E | .73 | .72 | .63 | .74 | | | | | | |
| F | .64 | .65 | **.58** | .65 | .68 | | | | | |
| G | .75 | .72 | .66 | .73 | .75 | .64 | | | | |
| H | .67 | .72 | .60 | .72 | .68 | .64 | .68 | | | |
| I | .75 | .74 | .64 | .77 | .76 | .67 | .79 | .73 | | |
| J | .72 | .75 | .62 | .77 | .77 | .67 | .76 | .73 | **.80** | |
| ∅ | .70 | .71 | **.62** | .72 | .72 | **.65** | .72 | .69 | .74 | .73 |

Table 6.3: Pairwise $\kappa$ of our annotation study

equal, ⅓ if one is a subset of the other, ⅔ if their intersection is non-empty, or 1 otherwise. The annotation shown in table 6.2 yields, for example, the set {(2), (3)}, which would result in a distance of ⅓ to the annotation {(1), (2), (3), (4)} of another rater, who annotated each row as class 1. Calculating the inter-rater agreement at the level of lemmas puts us in a position to assess whether the disagreement originates from the annotation of only a few lemmas or is spread over the whole dataset.

Table 6.3 shows the pairwise (Cohen's) $\kappa$ for each pair of raters. The annotators C and F have the lowest inter-rater agreement between each other (.58) and with all other raters ( .62 and .65). These two raters are thus on the opposite sides of the scale. Further analysis reveals that C is biased towards class 0 (different meaning) and F is biased towards class 1 (same meaning). Since C and F systematically deviate from the remaining group of annotators, we remove their annotations in order to obtain a more reliable dataset. Removing the annotations of certain raters is not unproblematic, because this could imply removing the hard cases an automatic system should learn to tackle. (Krippendorff, 1980, p. 150), however, argues that they should be "removed, checked, or recoded" if they deviate systematically, as it is the case for C and F. Without considering C and F, we obtain an inter-rater agreement of $\kappa = .74$ and $\alpha = .78$.

A dataset with such an agreement is considered reliable and allows to draw tentative conclusions (Krippendorff, 1980) – although its agreement is lower than reported for WordNet–Wikipedia alignment datasets. More precisely, Niemann and Gurevych (2011) report $\kappa = .87$ and Navigli and Ponzetto (2010) measure $\kappa = .90$. Since even the two skilled annotators I and J only obtained an agreement of $\kappa = .80$, we conclude that the task of aligning Wiktionary and WordNet is more difficult than aligning Wikipedia and WordNet. This does not come as a surprise, because Wikipedia contains encyclopedic knowledge, which is largely complementary to the linguistic knowledge found in WordNet. In contrast, Wiktionary and WordNet are both

| Group | Value | Items | $\kappa$ | Error($\kappa$) |
|---|---|---|---|---|
| Commented | no | 1,989 | .76 | ± .000 |
| | yes | 434 | .51 | ± .001 |
| Frequency | ≤ 5 | 322 | .71 | ± .001 |
| | 6…100 | 461 | .71 | ± .001 |
| | > 100 | 1,640 | .67 | ± .000 |
| Sense pairs | 1…5 | 222 | .66 | ± .002 |
| | 6…15 | 265 | .59 | ± .002 |
| | 16…30 | 309 | .71 | ± .001 |
| | 31…50 | 276 | .60 | ± .002 |
| | 51…63 | 233 | .67 | ± .002 |
| | 64…70 | 268 | .68 | ± .002 |
| | 71…100 | 323 | .80 | ± .001 |
| | 101…150 | 367 | .57 | ± .001 |
| | > 150 | 160 | .66 | ± .003 |
| Specificity | top-level | 62 | .52 | ± .008 |
| | inner node | 1,049 | .68 | ± .000 |
| | leaf node | 1,312 | .72 | ± .000 |

Table 6.4: Inter-rater agreement per group of annotation items

dictionaries and thus require the distinction of very subtle differences in the sense definitions. Nevertheless, we carefully study the sources for disagreement of our raters below.

**Sources of disagreement.** From table 6.3 we observe the highest pairwise agreement between the annotators I and J, which both had previous annotation experience. I and J also obtained the highest average agreement with the other annotators. Previous annotation experience can thus be recorded as helpful for this annotation task.

In order to diagnose groups of annotation items with a particular high or low agreement, we divide our dataset according to the use of comments, the occurrence frequency of the lemma, the number of sense pairs per lemma, and the specificity of the synset. Table 6.4 shows the number of annotation items, the $\kappa$ statistics, and its standard error for each group. We find a strong influence of the comment field: While the inter-rater agreement is $\kappa = .76$ for the 1,989 items which have not been commented by any rater, it is only $\kappa = .52$ for the remaining 434 items with at least one comment. Thus, the raters tend to comment those items that they are uncertain of. We can make use of this fact to identify hard cases in future annotation studies, which should be annotated by a higher number of raters or critically discussed and analyzed in order to obtain datasets of higher quality and reliability.

For each lemma, we count its occurrence frequency in the one million English words corpus of the *Leipzig Corpora Collection* (Quasthoff et al., 2006). We then measure the inter-rater agreement for the groups of annotation items having a lemma with low, medium, or high frequency. Although we observe a slight trend towards a lower agreement for high frequent lemmas, the difference between the groups is not very prominent. This lets us conclude that the occurrence frequency of the lemma has only a minor influence on the total agreement.

The lemma *hedonism* shown in table 6.2 has two senses in WordNet and two senses in Wiktionary. Hence, there are $2 \times 2 = 4$ sense pairs in our dataset that are to be annotated. We divide the number of sense pairs per lemma into groups of comparable size and calculate the correlation between the number of sense pairs and the inter-rater agreement. A high (negative) correlation would imply that the raters lose the overview if too many sense pairs are to be considered for a single lemma. Our results, however, show a coefficient of determination of merely $R^2 = .10$ indicating that there is no correlation between the number of annotation items and the inter-rater agreement. This confirms the observation by Brown et al. (2010), who analyze the influence of the number and the granularity of word senses on the inter-rater agreement in a word sense disambiguation setting. They find that the number of word senses has no or only a small influence.

We finally group the annotation items by their specificity. The term *automobile* is, for example, more specific than the term *vehicle* or the abstract notion of a *physical entity*. To estimate the specificity of a word sense, we evaluate its position within the WordNet taxonomy: we consider word senses represented as a top-level node (e.g., *agent*), as an inner node (e.g., *plant*), and as a leaf node (e.g., *cashew*) within the taxonomy graph. We observe the best inter-rater agreement for the leaf nodes. The agreement is slightly lower for the inner nodes. For the top-level nodes, we find, however, a substantially lower agreement compared to the more specific word senses. Deciding on the alignment of these very general concepts is thus more difficult for human raters. Future alignment approaches should take this finding into account and clarify how top-level nodes are to be aligned, before carrying out an annotation study or using automatic alignment methods.

## 6.5 Alignment Evaluation

**Gold standard dataset**. We create a *gold standard* from our annotated dataset, which we use to evaluate the quality of our automatic alignment approach. One way of compiling this gold standard would be removing those items the raters disagreed on. But we refrained from doing so, as these cases might be the most interesting ones, which should be properly tackled by our system. Instead, we remove only the annotations of the systematically deviating raters C and F as discussed in the previous section and rely on the majority vote of the remaining eight annotators. For breaking the 27 ties, we ask an additional adjudicator. The adjudicator has previous experience with annotation studies and receives the annotation guidebook.

| Method | $A$ | $P$ | $R$ | $F_1$ |
|---|---|---|---|---|
| RAND | .662 | .212 | .594 | .313 |
| MFS | .802 | .329 | .508 | .399 |
| COS | .901 | .598 | **.703** | .646 |
| PPR | **.915** | **.684** | .636 | .659 |
| COS&PPR | **.914** | .674 | .649 | **.661** |
| HUMAN | .937 | – | – | .775 |

Table 6.5: Evaluation results of our alignment algorithm

**Alignment quality**. Following Navigli and Ponzetto (2010), we compare our automatic word sense alignment with the gold standard and measure accuracy $A$, precision $P$, recall $R$, and the $F_1 = \frac{2PR}{P+R}$ score. As baseline approaches, we implement a first sense heuristic (MFS) and a method making a random selection (RAND). Table 6.5 shows the results of these baselines in comparison to our COS and PPR measures as well as their combination (COS&PPR). As an upper bound, we use the human performance by providing the observed agreement $A_O$ and the $F_1$ score among our raters (HUMAN) following Hripcsak and Rothschild (2005).

As noted in section 6.3, there are multiple options for representing the word senses as a bag-of-words. For WordNet, we try the synonyms and the definition of the synset, as well as its direct hypernym and hyponyms. For Wiktionary, we try the lemma, the sense definition, example sentences, and the encoded synonyms. We evaluate all possible combinations and find the best results when using the synonyms and the definition of the WordNet synset and of its direct hypernym together with all four Wiktionary features.

Our COS, PPR, and COS&PPR methods outperform the baselines by far. The improvement is statistically significant using McNemar's test ($p < .01$). While COS has the highest recall and PPR has the highest precision, COS&PPR is a reasonable trade-off yielding the highest $F_1$ score. The difference of PPR and COS&PPR over COS is again statistically significant ($p < .01$). The difference between PPR and COS&PPR is not statistically significant, which leads us to the conclusion that the PPR and COS&PPR methods perform equally well for our alignment task.

In the preceding section, we observed a lower inter-rater agreement for our dataset than for WordNet–Wikipedia alignments. This effect also becomes visible in our evaluation results: While Niemann and Gurevych (2011) measure an $F_1$ score of .53 for their MFS baseline and .78 for their COS&PPR method, the results are between .12 to .14 lower for the WordNet–Wiktionary alignment. For this reason, we hypothesize that aligning two dictionaries at the word sense level (e.g., WordNet and Wiktionary) is a more complex task than aligning a dictionary with an encyclopedia (e.g., WordNet and Wikipedia).

**Error analysis.** We carry out a detailed error analysis to identify the main types of errors made by our algorithm. Of the 2,423 sense pairs in the dataset, the COS&PPR method yields 98 false positives and 110 false negatives. Regarding the *false negatives* (i.e., the sense pairs that the method could not align, although they represent the same meaning), we identify three main error classes:

(1) The sense definitions are very different in their choice of words, such as in "good discernment" and "ability to notice what others might miss" for the lemma *eye*. These errors are hard to resolve, as they require deep understanding and world knowledge.
(2) The sense definitions are rather similar (e.g., "any of various plants of the genus Centaurea [...]" and "any of various common weeds of the genus Centaurea" for the lemma *knapweed*), but the similarity score of one of the two measures is slightly below the chosen threshold. These errors are caused by using fixed similarity thresholds, which could, for instance, be improved by using machine learning for aligning the sense pairs.
(3) References to derived words occur in the sense definitions. An example is the lemma *pacification*, which is described as "the process of pacifying" and thus refers to the definition of *pacifying*. This kind of error can be alleviated by taking the definitions of the derived words into account, which, however, raises the question of disambiguating the derived word.

Amongst the *false positives* (i.e., the automatically aligned sense pairs with different meanings), we mainly find two error classes:

(1) There are highly related word senses, such as "a computer that provides client stations with access to files and printers as shared resources to a computer network" and "any computer attached to a network" for *host*, which are clearly related, but differ in their extension: The former requires the *host* to provide shared resources; the latter does not. Although these word senses do not represent exactly the same meaning, their alignment is very useful for many natural language processing applications, for instance, for a semantic information retrieval system, which often does not require to make subtle sense distinctions when searching relevant documents. Future work could distinguish between sense alignments sharing the same meaning and sharing a highly related meaning, for example, by using a graded scale for the alignment annotations (cf. Erk et al., 2009; Eom et al., 2012).
(2) Another major class of errors is due to an erroneous interpretation of a definition's meaning. Consider again the computing-related word sense of *host*. This sense is also aligned to "any organization that provides resources and facilities for a function or event", because the words *resource, facility, function*, and *event* also frequently occur in the computer science domain. These errors are hard to resolve, but future work could investigate the influence of a sense's position in the taxonomy of a thesaurus.

## 6.6 Discussion and Further Perspectives

Aligning computational dictionaries is only one side of the coin. The other side is the question, how natural language processing applications can benefit from the aligned resource. The main advantages of our alignment of Wiktionary and WordNet are the increased coverage and the enriched representation of word senses.

**Increased coverage**. Coverage is crucial for almost every natural language processing task. Our final Wiktionary–WordNet alignment consists of 315,583 candidates, of which 56,970 sense pairs are marked as alignments. For 60,707 WordNet synsets there has been no corresponding word sense found in Wiktionary, and, vice versa, there are 371,329 Wiktionary word senses that have not been aligned with any WordNet synset. The term *devisor* is, for instance, only found within WordNet, and *libero* merely has an entry in Wiktionary. Our newly aligned resource contains 488,988 word senses.

Table 6.6 shows the number of word senses per part of speech, which are shared by the two dictionaries and which have no alignment with the other one. The extremely high number of word senses only occurring in Wiktionary can be explained by the 106,328 inflected word forms that are not encoded by WordNet. While the vast majority of encoded senses are nouns, also the coverage of other parts of speech benefits from the alignment of the two dictionaries. This is a clear advantage over Wikipedia–WordNet alignments, which usually focus on nouns only. Besides verbs, adjectives, and adverbs that are also encoded by WordNet, Wiktionary additionally contains pronouns, phrases, idioms, sayings, etc. (see section 4.4).

Pantel and Lin (2002, p. 613) note that manually compiled computational dictionaries "miss many domain specific senses". In section 4.6, we have observed that Wiktionary and WordNet differ largely in the domains covered by the encoded word senses. Thus, our aligned resource profits from combining the domain-specific word senses from either dictionary. This allows a natural language processing application to quickly adapt to a new domain or being applied to a cross-domain setting.

**Enriched sense representation**. Wiktionary is rich in its variety of information types including etymology, alternative spellings, pronunciations, sense definitions, related words, translations, and many more (see section 2.6). De Melo and Weikum (2010) extract, for example, alternative spellings and etymologies from Wiktionary for enriching their lexical database. However, they do not align their resource with Wiktionary at the word sense level and thus cannot make use of the semantic information types found therein. In combination with the information types encoded in WordNet, such as the large number of synonyms and the rigid subsumption hierarchy, our aligned resource yields an enriched representation of word senses.

| | Overlap | only Wiktionary | only WordNet |
|---|---|---|---|
| Nouns | 34,464 | 158,085 | 47,651 |
| Verbs | 8,252 | 29,119 | 5,515 |
| Adjectives/Adverbs | 14,236 | 60,977 | 7,541 |
| Other parts of speech | 0 | 16,778 | 0 |
| Inflected word forms | 0 | 106,328 | 0 |
| Total | 56,952 | 371,329 | 60,707 |

Table 6.6: Coverage of our aligned resource

**Applications**. The potential of aligned resources has been previously shown by many researchers: Shi and Mihalcea (2005), for instance, align *FrameNet* (2005) and *VerbNet* (2000 f.) with *WordNet* (2003) and obtain improved results for semantic parsing. A similar approach has been followed by Loper et al. (2007), who align *VerbNet* (2000 f.) and *PropBank* (2004) for improving semantic role labeling. Recently, Ponzetto and Navigli (2010) have used their Wikipedia–WordNet alignment to improve a knowledge-based word sense disambiguation system beyond the results of state-of-the-art supervised systems. Automatic word sense alignments can also be very valuable for extending computational dictionaries. Niemi et al. (2012, p. 231) note, for instance, that even "if imperfect, such methods can speed up the manual verification by often providing good suggestions".

Our alignment of Wiktionary and WordNet now allows for further work in these directions by (1) exploiting the high coverage of our aligned resource, and (2) using the enriched representation of senses. Apart from semantic parsing and word sense disambiguation noted above, also semantic relatedness is an promising task, since Zesch et al. (2008b) found very good results using Wiktionary in isolation. We discuss this in more detail in section 8.3.

**Extensions and outlook**. In a joint work with Gurevych et al. (2012a), we have generalized the method described in section 6.3 as a flexible, modular alignment framework, which we used to align *WordNet* (2006) and *OmegaWiki* (2006 f.) and, later on, with Matuschek et al. (2013) to align Wiktionary and *OmegaWiki* (2006 f.). This framework extends the method we described by a machine learning module, which alleviates the issue of choosing fixed thresholds for the similarity methods and simplifies the integration of additional features and methods.

Subsequent to our work, multiple other researchers proposed word sense alignments involving Wiktionary: Henrich et al. (2011) align *GermaNet 6.0* (2011) with the German Wiktionary edition using a method based on word overlap. Hartmann and Gurevych (2013) align Wiktionary and *FrameNet* (2010) using manually defined filters and the generalized version of our method. Recently, Matuschek and Gurevych (2013) proposed a graph-based alignment

algorithm with which they achieved an improved performance of $F_1 = .69$ on our evaluation dataset for aligning Wiktionary and *WordNet* (2006).

Eom et al. (2012) create an annotated dataset for aligning WordNet and *Collins COBUILD Advanced Learner's English Dictionary* (2006). Based on our finding that human raters tend to disagree on subtle differences of meanings, they use a graded scale for judging the word sense alignments and carry out multiple experiments revealing that alignment annotations cannot be done by non-experts.

## 6.7 Chapter Summary

In this chapter, we have described and evaluated an automatic word sense alignment method for combining two computational dictionaries at the level of word senses. We exemplified this method by aligning the English Wiktionary with *WordNet 3.0* (2006), for which our method based on cosine similarity and the Personalized PageRank outperformed the baseline systems by a large margin. In an extensive annotation study, we found that aligning two computational dictionaries is a more complex task than aligning a dictionary with an encyclopedia, and we gave recommendations for future annotation studies based on the analysis of the disagreements among the human raters. For the resulting aligned resource, we concluded that natural language processing applications can benefit from an increased coverage and an enriched representation of word senses. Both our aligned resource and the evaluation dataset are publicly available for future research.

Chapter 7

# Standardized Representation of Language Resources

This chapter aims at creating a standardized representation of Wiktionary using the Lexical Markup Framework (LMF). We first introduce the motivation for creating standardized resources in section 7.1, discuss related work in this area in section 7.2, and then describe the LMF standard in section 7.3. Our approach is based on the lexicon model UBY-LMF, which we present in section 7.4 along with our extensions to this model for representing Wiktionary and OntoWiktionary. Finally, we integrate our LMF compliant dictionaries into the large-scale lexical resource UBY (section 7.5) and conclude the natural language processing perspective of the thesis (section 7.6). In section 7.7, we summarize the chapter.

## 7.1 Motivation

We have discussed the benefits and synergies that arise when integrating a variety of computational dictionaries in the scope of the previous chapter. While we have already proposed a method for obtaining word sense alignments for a pair of dictionaries, a major challenge for their integration is the heterogeneity of their representation, which includes differences in:

- the *organization* and the *macrostructure*: the primary building blocks of Wiktionary are wiki pages describing a certain lexical item, which is different from, for example, the synset-based organization of WordNet (cf. section 2.3).
- the encoded *information types* and the *microstructure*: Wiktionary encodes pronunciations and etymologies, whereas WordNet encodes, for instance, troponymy and entailment relations and FrameNet encodes semantic arguments (cf. section 2.6).
- the *coverage* and *granularity* of the lexicographic descriptions: besides different languages and parts of speech being encoded, we found separate language varieties, which are predominantly covered by a certain dictionary (cf. chapter 4).

- the *terminology*: different terms are being used to refer to the same things. We introduced, for example, the notion of *sense definition*, which corresponds to the term *gloss* used for WordNet and the term *paraphrase* used for GermaNet.
- the *data format*: Wiktionary is released as XML data dumps in a MediaWiki format, while WordNet is shipped in a vendor-specific database, and OmegaWiki is available as an SQL database dump (cf. section 5.3).
- the *access paths* and *software tools*: Different web interfaces offer different search options for human users (cf. section 2.5) and multiple software tools are available for accessing a computational dictionary from a natural language processing system (cf. section 5.3).

This heterogeneity prevents us from employing a variety of computational dictionaries effectively, because each task and resource requires multiple changes to an application's source code and data model. Our goal is therefore to establish *interoperable* resources that are capable of exchanging their data across system boundaries. Ide and Pustejovsky (2010) distinguish between *syntactic interoperability* addressing data formats, terminology, and access paths and *semantic interoperability*, which ensures a common interpretation of the encoded data. In terms of computational dictionaries, this means that we need to define a shared *lexicon model*, which allows for switching between different computational dictionaries without the necessity of changing the source code of a natural language processing application and for combining the information found in different dictionaries by means of alignments between them.

To achieve this goal for Wiktionary and OntoWiktionary, we propose using the lexicon model UBY-LMF (Eckle-Kohler et al., 2012), which is based on the *Lexical Markup Framework* (LMF), an international standard described in ISO 24613 (2008). To this end, we make use of the shared terminology (e.g., `LexicalEntry`, `SenseExample`) and data formats defined by LMF yielding syntactic interoperability, and we select a set of *data categories* describing the individual information types and their values (e.g., `partOfSpeech`), which results in a semantically interoperable lexicon model. We then populate the lexicon model using data from Wiktionary and OntoWiktionary, and we integrate our standardized resource into *UBY* (2012), which allows us to interlink them with other computational dictionaries. The unified and standardized format of UBY as well as the available web interface and software libraries puts us in a position to quickly adapt a natural language processing application to make use of Wiktionary data easily. Our contributions described in this chapter can be summarized as:

**Contribution 6.1**: *We develop a standardized lexicon model for Wiktionary and OntoWiktionary as a part of UBY-LMF (section 7.4).*

**Contribution 6.2**: *We populate this lexicon model with the information extracted from Wiktionary and OntoWiktionary (section 7.5).*

**Contribution 6.3**: *We integrate the standardized representation of Wiktionary and OntoWiktionary into the large-scale lexical resource UBY (section 7.5).*

## 7.2 Related Work

**Resource standards.** In recent years, multiple standards have been proposed to represent different types of dictionaries and language resources. In addition to LMF, which we describe in detail in the next section, the most prominent ones are the *Text Encoding Initiative* (TEI), and the *Resource Description Framework* (RDF).

The TEI standard is being developed since 1987 as a set of formats and guidelines for exchanging texts. It is based on the XML format and is mainly used in the humanities for modeling common annotation schemes for a variety of texts including dictionaries. The current version of the standard is described in TEI-P5 (2013). The *Wörterbuchnetz* (2007 f.) published by the Trier Center for Digital Humanities currently contains over 20 TEI-compliant, digitized dictionaries (see Burch and Rapp, 2007). Although a large part of TEI specifically addresses the standardization of dictionaries, its recommendations are closely linked to running text as it is found in printed dictionary articles. This limits the modeling possibilities for computational dictionaries such as WordNet, which have a graph-based, non-linear structure.

The RDF (2004) standard is based on the notion of triples of the form (subject, predicate, object) which encode formal statements about a universe. Such statements are especially useful for modeling ontologies in the context of the Semantic Web and Linked Data (Berners-Lee et al., 2001; Bizer et al., 2009a). While most RDF resources encode world knowledge, the recently established *Open Linguistics Working Group*[69] aims at representing computational dictionaries and language resources in RDF to constitute a *Linguistic Linked Open Data cloud* (Chiarcos et al., 2012). A particular challenge is the definition of a common lexicon model that ensures the semantic interoperability of the resources, because RDF does not provide a fixed set of predicates or literals for representing the information types usually found in dictionaries. This has caused recommendations such as SKOS (2009) and the *LexInfo* model (Buitelaar et al., 2009), but none of them makes use of an extensible set of *data categories* to standardize their information types. In the context of the *lemon* project (McCrae et al., 2011), *LexInfo* has been extended by standardized data categories from the LMF data category registry *ISOcat*. This makes *LexInfo* and *lemon* highly similar to the UBY-LMF model we use, whereas the former explicitly separate between lexical and ontological information (cf. Buitelaar et al., 2009) and the latter pursues a representation of the entire dictionary in a single lexicon model.

**LMF representations.** So far, LMF representations primarily focus on individual dictionary projects or certain information types (e.g., Quochi et al., 2008; Hayashi, 2011). Soria et al. (2009) describe Wordnet-LMF, a lexicon model used in the context of the KYOTO project for standardizing wordnets, such as the LMF version of the *Italian WordNet* (Toral et al., 2010). During the conversion of *GermaNet* into this format, Henrich and Hinrichs (2010, p. 461) note, however, that "a number of modifications to Wordnet-LMF are needed if this conversion is to preserve

[69] `http://linguistics.okfn.org/` (24 April 2013)

all information present in the original resource". Because different lexicon models hamper the interoperability of resources, this motivates the development of a large-scale instantiation of the LMF standard, which is capable of preserving the information of the original resources but still provides a common representation for all of them. Another important aspect that has yet been largely ignored is the availability of documentation, tutorials, web interfaces, and application programming interfaces, which essentially determine if a standardized resource can be employed in a natural language processing system. The UBY-LMF lexicon model (Eckle-Kohler et al., 2012), to which we contribute our standardized representation of Wiktionary, fulfills these requirements, as it is instantiated in the large lexical resource *UBY* (2012). UBY provides ten LMF-compliant computational dictionaries and is accessible by means of a web interface and an application programming interface (Gurevych et al., 2012a, 2012b).

**Standardization of Wiktionary.** Each of the three standards RDF, TEI, and LMF has been previously used to represent Wiktionary. The earliest work has been carried out by de Melo and Weikum (2008), who integrate a set of information types extracted from Wiktionary in their RDF web service *Lexvo*. Because of the large differences between the different Wiktionary language editions, Lexvo is, however, limited to a small set of information types, mostly focusing on translations. This issue has been recently addressed by Hellmann et al. (2013), who are working on a more comprehensive RDF representation of Wiktionary. As an open research question, they note "how lexical resources with differing schemata can be linked" (ibid., p. 205). This question arises because they make use of a newly defined set of RDF predicates and labels. As opposed to their approach, we rely on standardized data categories from UBY-LMF which facilitates representing a large number of lexical resources using the same shared lexicon model.

Declerck et al. (2012) represent Wiktionary based on the TEI standard. Apart from their main goal of assisting the translation of taxonomic labels in canonical folk literature catalogs, they note that their standardized representation of Wiktionary can be beneficial for other natural language processing applications as well – in particular in the area of Digital Humanities. Our work differs in using LMF to represent the Wiktionary data, which allows us to explicitly model semantic relations and to incorporate word sense alignments easily.

Sérasset (2012) proposes a standardized version of Wiktionary based on LMF. He addresses a number of challenges in the extraction and standardization of the Wiktionary data, such as the distinction of homonymy and polysemy (which we also discuss below), differences in the encoding of the article constituents and ill-formed markup. An important difference to our LMF model is that Sérasset does not represent semantic relations at the level of word senses, but at the level of lexical entries, which yields close relationships between actually unrelated words. We propose a solution for that and represent the semantic relations between a word sense and a word form (Wiktionary) and between two word senses (OntoWiktionary). McCrae et al. (2012) describe a representation of Wiktionary within the *lemon* model, which they

use to bridge the gap between the large amount of data available as RDF and the fine-grained modeling possibilities of LMF. Similar to the integrated resource UBY, which we use for our work, the *lemon* model is able to represent multiple computational dictionaries using the same lexicon model. The *lemon* representation differs from UBY by focusing on the lexical information and linking to an external RDF-based ontology representing the semantic information of Wiktionary. Instead, our goal is modeling the entire dictionary using a single model.

## 7.3 Lexical Markup Framework

The *Lexical Markup Framework* (LMF) defines an abstract model for lexical resources. It has emerged as a result of multiple projects in the area of language resources, such as ACQUILEX, EAGLES/ISLE, MILE, and PAROLE (cf. Calzolari et al., 2013). The technical committee "terminology and other language and content resources" (ISO/TC 37) decided in 2004 to initiate the standardization of lexical resources, which yielded the publication of LMF as the international standard ISO 24613 (2008). We give a brief introduction to LMF below. A more detailed description has recently been presented by Francopoulo and George (2013).

**Lexicon model**. LMF considers itself an "abstract metamodel". In order to compile a standardized dictionary, we need to instantiate this metamodel in what we call a *lexicon model*. The descriptions of the LMF standard are based on the *Unified Modeling Language* (UML) described in ISO 19501 (2005). Using the UML notion of *packages*, the standard distinguishes between the *core package*, which is a mandatory part of every lexicon model, and multiple *extensions* that may be optionally selected, including

- the *Morphology extension*,
- the *Machine readable dictionary extension*,
- the *NLP syntax extension*,
- the *NLP semantics extension*,
- the *NLP multilingual notations extension*,
- the *NLP morphological patterns extension*,
- the *NLP multiword expression patterns extension*,
- the *Constraint expression extension*.

Each package provides a number of predefined *classes* (e.g., `LexiconEntry`, `SenseExample`) and *UML relations* (aggregation, association, generalization) between them. As part of the instantiation in a lexicon model, each LMF class is equipped with one or more *data categories* modeling the information types described by this class. A data category consists of a name, a unique persistent identifier, a specification of its data type and value domain, as well as a natural language description of its meaning. Among others, we can distinguish between *closed data*

*categories* for which all possible values can be enumerated and *open* or *constrained data categories*, which can take arbitrary values (possibly with certain constraints). The `writtenForm` of a `FormRepresentation` instance is an example for an open data category (allowing an arbitrary Unicode string), and the `partOfSpeech` of a `LexicalEntry` is a closed data category, which can only take a value of a predefined list (such as `noun`). To ensure the semantic interoperability of a lexicon model, all data categories are to be registered in a *data category registry*, which is specified by ISO 12620 (2009). An implementation of this standard is *ISOcat*[70] providing a web interface for browsing and looking up the defined data categories by their unique identifiers as well as for defining new data categories in a collaborative way. The `writtenForm` data category is, for example, registered in ISOcat with ID 1836, and the `partOfSpeech` data category with ID 396.[71]

**LMF process.** According to the LMF standard, the first step towards compiling a standardized dictionary is defining the lexicon model. That is, selecting the extension packages, which are used together with the mandatory core package. Based on the chosen extensions, the LMF classes of the lexicon model can then be chosen. Finally, we define a data category selection including all data categories from a data category registry, which are required to model the information types of the dictionary. The final lexicon model can be documented in the form of a UML class diagram, a Document Type Definition (DTD), or a similar form of schema description. We discuss our lexicon model for Wiktionary in section 7.4.

The second step of the LMF process addresses the population of the lexicon model. That is, the transformation of the original resource into the classes and data categories of the defined model. The populated, standardized resource can be made available, for example, as a database dump, in a triplestore as part of the *Linked Data cloud* (thus using the RDF standard to define the data format), or in a generic XML format. We discuss the population of our lexicon model in section 7.5.

## 7.4 Modeling Wiktionary in LMF

In a joint work with Eckle-Kohler et al. (2012), we developed UBY-LMF, which is a lexicon model for representing a variety of computational dictionaries in an LMF-compliant format. More specifically, UBY-LMF is the lexicon model of the large-scale unified lexical resource *UBY* (2012) introduced by Gurevych et al. (2012a). UBY currently contains an LMF-compliant representation of ten computational dictionaries in two languages: the English *WordNet* (2006), *Wikipedia* (2001 f.), *Wiktionary* (2002 f.), *FrameNet* (2010), *VerbNet* (2009), the German *GermaNet* (2011), *Wikipedia* (2001 f.), *Wiktionary* (2002 f.), *IMSLex-Subcat* (1999), and the multilingual *OmegaWiki* (2006 f.). A particular aspect of UBY-LMF is that it includes both

[70] `http://www.isocat.org` (23 April 2013)

[71] `http://www.isocat.org/datcat/DC-1836` and `http://www.isocat.org/datcat/DC-396` (26 April 2013)

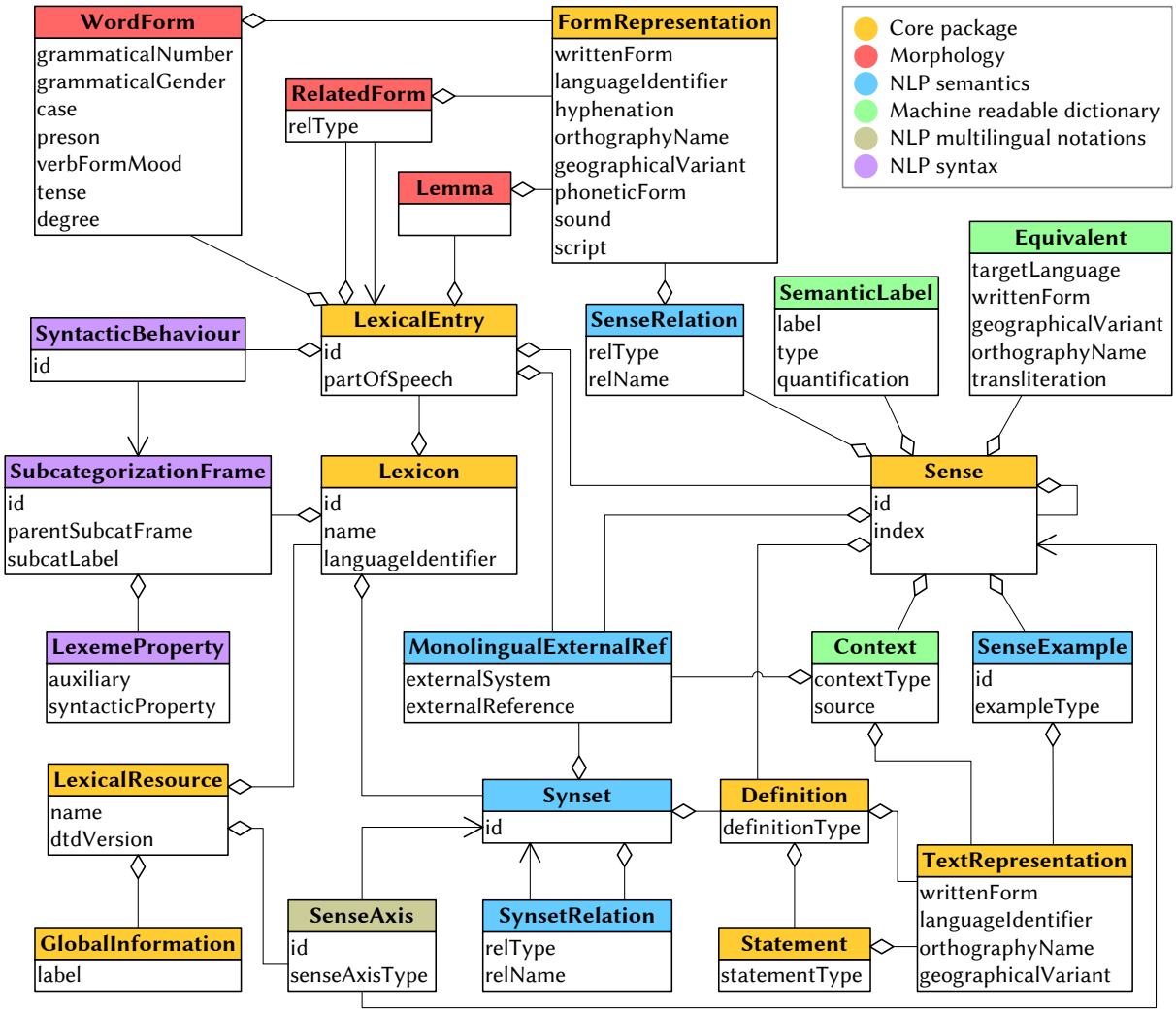



Figure 7.1: Overview of classes and data categories in our derived lexicon model

editorial and collaborative dictionaries. In this section, we describe the LMF packages, classes, and data categories we use for representing Wiktionary and OntoWiktionary in UBY-LMF. We particularly focus on the challenges we face when modeling collaborative dictionaries and discuss our solutions for them. Figure 7.1 shows the excerpt of UBY-LMF required for modeling Wiktionary and OntoWiktionary.

**Core package.** The centerpiece of the UBY-LMF architecture is defined by the LMF classes of the mandatory core package. A unique instance of the `LexicalResource` class represents the entire resource that is being standardized (e.g., UBY). It is accompanied by a descriptive `GlobalInformation` instance containing administrative information (such as the character encoding), and it consists of a number of separate `Lexicon` instances. Each `Lexicon` represents a specific Wiktionary language edition, which is identified by its ISO 639-3 language code. The `Lexicon`

essentially consists of a set of lexical entries, which are represented by the `LexicalEntry` class. A `LexicalEntry` may be connected to multiple instances of the `Sense` class modeling a certain meaning of the lexical entry. The lexical entries and word senses we extracted from Wiktionary can directly be used to populate the respective instances of `LexicalEntry` and `Sense`. Attached to the `Sense` instances, we store the sense definition using the `Definition` class and the etymology using the `Statement` class. Both of them aggregate a `TextRepresentation` instance holding the text of the sense definition and etymology.

Wiktionary distinguishes between homonymy and polysemy. It is important to note this distinction here, since neither expert-built wordnets nor any of the computational dictionaries modeled in UBY separate between these two concepts. Homonymy denotes a relation between words that share the same form, but originate from different etymologies. It can be represented in our LMF model by creating separate instances of `LexicalEntry` for the homonymous entries. As opposed to that, polysemy denotes a relation between different word meanings that share the same etymology. In this case, only one `LexicalEntry` is required, which distinguishes multiple `Sense` instances. Consider the English noun *post* as an example: There are separate lexical entries in Wiktionary to describe the homonyms originating from the Latin *postis* (i.e., the meaning of a *doorpost*, *pillar*) and from the Middle French *poste* (i.e., the meaning of a mail delivery system). Hence, there are two instances of `LexicalEntry` representing the two different etymologies. Each of the lexical entries has multiple word senses to model polysemy, such as the distinction between a regular mail system (e.g., "a letter sent by post") and a message in an electronic forum (e.g., "she read his post in the internet forum of the university").

Another issue is the representation of a lexical entry's part of speech. We have noted in section 4.4 that Wiktionary uses a large number of different part of speech labels of which some are more fine-grained than others. We find, for example, lexical entries marked as "family name", "proper noun", or simply "noun". Since fine-grained and coarse-grained labels are mixed, one option for standardizing the parts of speech would be to choose the most general level (i.e., "noun"). However, this would imply a loss of specificity with regard to the original resource. We therefore propose using a prefix notation for representing the parts of speech. That is to say, we introduce the data category values `noun`, `nounProper`, `nounProperFamilyName`, which allow for querying the encoded information at different levels of granularity: Searching for all lexical entries whose part of speech labels starts with "noun" also returns lexical entries tagged as family names, while it remains possible to restrict the search to lexical entries specifically labeled with a certain part of speech. Table 7.1 shows the part of speech tags of UBY-LMF in prefix notation.

**Morphology extension.** From the LMF morphology extension, we select the `Lemma`, `RelatedForm`, and `WordForm` classes. Using the latter, we represent the inflected word forms encoded in Wiktionary's inflection tables by creating an instance of this class for each word form and defining its grammatical number, gender, case, person, tense, etc. The citation form is specially

| Part of speech | ID | Part of speech | ID | Part of speech | ID |
|---|---|---|---|---|---|
| abbreviation | 329 | determinerDefinite | 1430 | interjection | 1318 |
| abbreviationAcronym | 334 | determinerDemonstrative | 1269 | pronoun | 1370 |
| abbreviationInitialism | 333 | determinerIndefinite | 1307 | pronounDemonstrative | 1270 |
| adjective | 1230 | determinerInterrogative | 1320 | pronounIndefinite | 1309 |
| adpositionCircumposition | 1906 | determinerPossessive | 1357 | pronounInterrogative | 1321 |
| adpositionPostposition | 1360 | noun | 3347 | pronounPersonal | 3013 |
| adpositionPreposition | 1366 | nounCommon | 385 | pronounPersonalIrreflexive | 3013 |
| adverb | 1232 | nounProper | 1371 | pronounPersonalReflexive | 3014 |
| adverbPronominal | 2998 | nounProperGivenName | 4194 | pronounPossessive | 1359 |
| affix | 1234 | nounProperFamilyName | 4195 | pronounRelative | 1380 |
| affixPrefix | 1365 | numeral | 1334 | symbol | 1398 |
| affixSuffix | 1395 | particle | 3372 | verb | 1424 |
| conjunction | 3132 | particleAnswer | 2891 | verbAuxiliary | 1244 |
| conjunctionCoordinating | 1262 | particleComparative | 1922 | verbMain | 1400 |
| conjunctionSubordinating | 1393 | particleInfinitive | 1896 | verbModal | 1329 |
| contraction | 354 | particleNegative | 1894 | | |
| determiner | 1272 | phraseme | 339 | | |

Table 7.1: Label and ISOcat data category identifier of the parts of speech used in UBY-LMF

marked using the Lemma class. Wiktionary contains a large number of phonetic representations explaining how a word is pronounced (see section 2.6). We use the phoneticForm data category of the FormRepresentation class to represent the pronunciation information.

For representing the form-based relations introduced in section 4.7, we use instances of the RelatedForm class. From a modeling perspective, it would make sense to define the RelatedForm class as an association between two lexical entries (e.g., defining a derivation relation between the verb *(to) drive* and the noun *driver*). Form-based relations are, however, encoded as lemma-oriented rather than entry-oriented links in Wiktionary (see section 2.4). An additional problem is that the target article of the relation (and thus the target lexical entry) might not be encoded in Wiktionary yet. We therefore model the form-based relations between an instance of LexicalEntry and a FormRepresentation, which allows us to represent the target of a form-based relation as it is specified by the Wiktionary community.

**NLP semantics extension.** From the NLP semantics package, we use the SenseExample class to model example sentences provided by the Wiktionary community. For modeling semantic relations, such as synonyms, hyponyms, antonyms, etc., we face a similar issue as for the form-based relations discussed above: They are encoded as lemma-oriented rather than sense-oriented links. In the original LMF standard, the target of a SenseRelation is an instance of the Sense class. Since we have disambiguated the semantic relations for OntoWiktionary (see section 5.4), we can generally provide the target in form of a Sense instance. However, there are

cases in which the target word sense has not been described by the community or in which the entire article is still missing from the dictionary. This is why we want to preserve the original information encoded in Wiktionary and hence add a new aggregation relationship between the `SenseRelation` class and the `FormRepresentation` class. Although it would be possible to store the target word form directly as a data category within the `SenseRelation`, we have chosen the `FormRepresentation` class to allow for storing additional information given for a relation target, such as the geographical variant. Consider, for instance, the Wiktionary article on *dual carriageway*, which encodes a synonymy relation to the article *divided highway*. This synonymy link is marked with "(US)" to indicate that the word is usually used in American English. Using our new aggregation with the `FormRepresentation` class, we get in a position to represent this kind of information in our standardized model.

For OntoWiktionary, we additionally employ the `Synset` class to represent the concepts, and the `SynsetRelation` class to represent the conceptual relations of our ontology. To ensure the compatibility of our standardized representation and the original Wiktionary data, we use the `MonolingualExternalRef` class to store the unique identifiers for word senses and synsets from the original resources, as it is recommended by the UBY-LMF model. This allows us to adapt to future changes in the resources.

**Machine-readable dictionary extension.** In section 4.6, we found that Wiktionary provides a broad range of pragmatic labels, which describe different varieties of language. Unfortunately, there is no direct correspondence for modeling such labels according to the original LMF standard. Although the `SubjectField` class of the machine-readable dictionary extension is highly similar, it is designed for representing domain or status information, but does not cover register, style, or temporal labels (cf. ISO 24613, 2008, § C.2.6). For UBY-LMF, we therefore decide to replace the `SubjectField` class by a `SemanticLabel` class. This new class allows us to standardize the different kinds of pragmatic labels found in Wiktionary.[72]

In addition to that, we utilize the `Context` class for modeling quotations and the `Equivalent` class for translations. Unlike previous suggestions of using the `Equivalent` class for Wiktionary translations by Sérasset (2012), we represent the instances of this class as an aggregation of the `Sense` class rather than of `LexicalEntry`, in order to facilitate sense-disambiguated translations. We also use the data categories `transliteration` to encode different scripts (e.g., Cyrillic), `geographicalVariant` to represent a certain region in which the translated word is predominantly used (e.g., Moscow), and `orthographyName` to store a certain orthographic variant, such as the German orthography reform of 1996.

It is interesting to note that the `Equivalent` class models the target of a translation as a word form (e.g., using the `writtenForm` data category with the ISOcat ID 1836) rather than a

[72]In addition to the pragmatic labels encoded in Wiktionary, the new class was also motivated by selectional restrictions and sentiment information from *VerbNet* (2009) and *FrameNet* (2010), as well as the necessity of associating instances of `SemanticPredicate` and `SemanticArgument` with semantic labels (see Gurevych et al., 2012a).

word sense, as it is the case for SenseRelation instances. Choosing a Sense instance would be reasonable, since we have seen that translations hold between word senses rather than word forms (recall the example between *(to) hang* and *ausstellen* discussed in section 5.4). From the modeling perspective, this has three reasons: First, each Lexicon instance is designed to represent only lexical entries and word senses of one language. The rationale behind this is dealing with language-specific differences, in particular syntactic information (cf. Eckle-Kohler et al., 2013). Second, machine-readable dictionaries usually omit the target word sense of a relation, since human users can easily disambiguate them based on the context (see section 5.4). Third, using a word form allows for representing translations to many languages without modeling detailed linguistic knowledge about them. We can, for instance, represent the Greenlandic translation *paarlaaffik* for the English *ski lift* without defining a separate Lexicon instance for Greenlandic data.

**NLP multilingual notations extension.** Nevertheless, we can additionally represent the target word sense for those translations that are encoded in another Lexicon. This is achieved by the SenseAxis class, which is designed to relate two closely related word senses from different languages. To this end, the SenseAxis class is associated with two instances of the Sense class that belong to different Lexicon instances. We extend the original notion of the SenseAxis class to relate instances from two different lexicons (regardless of their language) in order to store the word sense alignments between two computational dictionaries in our standardized model. The alignment can be defined either between two word senses (i.e., Sense instances) or synsets (i.e., Synset instances).

**NLP syntax extension.** Although there is no detailed lexico-syntactic knowledge encoded in Wiktionary, such as subcategorization frames, we have seen in section 4.6 that some of the pragmatic labels do not describe pragmatic language varieties, but grammatical properties of a lexical entry (e.g., the transitivity of a verb). These labels should not be represented as instances of the SemanticLabel class, but rather be modeled according to the recommendations of the NLP syntax extension. For each of these syntactic labels, we create an instance of the SyntacticBehaviour class, which is attached to the LexicalEntry instance. The SyntacticBehaviour class acts as a mediator between the lexical entry, an optional reference to a word sense, and an instance of SubcategorizationFrame. The actual grammatical information is stored within the SubcategorizationFrame instance (if addressing the composition of sentences) or its adjunct LexemeProperty instance (if limited to the lexeme itself).

## 7.5 Integrating Wiktionary into UBY

**Populating the model.** Having defined our lexicon model, we now populate the model with data from Wiktionary and OntoWiktionary. Therefore, we create a software tool that trans-

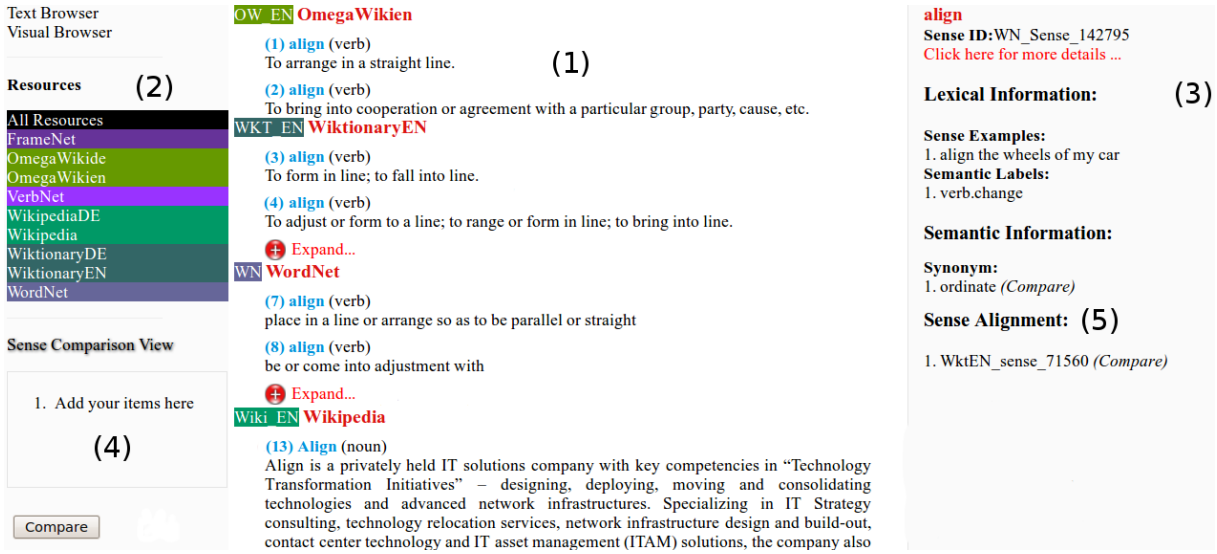



Figure 7.2: Textual browser of the UBY web interface: (1) list of word senses for the lemma *align*, grouped by resource, (2) selection of resources, (3) detail pane for the currently selected word sense, (4) drag&drop region for selecting two word senses for direct comparison, (5) links to related and aligned word senses (the figure has been taken from Gurevych et al., 2012b, p. 196).

forms the information types into their respective LMF classes and data categories. The Java source code is based on the Hibernate framework[73] and is also the basis for the web interface and the application programming interface (API) described below.

We integrate our LMF-compliant versions of Wiktionary and OntoWiktionary into UBY by modeling each Wiktionary language edition as a separate `Lexicon` instance. These `Lexicon` instances are interlinked with *WordNet* (2006) based on our work described in chapter 6 and with *OmegaWiki* (2006 f.) and *FrameNet* (2010) using the alignments by Matuschek et al. (2013) and Hartmann and Gurevych (2013). The UBY-LMF model, the *UBY* (2012) resource, and the accompanying software are freely available from the UBY homepage.[74]

**Web interface.** To facilitate the use of UBY by human users, we contributed to a web interface allowing them to navigate the encoded information across resource boundaries (Gurevych et al., 2012b). The web interface consists of a textual browser presenting the encoded lexical entries and word senses in a dictionary-like format as well as a visual browser, which enables exploring the word sense alignment between the different language resources. Figure 7.2 shows an excerpt of the textual browser as of 2012. The web interface is publicly accessible.[75]

[73] `http://www.hibernate.org` (25 April 2013)
[74] `http://code.google.com/p/uby` (25 April 2013)
[75] `https://uby.ukp.informatik.tu-darmstadt.de` (25 April 2013)

```
1  Uby uby = new Uby(dbConfig);
2  Lexicon lexicon = uby.getLexiconByName("WiktionaryEN");
3  String lemma = "boat";
4  EPartOfSpeech partOfSpeech = EPartOfSpeech.noun;
5  List<LexicalEntry> lexEntries = uby.getLexicalEntries(
6          lemma, partOfSpeech, lexicon);
7  for (LexicalEntry lexEntry : lexEntries) {
8      System.out.println(lexEntry.getLemmaForm());
9      for (Sense sense : lexEntry.getSenses())
10         System.out.println(sense.getDefinitionText());
11 }
```

Figure 7.3: Java code example for using the UBY-API

**Application programming interface**. The Hibernate-based transformation software also provides the means to access each LMF class and its data categories. We extend these access functions by a number of search possibilities yielding a Java-based API for accessing the LMF-compliant versions of Wiktionary and OntoWiktionary. The key design paradigm of the UBY-LMF lexicon model and the API is providing access to the information from multiple computational dictionaries using the same source code, because this solves the two issues we introduced in the beginning of this chapter: switching between different resources without changing the source code and combining the information from different resources using the word sense alignments between them. Figure 7.3 shows a simple code example for printing the lemma form and the sense definitions for the noun *boat* encoded in the English Wiktionary. By using "WordNet" as the lexicon name in line 2, the code can be easily adapted to printing the descriptions from *WordNet* (2006).

## 7.6 Discussion and Further Perspectives

The second research question we formulated for this thesis is targeted towards harvesting linguistic knowledge from Wiktionary (see section 1.3). Previous works in this direction mainly focused on extracting the encoded information items by means of text mining methods. Although being a fundamental step, we argue that additional methods are required in order to effectively make use of Wiktionary data. This is why we propose:

(1) disambiguating and ontologizing the extracted knowledge,
(2) aligning Wiktionary with other dictionaries at the level of word senses, and
(3) using standards for representing language resources in a unified format.

The discussed contributions towards this goal lead to the compilation of OntoWiktionary and the integration of Wiktionary data into the large-scale resource project UBY. As a result of our

metalexicographic study, we already found that Wiktionary is continually growing and that it covers a large number of languages and especially descriptions on non-standard language varieties. The integration of Wiktionary into UBY facilitates using this kind of information. In addition to that, we get in a position to easily combine the knowledge mined from Wiktionary with descriptions found in other dictionaries, which are aligned at the level of word senses. We identified a fine-grained lexicon model to be essential for modeling the Wiktionary data as close as possible to its original representation. This applies in particular to semantic information types, such as semantic relations, which have been modeled at the lexical level in previous works.

In summary, we see most potential in using the harvested Wiktionary knowledge for natural language processing applications that make use of the sense-disambiguated information provided by OntoWiktionary, rely on domain-specific or variety-specific information, and combine information from multiple language resources. The following chapter is targeted at discussing actual natural language processing applications that utilized Wiktionary previously, and we point out how they can benefit from our analysis and resources, which we exemplify in multiple case studies.

## 7.7 Chapter Summary

In this chapter, we described the standardization of Wiktionary and OntoWiktionary based on the Lexical Markup Framework (ISO 24613, 2008). To this end, we modeled the extracted information types according to the comprehensive lexicon model UBY-LMF by selecting the extensions, classes and data categories from LMF and the data category registry ISOcat. In a second step, we populated this model and integrated our standardized representation with *UBY* (2012). The unified LMF model allows us to access a broad range of different computational dictionaries using a common web interface and API, which spares the effort of adapting a natural language processing system to a specific language resource. In addition to that, we get in a position to represent the word sense alignments introduced in chapter 6 and hence to combine the information items found in different resources. In the next chapter, we describe actual use cases for the harvested Wiktionary knowledge.

Chapter 8

# Natural Language Processing Applications

This chapter aims at exploring how Wiktionary can be employed in a broad range of natural language processing tasks. After giving a brief overview of the chapter in section 8.1, we present a survey of applications that make use of Wiktionary data (section 8.2). Then, we discuss our two case studies in the context of measuring verb similarity (section 8.3) and detecting marketing blunders (section 8.4), which make use of the Wiktionary-related resources and the findings described in this thesis. We summarize the chapter in section 8.5.

## 8.1 Overview

In the course of the previous chapters, we have introduced Wiktionary as a valuable resource for computational linguistics. Now, we employ Wiktionary in actual natural language processing applications. Due to the large variety of information types and possible applications, we first provide a survey of previous works that successfully exploit the data encoded in Wiktionary. The survey considers all major information types and comments on how Wiktionary can effectively be used.

In the subsequent part of the chapter, we describe our own work towards using Wiktionary and OntoWiktionary to exemplify the viability and usefulness of our resources based on the findings described in the thesis. Our first experiment employs OntoWiktionary for computing monolingual and cross-lingual verb similarity. The former has been previously discussed by Zesch et al. (2008b), who did not disambiguate the knowledge extracted from Wiktionary. We argue that this is crucial for dealing with verbs, and we find that our work significantly outperforms previous work relying on *Wikipedia* (2001 f.) or undisambiguated information from Wiktionary. In comparison to expert-built wordnets, we find OntoWiktionary to be

competitive to them when computing monolingual verb similarity and to outperform them when computing cross-lingual verb similarity.

The second experiment is targeted at detecting cross-lingual marketing blunders. Such blunders are the result of an inappropriate translation of product or company names causing offense or embarrassment to potential customers while opening up a new market. To date, there are no tools for assisting the detection of marketing blunders that go beyond manual research using dictionaries, interviews, or similar resources. Wiktionary is continually updated, and we find a broad diversity of non-standard language varieties across many languages, including slang and offensive expressions. This makes Wiktionary an ideal resource for detecting cross-lingual marketing blunders. We propose a system based on the multilingual lexicalizations and pragmatic labels of OntoWiktionary. Our evaluation shows that our system is able to detect 71 % of the cross-lingual marketing blunders without suffering from too many irrelevant clues. The contributions described in this chapter can be summarized as:

**Contribution 7.1:** *We survey previous natural language processing applications that make use of Wiktionary data (section 8.2).*

**Contribution 7.2:** *We employ OntoWiktionary for computing monolingual and cross-lingual verb similarity (section 8.3).*

**Contribution 7.3:** *We employ OntoWiktionary for detecting cross-lingual marketing blunders (section 8.4).*

## 8.2 Survey of Wiktionary-based Applications

**Phonetic information**. Phonetic transcriptions in IPA or SAMPA notation can be used in speech recognition and speech synthesis research. He (2009) notes that Wiktionary is a good basis for the rapid creation of pronunciation dictionaries and exemplifies this for the English, French, German, Spanish, and Vietnamese languages. In a subsequent work, Schlippe et al. (2010) evaluate these pronunciation dictionaries in an automatic speech recognition setting and find that the French language edition has the best pronunciation coverage, while the Spanish Wiktionary yields the highest relative improvement over the baseline. The authors especially note the advantage of finding pronunciations for proper nouns within Wiktionary, as well as the inclusion of pronunciation variants. However, they also observe large differences between the quality and quantity of pronunciations from different Wiktionary language editions. Since their system yet only makes use of the first IPA notation of a page, this strand of research can benefit from improved text mining tools and a standardized representation of the harvested data. This holds particularly for the different geographical variants of a pronunciation, which is often described by supplementary Wiktionary labels.

Jouvet et al. (2011) likewise construct a pronunciation dictionary from Wiktionary. They propose a grapheme-to-phoneme conversion to also provide phonetic transcriptions for en-

tries lacking an explicitly encoded pronunciation constituent. They observe a reasonable, but slightly worse performance in comparison to a state-of-the-art system for speech recognition in French, but note that the Wiktionary data is freely available and that Wiktionary covers languages, for which no other pronunciation dictionaries exist.

The pronunciation-related audio files encoded in Wiktionary have, to our knowledge, not been used in our community so far.

**Morphological information**. Inflected word forms and word formation information are useful for a variety of preprocessing tasks, such as automatic stemming, lemmatization, and compound splitting. Perera and Witte (2005, p. 636) argue that "accurate lemmatization of German nouns mandates the use of a lexicon. Comprehensive lexicons, however, are expensive to build and maintain." They consider the use of Wiktionary, which would solve the problem of high construction and maintenance cost of expert-built lexicons. By that time (May 2005), Wiktionary had, however, less than 5,000 entries for the German language, which has caused them to propose a self-learning lemmatizer trained on German documents. Our analysis in chapter 4 shows that Wiktionary has largely grown since then, but no further attempts have been made in using Wiktionary for lemmatization or related tasks.

Chu et al. (2012) describe a rule-based method for mapping the graphemes of related logographic languages. They compare their work on mapping *Kanji* (Japanese) to *Hanzi* (Chinese) characters with mappings explicitly encoded in Wiktionary and find a higher coverage in three out of seven categories. This is why they propose using a combination of Wiktionary data and their method.

**Grammatical information**. Stochastic part of speech taggers usually require a large amount of training data for creating a statistical model. Walther et al. (2010) use the part of speech tags of the Kurmanji Wiktionary (i.e., the main dialect of the Kurdish language) for this task. They note that freely available resources like Wiktionary can be of great help for small, resource-poor languages, which we also observed in section 4.3. Nguyen and Ock (2012) employ part of speech tags from the English, Korean, and Vietnamese Wiktionary editions to train a classifier based on support vector machines, with which they obtain improved results over a baseline method for part of speech tagging. However, they do not use the actual information items encoded in the Wiktionary articles pages, but solely rely on the category system. Since we discussed the heterogeneity of Wiktionary's part of speech tags (see section 7.4), our proposed prefix notation can help building better models in the future, when taking the hierarchical nature of the part of speech tags into account.

In another strand of research, Richman and Schone (2008) propose using Wiktionary's part of speech tags for named entity recognition. They focus on freely available resources in several languages and find *Wikipedia* (2001 f.) and Wiktionary to be very valuable for this task. The

proposed system obtains an $F_1$ score of up to .84 for the languages French, Polish, Portuguese, Russian, Spanish, and Ukrainian.

**Semantic information.** Extensive research has been carried out on using Wiktionary for assessing vocabulary difficulty and readability. Hauff and Trieschnigg (2010) consider a word to be difficult if it is not part of Wiktionary's basic words category. Their goal is to assist the text reception of children. To this end, their proposed system adds hyperlinks to articles of the *Simple Wiktionary* for difficult words occurring in children's books of the Project Gutenberg. Eickhoff et al. (2010) pursue a similar goal by assessing suitability of online texts for children. They use the degree of polysemy and the length of the sense definition along with a language model of the *Simple Wiktionary* and a number of other features. Although they do not perform a separate evaluation for each feature, Wiktionary-based features are part of their best-performing feature set. Medero and Ostendorf (2009, 2011) propose similar features, but also take into account the number of lexical entries and translations of a lemma. They find Wiktionary to outperform standard measures based on word length and corpus frequency.

Chesley et al. (2006) employ Wiktionary to detect the polarity of adjectives in blog posts. Their polarity judgments are based on a manually compiled list of polarity cues occurring in the sense definitions of adjectives. They measure an accuracy of over 80 % and note that the continually growing Wiktionary has the potential to further improve the recall of their system.

Zesch et al. (2008b) evaluate different measures for calculating the semantic relatedness of words in English and German. They distinguish between measures based on the path length within a semantic network and using concept vectors of *WordNet* (2006), *Wikipedia* (2001 f.), and Wiktionary. In their evaluation, they observe the best results when using concept vectors from Wiktionary. We discuss this work in more detail in section 8.3. In subsequent work, Müller and Gurevych (2009) observe an improvement in the mean average precision of a domain-specific information retrieval task when using this semantic relatedness method.

Bernhard and Gurevych (2009) create a parallel corpus using pairs of sense definitions from *WordNet* (2006), *Wikipedia* (2001 f.), the *Simple Wikipedia* (2003 f.), and Wiktionary. Their idea is that the sense definitions from different computational dictionaries are near-paraphrases of each other, which they utilize to train a monolingual translation model. They evaluate this translation model in an information retrieval task for finding question–answer pairs that are related to a given query and measure better results using the lexical resources than using task-specific corpora of question–answer pairs. For creating their corpus, Bernhard and Gurevych consider all possible pairs of word senses whose descriptions have at least one word in common (omitting stop words). In chapter 6, we found that more elaborate measures (e.g., the Personalized PageRank algorithm) can yield better results than word-overlap-based methods for compiling a word sense alignment. In future research, these insights from word sense alignment research should be applied to the creation of parallel corpora and monolingual translations models to support information retrieval tasks.

The sense definitions of Wiktionary are also used by Henrich and Hinrichs (2012), who present a comparative evaluation of word sense disambiguation algorithms for the German language. In their best configuration, they combine sense definitions from Wiktionary with lexical fields from *GermaNet* (2011). Based on our finding of a good coverage of language varieties (section 4.6), we particularly expect improvements in domain-specific word sense disambiguation, which is still an important open research question (cf. Navigli, 2009).

Besides domain labels, the extensive use of register and style labels in Wiktionary provide promising research avenues. Burfoot and Baldwin (2009) propose a statistical model for determining whether a news article is "true" or satirical. As one feature of their system, they integrate the slang-related pragmatic labels from Wiktionary. They find that their model outperforms the majority class baseline and simple *n*-gram models.

**Cross-lingual information**. The vast number of languages and translations in Wiktionary has motivated multiple Wiktionary-based bi- and multilingual natural language processing applications. Otte and Tyers (2011) describe a rule-based machine translation system for Dutch and Afrikaans, which is based on freely available resources, including Wiktionary. Their system yields promising results for translating from Dutch to Afrikaans in comparison to the translation quality of a previously proposed system. However, they did not observe an improvement for the opposite direction from Afrikaans to Dutch. One reason for this are the large differences in the coverage of the language resources they use.

Lin and Krizhanovsky (2011) present a system for cross-lingual ontology matching, which makes use of translations extracted from the English and French Wiktionary editions. They obtain a high precision on the OAEI[76] benchmark dataset, but observe a lower recall than a statistical machine translation system. Similar to Otte and Tyers, they note that several Wiktionary editions should be combined to a single dictionary.

Etzioni et al. (2007) build a graph of over 2.3 million word translations from Wiktionary and other bilingual dictionaries. They employ this translation graph for cross-language image search by translating a user's query into multiple languages and obtaining the corresponding *Google Images*[77] search results, which are then combined by their *PanImages* system. They report an increase in the number of correct images on the first fifteen pages by 75 % and a 27 % increase in the overall precision compared to using the original query in isolation.

Another application in the context of information retrieval has been discussed by Müller and Gurevych (2009), who utilize semantic relatedness measures in a cross-lingual information retrieval setting. They obtain promising results using *Wikipedia* (2001 f.) as a source of lexical translations and note that Wiktionary contains similar translation links, but they do not report any experiments for them.

[76] `http://oaei.ontologymatching.org` (6 May 2013)
[77] `https://www.google.com/imghp` (6 May 2013)

From a different point of view, Bouchard-Côté et al. (2007) extract Italian, Latin, Portuguese, and Spanish cognates from Wiktionary with which they learn a probabilistic model for reconstructing ancient and modern word forms. Their goal is to describe phonological changes and to select between different phylogenies (i.e., tracing the evolution of a word over multiple languages). To achieve this goal, they propose an expectation-maximization algorithm based on Wiktionary translations.

## 8.3 Measuring Verb Similarity

Measuring *semantic similarity* (or *semantic relatedness*) is a foundational task of many natural language processing applications. Given two words $w_1$ and $w_2$, the task is to quantify their degree of similarity, such that highly similar words (e.g., synonyms, co-hyponyms, derived terms) receive a higher score than unrelated words. A *canoe* is, for instance, more similar to a *boat* than to a *stock option*.

A specific subproblem is measuring *verb similarity*. That is, the task of quantifying the similarity of a pair of verbs $(v_1, v_2)$. Verbs are known to be highly polysemous, which has caused rather mediocre results in previous approaches to this problem. Judging verb similarity is, however, of particular interest for applications such as (monolingual and cross-lingual) word sense disambiguation (Lefever and Hoste, 2010), lexical substitution (Mihalcea et al., 2010), or question answering (Magnini et al., 2005).

State of the art knowledge-based systems rely heavily on *Wikipedia* (2001 f.), which works well for measuring semantic similarity in general, but since it is focused on encyclopedic knowledge about nouns, it is not suitable for computing verb similarity. The large amount of linguistic knowledge in Wiktionary let us expect good results not only for nouns, but also for other parts of speech and verbs in particular.

Zesch et al. (2008b) observe better results when using Wiktionary for measuring semantic similarity than using Wikipedia and expert-built wordnets. For measuring verb similarity, however, they find the performance of Wiktionary to be way behind the wordnets. We argue that this is because they did not disambiguate the extracted Wiktionary knowledge. This is why we employ our sense-disambiguated ontology OntoWiktionary for this task using the same experimental setup as Zesch et al. (2008b). In addition to that, we discuss measuring cross-lingual verb similarity using OntoWiktionary, which becomes possible with our disambiguated translations.[78]

**Monolingual verb similarity**. Yang and Powers (2006) introduce an evaluation dataset for verb similarity that consists of 130 English verb pairs taken from TOEFL (Test of English as a Foreign Language) and ESL (English as a second language) questions. They have asked six

[78]The results reported in this section are based on Wiktionary data from April 2011.

human raters to annotate the similarity of each verb pair on a graded scale from 0 (not at all related) to 4 (inseparably related). Yang and Powers report a correlation of $r = .866$ between the raters and calculate the average of their scores as the gold standard annotation for each verb pair. An example from this dataset is the pair (*approve*, *support*) with a gold standard score of 3.

To the best of our knowledge, Zesch et al. (2008b) report the latest evaluation results on this dataset, which are depicted in table 8.1. They use *Explicit Semantic Analysis* (ESA), a method based on *concept vectors* (Gabrilovich and Markovitch, 2007) built from *WordNet* (2006), *Wikipedia* (2001 f.), and the undisambiguated Wiktionary. Each entry from these resources (synsets in WordNet and wiki pages in Wikipedia and Wiktionary) is regarded as one ESA concept. For a given word pair, they create two concept vectors that consist of the word's term frequency $*$ inverse document frequency (tf.idf) scores over the ESA concepts. The similarity for this word pair is then expressed by the cosine of the two concept vectors (cf. section 5.4).

We reproduce the results of Zesch et al. (2008b) using the same dictionaries and show them in the column (en:en) of table 8.1. Note that Zesch et al. compare their concept-vector-based method with a method based on path lengths. This is why they take only the 80 verb pairs (62 % of the dataset) into account for which a path length could be determined. Since a non-empty concept vector exists for each verb, we use all 130 verb pairs instead, which makes our scores slightly different. In addition to the three dictionaries used previously, we report the performance when using the sense-disambiguated OntoWiktionary. While each article page encoded in the undisambiguated Wiktionary represents one ESA concept (with all word senses flattened) for the approach by Zesch et al., we regard each word sense in OntoWiktionary as a separate ESA concept.

To measure performance, we calculate Spearman's rank correlation coefficient with Horn's correction for tied ranks (Horn, 1942). All methods reported hereafter and in table 8.1 correlate significantly with the human judgments as verified by a two-tailed, paired $t$-test ($p < .05$). From the table, we observe that using OntoWiktionary yields better results for the (en:en) dataset than using Wikipedia or the undisambiguated Wiktionary. The previously best resource WordNet is slightly outperformed by our resource. This difference is, however, not statistically significant. All four concept-vector-based methods cover 100 % of the (en:en) dataset and are thus directly comparable.

We also study German verbs and therefore manually translate the (en:en) dataset. The verb pair (*approve*, *support*) is, for instance, translated to (*annehmen*, *unterstützen*) keeping its similarity score of 3 (strongly related). Table 8.1 shows the results for this new (de:de) dataset. To create the concept vectors, we use *GermaNet* (2011) instead of WordNet as well as the German editions of Wikipedia and Wiktionary. We take only the 120 verb pairs (92 % of the dataset) into account that are covered by all four resources to allow for a fair comparison. OntoWiktionary is again able to outperform Wikipedia and the undisambiguated Wiktionary by a wide margin. It competes with the performance of the expert-built GermaNet, but yields

| Resource | Zesch et al. | en:en | de:de | en:de | de:en |
|---|---|---|---|---|---|
| WordNet/GermaNet | **.71** | .69 | **.57** | .31 | .23 |
| Wikipedia | .29 | .27 | .33 | .23 | .28 |
| Wiktionary | .65 | .63 | .36 | .34 | .37 |
| OntoWiktionary | — | **.73** | .52 | **.53** | **.51** |
| Coverage of verb pairs | 62 % | 100 % | 92 % | 95 % | 97 % |

Table 8.1: Evaluation results on the four verb similarity datasets when using different dictionaries

slightly lower results. This can be explained by the comparable coverage of GermaNet and the German Wiktionary, which we observed in section 4.4. As Wiktionary is continually growing, this lets us expect better results in the future. Given the large number of Wiktionary language editions (cf. section 4.3), we also note that Wiktionary can be a promising alternative for measuring verb similarity in languages with less developed expert-built resources.

**Cross-lingual verb similarity**. Based on the English and the German verb pairs, we create two cross-lingual verb similarity datasets that use the first English verb together with the second German verb from each corresponding verb pair (en:de) and, vice versa, the first German verb together with the second English verb (de:en). For the example introduced above, this yields the two verb pairs (*approve*, *unterstützen*) and (*annehmen*, *support*). Both translated verb pairs keep their original score of 3.

Table 8.1 shows the evaluation results for these two datasets. To create the cross-lingual concept vectors, we use the inter-lingual index between WordNet and GermaNet provided by *EuroWordNet* (1999), the interwiki links from Wikipedia, and the translation links from Wiktionary. Note that the translation links are regarded at the word level for the undisambiguated Wiktionary, but at the word sense level for OntoWiktionary.

We have already discussed in section 4.7 that the inter-lingual index of WordNet and GermaNet is very small. Consequently, we observe that the expert-built wordnets yield a substantially lower performance for (en:de) and (de:en) than in the monolingual setting. Wikipedia likewise yields low scores because of its lack of the knowledge about verbs. As opposed to that, OntoWiktionary significantly outperforms both the expert-built wordnets and Wikipedia ($p < .01$). The undisambiguated Wiktionary is up to the performance measured for the German dataset, which shows that our disambiguation of Wiktionary translations is useful for this task.

**Discussion and error analysis**. Wikipedia is not appropriate for computing verb similarity, as it focuses on encyclopedic knowledge about nouns. Although the encyclopedic descriptions also contain a vast number of verbs, they are highly scattered throughout this resource. The verb *(to) accentuate* is, for instance, frequently used within the articles *Self-categorization the-*

*ory*, *Accentuate the Positive (album)*, *Character displacement*, *Religion in Nigeria*, and *Cameo lighting*, which have little in common to describe the verb's meaning.

Expert-built wordnets work well for computing monolingual verb similarity, because they have a sufficient coverage and encode thoroughly elaborated linguistic knowledge. Our sense-disambiguated ontology OntoWiktionary competes with their quality. Since Wiktionary is available in over 170 languages, our approach is, however, also applicable to those languages lacking large expert-built resources. In a cross-lingual setting, we observe a different picture: Expert-built multilingual wordnets suffer from their small size, which yields a very low correlation with the human judges. Since the disambiguated translations in OntoWiktionary allow us to build cross-lingual concept vectors, they can be effectively utilized in this task and lead to significantly better results.

Our analysis of the results when using OntoWiktionary shows that the main source of error is due to low similarity scores for highly similar verb pairs. The similarity of the English–German verb pair (*concoct*, *ausarbeiten*) is, for instance, zero (i.e., the corresponding concept vectors are orthogonal), because Wiktionary is yet missing many translations between verbs related to this verb pair. This issue will supposedly be alleviated with increasing growth of Wiktionary, whereas future work should also consider combining information from multiple resources.

## 8.4 Detecting Cross-lingual Marketing Blunders

Large companies increasingly advertise and sell their products in international markets. Developing a marketing campaign for a new country requires tremendous translation efforts in order to bridge language-related and cultural boundaries. A particular problem often occurs if an established product or company name is used in a new foreign market without being adapted to local habits and language use. This regularly yields offensive, embarrassing, or (at best) funny results causing excessive remedial cost and maybe even the withdrawal of the product from the new market.

Such a *marketing blunder* can have different reasons. At the uppermost level, we can distinguish between *cross-cultural marketing blunders* and *cross-lingual marketing blunders*. The former addresses different customs, perception, and values of a society, which lead to the refusal of a certain product or company. A commercial for a men's fragrance showing a man with his dog, for instance, failed in Islamic countries where dogs are considered unclean. Instead, cross-lingual marketing blunders are a result of using inappropriate or negatively connotated expressions. A common example is the word *mist*, which is used to describe fabulous, enigmatic, lightweight, or mystic things in English. The word has, for instance, been used by a British car manufacturer to advertise their *Silver Mist* model. In German, the homograph *Mist* means, however, dung or manure, and it is a frequently used slang expression to describe a futile, cheap, or broken product, nonsense, or an annoying, tedious situation. This

pejorative meaning has caused the car manufacturer to rename its product. Hereafter, we will concentrate on such cross-lingual marketing blunders.

Spotting a marketing blunder can be very time-consuming and expensive for companies, especially if they do not operate local branches in all their target countries. They face two major problems:

(1) The absence of large-scale resources yielding *clues* for potential marketing blunders. Since many blunders are caused by false friends and predominantly by words used in colloquial speech, bilingual and multilingual dictionaries covering standard language are of limited help, and although there are specialized monolingual slang dictionaries such as the *McGraw-Hill's American Slang Dictionary* (2007) or Küpper's *Illustriertes Lexikon der deutschen Umgangssprache* (1982–1984), it is very challenging to keep these dictionaries up-to-date and provide them for a large number of languages.
(2) The absence of tools that assist the process of identifying the *relevant clues* from these resources. Obviously, not every word of a target language is problematic for marketing a product. The English word *fog* is, for instance, a false friend of the Hungarian *fog*, which translates to the English *tooth*. Neither meaning has a negative connotation that would impede the use of *fog* in a marketing campaign within those countries. A tool assisting the detection of marketing blunders thus needs to separate relevant clues from irrelevant ones in order to reduce the manual research effort. To the best of our knowledge, we are not aware of any natural language processing tool supporting this work.

In our metalexicographic analysis, we found the vast coverage of translations and the broad diversity of technical domains and non-standard language varieties to be particular strengths of Wiktionary, including slang and expressions with negative connotation. This is why we propose using Wiktionary to identify cross-lingual marketing blunders. In the course of this section, we describe and evaluate an automatic method for searching for relevant clues for potential blunders.

**Method**. As the example of the *Silver Mist* car suggests, a large share of cross-lingual marketing blunders is due to *false friends* (i.e., two words with the same pronunciation or written form, but a different meaning). Based on the multilingual lexicalizations of the English and German concepts of OntoWiktionary, we create a *homograph index* of words sharing the same form. For the word form *mist*, our index contains, for example, a lexicalization in Dutch, English, German, Old English, and Swedish. Querying this index can already reveal if an expression has a meaning in any language. Based on a given text $t$ representing the product or company name, its teaser or slogan, we define the LOOKUP method, which first normalizes each token of $t$ by converting it to lower case and removing special characters and then looks up each possible $n$-gram of tokens in the homograph index. We call each returned entry of the homograph index a *clue* for a potential marketing blunder.

```
1  function SOUNDEX(word)
2      result := upperCase(word_1);
3      for i ∈ {2, ... , length(word)} do
                          ⎧ 1   if word_i ∈ {b, f, p, v},
                          ⎪ 2   if word_i ∈ {c, g, j, k, q, s, x, z},
                          ⎪ 3   if word_i ∈ {d, t},
4          code :=        ⎨ 4   if word_i ∈ {l},
                          ⎪ 5   if word_i ∈ {m, n},
                          ⎪ 6   if word_i ∈ {r},
                          ⎩ ε   otherwise;
5          if result_{length(result)} ≠ code then
6              result := result ∘ code;
7      while length(result) < 3 do
8          result := result ∘ 0;
9      return result;
10 end.
```

Figure 8.1: Pseudo code of the Soundex algorithm ($\varepsilon$ represents an empty word)

The goal of our tool is separating *relevant* from *irrelevant clues*. The Dutch and Swedish forms of *mist* are, for instance, *cognates* of the English word form and therefore carry the same, unproblematic meaning as the English form. Likewise, false friends without any negative connotation, such as the English and Hungarian *fog* discussed above, yield irrelevant clues that should potentially be ignored by our tool. In a dictionary, the corresponding lexical entries for these words are usually *unmarked* – i.e., they are not associated with any pragmatic label but attributed to the standard variety of language. We extend our homograph index by the sociolectal, register, style, and evaluative labels (see section 4.6) from OntoWiktionary in order to mark non-standard usages that potentially contain slang, pejorative, rude, vulgar, or offensive expressions. The German *Mist* is, for instance, marked as "umgangssprachlich" (English: slang) and as "verärgerte Äußerung" (English: annoyed utterance), which are good indicators to avoid using *Mist* in a product or company name or at least initiate a careful study of the word's meaning in the target language.

Besides the actual word form that is being marked by such a pragmatic label, we also mark all translations of the marked word sense with the same pragmatic labels. The English translation *crap*, the Italian *scemenza*, and the Russian *дерьмо* are, for instance, marked as being slang expressions based on the labels encoded for the German *Mist*. Our final homograph index contains 2.4 million word forms from 1,867 languages, of which about 56,000 word forms are marked. Based on this index, we define a second method MARKED that looks up each normalized token of $t$ (equivalent to LOOKUP), but only returns those entries that are marked by one of the selected pragmatic labels.

In addition to that, alternative forms might yield a potential blunder. The word forms *misd*, *misth*, or *miest* would, for instance, cause similar reactions in German-speaking markets as their pronunciation is highly similar to *Mist*. This is why we propose a third method SOUNDEX, which queries the homograph index for forms starting with the same three letters and having the same *Soundex* score (Russell, 1918) as $t$, but not being identical to $t$. Soundex is an algorithm that returns a pseudo-phonetic representation of a given English word. The main idea of the algorithm is that two words with a similar pronunciation return similar Soundex scores. A Soundex score consists of the initial letter of the given word followed by at least three digits, which represent groups of similar consonants. Figure 8.1 shows this algorithm in pseudo code. The Soundex score of *boat* is B300, and the Soundex score of *lexicography* is L261. Although the Soundex algorithm is designed for the English language, we apply it to any token of $t$ and leave the development of a language-independent phonetic model to future work.

Finally, we combine the three methods by first querying the homograph index using MARKED. If this search does not yield relevant clues, we query the index using SOUNDEX and, analogously, we query the index using the LOOKUP method if there are still no relevant clues. The rationale behind this is to model an actual usage situation of our query tool: Since only a fraction of the entries are marked by pragmatic labels, we first present those to a user (MARKED). If she or he finds evidence for a potential marketing blunder, no further lookup is required. Otherwise, the user can check similar phonetic forms (SOUNDEX) and only turn towards reading all entries of the homograph index (LOOKUP) if there are still no relevant clues. We expect this procedure – which we call COMBINE – to enhance and speed up the lookup process, since less clues need to be examined.

**Evaluation**. To evaluate our approach, we create a dataset of previously occurred cross-lingual marketing blunders. To this end, we use the examples discussed by Ricks (2006, § 3) and provided on the homepage of the British translation agency *Kwintessential*.[79] We remove cross-cultural marketing blunders and examples for which the exact translation in the target language is not explicitly provided. Ricks notes, for instance, that *General Mills* wrongly translated the name of their mascot "Jolly Great Giant" into Arabic as "intimidating green ogre", but does not provide the exact Arabic form, which would be required to properly simulate the tool-assisted detection of this blunder.

For each marketing blunder in our dataset, we consider the problematic text $t$, a remark on the vendor or type of product, and a short textual explanation of the blunder. Since especially Ricks describes the blunders in prose, we manually extract the individual examples and bring them into a tabular form. In addition to that, we group the blunders according to the following types:

[79] `http://kwintessential.co.uk/cultural-services/articles/crosscultural-marketing.html`; `http://kwintessential.co.uk/cultural-services/articles/results-of-poor-cross-cultural-awareness.html`; `http://kwintessential.co.uk/cultural-services/articles/crosscultural-blunders.html` (30 May 2013)

- *vulgar* marketing blunders make use of an expression that is considered vulgar, rude, or offensive in the target language,
- *sexual* marketing blunders contain sexual innuendos in the target language,
- *negative* marketing blunders use expressions with a negative connotation or suggest negative properties, and
- marketing blunders of the *intent* group contain an expression that has a different, unrelated meaning in the target language, which causes astonishment and distraction among potential customers.

The final dataset consists of the 45 cross-lingual marketing blunders shown in table 8.2. Besides the information on the blunder itself, the table contains a separate column for each of the four methods we propose. We report the ratio of relevant clues to the total number of retrieved clues and whether the method could detect the corresponding marketing blunder (✓ if yes; ✘ if no), which is the case if there is at least one relevant clue in the method's search result. The notation "✓ 5 / 20" indicates, for instance, that the marketing blunder could successfully be detected and that there are five relevant clues among the 20 clues retrieved.

In order to judge between relevant and irrelevant clues, we asked two human raters to annotate the set of 1,494 total clues retrieved by any of the four methods for being relevant or irrelevant to detect the specified marketing blunder. The raters agreed on 95 % of the judgments, which yields an inter-rater agreement of $\kappa = .88$ (using Cohen's kappa). Based on this agreement, we consider the annotations reliable (cf. Artstein and Poesio, 2008). For obtaining a gold standard dataset, we asked an additional adjudicator to decide on the 76 ties.

At the bottom of table 8.2, we report the number of detected marketing blunders over the total number of blunders both for the whole dataset and separately for each blunder type. In addition to that, the table provides the total number of relevant clues and the total number of retrieved clues as well as the precision, recall, and $F_1$ scores. We define the precision as the ratio of relevant clues to the total number of retrieved clues (i.e., a method is more precise if it returns more relevant clues) and recall as the proportion of detected marketing blunders in the dataset (i.e., a method has a higher recall if it is able to detect more marketing blunders). The $F_1$ score follows the standard definition of being the harmonic mean between precision and recall.

Our primary objective is obtaining a high recall, since it is more important to retrieve potentially relevant clues for a marketing blunder than to minimize the number of irrelevant clues in the system result (which is a secondary objective). The basic LOOKUP method yields a recall of .64 indicating that Wiktionary is able to effectively detect cross-lingual marketing blunders. As the low precision indicates, there are, however, a large number of irrelevant clues that are retrieved by the simple index lookup.

As opposed to that, we find a high precision for the MARKED method, as the lexicalizations marked with pragmatic labels yield relevant clues in over 70 % of the cases. The MARKED method is, however, not applicable to detect marketing blunders of the type *intent* and the

| Marketing blunder | Type | MARKED | SOUNDEX | LOOKUP | COMBINE |
|---|---|---|---|---|---|
| Olympia Roto (copier)<br>*roto* means *broken* in Spanish | negative | ✓ 2 / 2 | ✘ 0 / 5 | ✓ 5 / 20 | ✓ 2 / 2 |
| Kinki Nippon Tourist<br>*kinki* sounds similar to the English *kinky* | sexual | ✘ 0 / 1 | ✓ 8 / 19 | ✘ 0 / 8 | ✓ 8 / 20 |
| Matador (US car)<br>*matador* means *killer* in Portuguese | negative | ✘ 0 / 0 | ✓ 3 / 10 | ✓ 9 / 10 | ✓ 3 / 10 |
| Toyota MR2<br>*MR2* sounds similar to the French *merde* (excrement) | vulgar | ✘ 0 / 0 | ✘ 0 / 1 | ✘ 0 / 11 | ✘ 0 / 12 |
| Studebaker Dictator<br>*dictatorship* has a negative connotation | negative | ✘ 0 / 0 | ✘ 0 / 1 | ✓ 10 / 12 | ✓ 10 / 13 |
| Buick LaCrosse (US car)<br>*la crosse* means *masturbation* in French | sexual | ✘ 0 / 0 | ✘ 0 / 3 | ✘ 0 / 12 | ✘ 0 / 15 |
| Fiera (US car)<br>*fiera* means *ugly* in Spanish | negative | ✘ 0 / 0 | ✓ 1 / 1 | ✘ 0 / 16 | ✓ 1 / 1 |
| Mercury Caliente (US car)<br>*caliente* means *sexually aroused* in Spanish | sexual | ✓ 1 / 2 | ✘ 0 / 5 | ✓ 22 / 40 | ✓ 1 / 2 |
| Pinto (US car)<br>*pinto* means *small appendage* in Brazilian Portuguese | sexual | ✓ 3 / 3 | ✘ 0 / 8 | ✓ 4 / 16 | ✓ 3 / 3 |
| Silver Mist (UK car)<br>*Mist* means *manure* and is used as a vulgarity in German | vulgar | ✓ 3 / 3 | ✘ 0 / 2 | ✓ 9 / 48 | ✓ 3 / 3 |
| Mist Stick (hair curling iron)<br>*Mist* means *manure* and is used as a vulgarity in German | vulgar | ✓ 6 / 11 | ✓ 1 / 14 | ✓ 28 / 100 | ✓ 6 / 11 |
| Bundh (UK sauce)<br>*bundh* means *arse* in Punjabi | vulgar | ✘ 0 / 0 | ✓ 7 / 11 | ✘ 0 / 0 | ✓ 7 / 11 |
| Pavian (fruit drink)<br>*Pavian* means *baboon* in German | intent | ✘ 0 / 0 | ✘ 0 / 0 | ✓ 6 / 6 | ✓ 6 / 6 |
| Grab Bucket (excavator part)<br>*Grab* means *grave*, and *bucket* sounds like *bouquet* in German | negative | ✓ 2 / 6 | ✘ 0 / 2 | ✓ 5 / 54 | ✓ 2 / 6 |
| Vicks (cough drops)<br>*vicks* sounds similar to the German vulgarity *fick* (sexual intercourse) | sexual | ✘ 0 / 0 | ✘ 0 / 2 | ✘ 0 / 0 | ✘ 0 / 2 |
| Puff tissues (facial tissues)<br>*Puff* means *whorehouse* in German and *homosexual* in the UK | sexual | ✓ 3 / 10 | ✓ 10 / 12 | ✓ 4 / 40 | ✓ 3 / 10 |
| Probe (US car)<br>*Probe* means *test* or *rehearsal* in German | negative | ✘ 0 / 0 | ✘ 0 / 5 | ✓ 12 / 39 | ✓ 12 / 44 |
| Bardak (machine)<br>*бардак* means *whorehouse* in Russian | sexual | ✘ 0 / 0 | ✘ 0 / 2 | ✘ 0 / 3 | ✘ 0 / 5 |
| Bran Buds (cereal)<br>*bran* means *burned* in Swedish | intent | ✘ 0 / 0 | ✘ 0 / 6 | ✓ 4 / 15 | ✓ 4 / 21 |
| Jotter (US pen)<br>*jotter* means *jockstrap* in some Latin America markets | sexual | ✘ 0 / 0 | ✘ 0 / 1 | ✘ 0 / 3 | ✘ 0 / 4 |
| Zit! (chocolate)<br>*zit* is a slang word for *pimple* in English | negative | ✘ 0 / 0 | ✓ 1 / 3 | ✘ 0 / 0 | ✓ 1 / 3 |
| Sic (French soft drink)<br>*sic* sounds similar to *sick* in English | negative | ✘ 0 / 0 | ✓ 5 / 14 | ✓ 2 / 19 | ✓ 5 / 14 |
| Super Piss (Finnish de-icer)<br>*piss* is a slang word for *urine* in English | vulgar | ✓ 14 / 33 | ✘ 0 / 4 | ✓ 15 / 75 | ✓ 14 / 33 |
| Bum (Spanish potato chips)<br>*bum* is a slang word for *buttocks* in English | vulgar | ✓ 16 / 17 | ✓ 1 / 2 | ✓ 23 / 33 | ✓ 16 / 17 |
| Polio (Czech detergent)<br>*Polio* is a short form of the disease *Poliomyelitis* | negative | ✘ 0 / 0 | ✓ 6 / 19 | ✓ 3 / 4 | ✓ 6 / 19 |
| Homo (Asian fish sausage)<br>*homo* is a short form of *homosexual* | sexual | ✓ 11 / 11 | ✘ 0 / 3 | ✓ 41 / 66 | ✓ 11 / 11 |

| Marketing blunder | Type | MARKED | SOUNDEX | LOOKUP | COMBINE |
|---|---|---|---|---|---|
| Swine (Chinese chocolate)<br>*swine* is used pejoratively in English and considered dirty | negative | ✓ 9 / 9 | ✓ 6 / 6 | ✓ 16 / 16 | ✓ 9 / 9 |
| Ass Glue (Chinese glue)<br>*ass* is a slang word for *buttocks* in English | vulgar | ✓ 13 / 14 | ✓ 11 / 11 | ✓ 15 / 71 | ✓ 13 / 14 |
| Last Climax (Japanese tissues)<br>*climax* is a slang word for *orgasm* in English | vulgar | ✓ 1 / 1 | ✘ 0 / 5 | ✓ 20 / 149 | ✓ 1 / 1 |
| Creap (Japanese coffee creamer)<br>*creap* sounds similar to *crap* and *creep* in English | negative | ✘ 0 / 0 | ✓ 5 / 10 | ✘ 0 / 0 | ✓ 5 / 10 |
| Maxipuke (Chinese playing cards)<br>*puke* is a slang word for *vomit* in English | vulgar | ✘ 0 / 0 | ✘ 0 / 0 | ✘ 0 / 0 | ✘ 0 / 0 |
| Pansy (Chinese underwear)<br>*pansy* is a slang word for *homosexual* in English | vulgar | ✓ 2 / 2 | ✓ 2 / 16 | ✓ 2 / 7 | ✓ 2 / 2 |
| Skintababe (Japanese soap)<br>sounds similar to *skin a babe* in English | negative | ✘ 0 / 0 | ✘ 0 / 0 | ✘ 0 / 0 | ✘ 0 / 0 |
| PET (US dairy products)<br>*pet* means *fart* in French | vulgar | ✓ 3 / 4 | ✓ 4 / 10 | ✓ 12 / 53 | ✓ 3 / 4 |
| Cue (US toothpaste)<br>*Cue* is an infamous pornographic magazine in France | sexual | ✘ 0 / 0 | ✘ 0 / 1 | ✘ 0 / 13 | ✘ 0 / 14 |
| Fesca (US soda pop)<br>*fresca* means *lesbian* in Mexican slang | sexual | ✘ 0 / 0 | ✘ 0 / 9 | ✘ 0 / 0 | ✘ 0 / 9 |
| Misair (Egyptian airline)<br>*misair* sounds similar to the French *misère* (misery) | negative | ✘ 0 / 0 | ✓ 5 / 10 | ✘ 0 / 0 | ✓ 5 / 10 |
| EMU Airways (Australian airline)<br>an *emu* is a bird that can't fly | negative | ✘ 0 / 0 | ✘ 0 / 0 | ✓ 36 / 51 | ✓ 36 / 51 |
| AMF Corporation (Australian airline)<br>*AMF* is a short form of *Australian military forces* | intent | ✘ 0 / 3 | ✘ 0 / 0 | ✓ 3 / 15 | ✓ 3 / 15 |
| Gift (giftware magazine teaser)<br>*Gift* means *poison* in German | negative | ✘ 0 / 0 | ✘ 0 / 0 | ✓ 12 / 61 | ✓ 12 / 61 |
| Touch Woody – the Internet Pecker<br>*woody* and *pecker* are slang words for the male genitals in English | sexual | ✓ 3 / 6 | ✘ 0 / 3 | ✓ 4 / 75 | ✓ 3 / 6 |
| FARTFULL (Swedish furniture)<br>sounds like *full of farts* in English | vulgar | ✘ 0 / 0 | ✘ 0 / 0 | ✘ 0 / 0 | ✘ 0 / 0 |
| Wang Cares (company slogan)<br>sounds similar to the British *wanker*; *wang* is also an English slang term for the male genitals | sexual | ✓ 2 / 2 | ✓ 1 / 1 | ✓ 3 / 20 | ✓ 2 / 2 |
| Honda Fitta (Japanese car)<br>*fitta* is a slang word for women's genitals in Swedish | sexual | ✓ 11 / 11 | ✓ 8 / 10 | ✓ 16 / 21 | ✓ 11 / 11 |
| table (agenda item)<br>means *propose for discussion* in the UK, but *postpone* in the US | intent | ✘ 0 / 0 | ✘ 0 / 0 | ✓ 2 / 45 | ✓ 2 / 45 |
| Detected marketing blunders: | | 18 / 45 | 18 / 45 | 29 / 45 | 35 / 45 |
| *intent* | | 0 / 4 | 0 / 4 | 4 / 4 | 4 / 4 |
| *negative* | | 3 / 15 | 8 / 15 | 10 / 15 | 14 / 15 |
| *sexual* | | 7 / 14 | 4 / 14 | 7 / 14 | 8 / 14 |
| *vulgar* | | 8 / 12 | 6 / 12 | 8 / 12 | 9 / 12 |
| Retrieved clues: | | 105 / 151 | 85 / 247 | 343 / 1247 | 231 / 562 |
| Precision: | | .70 | .34 | .28 | .41 |
| Recall: | | .40 | .40 | .64 | .78 |
| $F_1$ score: | | .51 | .37 | .39 | .54 |

Table 8.2: Wiktionary-based detection of cross-lingual marketing blunders

majority of the type *negative*. This yields a low recall for the MARKED method. It should be noted that it is not surprising to find the recall of MARKED lower than that of LOOKUP, because the former always returns a subset of the latter. This is different for the SOUNDEX method, which facilitates the detection of marketing blunders that remain unseen by the other methods, for example, in the case of the *Kinki Nippon Tourist* agency, whose detection requires a replacement of the letter 'i' with 'y'.

We find that our COMBINE procedure yields the most reasonable trade-off between precision and recall, since it achieves the highest recall of the three methods and a higher precision than LOOKUP and SOUNDEX. With this method, we are able to detect 35 of the 45 cross-lingual marketing blunders (recall .78). For the corresponding marketing campaigns, the copywriters would have to examine a total of 562 clues (on average 12 per blunder). Each clue consists of the word form, a language code, and a short sense definition. 231 of the clues are relevant for detecting the blunder (on average 5 per blunder; precision .41).

**Discussion and error analysis**. The MARKED method works well for detecting marketing blunders of type *vulgar* or *sexual*, whereas the LOOKUP method predominantly retrieves clues for the blunders of type *intent* and *negative*. Since we designed our SOUNDEX method to only return marked entries from the homograph index, it is likewise less suitable for the blunder types *intent* and *negative*. Future work should incorporate research results on sentiment and polarity analysis for detecting negatively connotated expressions.

None of our methods is able to detect marketing blunders whose text contains a problematic word as a substring. The product name FARTFULL, for instance, needs to be separated into *fart* and *full*, before an index lookup can yield relevant clues. The large number of substring combinations, however, also yields a larger number of potentially irrelevant clues. The English words *farther* or *penny-farthing* both contain the substring *fart*, but do not lead to a vulgar interpretation right away.

A similar problem occurs for inflected word forms, which are yet only present for English, German, and Russian within OntoWiktionary, because Wiktionary translations are usually limited to providing the lemma in the target language. Using automatic lemmatizing or morphologic analysis tools is difficult, since they would essentially need to cover any language spoken in the world. A viable option is developing heuristics or enabling the extraction of additional Wiktionary language editions, which we have envisaged in section 5.6.

Apart from that, we observe that marketing blunders whose source and target language make use of different scripts are not handled well by our methods. Although the *Bardak* machines of a US manufacturing company caused sexual innuendos in Russia, our method would need to automatically transliterate the Latin spelling to the Cyrillic *бардак*. Future work should therefore incorporate state-of-the-art transliteration systems.

## 8.5 Chapter Summary

In this chapter, we discussed how Wiktionary data and the resources developed in this thesis can be employed in actual natural language processing applications. We first presented a detailed survey of previous works in this direction and commented on improving these existing strands of research by the findings of the preceding chapters. We then exemplified the usefulness of our resources by describing two case studies. By applying OntoWiktionary to the computation of monolingual and cross-lingual verb similarity, we obtained competitive results to expert-built wordnets in the monolingual case and outperformed them in the cross-lingual case. As part of the second case study, we took advantage of the high language coverage, the continual growth, and the vast amount of information on non-standard language varieties in order to detect cross-lingual marketing blunders. We found a high recall and a reasonable precision for our system, which, to the best of our knowledge, is the first tool-assisted approach to this task.

Chapter 9

# Conclusion

Collaborative lexicography denotes the compilation of dictionaries based on user contributions without being controlled by professional editors. In this thesis, we have researched this new lexicographic paradigm both from a metalexicographic and a natural language processing perspective using the example of *Wiktionary* (2002 f.), which is currently the largest collaborative dictionary available.

**The metalexicographic perspective**. For the metalexicographic study, we formulated the research question as investigating the differences and implications of collaborative lexicography to pertinent methods and theories in dictionary making. To this end, we gave a detailed description of Wiktionary in chapter 2, which comments on the various structures of the dictionary. We described the organization of multilingual knowledge in multiple language editions and foreign language entries, the use of hypertext to interconnect the dictionary articles, the different access paths with which users can find the encoded information, as well as the dictionary microstructure, which defines the order and the format of the individual lexicographic descriptions (e.g., usage examples, pragmatic labels, semantic relations). Beyond the actual dictionary articles, we also discussed the composition of the Wiktionary user community and the lexicographic metatexts, which are used to organize the lexicographic work and to explain the dictionary usage.

In chapter 3, we focused on the collaborative compilation approach of Wiktionary and compared its dictionary conception to expert-built dictionaries. We observed that Wiktionary does not specify its dictionary functions and target audience, which is stipulated by state-of-the-art lexicographic theories. We also discussed multiple quality flaws originating from unspecific or too general definitions, spelling errors, and old-fashioned descriptions yielding a dictionary in which inconsistencies predominate. By comparing the English and the German Wiktionary, we noticed opposing cultures of providing references to verify the encoded descriptions: for the German language, we identified many references to contemporary online resources, while the English Wiktionary encodes only a few references, which mainly

point to archaic copyright-expired dictionaries. Wiktionary does not make use of a specific lexicographic corpus yielding a lack of corpus evidence. Rather, it is strongly based on other dictionaries and reference works, but also provides evidence based on the intuition and language feeling of the voluntary authors. Wiktionary saves each modification to a dictionary article in a revision history and provides separate discussion pages. This kind of information on the lexicographic process is usually either not documented or kept private in expert-built dictionaries. Our analysis showed that the lexicographic process of collaborative dictionaries cannot be described with established models. This is why we proposed a novel process description addressing the frequent revision of the dictionary articles, the inseparable phases of collecting, editing, analyzing, and typesetting the lexicographic descriptions, as well as the opportunity to discuss individual articles, instructions, and the dictionary as a whole.

In chapter 4, we carried out a large quantitative comparison between Wiktionary and a number of publicly available dictionaries. We argued that the different levels of detail and developmental stages of the dictionary articles raise the need for quantitative methods, as the previously used qualitative approaches can only yield limited conclusions. We observed that Wiktionary is continually and rapidly growing due to the division of labor. It covers a large number of languages from almost any populated region of the world, and we envisaged Wiktionary's potential for providing linguistic knowledge on resource-poor languages in a small case study on Greenlandic. Among the encoded lexical entries, we predominantly found basic vocabulary, neologisms, phrasemes, translations, technical terms from natural and structural sciences and sports, as well as expressions from non-standard language varieties, such as slang, jargon, and dialect. Underrepresented were the domains of humanities and social sciences, and we observed that especially the English Wiktionary is very sparse in the encoding of semantic relations (i.e., synonyms, hypernyms, etc.). The comparison of the lexical overlap between Wiktionary and expert-built wordnets and thesauri showed that it contains largely complementary entries, which motivated us to explore ways of combining the knowledge of experts and collaborative authors in the subsequent parts of the thesis.

We concluded the metalexicographic perspective by finding that the collaborative Wiktionary is not an appropriate replacement for expert-built dictionaries due to its inconsistencies, quality flaws, one-fits-all-approach, and strong dependence on expert-built dictionaries. However, Wiktionary's rapid and continual growth, high coverage of languages, technical domains, and non-standard language varieties, as well as the kind of evidence based on the authors' intuition provide promising opportunities for both lexicography and natural language processing.

**The natural language processing perspective**. From the natural language processing perspective, we targeted the research question of harvesting linguistic knowledge from Wiktionary, such that it can be employed for various natural language processing applications. In chapter 5, we surveyed multiple text mining software tools for obtaining and extracting

Wiktionary knowledge. Unlike most previous works, we argued that extracting sense-disambiguated knowledge is crucial for downstream applications. This is why we propose a rule-based method for disambiguating semantic relations and translations. Our method significantly outperforms the random and most frequent sense baselines as well as an off-the-shelf text similarity method. We observed that machine learning was not able to improve our rule-based method much, because of the large heterogeneity of the training data. Based on the disambiguated synonymy relations, we inferred synsets and used them as the concepts of our newly derived multilingual lexical ontology OntoWiktionary. To overcome the previously observed sparseness of the semantic relations, we generated bidirectional links and inferred yet missing relations based on the disambiguated translations. Our lexical ontology is publicly available, and we found that it fills the gap between the small expert-built wordnets and the large amount of encyclopedic knowledge from Wikipedia.

In chapter 6, we discussed that relying on only one specific dictionary limits the performance of a natural language processing system, because of complementary contents and individual advantages of each dictionary. Integrated resources can instead benefit from synergies, which motivated us to combine Wiktionary with other dictionaries. We identified the alignment of word senses as a particular challenge requiring subtle distinctions of meaning. For this purpose, we described an automatic method based on the cosine similarity and the Personalized PageRank, with which we achieved better results than baseline approaches and purely word-overlap-based methods for our experiments on aligning Wiktionary with *WordNet 3.0* (2006). Since there were previously no evaluation datasets for Wiktionary and since other existing datasets were unbalanced and rather small as well, we created a novel annotated dataset for evaluating our alignment approach. We asked ten human raters to annotate this dataset, and we carefully studied their inter-rater agreement and the sources of disagreement. Our final aligned resource is characterized by an increased coverage including any parts of speech and the largely complementary technical domains encoded in Wiktionary and WordNet. In addition to that, we obtained an enriched representation of word senses by combining the various information types from either dictionary.

In chapter 7, we addressed the issue of using international standards to achieve syntactic and semantic interoperability of language resources. This is important because dictionaries largely differ in their macrostructure, microstructure, coverage and granularity, terminology, data format, and access paths. To obtain a standardized representation of Wiktionary and OntoWiktionary, we described the lexicon model UBY-LMF, which is an implementation of the Lexical Markup Framework (LMF). Therefore, we selected extension packages and classes from the LMF standard, and we defined data categories for Wiktionary's information items, which we registered at the data category registry ISOcat. In a second step, we populated this lexicon model with the knowledge harvested from Wiktionary and OntoWiktionary as well as our alignments between different pairs of dictionaries. We then integrated the standardized representations into the lexical resource project *UBY* (2012). This facilitates a unified access

to a number of different lexical resources by means of a shared web interface for human users and an application programming interface for natural language processing applications. A user can, in particular, switch between and combine information from Wiktionary and other dictionaries without completely changing the software.

In chapter 8, we researched employing Wiktionary and OntoWiktionary for natural language processing applications. We first presented a survey of previous works utilizing multiple information types from Wiktionary to improve their applications. Besides reporting the goals, contexts, and results of the individual approaches, we commented on how our findings can enable future research avenues in these directions. In two case studies, we then showed the usefulness of our resources. For measuring verb similarity, we made use of the disambiguated semantic relations and translations of OntoWiktionary, with which we significantly outperformed previously reported results using undisambiguated Wiktionary data. We particularly achieved competitive results to expert-built wordnets in a monolingual setting, and we exceeded their performance by a large margin in a cross-lingual setting. In the second study, we exploited the vast number of translations and pragmatic labels covering non-standard language varieties in order to detect cross-lingual marketing blunders. That is to say, we developed a tool that retrieves clues for pejorative, offensive, or vulgar expressions, as well as words with a negative connotation in a large number of languages. Sales promoters can use this tool to detect naming problems of their product or company in a target language, which would potentially cause the refusal of a product due to embarrassment or offense. We found that Wiktionary can effectively be used for this task and that our method represents a reasonable trade-off between the number of detected marketing blunders and the number of clues that have to be examined by the copywriters.

We concluded the natural language processing perspective by finding that it is crucial to develop methods for extracting and disambiguating Wiktionary data to make proper use of it. Following the results of our metalexicographic study, we identified the numerous languages, the high coverage, and the information on non-standard language varieties as particularly beneficial for natural language processing applications and as a complement to expert-built lexical resources.

To stipulate further work on Wiktionary in the scientific community, we publish the software, resources, and datasets described in appendix A on our homepage:

```
http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/
http://www.christian-meyer.org/research/publications/dissertation/data/
```

**Limitations and outlook.** The present thesis allows for multiple directions of future research. Our study of collaborative lexicography is largely based on Wiktionary, although there are multiple other collaboratively built dictionaries that we could only briefly discuss. Further efforts are required in order to compare them (cf. Mann, 2010), to classify them (cf. Abel and Meyer, 2013), and to study the interface between professional lexicographers and lay authors

(cf. Lew, 2011). Especially the latter appears to bear much unused potential, because of the different types of information and evidence encoded in expert-built and collaborative dictionaries. A major challenge for bridging the gap between professional and voluntary authors is coordinating their different goals: Fully collaboratively built dictionaries (such as Wiktionary) manage to attract large web communities that are motivated by altruism, socializing with others, and providing open-licensed knowledge. Instead, collaborative-institutional dictionaries often suffer from a small number of contributors. This might be due to complicated instructions, the fear of hostile corrections, or the commercial interest of the publisher and requires further investigation.

For the study in chapter 3 and chapter 4, we mainly focused on the English, German, and Russian Wiktionary editions, since there is currently no software for extracting information from all language editions. Future work should concentrate on flexible text mining software, for example, based on machine learning, pattern recognition, and information extraction technologies such as wrapper induction. Developing this kind of software is not only useful for accessing Wiktionary data, but also facilitates the acquisition of structured data from other human-oriented dictionaries. Another promising option is community-based approaches that build on the work of many voluntary developers who also actively contribute to the specific dictionary projects (cf. Hellmann et al., 2013).

We found that extracting sense-disambiguated information is of particular importance, because otherwise, unrelated words get connected by means of semantic relations (see section 5.4). Future developments in word sense disambiguation and word sense induction research can improve the automatic disambiguation of relations and the inference of synsets. Recent works by Flati and Navigli (2012) and Matuschek and Gurevych (2013) suggest that using graph-based methods is a very promising avenue of research. A different possibility is relying on manual disambiguation. This might become possible in the context of the *Wikidata* project, which strives for providing rigidly structured data for Wiktionary and other Wikimedia projects. Since this would also alleviate the noise in the extracted information items, *Wikidata* certainly presents an interesting starting point for future research.

Another strand of research is the combination of multiple complementary resources. In chapter 6, we have proposed and discussed methods for aligning two dictionaries at the level of word senses. Though being an important step towards providing combined resources, we envisage aligning other information types than word senses (cf. Eckle-Kohler and Gurevych, 2012), aligning more than two lexical resources at the same time (cf. Kirschner, 2012), and dealing with different granularities of word senses as three major open research topics.

Besides developing alignment algorithms, we consider the work on internationally recognized resource standards and the corresponding documentation and education to be of critical importance, because the lack of syntactic and semantic interoperability often hinders the combination and the use of lexical resources in natural language processing applications. Our representation of Wiktionary according to the Lexical Markup Framework (LMF) is an im-

portant step in modeling collaborative dictionaries. However, consolidating the efforts of the major lexicon models *lemon*, UBY-LMF, Wordnet-LMF, the *Wörterbuchnetz*, and many others is definitely a desideratum for future research targeted towards the development of a unified model.

Finally, employing collaborative dictionaries (and Wiktionary in particular) for natural language processing provides much room for continuative works. We have discussed multiple tasks that can further benefit from using Wiktionary and OntoWiktionary in our survey of Wiktionary-based applications (see section 8.2). The two case studies we have presented let us expect promising results when utilizing collaborative dictionaries for tasks requiring multilingual information or information on non-standard language varieties.

# List of Tables

# List of Figures

# Bibliography

## Dictionaries and Resources


*An American Dictionary of the English Language* by Noah Webster, New York: S. Converse, 1828.

*The American Heritage dictionary of the English language* edited by William Morris, New York: American Heritage Publishing, 1969.

*The American Heritage dictionary of the English language*, 4th edition, Boston: Houghton Mifflin, 2000.

*BabelNet*, Rome: Sapienza Università di Roma, 2012 f. [cf. Navigli and Ponzetto (2010)]. Online: `http://lcl.uniroma1.it/babelnet`.

*BalkaNet*, Patras: University of Patras, August 2004. [cf. Tufiș et al. (2004)]. Online: `http://www.dblab.upatras.gr/balkanet`.

*British National Corpus*, Oxford: BNC Consortium, 1991–1994.

*Brockhaus*, 20th edition, Leipzig/Mannheim: Brockhaus, 1996–1999.

*Cambridge Dictionaries Online*, Cambridge: Cambridge University Press, 1999 f. Online: `http://dictionary.cambridge.org`.

*Cambridge International Dictionary of English* edited by Paul Procter, Cambridge: Cambridge University Press, 1995.

*Canadian Oxford Dictionary* edited by Katherine Barber, 2nd edition, Toronto: Oxford University Press, 2004.

*canoonet*, Basel: Canoo Engineering, 2000 f. Online: `http://www.canoo.net`.

*The Century Dictionary and Cyclopedia* edited by William Dwight Whitney and Benjamin Eli Smith, New York: The Century Company, 1911.

*Collins COBUILD English Language Dictionary* edited by John M. Sinclair, London: Collins, 1987.

*Collins COBUILD English Dictionary* edited by John M. Sinclair, 2nd edition, London: HarperCollins, 1995.

*Collins COBUILD Advanced Learner's English Dictionary*, 5th edition, Glasgow: HarperCollins, 2006.

*Collins English Dictionary* edited by Patrick Hanks, London: Collins, 1986.

*Cyc*, Austin: Cycorp, 1995 f. [cf. Lenat (1995)]. Online: `http://www.cyc.com`.

*DBpedia*, 2007 f. [cf. Bizer et al. (2009b)]. Online: `http://dbpedia.org`.

*Deutsches Wörterbuch* by Jacob Grimm and Wilhelm Grimm, Leipzig: S. Hirzel, 1854–1961.

*Dictionary of the West Greenland Eskimo language* by Christian Wilhelm Schultz-Lorentzen, Copenhagen: C. A. Reitzel, 1927.
*Dictionary.com*, Oakland, CA: Dictionary.com, 1995 f. Online: `http://www.dictionary.com`.
*Digitales Wörterbuch der deutschen Sprache*, Berlin: Berlin-Brandenburgischen Akademie der Wissenschaften, 2004 f. Online: `http://www.dwds.de`.
*Duden: Die deutsche Rechtschreibung* edited by Dudenredaktion, 25th edition, Mannheim: Bibliographisches Institut, 2009.
*Duden: Deutsches Universalwörterbuch* edited by Dudenredaktion, 7th edition, Mannheim: Bibliographisches Institut, 2011.
*Duden: Das große Fremdwörterbuch* edited by Wissenschaftlichen Rat der Dudenredaktion, 4th edition, Mannheim: Bibliographisches Institut & F. A. Brockhaus, 2007.
*Duden online*, Mannheim: Bibliographisches Institut, April 2011 f. Online: `http://www.duden.de`.
*The New Encyclopædia Britannica*, 15th edition, Chicago: Encyclopædia Britannica, 1985.
*Encyclopædia Britannica Online*, Chicago: Encyclopædia Britannica, 1994 f. Online: `http://www.britannica.com`.
*EuroWordNet*, Paris: ELRA Distribution Agency, 1999. [cf. Vossen (1998)]. Online: `http://www.illc.uva.nl/EuroWordNet`.
*FrameNet*, Berkeley, CA: International Computer Science Institute, 1997 f. Online: `http://framenet.icsi.berkeley.edu`.
— Version 1.2, Berkeley, CA: International Computer Science Institute, July 2005.
— Version 1.3, Berkeley, CA: International Computer Science Institute, December 2006.
— Version 1.5, Berkeley, CA: International Computer Science Institute, September 2010.
*The New Geordie Dictionary* edited by Frank Graham, Rothbury: Butler, 1987.
*GermaNet*, Tübingen: Seminar für Sprachwissenschaft, 1997 f. [cf. Kunze and Lemnitzer (2002)]. Online: `http://www.sfs.uni-tuebingen.de/GermaNet`.
— Version 5.0, Tübingen: Seminar für Sprachwissenschaft, May 2006.
— Version 6.0, Tübingen: Seminar für Sprachwissenschaft, April 2011.
*Goethe-Wörterbuch* edited by Berlin-Brandenburgische Akademie der Wissenschaften, Akademie der Wissenschaften zu Göttingen, and Heidelberger Akademie der Wissenschaften, Stuttgart: Kohlhammer, 1966 f.
*Greenlandic English Dictionary*, Nuuk: Greenland Language Secretariat Oqaasileriffik and Nuuk: Inerisaavik, University of Greenland, 2007 f. Online: `http://www.oqaasileriffik.gl/en/resources/greenlandicenglishdictionary`.
*Den Grønlandske Ordbog: Grønlandsk–Dansk* by Christian Wilhelm Schultz-Lorentzen, København: A. Rosenbergs bogtr., 1926.
*Großes Abkürzungsbuch. Abkürzungen, Kurzwörter, Zeichen, Symbole* by Heinz Koblischke, 2nd edition, Leipzig: Bibliographisches Institut, 1980.
*HECTOR*, Oxford: Oxford University Press and Palo Alto: Digital Equipment Corporation, 1993. [cf. Atkins (1992)].
*Der Neue Herder*, Freiburg im Breisgau: Herder, 1949.
*Illustriertes Lexikon der deutschen Umgangssprache* by Heinz Küpper, Stuttgart: Klett, 1982–1984.
*IMSLex-Subcat* by Judith Eckle-Kohler, Stuttgart: Universität Stuttgart, 1999. [cf. Eckle-Kohler (1999)].

*LEO Deutsch–Italienisch*, München: LEO, 2008 f. Online: http://dict.leo.org/itde.
*Lexikon der Sprachwissenschaft* edited by Hadumod Bußmann, 3rd edition, Stuttgart: Kröner, 2002.
*Longman Dictionary of Contemporary English* edited by Paul Procter, Harlow: Longman, 1978.
*Longman Dictionary of Contemporary English* edited by Della Summers, 3rd edition, Harlow: Longman, 1995.
*Louvain EAP Dictionary*, Louvain-la-Neuve: Université catholique de Louvain, 2010 f. [cf. Granger and Paquot (2010)]. Online: http://www.uclouvain.be/en-322619.html.
*Macmillan Dictionary Online*, Houndmills: Macmillan, 2009 f.
Online: http://www.macmillandictionary.com.
*Macmillan Open Dictionary*, Houndmills: Macmillan, 2009 f.
Online: http://www.macmillandictionary.com/open-dictionary.
*McGraw-Hill's American Slang Dictionary* by Richard A. Spears, 2nd edition, Chicago: McGraw-Hill, 2007.
*Merriam-Webster Open Dictionary*, Springfield: Merriam-Webster, 2005 f.
Online: http://nws.merriam-webster.com/opendictionary.
*Metzler-Lexikon Sprache* edited by Helmut Glück, 3rd edition, Stuttgart/Weimar: Metzler, 2005.
*Mineralogy Database* edited by David Barthelmy, 1997 f. Online: http://www.webmineral.com.
*Moby Thesaurus* by Grady Ward, Salt Lake City: Project Gutenberg Literary Archive Foundation, 1996.
Online: http://www.gutenberg.org/ebooks/3202.
*MultiWordNet*, Povo-Trento: Fondazione Bruno Kessler, 2002 f. [cf. Pianta et al. (2002)].
Online: http://multiwordnet.fbk.eu.
*Neues Lexikon der Vornamen* by Walter Burkart, Bergisch Gladbach: Lübbe, 1993.
*NULEX*, Evanston, IL: Northwestern University, 2011. [cf. McFate and Forbus (2011)].
Online: http://www.qrg.northwestern.edu/resources/nulex.html.
*OmegaWiki*, 2006 f. [cf. Meijssen (2009)]. Online: http://www.omegawiki.org.
*OneLook Dictionary Search*, Hilton Head Island, SC: Datamuse, April 1996 f.
Online: http://www.onelook.com.
*Online Etymology Dictionary* by Douglas Harper, 2001 f. Online: http://www.etymonline.com.
*OntoWordNet*, Trento: Laboratory for Applied Ontology, Consiglio Nazionale delle Ricerche, 2003. [cf. Gangemi et al. (2003)]. Online: http://www.loa.istc.cnr.it/DOLCE.html#OntoWordNet.
*Open Mind Common Sense*, Cambridge: MIT Media Lab, September 1999–2012. [currently offline].
Online: http://openmind.media.mit.edu.
*OpenCyc*, Version 4.0, Austin: Cycorp, June 2012.
*OpenThesaurus*, 2004 f. [cf. Naber (2005)]. Online: http://www.openthesaurus.de.
*ordnet.dk*, København: Det Danske Sprog- og Litteraturselskab, 2004 f. Online: http://ordnet.dk.
*OWID – Wortschatzinfomrationssystem Deutsch*, Mannheim: Institut für Deutsche Sprache, 2008 f.
Online: http://www.owid.de.
*Oxford Advanced Learner's Dictionary of Current English* edited by Anthony Paul Cowie, 4th edition, Oxford: Oxford University Press, 1989.
*Oxford Advanced Learner's Dictionary of Current English* edited by Jonathan Crowther, 5th edition, Oxford: Oxford University Press, 1995.

*Oxford Dictionaries Online*, Oxford: Oxford University Press, 2010 f.
Online: http://www.oxforddictionaries.com.
*Oxford Dictionary of English* edited by Catherine Soanes and Angus Stevenson, 2nd edition, Oxford: Oxford University Press, 2003.
*Oxford English Dictionary Online*, Oxford: Oxford University Press, 2002 f.
Online: http://dictionary.oed.com.
*Oxford English Dictionary* edited by James Murray, John A. Simpson, and Edmund S. Weiner, 2nd edition, Oxford: Clarendon Press, 1989.
*Oxford Student's Dictionary of Current English* by Albert Sydney Hornby and Christina Ruse, 2nd edition, Oxford: Oxford University Press, 1988.
*Princeton WordNet Gloss Corpus*, Version 3.0, Princeton: Princeton University, January 2008.
Online: http://wordnet.princeton.edu/glosstag.shtml.
*PropBank*, Philadelphia: Linguistic Data Consortium, September 2004.
*Random House Webster's Unabridged Dictionary*, New York: Random House, 1997.
*Redensarten-Index* edited by Peter Udem, 2001 f. Online: http://www.redensarten-index.de.
*Roget's International Thesaurus* edited by C. O. Sylvester Mawson, New York: Crowell, 1911.
*Roget's Thesaurus of English Words and Phrases* edited by Betty Kirkpatrick, Harlow: Longman, 1987.
*Roget's Thesaurus of English Words and Phrases* edited by Betty Kirkpatrick, London: Penguin, 1998.
*Russian WordNet*, Version 3.0, August 2008. [cf. Гельфейнбейн et al. (2003)].
Online: http://www.wordnet.ru.
*SemCor*, Princeton: Princeton University, 1993 f.
*Simple English Wikipedia*, San Francisco: Wikimedia Foundation, November 2003 f.
Online: http://simple.wikipedia.org.
*Taber's Encyclopedic Medical Dictionary* edited by Clayton L. Thomas, 5th edition, Philadelphia: F. A. Davis, 1993.
*The New Oxford Dictionary of English* edited by Judy Pearsall, Oxford: Oxford University Press, 1998.
*TheFreeDictionary.com*, Huntingdon Valley, PA: Farlex, 2003 f.
Online: http://www.thefreedictionary.com.
*UBY*, Version 1.0, Darmstadt: Ubiquitous Knowledge Processing Lab, April 2012. [cf. Gurevych et al. (2012a)]. Online: http://www.ukp.tu-darmstadt.de/uby.
*Unified Medical Language System – Knowledge Sources*, Bethesda: U. S. National Library of Medicine, 2001 f. Online: http://umlsks.nlm.nih.gov.
*Universal WordNet*, Saarbrücken: Max-Planck-Institut Informatik, 2010 f. [cf. de Melo and Weikum (2009)]. Online: http://www.mpi-inf.mpg.de/yago-naga/uwn.
*Urban Dictionary*, San Francisco: Urban Dictionary, 1999 f. Online: http://www.urbandictionary.com.
*VerbNet*, Boulder: University of Colorado, 2000 f.
Online: http://verbs.colorado.edu/~mpalmer/projects/verbnet.html.
— Version 1.5, Boulder: University of Colorado, 2005.
— Version 3.1, Boulder: University of Colorado, 2009.
*Webster's New World Dictionary* edited by Victoria E. Neufeldt, New York: Simon & Schuster, 1988.
*Webster's Revised Unabridged Dictionary* edited by Noah Porter, Springfield: G. & C. Merriam, 1913.

*Webster's Seventh New Collegiate Dictionary* edited by Philip B. Gove, Springfield: G. & C. Merriam, 1965.

*Webster's Third New International Dictionary* edited by Philip B. Gove, Unabridged edition, Springfield: G. & C. Merriam, 1961.

*WikiNet*, Heidelberg: HITS, 2010 f. [cf. Nastase et al. (2010)].
Online: `http://www.h-its.org/english/research/nlp/download/wikinet`.

*Wikipedia*, San Francisco: Wikimedia Foundation, January 2001 f. Online: `http://www.wikipedia.org`.

*Wiktionary*, San Francisco: Wikimedia Foundation, December 2002 f.
Online: `http://www.wiktionary.org`.

*wissen.de*, Gütersloh/München: wissenmedia, 2000 f. Online: `http://www.wissen.de`.

*WordNet*, Princeton: Princeton University, 1985 f. [cf. Fellbaum (1998)].
Online: `http://wordnet.princeton.edu`.

— Version 1.5, Princeton: Princeton University, March 1995.

— Version 1.6, Princeton: Princeton University, February 1998.

— Version 1.7.1, Princeton: Princeton University, July 2002.

— Version 2.0, Princeton: Princeton University, July 2003.

— Version 2.1, Princeton: Princeton University, March 2005.

— Version 3.0, Princeton: Princeton University, December 2006.

— Version 3.1, Princeton: Princeton University, February 2011.

*WordNet++*, Rome: Sapienza Università di Roma, August 2010. [cf. Ponzetto and Navigli (2010)].
Online: `http://lcl.uniroma1.it/wordnetplusplus`.

*WordNet Domains*, Version 3.2, Povo-Trento: Fondazione Bruno Kessler, February 2007. [cf. Bentivogli et al. (2004)]. Online: `http://wndomains.fbk.eu`.

*Wordnik*, San Mateo, CA: Wordnik, 2009 f. Online: `http://www.wordnik.com`.

*Wortschatz-Lexikon*, Leipzig: Universität Leipzig, 1998 f. Online: `http://dict.uni-leipzig.de`.

*Wörterbuchnetz*, Trier: Trier Center for Digital Humanities, 2007 f. Online: `http://woerterbuchnetz.de`.

*YAGO*, Saarbrücken: Max-Planck-Institut Informatik, 2007 f. [cf. Suchanek et al. (2007)].
Online: `http://www.mpi-inf.mpg.de/yago-naga/yago`.

漢語大詞典, 3.0, 香港：商務印書館, 2007. [Han yu da ci dian 3.0, Hong Kong: The Commercial Press, 2007].

## Scientific Literature

Andrea Abel: 'Elektronische Wörterbücher: Neue Wege und Tendenzen', in Felix San Vincente (Ed.): *Lessicografia bilingue e Traduzione: metodi, strumenti e approcci attuali*, pp. 35–56, Monza: Polimetrica, 2006.

Andrea Abel: 'Dictionary Writing Systems and Beyond', in: Granger and Paquot (2012), chapter 5, pp. 83–106.

Andrea Abel and Christian M. Meyer: 'The dynamics outside the paper: user contributions to online dictionaries', in: *Proceedings of the 3rd Biennial Conference on Electronic Lexicography* (eLex), Tallinn, Estonia, October 2013. *(to appear)*.

Eneko Agirre and Aitor Soroa: 'Personalizing PageRank for Word Sense Disambiguation', in: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), pp. 33–41, Athens, Greece, March 2009.

Melina Alexa, Bernd Kreissig, Martina Liepert, Klaus Reichenberger, Lothar Rostek, Karin Rautmann, Werner Scholze-Stubenrecht, and Sabine Stoye: 'The Duden Ontology: An Integrated Representation of Lexical and Ontological Information', in: *Proceedings of the LREC Workshop 'The Ontologies and Lexical Knowledge Bases'*, pp. 1–8, Las Palmas, Canary Islands, Spain, May 2002.

Robert A. Amsler: 'Computational Lexicology: A Research Program', in: *Proceedings of the American Federation of Information Processing Societies* (AFIPS), Vol. 51, pp. 657–663, Houston, TX, USA, June 1982.

Ron Artstein and Massimo Poesio: 'Inter-Coder Agreement for Computational Linguistics', *Computational Linguistics* 34 (4): 555–596, December 2008.

B. T. Sue Atkins: 'Building a Lexicon. The Contribution of Lexicography', *International Journal of Lexicography* 4 (3): 167–204, Autumn 1991.

B. T. Sue Atkins: 'Tools for computer-aided lexicography: the Hector project', in: *Papers in Computational Lexicography: COMPLEX '92. Proceedings of the 2nd International Conference*, pp. 1–59, Budapest, Hungary, October 1992.

B. T. Sue Atkins and Michael Rundell: *The Oxford Guide to Practical Lexicography*, Oxford, UK: Oxford University Press, 2008.

Satanjeev Banerjee and Ted Pedersen: 'Extended Gloss Overlaps as a Measure of Semantic Relatedness', in: *Proceedings of the 18th International Joint Conference on Artificial Intelligence* (IJCAI), pp. 805–810, Acapulco, Mexico, August 2003.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni: 'Open Information Extraction from the Web', in: *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (IJCAI), pp. 2670–2676, Hyderabad, India, January 2007.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta: 'Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing', in: *Proceedings of the COLING Workshop on 'Multilingual Linguistic Resources'*, pp. 101–108, Geneva, Switzerland, August 2004.

Henning Bergenholtz and Rufus Gouws: 'A New Perspective on the Access Process', *Hermes – Journal of Language and Communication Studies* 44: 103–128, 2010.

Henning Bergenholtz, Sandro Nielsen, and Sven Tarp (Eds.): *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, Linguistic Insights: Studies in Language and Communication Vol. 90, Bern: Peter Lang, 2009.

Henning Bergenholtz and Sven Tarp: 'Two opposing theories: On H. E. Wiegand's recent discovery of lexicographic functions', *Hermes – Journal of Language and Communication Studies* 31: 171–196, 2003.

Tim Berners-Lee, James Hendler, and Ora Lassila: 'The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities', *Scientific American* 284 (5): 34–43, May 2001.

Delphine Bernhard and Iryna Gurevych: 'Combining Lexical Semantic Resources with Question & Answer Archives for Translation-Based Answer Finding', in: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*

*Processing of the AFNLP* (ACL/IJCNLP), Vol. 2, pp. 728–736, Singapore, August 2009.

Herman L. Beyer: 'n Algemene tipologie van leksikografiese etikette', *Tydskrif vir Geesteswetenskappe* 51 (3): 419–446, September 2011.

Christian Bizer, Tom Heath, and Tim Berners-Lee: 'Linked Data – The Story So Far', *International Journal on Semantic Web and Information Systems: Special Issue on Linked Data* 5 (3): 1–22, July–September 2009a.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann: 'DBpedia – A Crystallization Point for the Web of Data', *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 7 (3): 154–165, September 2009b.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio: 'Learning Structured Embeddings of Knowledge Bases', in: *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 301–306, San Francisco, CA, USA, August 2011.

Alexandre Bouchard-Côté, Percy Liang, Thomas L. Griffiths, and Dan Klein: 'A Probabilistic Approach to Diachronic Phonology', in: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP/CoNLL), pp. 887–896, Prague, Czech Republic, June 2007.

Susan Windisch Brown, Travis Rood, and Martha Palmer: 'Number or Nuance: Which Factors Restrict Reliable Word Sense Annotation?', in: *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC), pp. 3237–3243, Valletta, Malta, May 2010.

Alexander Budanitsky and Graeme Hirst: 'Evaluating WordNet-based Measures of Lexical Semantic Relatedness', *Computational Linguistics* 32 (1): 13–47, March 2006.

Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek: 'Towards Linguistically Grounded Ontologies', in Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl (Eds.): *The Semantic Web: Research and Applications, 6th European Semantic Web Conference* (ESWC), Lecture Notes in Computer Science Vol. 5554, pp. 111–125, Berlin/Heidelberg: Springer, June 2009.

Thomas Burch and Andrea Rapp: 'Das Wörterbuch-Netz: Verfahren – Methoden – Perspektiven', in: *Geschichte im Netz: Praxis, Chancen, Visionen. Beiträge der Tagung .hist 2006*, Historisches Forum 10, Teilband I, pp. 607–627, Berlin: Humboldt-Universität zu Berlin, 2007.

Clint Burfoot and Timothy Baldwin: 'Automatic Satire Detection: Are You Having a Laugh?', in: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (ACL/IJCNLP), pp. 161–164, Singapore, August 2009.

Anita Burgun and Olivier Bodenreider: 'Comparing Terms, Concepts and Semantic Classes in WordNet and the Unified Medical Language System', in: *Proceedings of the NAACL Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations'*, pp. 77–82, Pittsburgh, PA, USA, June 2001.

Nicoletta Calzolari: 'Detecting Patterns in a Lexical Data Base', in: *Proceedings of the 10th International Conference on Computational Linguistics and the 22nd Annual Meeting of the Association for Computational Linguistics* (COLING/ACL), pp. 170–173, Stanford, CA, USA, July 1984.

Nicoletta Calzolari, Monica Monachini, and Claudia Soria: 'LMF – Historical Context and Perspectives', in: Francopoulo (2013), chapter 1, pp. 1–18.

Michael Carr: 'Internet Dictionaries and Lexicography', *International Journal of Lexicography* 10 (3): 209–230, September 1997.

Imogen Casebourne, Chris Davies, Michelle Fernandes, and Naomi Norman: 'Assessing the accuracy and quality of Wikipedia entries compared to popular online encyclopaedias: A preliminary comparative study across disciplines in English, Spanish and Arabic', August 2012, Online: http://commons.wikimedia.org/wiki/File:EPIC_Oxford_report.pdf.

Alice Yin Wa Chan and Andrew Taylor: 'Evaluating Learner Dictionaries: What the Reviews Say', *International Journal of Lexicography* 14 (3): 163–180, September 2001.

Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari: 'Using Verbs and Adjectives to Automatically Classify Blog Sentiment', in: *Proceedings of the AAAI Spring Symposium 'Computational Approaches to Analysing Weblogs'*, pp. 27–29, Palo Alto, CA, USA, March 2006.

Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer: 'The Open Linguistics Working Group', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp. 3603–3610, Istanbul, Turkey, May 2012.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi: 'Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp. 2149–2152, Istanbul, Turkey, May 2012.

Jeff Conklin: 'Hypertext: An Introduction and Survey', *Computer* 20 (9): 17–41, September 1987.

John Considine: *Dictionaries in Early Modern Europe: Lexicography and the Making of Heritage*, Cambridge: Cambridge University Press, March 2008.

Johannes Daxenberger and Iryna Gurevych: 'A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles', in: *Proceedings of the 24th International Conference on Computational Linguistics* (COLING), Vol. 2, pp. 711–726, Mumbai, India, December 2012.

Thierry Declerck, Karlheinz Mörth, and Piroska Lendvai: 'Accessing and standardizing Wiktionary lexical entries for the translation of labels in Cultural Heritage taxonomies', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp. 2511–2514, Istanbul, Turkey, May 2012.

Don E. Descy: 'The Wiki: True Web Democracy', *TechTrends* 50 (1): 4–5, January 2006.

Sebastian Drude, Daan Broeder, Paul Trilsbeek, and Peter Wittenburg: 'The Language Archive – a new hub for language resources', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp. 3264–3267, Istanbul, Turkey, May 2012.

Isidore Dyen, Joseph B. Kruskal, and Paul Black: *An Indoeuropean Classification: A Lexicostatistical Experiment*, Transactions of the American Philosophical Society Vol. 82 part 5, Philadelphia: The American Philosophical Society, 1992.

Judith Eckle-Kohler: *Linguistisches Wissen zur automatischen Lexikon-Akquisition aus deutschen Textcorpora*, Berlin: Logos, 1999. (Ph.D. Thesis, Universität Stuttgart, 1999).

Judith Eckle-Kohler and Iryna Gurevych: 'Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability', in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), pp. 550–560, Avignon, France, April 2012.

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer: 'UBY-LMF – A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp. 275–282, Istanbul, Turkey, May 2012.

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer: 'UBY-LMF – Exploring the Boundaries of Language-Independent Lexicon Models', in: Francopoulo (2013), chapter 10, pp. 145–156.

Carsten Eickhoff, Pavel Serdyukov, and Arjen P. de Vries: 'Web Page Classification on Child Suitability', in: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (CIKM), pp. 1425–1428, Toronto, ON, Canada, October 2010.

William Emigh and Susan C. Herring: 'Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias', in: *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (HICSS), Vol. 4, p. 99a, Honolulu, HI, USA, January 2005.

Soojeong Eom, Markus Dickinson, and Graham Katz: 'Using semi-experts to derive judgments on word sense alignment: a pilot study', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp. 605–611, Istanbul, Turkey, May 2012.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord: 'Investigations on Word Senses and Word Usages', in: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (ACL/IJCNLP), Vol. 1, pp. 10–18, Singapore, August 2009.

Oren Etzioni, Kobi Reiter, Stephen Soderland, and Marcus Sammer: 'Lexical Translation with Application to Image Search on the Web', in: *Proceedings of Machine Translation Summit XI*, pp. 175–182, Copenhagen, Denmark, September 2007.

Christiane Fellbaum (Ed.): *WordNet: An Electronic Lexical Database*, Language, Speech, and Communication, Cambridge, MA: MIT Press, May 1998.

Samuel Fernando and Mark Stevenson: 'Aligning WordNet Synsets and Wikipedia Articles', in: *Proceedings of the AAAI Workshop 'Collaboratively-Built Knowledge Sources and Artificial Intelligence'*, pp. 48–50, Atlanta, GA, USA, July 2010.

Oliver Ferschke, Johannes Daxenberger, and Iryna Gurevych: 'A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia', in Iryna Gurevych and Jungi Kim (Eds.): *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, Theory and Applications of Natural Language Processing, chapter 5, pp. 121–160, Berlin/Heidelberg: Springer, April 2013.

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar: 'Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages', in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), pp. 777–786, Avignon, France, April 2012.

Tiziano Flati and Roberto Navigli: 'The CQC Algorithm: Cycling in Graphs to Semantically Enrich and Enhance a Bilingual Dictionary', *Journal of Artificial Intelligence Research* 43: 135–171, February 2012.

Joseph L. Fleiss: 'Measuring Nominal Scale Agreement among many Raters', *Psychological Bulletin* 76 (5): 378–381, November 1971.

Gil Francopoulo (Ed.): *LMF: Lexical Markup Framework*, Computer Engineering and IT, London: Wiley-ISTE, March 2013.

Gil Francopoulo and Monte George: 'Model Description', in: Francopoulo (2013), chapter 2, pp. 19–40.

Pedro A. Fuertes-Olivera: 'The Function Theory of Lexicography and Electronic Dictionaries: Wiktionary as a Prototype of Collective Free Multiple-Language Internet Dictionary', in: Bergenholtz et al. (2009), pp. 99–134.

Evgeniy Gabrilovich and Shaul Markovitch: 'Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis', in: *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (IJCAI), pp. 1606–1611, Hyderabad, India, January 2007.

Aldo Gangemi, Roberto Navigli, and Paola Velardi: 'The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet', in Robert Meersman, Zahir Tari, and Douglas C. Schmidt (Eds.): *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, Lecture Notes in Computer Science Vol. 2888, pp. 820–838, Berlin/Heidelberg: Springer, November 2003.

Konstantina Garoufi, Torsten Zesch, and Iryna Gurevych: 'Graph-Theoretic Analysis of Collaborative Knowledge Bases in Natural Language Processing', in: *Proceedings of the Poster And Demonstration Session of the 7th International Semantic Web Conference* (ISWC), CEUR Workshop Proceedings Vol. 401, Karlsruhe, Germany, October 2008.

Jim Giles: 'Internet encyclopaedias go head to head', *Nature* 438 (7070): 900–901, December 2005.

Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, and Herbert Ernst Wiegand (Eds.): *Dictionaries. International Encyclopedia of Lexicography: Supplementary Volume. Recent Developments with a Special Focus on Computational Lexicography*, Handbooks of Linguistics and Communication Science Series Vol. 5.4, Berlin/New York: de Gruyter, October 2013. *(to appear)*.

Sylviane Granger: 'Introduction: Electronic lexicography—from challenge to opportunity', in: Granger and Paquot (2012), chapter 1, pp. 1–11.

Sylviane Granger and Magali Paquot: 'The Louvain EAP Dictionary (LEAD)', in: *Proceedings of the 14th EURALEX International Congress*, pp. 321–326, Leeuwarden, The Netherlands, July 2010.

Sylviane Granger and Magali Paquot (Eds.): *Electronic Lexicography*, Oxford: Oxford University Press, November 2012.

Nicola Guarino, Daniel Oberle, and Steffen Staab: 'What Is an Ontology?', in Steffen Staab and Rudi Studer (Eds.): *Handbook on Ontologies*, International Handbooks on Information Sciences, pp. 1–7, Berlin/Heidelberg: Springer, 2nd edition, 2009.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Tri Duc Nghiem: 'UBY – A Large-Scale Unified Lexical-Semantic Resource', in: *Book of Abstracts of the 23rd Meeting of Computational Linguistics in the Netherlands* (CLIN), p. 81, Enschede, The Netherlands, January 2013.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth: 'UBY – A Large-Scale Unified Lexical-Semantic Resource Based on LMF', in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), pp. 580–590, Avignon, France, April 2012a.

Iryna Gurevych, Michael Matuschek, Tri Duc Nghiem, Judith Eckle-Kohler, Silvana Hartmann, and Christian M. Meyer: 'Navigating Sense-Aligned Lexical-Semantic Resources: The Web Interface to

UBY', in: *Proceedings of the 11th Conference on Natural Language Processing: Empirical Methods in Natural Language Processing* (KONVENS), pp. 194–198, Vienna, Austria, September 2012b.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten: 'The WEKA Data Mining Software: An Update', *ACM SIGKDD Explorations Newsletter* 11 (1): 10–18, June 2009.

Patrick Hanks: 'Evidence and intuition in lexicography', in Jerzy Tomaszczyk and Barbara Lewandowska-Tomaszczyk (Eds.): *Meaning and Lexicography*, Linguistic and Literary Studies in Eastern Europe Vol. 28, pp. 31–41, Amsterdam: John Benjamins, 1990.

Patrick Hanks: 'Word Meaning and Word Use: Corpus evidence and electronic lexicography', in: Granger and Paquot (2012), chapter 4, pp. 57–82.

Sanda M. Harabagiu, George A. Miller, and Dan I. Moldovan: 'WordNet 2 – A Morphologically and Semantically Enhanced Resource', in: *Proceedings of the ACL Special Interest Group on the Lexicon Workshop on Standardizing Lexical Resources* (SIGLEX), pp. 1–7, College Park, MD, USA, June 1999.

Reinhard R. K. Hartmann (Ed.): *The History of Lexicography: Papers from the Dictionary Research Centre Seminar at Exeter*, Studies in the History of the Language Sciences Vol. 40, Amsterdam: John Benjamins, March 1986.

Reinhard R. K. Hartmann and Gregory James: *Dictionary of Lexicography*, London/New York: Routledge, 1998.

Silvana Hartmann and Iryna Gurevych: 'FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection', in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 1363–1373, Sofia, Bulgaria, August 2013.

Claudia Hauff and Dolf Trieschnigg: 'Enhancing Access To Classic Children's Literature', in: *Proceedings of the 3rd CIKM Workshop 'BooksOnline'*, Toronto, ON, Canada, October 2010.

Franz J. Hausmann, Oskar Reichmann, Herbert E. Wiegand, Ladislav Zgusta, Rufus Gouws, Ulrich Heid, and Wolfgang Schweickard (Eds.): *Wörterbücher/Dictionaries/Dictionnaires. Ein internationales Handbuch zur Lexikographie/An International Encyclopedia of Lexicography/Encyclopédie internationale de lexicographie*, Handbücher zur Sprach- und Kommunikationswissenschaft Vol. 5.1, Berlin/New York: de Gruyter, 1989.

Franz Josef Hausmann: 'Die gesellschaftlichen Aufgaben der Lexikographie in Geschichte und Gegenwart', in: Hausmann et al. (1989), chapter 1, pp. 1–18.

Yoshihiko Hayashi: 'A Representation Framework for Cross-lingual/Interlingual Lexical Semantic Correspondences', in: *Proceedings of the Ninth International Conference on Computational Semantics* (IWCS), pp. 155–164, Oxford, UK, January 2011.

Qingyue He: *Automatic Pronunciation Dictionary Generation from Wiktionary and Wikipedia*, Master's thesis, Cognitive Systems Lab, Karlsruhe Institute of Technology, August 2009.

Sebastian Hellmann, Jonas Brekle, and Sören Auer: 'Leveraging the Crowdsourcing of Lexical Resources for Bootstrapping a Linguistic Data Cloud', in Hideaki Takeda, Yuzhong Qu, Riichiro Mizoguchi, and Yoshinobu Kitamura (Eds.): *Semantic Technology: Second Joint International Conference*, Lecture Notes in Computer Science Vol. 7774, pp. 191–206, Berlin/Heidelberg: Springer, 2013.

Verena Henrich and Erhard Hinrichs: 'Standardizing Wordnets in the ISO Standard LMF: Wordnet-LMF for GermaNet', in: *Proceedings of the 23rd International Conference on Computational*

*Linguistics* (COLING), pp. 456–464, Beijing, China, August 2010.

Verena Henrich and Erhard Hinrichs: 'A Comparative Evaluation of Word Sense Disambiguation Algorithms for German', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp. 576–583, Istanbul, Turkey, May 2012.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova: 'Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary', in: *Proceedings of 5th Language & Technology Conference* (LTC), pp. 126–130, Poznań, Poland, November 2011.

Benjamin Herbert, György Szarvas, and Iryna Gurevych: 'Combining Query Translation Techniques to Improve Cross-Language Information Retrieval', in Paul Clough, Colum Foley, Cathal Gurrin, Gareth J.F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Murdoch (Eds.): *Advances in Information Retrieval: 33rd European Conference on IR Research* (ECIR), Lecture Notes in Computer Science Vol. 6611, pp. 712–715, Berlin/Heidelberg: Springer, April 2011.

Thomas Herbst: 'On the way to the perfect learners' dictionary: a first comparison of OALD5, LDOCE3, COBUILD2 and CIDE', *International Journal of Lexicography* 9 (4): 321–357, December 1996.

Daniel Horn: 'A Correction for the Effect of Tied Ranks on the Value of the Rank Difference Correlation Coefficient', *Journal of Educational Psychology* 33 (9): 686–690, December 1942.

Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto: 'Collaboratively built semi-structured content and Artificial Intelligence: The story so far', *Artificial Intelligence: Special Issue "Wikipedia and Semi-Structured Resources"* 194: 2–27, January 2013.

George Hripcsak and Adam S. Rothschild: 'Agreement, the F-Measure, and Reliability in Information Retrieval', *Journal of the American Medical Informatics Association* 12 (3): 296–298, May–June 2005.

Nancy Ide and James Pustejovsky: 'What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability for Language Technology', in: *Proceedings of the Second International Conference on Global Interoperability for Language Resources* (ICGL), Hong Kong, China, January 2010.

ISO 12620: *Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources*, ISO 12620:2009, International Organization for Standardization, Geneva, Switzerland, 2009.

ISO 19501: *Information technology – Open Distributed Processing – Unified Modeling Language (UML)*, ISO 19501:2005, International Organization for Standardization, Geneva, Switzerland, 2005.

ISO 24613: *Language resource management – Lexical markup framework (LMF)*, ISO 24613:2008, International Organization for Standardization, Geneva, Switzerland, 2008.

ISO 639-1: *Codes for the representation of names of languages – Part 1: Alpha-2 code*, ISO 639-1:2002, International Organization for Standardization, Geneva, Switzerland, 2002.

Mario Jarmasz and Stan Szpakowicz: 'Roget's Thesaurus and Semantic Similarity', in: *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP), pp. 212–219, Borovets, Bulgaria, September 2003.

Denis Jouvet, Dominique Fohr, and Irina Illina: 'Building a Pronunciation Lexicon for a Speech Transcription System from Wiktionary Pronunciations only', in: *Proceedings of the XIV International Conference "Speech and Computer"* (SPECOM), Kazan, Russia, September 2011.

Genichiro Kikui: 'Resolving Translation Ambiguity Using Non-parallel Bilingual Corpora', in: *Proceedings of the ACL Workshop on 'Unsupervised Learning in Natural Language Processing'*, pp.

31–36, College Park, MD, USA, June 1999.

Alan Kirkness: 'Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm', in Ulrike Haß (Ed.): *Große Lexika und Wörterbücher Europas: Europäische Enzyklopädien und Wörterbücher in historischen Porträts*, De Gruyter Lexikon, pp. 211–232, Berlin/Boston: de Gruyter, November 2011.

Christian Kirschner: *Kombination mehrerer lexikalisch-semantischer Ressourcen durch multiple Alignments von Wortbedeutungen*, Master's thesis, Technische Universität Darmstadt, December 2012.

Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi: 'He Says, She Says: Conflict and Coordination in Wikipedia', in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 453–462, San Jose, CA, USA, May 2007.

Annette Klosa: 'The Lexicographical Process II: Online Dictionaries', in: Gouws et al. (2013). *(to appear)*.

Kevin Knight and Steve K. Luk: 'Building a large-scale knowledge base for machine translation', in: *Proceedings of the Twelfth National Conference on Artificial Intelligence* (AAAI), pp. 773–778, Seattle, WA, USA, August 1994.

Henrik Køhler Simonsen: 'User Involvement in Corporate LSP Intranet Lexicography', in Henrik Gottlieb, Jens Erik Mogensen, and Arne Zettersten (Eds.): *Symposium on Lexicography XI: Proceedings of the Eleventh International Symposium on Lexicography*, Lexicographica: Series Maior Vol. 115, pp. 489–510, Tübingen: Niemeyer, 2005.

Klaus Krippendorff: *Content Analysis: An Introduction to Its Methodology*, Sage CommText Vol. 5, Beverly Hills, CA: Sage Publications, 1980.

Klaus Krippendorff: *Content Analysis: An Introduction to Its Methodology*, Thousand Oaks, CA: Sage Publications, 2nd edition, 2004.

Andrew A. Krizhanovsky: 'The comparison of Wiktionary thesauri transformed into the machine-readable format', arXiv:1006.5040v1 [cs.IR], June 2010,
Online: `http://arxiv.org/abs/1006.5040`.

Andrew A. Krizhanovsky and Feiyu Lin: 'Related Terms Search Based on WordNet / Wiktionary and its Application in Ontology Matching', in: *Proceedings of the 11th Russian Conference on Digital Libraries* (RCDL), pp. 363–369, Petrozavodsk, Russia, September 2009.

Robert Krovetz: 'Sense-Linking in a Machine Readable Dictionary', in: *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 330–332, Newark, DE, USA, June 1992.

Claudia Kunze and Lothar Lemnitzer: 'GermaNet — representation, visualization, application', in: *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC), Vol. 5, pp. 1485–1491, Las Palmas, Canary Islands, Spain, May 2002.

Zachary Kurmas: 'Zawilinski: A library for studying grammar in Wiktionary', in: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, Gdańsk, Poland, July 2010.

Oi Yee Kwong: 'Aligning WordNet with Additional Lexical Resources', in: *Proceedings of the COLING/ACL Workshop 'Usage of WordNet in Natural Language Processing Systems'*, pp. 73–79, Montreal, QC, Canada, August 1998.

Egoitz Laparra and German Rigau: 'eXtended WordFrameNet', in: *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC), pp. 1214–1219, Valletta, Malta, May 2010.

Els Lefever and Véronique Hoste: 'SemEval-2010 Task 3: Cross-Lingual Word Sense Disambiguation', in: *Proceedings of the 5th International Workshop on Semantic Evaluation* (SemEval), pp. 15–20, Uppsala, Sweden, July 2010.

Douglas B. Lenat: 'Cyc: A large-scale investment in knowledge infrastructure', *Communications of the ACM* 38 (11): 33–38, November 1995.

Jill Lepore: 'Noah's Mark: Webster and the original dictionary wars', *The New Yorker* LXXXII (36): 78–87, November 2006.

Michael Lesk: 'Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone', in: *Proceedings of the 5th Annual International Conference on Systems Documentation* (SIGDOC), pp. 24–26, Toronto, ON, Canada, June 1986.

Bo Leuf and Ward Cunningham: *The Wiki Way: Quick Collaboration on the Web*, Boston, MA: Addison-Wesley, 2001.

Robert Lew: 'Online dictionaries of English', in Pedro A. Fuertes-Olivera and Henning Bergenholtz (Eds.): *E-Lexicography: The Internet, Digital Initiatives and Lexicography*, chapter 11, pp. 230–250, London/New York: Continuum, 2011.

M. Paul Lewis (Ed.): *Ethnologue: Languages of the World*, Dallas, TX: SIL International, 16th edition, 2009. Online: `http://www.ethnologue.com`.

Feiyu Lin and Andrew A. Krizhanovsky: 'Multilingual Ontology Matching based on Wiktionary Data Accessible via SPARQL Endpoint', in: *Proceedings of the 13th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections"* (RCDL), pp. 1–8, Voronezh, Russia, October 2011.

Kenneth C. Litkowski: 'Towards a Meaning-Full Comparison of Lexical Resources', in: *Proceedings of the ACL Special Interest Group on the Lexicon Workshop on Standardizing Lexical Resources* (SIGLEX), pp. 30–37, College Park, MD, USA, June 1999.

Kenneth C. Litkowski: 'Computational Lexicons and Dictionaries', in Keith Brown (Ed.): *Encyclopedia of Language & Linguistics*, pp. 753–761, Oxford: Elsevier, 2nd edition, 2006.

Edward Loper, Szu-ting Yi, and Martha Palmer: 'Combining Lexical Resources: Mapping Between PropBank and VerbNet', in: *Proceedings of the Seventh International Workshop on Computational Semantics* (IWCS), pp. 118–128, Tilburg, The Netherlands, January 2007.

Bernardo Magnini, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Simov, and Richard Sutcliffe: 'Overview of the CLEF 2004 Multilingual Question Answering Track', in Carol Peters, Paul Clough, Julio Gonzalo, Gareth J.F. Jones, Michael Kluck, and Bernardo Magnini (Eds.): *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum* (CLEF), Lecture Notes in Computer Science Vol. 3491, pp. 371–391, Berlin/Heidelberg: Springer, September 2005.

Michael Mann: 'Internet-Wörterbücher am Ende der „Nullerjahre": Der Stand der Dinge. Eine vergleichende Untersuchung beliebter Angebote hinsichtlich formaler Kriterien', *Lexicographica* 26: 19–46, December 2010.

Michael Matuschek and Iryna Gurevych: 'Where the journey is headed: Collaboratively constructed multilingual Wiki-based resources', in: *Proceedings of the GSCL Workshop 'Sprachtechnologie für ein mehrsprachiges Europa'*, Hamburger Arbeiten zur Mehrsprachigkeit, Hamburg, Germany, September 2011.

Michael Matuschek and Iryna Gurevych: 'Dijkstra-WSA: A Graph-Based Approach to Word Sense Alignment', *Transactions of the Association for Computational Linguistics* 1: 151–164, May 2013.

Michael Matuschek, Christian M. Meyer, and Iryna Gurevych: 'Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Translation Applications', *Translation: Computation, Corpora, Cognition: Special Issue "Language Technology for a Multilingual Europe"* 3 (1): 87–118, July 2013.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Kobi Reiter, Michael Skinner, Marcus Sammer, and Jeff A. Bilmes: 'Panlingual Lexical Translation via Probabilistic Inference', *Artificial Intelligence* 174 (9–10): 619–637, June 2010.

John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano: 'Integrating WordNet and Wiktionary with lemon', in Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann (Eds.): *Linked Data in Linguistics*, pp. 25–34, Berlin/Heidelberg: Springer, 2012.

John McCrae, Dennis Spohr, and Philipp Cimiano: 'Linking Lexical Resources and Ontologies on the Semantic Web with Lemon', in Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan (Eds.): *The Semantic Web: Research and Applications: 8th Extended Semantic Web Conference* (ESWC), Lecture Notes in Computer Science Vol. 6643, pp. 245–259, Berlin/Heidelberg: Springer, June 2011.

Clifton McFate and Kenneth Forbus: 'NULEX: An Open-License Broad Coverage Lexicon', in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (ACL/HLT), pp. 363–367, Portland, OR, USA, June 2011.

Olena Medelyan, Catherine Legg, David Milne, and Ian H. Witten: 'Mining Meaning from Wikipedia', *International Journal of Human-Computer Studies* 67 (9): 716–754, September 2009.

Julie Medero and Mari Ostendorf: 'Analysis of Vocabulary Difficulty Using Wiktionary', in: *Proceedings of the ISCA International Workshop on Speech and Language Technology in Education* (SLaTE), Warwickshire, UK, September 2009.

Julie Medero and Mari Ostendorf: 'Identifying Targets for Syntactic Simplification', in: *Proceedings of the ISCA International Workshop on Speech and Language Technology in Education* (SLaTE), Venice, Italy, September 2011.

Gerard Meijssen: 'The Philosophy behind OmegaWiki and the Visions for the Future', in: Bergenholtz et al. (2009), pp. 91–98.

Luca Melchior: 'Halbkollaborativität und Online-Lexikographie. Ansätze und Überlegungen zu Wörterbuchredaktion und Wörterbuchforschung am Beispiel LEO Deutsch-Italienisch', *Lexicographica* 28 (1): 337–372, December 2012.

Gerard de Melo and Gerhard Weikum: 'Language as a Foundation of the Semantic Web', in: *Proceedings of the Poster And Demonstration Session of the 7th International Semantic Web Conference* (ISWC), CEUR Workshop Proceedings Vol. 401, Karlsruhe, Germany, October 2008.

Gerard de Melo and Gerhard Weikum: 'Towards a Universal Wordnet by Learning from Combined Evidence', in: *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (CIKM), pp. 513–522, Hong Kong, China, November 2009.

Gerard de Melo and Gerhard Weikum: 'Providing Multilingual, Multimodal Answers to Lexical Database Queries', in: *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC), pp. 348–355, Valletta, Malta, May 2010.

Christian M. Meyer and Iryna Gurevych: 'How Web Communities Analyze Human Language: Word Senses in Wiktionary', in: *Proceedings of the Second Web Science Conference* (WebSci), Raleigh, NC, USA, April 2010a.

Christian M. Meyer and Iryna Gurevych: 'Worth its Weight in Gold or Yet Another Resource – A Comparative Study of Wiktionary, OpenThesaurus and GermaNet', in Alexander Gelbukh (Ed.): *Computational Linguistics and Intelligent Text Processing: 11th International Conference* (CICLing), Lecture Notes in Computer Science Vol. 6008, pp. 38–49, Berlin/Heidelberg: Springer, March 2010b.

Christian M. Meyer and Iryna Gurevych: 'What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage', in: *Proceedings of the 5th International Joint Conference on Natural Language Processing* (IJCNLP), pp. 883–892, Chiang Mai, Thailand, November 2011.

Christian M. Meyer and Iryna Gurevych: 'OntoWiktionary – Constructing an Ontology from the Collaborative Online Dictionary Wiktionary', in Maria Teresa Pazienza and Armando Stellato (Eds.): *Semi-Automatic Ontology Development: Processes and Resources*, chapter 6, pp. 131–161, Hershey, PA: IGI Global, February 2012b.

Christian M. Meyer and Iryna Gurevych: 'To Exhibit is not to Loiter: A Multilingual, Sense-Disambiguated Wiktionary for Measuring Verb Similarity', in: *Proceedings of the 24th International Conference on Computational Linguistics* (COLING), Vol. 4, pp. 1763–1780, Mumbai, India, December 2012c.

Christian M. Meyer and Iryna Gurevych: 'Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography', in: Granger and Paquot (2012), chapter 13, pp. 259–291.

Christian M. Meyer and Iryna Gurevych: 'Der lexikographische Prozess im deutschen Wiktionary', in Vera Hildenbrandt (Ed.): *Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie"*, OPAL – Online publizierte Arbeiten zur Linguistik, Mannheim: Institut für Deutsche Sprache, 2013. *(to appear)*.

Rada Mihalcea: 'Using Wikipedia for Automatic Word Sense Disambiguation', in: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL), pp. 196–203, Rochester, NY, USA, April 2007.

Rada Mihalcea and Dan Moldovan: 'eXtended WordNet: Progress Report', in: *Proceedings of the NAACL Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations'*, pp. 95–100, Pittsburgh, PA, USA, June 2001.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy: 'SemEval-2010 Task 2: Cross-Lingual Lexical Substitution', in: *Proceedings of the 5th International Workshop on Semantic Evaluations* (SemEval), pp. 9–14, Uppsala, Sweden, July 2010.

David Milne and Ian H. Witten: 'An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links', in: *Proceedings of the AAAI Workshop 'Wikipedia and Artificial Intelligence: An Evolving Synergy'*, pp. 25–30, Chicago, IL, USA, July 2008.

Dan Moldovan and Adrian Novischi: 'Word sense disambiguation of WordNet glosses', *Computer Speech and Language* 18 (3): 301–317, July 2004.

Rosamund Moon: 'The Analysis of Meaning', in John M. Sinclair (Ed.): *Looking Up: An Account of the COBUILD Project in Lexical Computing*, chapter 4, pp. 86–103, London: Collins, 1987.

Karlheinz Mörth, Thierry Declerck, Piroska Lendvai, and Tamás Váradi: 'Accessing Multilingual Data on the Web for the Semantic Annotation of Cultural Heritage Texts', in: *Proceedings of the 2nd International Workshop on the Multilingual Semantic Web* (MSW), pp. 80–85, Bonn, Germany, October 2011.

Christopher Moseley (Ed.): *Atlas of the World's Languages in Danger*, Paris, UNESCO Publishing, 3rd edition, 2010. Online: `http://www.unesco.org/culture/en/endangeredlanguages/atlas`.

Christof Müller and Iryna Gurevych: 'Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval', in Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras (Eds.): *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum* (CLEF), Lecture Notes in Computer Science Vol. 5706, pp. 219–226, Berlin/Heidelberg: Springer, 2009.

Carolin Müller-Spitzer: 'Ordnende Betrachtungen zu elektronischen Wörterbüchern und lexikographischen Prozessen', *Lexicographica* 19: 140–168, August 2004.

Carolin Müller-Spitzer: 'Der Aufbau einer maßgeschneiderten XML-basierten Modellierung für ein Wörterbuchnetz', in Annette Klosa and Carolin Müller-Spitzer (Eds.): *Datenmodellierung für Internetwörterbücher: 1. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie"*, OPAL – Online publizierte Arbeiten zur Linguistik Vol. 2/2011, pp. 37–51, Mannheim: Institut für Deutsche Sprache, 2011.

Daniel Naber: 'OpenThesaurus: ein offenes deutsches Wortnetz', in Bernhard Fisseni, Hans-Christian Schmitz, Bernhard Schröder, and Petra Wagner (Eds.): *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung*, Sprache, Sprechen, Computer Vol. 8, pp. 422–433, Frankfurt: Peter Lang, March 2005.

Vivi Nastase, Michael Strube, Benjamin Börschinger, Cäcilia Zirn, and Anas Elghafari: 'WikiNet: A Very Large Scale Multi-Lingual Concept Network', in: *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC), pp. 1015–1022, Valetta, Malta, May 2010.

I. S. P. Nation: 'How large a vocabulary is needed for reading and listening?', *The Canadian Modern Language Review / La revue canadienne des langues vivantes* 63 (1): 59–82, September 2006.

Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Ivy Kuo, Pierre Magistry, and Chu-Ren Huang: 'Wiktionary and NLP: Improving synonymy networks', in: *Proceedings of the ACL/IJCNLP Workshop 'The People's Web Meets NLP: Collaboratively Constructed Semantic Resources'*, pp. 19–27, Singapore, August 2009.

Roberto Navigli: 'Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance', in: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (COLING/ACL), pp. 105–112, Sydney, Australia, July 2006.

Roberto Navigli: 'Word sense disambiguation: A survey', *ACM Computing Surveys* 41 (2): 1–69, February 2009.

Roberto Navigli and Simone Paolo Ponzetto: 'BabelNet: Building a Very Large Multilingual Semantic Network', in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 216–225, Uppsala, Sweden, July 2010.

Kiem-Hieu Nguyen and Cheol-Young Ock: 'Using Wiktionary to Improve Lexical Disambiguation in Multiple Languages', in Alexander Gelbukh (Ed.): *Computational Linguistics and Intelligent Text Processing: 13th International Conference, Part I* (CICLing), Lecture Notes in Computer Science Vol. 7181, pp. 238–248, Berlin/Heidelberg: Springer, March 2012.

Sandro Nielsen: 'The Evaluation of the Outside Matter in Dictionary Reviews', *Lexikos* 19: 207–224, October 2009.

Elisabeth Niemann and Iryna Gurevych: 'The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet', in: *Proceedings of the Ninth International Conference on Computational Semantics* (IWCS), pp. 205–214, Oxford, UK, January 2011.

Jyrki Niemi, Krister Lindén, and Mirka Hyvärinen: 'Using a Bilingual Resource to Add Synonyms to a Wordnet: FinnWordNet and Wikipedia as an Example', in: *Proceedings of the 6th Global WordNet Conference*, pp. 227–231, Matsue, Japan, January 2012.

Stephan Oepen, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger: 'Towards Hybrid Quality-Oriented Machine Translation: On Linguistics and Probabilities in MT', in: *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI), pp. 144–153, Skövde, Sweden, September 2007.

Charles K. Ogden: *Basic English: A General Introduction with Rules and Grammar*, London: Kegan Paul, Trench, Trubner & Co., 7th edition, 1938.

Mari Broman Olsen, Bonnie J. Dorr, and Scott Thomas: 'Enhancing Automatic Acquisition of Thematic Structure in a Large-Scale Lexicon for Mandarin Chinese', in David Farwell, Laurie Gerber, and Eduard Hovy (Eds.): *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas* (AMTA), Lecture Notes in Computer Science Vol. 1529, pp. 41–50, Berlin/Heidelberg: Springer, October 1998.

Noel Edward Osselton: 'The History of Academic Dictionary Criticism With Reference to Major Dictionaries', in: Hausmann et al. (1989), chapter 27, pp. 225–229.

Pim Otte and Francis M. Tyers: 'Rapid rule-based machine translation between Dutch and Afrikaans', in: *Proceedings of the 15th Conference of the European Association for Machine Translation* (EAMT), pp. 153–160, Leuven, Belgium, May 2011.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum: 'Making fine-grained and coarse-grained sense distinctions, both manually and automatically', *Natural Language Engineering* 13 (2): 137–163, June 2007.

Patrick Pantel and Dekang Lin: 'Discovering Word Senses from Text', in: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 613–619, Edmonton, AB, Canada, July 2002.

Patrick Pantel and Marco Pennacchiotti: 'Automatically Harvesting and Ontologizing Semantic Relations', in Paul Buitelaar and Philipp Cimiano (Eds.): *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pp. 171–198, Amsterdam: IOS Press, 2008.

Rebecca J. Passonneau: 'Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation', in: *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC), pp. 831–836, Genoa, Italy, May 2006.

Darrell J. Penta: 'The Wiki-fication of the Dictionary: Defining Lexicography in the Digital Age', in: *Proceedings of the Seventh Media in Transition Conference* (MiT), Cambridge, MA, USA, May 2011.

Praharshana Perera and René Witte: 'A Self-Learning Context-Aware Lemmatizer for German', in: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP), pp. 636–643, Vancouver, BC, Canada, October 2005.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi: 'MultiWordNet: Developing an aligned multilingual database', in: *Proceedings of the First International WordNet Conference* (GWC), pp. 293–302, Mysore, India, January 2002.

Simone Paolo Ponzetto and Roberto Navigli: 'Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia', in: *Proceedings of the 21st International Joint Conference on Artificial Intelligence* (IJCAI), pp. 2083–2088, Pasadena, CA, USA, July 2009.

Simone Paolo Ponzetto and Roberto Navigli: 'Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems', in: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 1522–1531, Uppsala, Sweden, July 2010.

Simone Paolo Ponzetto and Michael Strube: 'Deriving a Large-Scale Taxonomy from Wikipedia', in: *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, pp. 1440–1445, Vancouver, BC, Canada, July 2007.

Daniel J. Prinsloo: 'Internet dictionaries for African languages', *Lexicographica* 26: 183–194, December 2010.

Uwe Quasthoff, Matthias Richter, and Christian Biemann: 'Corpus Portal for Search in Monolingual Corpora', in: *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC), pp. 1799–1802, Genoa, Italy, May 2006.

Valeria Quochi, Monica Monachini, Riccardo Del Gratta, and Nicoletta Calzolari: 'A lexicon for biology and bioinformatics: the BOOTStrep experience', in: *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC), pp. 2285–2292, Marrakech, Morocco, May 2008.

RDF: *Resource Description Framework (RDF): Concepts and Abstract Syntax*, W3C Recommendation 10 February 2004, World Wide Web Consortium, Cambridge, MA, February 2004.

Stephen L. Reed and Douglas B. Lenat: 'Mapping Ontologies into Cyc', in: *Proceedings of the AAAI Workshop 'Ontologies and the Semantic Web'*, pp. 1–6, Edmonton, AB, Canada, August 2002.

Alexander E. Richman and Patrick Schone: 'Mining Wiki Resources for Multilingual Named Entity Recognition', in: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (ACL/HLT), pp. 1–9, Columbus, OH, USA, June 2008.

David A. Ricks: *Blunders in International Business*, Malden: Blackwell Publishing, 2006.

Merritt Ruhlen: *A Guide to the World's Languages*, Vol. 1: Classification, Stanford, CA: Stanford University Press, 1987.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells: 'Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets', in Piotr S. Szczepaniak, Janusz Kacprzyk, and Adam Niewiadomski (Eds.): *Advances in Web Intelligence: Third International Atlantic Web Intelligence Conference* (AWIC), Lecture Notes in Computer Science Vol. 3528, pp. 380–386, Berlin/Heidelberg: Springer, May 2005.

Michael Rundell: ''It works in practice but will it work in theory?' The uneasy relationship between lexicography and matters theoretical', in: *Proceedings of the 15th EURALEX International Congress*, pp. 47–92, Oslo, Norway, August 2012.

Robert C. Russell: *Index*, United States Patent 1,261,167, (filed October 25, 1917) April 2, 1918.

Franck Sajous, Emmanuel Navarro, Bruno Gaume, Laurent Prévot, and Yannick Chudy: 'Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary', in Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir (Eds.): *Advances in Natural Language Processing: Proceedings of the 7th International Conference on NLP* (IceTAL), Lecture Notes in Artificial Intelligence Vol. 6233, pp. 332–344, Berlin/Heidelberg: Springer, August 2010.

Tim Schlippe, Sebastian Ochs, and Tanja Schultz: 'Wiktionary as a Source for Automatic Pronunciation Extraction', in: *Proceedings of the 11th Annual Conference of the International Speech Communication Association* (Interspeech), pp. 2290–2293, Makuhari, Japan, September 2010.

Gilles-Maurice de Schryver: 'Lexicographers' Dreams in the Electronic-Dictionary Age', *International Journal of Lexicography* 16 (2): 143–199, June 2003.

Gilles-Maurice de Schryver and Daan J. Prinsloo: 'Dictionary-Making Process with 'Simultaneous Feedback' from the Target Users to the Compilers.', in: *Proceedings of the 9th EURALEX International Congress*, Vol. 1, pp. 197–209, Stuttgart, Germany, August 2000.

Dietrich Schüller: 'Safeguarding the Documentary Heritage of Cultural and Linguistic Diversity', *Language Archive Newsletter* 1 (3): 9–10, July 2004.

Hinrich Schütze: 'Automatic Word Sense Discrimination', *Computational Linguistics: Special Issue on Word Sense Disambiguation* 24 (1): 97–123, March 1998.

Gilles Sérasset: 'Recent Trends of Electronic Dictionary Research and Development in Europe', *Technical memorandum*, Tokyo: Japan Electronic Dictionary Research Institute, 1993.

Gilles Sérasset: 'Dbnary: Wiktionary as a LMF based Multilingual RDF network', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), pp. 2466–2472, Istanbul, Turkey, May 2012.

Lei Shi and Rada Mihalcea: 'Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing', in Alexander Gelbukh (Ed.): *Computational Linguistics and Intelligent Text Processing: 6th International Conference* (CICLing), Lecture Notes in Computer Science Vol. 3406, pp. 100–111, Berlin/Heidelberg: Springer, February 2005.

Push Singh: 'The public acquisition of commonsense knowledge', in: *Proceedings of the AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, pp. 47–52, Palo Alto, CA, USA, March 2002.

SKOS: *Simple Knowledge Organization System Primer*, W3C Working Draft 15 June 2009, World Wide Web Consortium, Cambridge, MA, June 2009.

Claudia Soria, Monica Monachini, and Piek Vossen: 'Wordnet-LMF: Fleshing out a Standardized Format for Wordnet Interoperability', in: *Proceedings of the 2009 International Workshop on Intercultural Collaboration*, pp. 139–146, Palo Alto, CA, USA, February 2009.

Christian Stegbauer: *Wikipedia: Das Rätsel der Kooperation*, Wiesbaden: VS Verlag für Sozialwissenschaften, 2009.

Angelika Storrer: 'Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher', in Herbert Ernst Wiegand (Ed.): *Wörterbücher in der Diskussion III*, Lexicographica Series Maior Vol. 84, pp. 107–135, Tübingen: Niemeyer, 1998.

Angelika Storrer: 'Wörterbücher im sozialen Netz: Konzepte – Prozesse – Perspektiven', Talk given at the Workshop 'Künftige Standards wissenschaftlicher Lexikographie', March 2012, Online: http://www.studiger.tu-dortmund.de/images/Storrer-lexikographie-12.pdf.

Angelika Storrer: 'Representing dictionaries in hypertextual form', in: Gouws et al. (2013). *(to appear)*.

Angelika Storrer and Katrin Freese: 'Wörterbücher im Internet', *Deutsche Sprache* 24 (2): 97–153, 1996.

Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser: 'Information Quality Work Organization in Wikipedia', *Journal of the American Society for Information Science and Technology* 59 (6): 983–1001, April 2008.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum: 'YAGO: A Core of Semantic Knowledge: Unifying WordNet and Wikipedia', in: *Proceedings of the 16th International World Wide Web Conference* (WWW), pp. 697–706, Banff, AB, Canada, May 2007.

James Surowiecki: *The Wisdom of Crowds*, New York, NY: Anchor Books, August 2005.

TEI P5: *Guidelines for Electronic Text Encoding and Interchange*, Version 2.3.0, TEI Consortium, Charlottesville, VA, 2013.

Antonio Toral, Stefania Bracale, Monica Monachini, and Claudia Soria: 'Rejuvenating the Italian WordNet: Upgrading, Standardising, Extending', in: *Proceedings of the 5th Global WordNet Conference* (GWC), Mumbai, India, February 2010.

Antonio Toral, Óscar Ferrández, Eneko Agirre, and Rafael Muñoz: 'A study on Linking Wikipedia categories to Wordnet synsets using text similarity', in: *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing* (RANLP), pp. 449–454, Borovets, Bulgaria, September 2009.

Antonio Toral, Rafael Muñoz, and Monica Monachini: 'Named Entity WordNet', in: *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC), pp. 741–747, Marrakech, Morocco, May 2008.

Takashi Tsunakawa and Hiroyuki Kaji: 'Augmenting a Bilingual Lexicon with Information for Word Translation Disambiguation', in: *Proceedings of the Eighth COLING Workshop on 'Asian Language Resources'*, pp. 30–37, Beijing, China, August 2010.

Dan Tufiș, Dan Cristea, and Sofia Stamou: 'BalkaNet: Aims, Methods, Results and Perspectives. A General Overview', *Romanian Journal of Information Science and Technology* 7 (1–2): 9–43, 2004.

Tony Veale, Nuno Seco, and Jer Hayes: 'Creative Discovery in Lexical Ontologies', in: *Proceedings of the 20th International Conference on Computational Linguistics* (COLING), pp. 1333–1338, Geneva, Switzerland, August 2004.

Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave: 'Studying Cooperation and Conflict between Authors with History Flow Visualizations', in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 575–582, Vienna, Austria, April 2004.

Piek Vossen: 'Introduction to EuroWordNet', *Computers and the Humanities* 32 (2–3): 73–89, March 1998.

Géraldine Walther, Benoît Sagot, and Karën Fort: 'Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish', in: *Proceedings of the 29th International Conference on Lexis and Grammar*, Belgrade, Serbia, September 2010.

Tong Wang and Graeme Hirst: 'Exploring Patterns in Dictionary Definitions for Synonym Extraction', *Natural Language Engineering* 18 (3): 313–342, July 2012.

Michael West: *A General Service List of English Words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*, London: Longman, Green & Co., 1953.

Herbert Ernst Wiegand: 'Aspekte der Makrostruktur im Allgemeinen einsprachigen Wörterbuch: Alphabetische Anordnungsformen und ihre Probleme', in: Hausmann et al. (1989), chapter 38, pp. 371–409.

Herbert Ernst Wiegand: 'Altes und Neues zur Mediostruktur in Printwörterbüchern', *Lexicographica* 18: 168–252, March 2003.

Herbert Ernst Wiegand: 'Angaben, funktionale Angabezusätze, Angabetexte, Angabestrukturen, Strukturanzeiger, Kommentare und mehr', *Lexicographica* 21: 202–237, December 2006.

Herbert Ernst Wiegand, Michael Beißwenger, Rufus Hjalmar Gouws, Matthias Kammerer, Angelika Storrer, and Werner Wolski (Eds.): *Wörterbuch zur Lexikographie und Wörterbuchforschung / Dictionary of Lexicography and Dictionary Research*, Vol. 1 (Systematische Einführung / Systematic Introduction, A–C), Berlin/New York: de Gruyter, 2010.

Dennis M. Wilkinson and Bernardo A. Huberman: 'Cooperation and Quality in Wikipedia', in: *Proceedings of the 2007 International Symposium on Wikis* (WikiSym), pp. 157–164, Montreal, QC, Canada, October 2007.

Elisabeth Wolf and Iryna Gurevych: 'Aligning Sense Inventories in Wikipedia and WordNet', in: *Proceedings of the First Workshop on Automated Knowledge Base Construction* (AKBC), pp. 24–28, Grenoble, France, May 2010.

Dongqiang Yang and David M. W. Powers: 'Verb Similarity on the Taxonomy of WordNet', in: *Proceedings of the Third International WordNet Conference* (GWC), pp. 121–128, Jeju Island, Korea, January 2006.

Torsten Zesch, Christof Müller, and Iryna Gurevych: 'Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary', in: *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC), pp. 1646–1652, Marrakech, Morocco, May 2008a.

Torsten Zesch, Christof Müller, and Iryna Gurevych: 'Using Wiktionary for Computing Semantic Relatedness', in: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pp. 861–867, Chicago, IL, USA, July 2008b.

Илья Г. Гельфейнбейн, Артем В. Гончарук, Влад П. Лехельт, Антон А. Липатов, and Виктор В. Шило: 'Автоматический перевод семантической сети WordNet на русский язык', in: *Труды Международного семинара Диалог по компьютерной лингвистике и её приложениям*, Протвино, Россия, 2003. [Ilya G. Gelfenbeyn, Artem B. Goncharuk, Vladislav P. Lehelt, Anton A. Lipatov, and Victor B. Shilo: Automatic translation of WordNet's semantic network into Russian, in: Proceedings of the International Dialog Conference, Protvino, Russia, 2003].

Андрей А. Крижановский: 'Количественный анализ лексики английского языка в викисловарях и Wordnet', *Труды СПИИРАН* 19: 87–101, 2011. [Andrew A. Krizhanovsky: 'A quantitative analysis of the English lexicon in Wiktionaries and Wordnet,' Trudy SPIIRAN 19: 87–101, 2011].

Э. Штейнфельдт: *Частотный словарь современного русского литературного языка*, Москва: Прогресс, 1963. [E. Steinfeldt: Frequency dictionary for the modern Russian literary language, Moscow: Progress, 1963].

# Appendix

## A Software and Data

We have created, extended, or contributed to the following software projects, resources, and datasets.

### A.1 Open source software

*JWKTL (Java-based Wiktionary Library)* is an application programming interface for Wiktionary. The software has initially been created by Christof Müller, Lizhen Qu, and Torsten Zesch (cf. Zesch et al., 2008a). Together with Yevgen Chebotar, we have extended the software by developing a novel adapter to the *Wikokit* library[80] for parsing the Russian Wiktionary and by adding support for extracting pronunciations, inflected word forms, Wikisaurus, transliterations. We have continually adapted the software to changes in Wiktionary yielding 18 releases between version 0.13.1 and version 1.0.0 (see also section 5.3).

▶ `http://www.ukp.tu-darmstadt.de/software/jwktl/`

▶ `http://code.google.com/p/jwktl/`

*DKPro Statistics* is a collection of open-licensed statistical tools, including correlation and inter-rater agreement methods. We have implemented eleven commonly used inter-rater agreement measures, and we have prepared the open source release of this statistics package. The remaining classes of DKPro Statistics have been contributed by Torsten Zesch. We utilized the software library for our dataset analyses in the chapters 5, 6, and 8.

▶ `http://code.google.com/p/dkpro-statistics/`

### A.2 Lexical resources

*OntoWiktionary* is a multilingual lexical ontology based on information extracted from Wiktionary (see chapter 5). OntoWiktionary consists of concepts, lexicalizations, and ontological

[80] `http://code.google.com/p/wikokit/` (2 August 2013)

relations in English, German, and Russian, as well as translations of the concepts into over 1,000 languages. The data is available in a simple XML format and as part of UBY (see below). In addition to that, we provide an XSLT-based user interface for browsing the data. Yevgen Chebotar has contributed to this resource.

▶ http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/ontowiktionary/

▶ http://www.christian-meyer.org/research/publications/igi-saod2012/data/

*Sense alignment between Wiktionary and WordNet:* an automatically computed alignment of the English Wiktionary (as of April 3, 2010) and WordNet 3.0 (2006) at the word sense level (see chapter 6). The data is available as a list of identifier pairs referencing the word senses/synsets encoded by the two dictionaries. Christian Kirschner and Elisabeth Niemann have contributed to this resource.

▶ http://www.ukp.tu-darmstadt.de/data/lexical-resources/wordnet-wiktionary-alignment/

▶ http://www.christian-meyer.org/research/publications/ijcnlp2011/data/

*UBY* is a large-scale unified lexical-semantic resource based on the Lexical Markup Framework (LMF). The development of UBY has been a shared effort by Yevgen Chebotar, Judith Eckle-Kohler, Richard Eckart de Castilho, Iryna Gurevych, Than-Le Ha, Silvana Hartmann, Zijad Maksuti, Michael Matuschek, Christian M. Meyer, Tri Duc Nghiem, and Christian Wirth. We have contributed to the development of the lexicon model UBY-LMF, and we have implemented the conversion of Wiktionary and OntoWiktionary to this lexicon model (see chapter 7).

▶ http://www.ukp.tu-darmstadt.de/uby/

▶ http://code.google.com/p/uby/

## A.3 Evaluation data

*Relation disambiguation gold standards:* four datasets with 1,117 (en:en), 1,119 (de:de), 614 (en:de), and 656 (de:en) human judgments for evaluating the disambiguation of the targets of semantic relations and translations in the English and German Wiktionary editions (see section 5.4). Daniel Bär, Yevgen Chebotar, Christian Kirschner, Bastian Laur, and Elisabeth Niemann have contributed to these datasets.

▶ http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/disambiguation/

▶ http://www.christian-meyer.org/research/publications/coling2012/data/

*Concept formation gold standards:* two datasets with human judgments on the consistency of 100 English and 100 German concepts of OntoWiktionary (see section 5.5). Yevgen Chebotar has contributed to these datasets.

▶ http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/ontowiktionary/

▶ http://www.christian-meyer.org/research/publications/igi-saod2012/data/

*Word sense alignment gold standard:* dataset of 2,423 human judgments on the correspondence of Wiktionary word senses and WordNet 3.0 synsets for evaluating word sense alignment

methods (see chapter 6). Yevgen Chebotar, Elisabeth Niemann, and the students of the 2010 *Lexical-Semantic Methods for Language Understanding* course have contributed to this dataset.
▶ http://www.ukp.tu-darmstadt.de/data/lexical-resources/wordnet-wiktionary-alignment/
▶ http://www.christian-meyer.org/research/publications/ijcnlp2011/data/

*Verb similarity gold standards:* three datasets judging the similarity of German–German, English–German, and German–English verb pairs (see section 8.3). The datasets are translations of the 130 English verb similarity dataset created by Yang and Powers (2006). The datasets consist of verb pairs with a corresponding similarity score between 0 and 4.
▶ http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/disambiguation/
▶ http://www.christian-meyer.org/research/publications/coling2012/data/

*Marketing blunder dataset:* collection of 45 cross-lingual marketing blunders taken from Ricks (2006) and from the homepage of the British translation agency *Kwintessential*.[81] The dataset consists of the product or company name yielding the marketing blunder, a classification of the blunder type, and an explanation of the reasons of the marketing blunder. In addition to that, 1,494 clues for such blunders have been annotated as being relevant or irrelevant (see section 8.4). Richard Steuer has contributed to this dataset.
▶ http://www.ukp.tu-darmstadt.de/data/marketing-blunders/
▶ http://www.christian-meyer.org/research/publications/dissertation/data/

## A.4 Supplementary data

*Quantitative analyses:* a collection of tables and figures providing quantitative statistics on Wiktionary in comparison to a number of other dictionaries (see chapter 3 and chapter 4).
▶ http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/comparative-study/
▶ http://www.christian-meyer.org/research/publications/dissertation/data/

*Pragmatic label groups:* a classification of 2,969 pragmatic labels from the English, German, and Russian Wiktionary editions into 75 label groups (see section 4.6).
▶ http://www.ukp.tu-darmstadt.de/data/lexical-resources/wiktionary/pragmatic-labels/
▶ http://www.christian-meyer.org/research/publications/dissertation/data/

# B Annotation Guidelines

In the following sections, we reproduce the annotation guidelines given to the human raters of the annotation studies carried out for this thesis. The layout of the original guideline documents has been reformatted. Note that – depending on the annotation study – the annotators received additional explanations and training.

[81] http://kwintessential.co.uk/cultural-services/articles/crosscultural-marketing.html; http://kwintessential.co.uk/cultural-services/articles/results-of-poor-cross-cultural-awareness.html; http://kwintessential.co.uk/cultural-services/articles/crosscultural-blunders.html (30 May 2013)

## B.1 Relation disambiguation gold standards

**Introduction**. Many dictionaries contain links to other words. You might find a link to the adjective *hard* within the lexical entry *difficult*, as these two words can be considered synonymous. While it is inherently clear for humans that the relation between *hard* and *difficult* refers to something that requires a lot of skills and effort (e.g., a *difficult*/*hard* problem), this is a great challenge for computer programs. The reason is that there are other meanings for the word *hard* such as being tough (e.g., only a few people are hard enough) that are not synonymous to *difficult* (∗ only a few people are *difficult* enough).

The same applies to translations. There is, for example, the German translation *schwierig* for the English *hard*, however only in its sense of requiring skills/effort (e.g., ein *schwieriges* Problem [a hard problem]). The meaning of being tough cannot be translated to *schwierig* (∗ nur wenige Menschen sind schwierig genug [∗ only a few people are difficult enough]). The goal of this annotation study is the creation of a gold standard dataset of sense disambiguated lexical relations (more precisely: synonyms, antonyms, hypo-/hypernyms, mero-/holonyms) and translations.

**Task description**. Along with this annotation guidebook, you'll receive one of four spreadsheets that can be opened in Microsoft Excel or OpenOffice (please contact the authors in case of any problems with opening the file). The spreadsheet contains a list of

- lexical relations between English words (dataset en–en),
- lexical relations between German words (dataset de–de),
- lexical translations from English to German (dataset en–de), or
- lexical translations from German to English (dataset de–en).

Each relation/translation is shown in a separate row. For each relation/translation, the source and target word is given – i.e., the two endpoints of the relation/translation. The meaning of the source and the target is described by a short textual description (a gloss). Additionally, the type of relationship (synonymy, antonymy, etc.) or the language of the translation is provided. In case the target word has multiple word senses, all of them are listed in a separate row.

Your task is to judge the appropriateness of each relation/translation. That is, we ask you to fill the "iscorrect" column for each row with a

0 if the word senses of the source and target cannot be connected by this type of relation/translation, because they represent a wrong meaning in this context, or a

1 if the word senses of the source and target are appropriate for this kind of relation/translation.

**Examples:**

| source definition | source | is correct | target | target definitions | relation type | comment |
|---|---|---|---|---|---|---|
| The quality of a confident character not to be afraid or intimidated easily but without being incautious or inconsiderate | courage | 1 | cowardice | Lack of courage | antonymy | clearly opposite meanings |

| source definition | source | is correct | target | target definition | relation type | comment |
|---|---|---|---|---|---|---|
| The highest point of something | apex | 1 | acme | The top or highest point; pinnacle; culmination | synonymy | |
| The highest point of something | apex | 0 | acme | (medicine) The crisis or height of a disease | synonymy | |
| The highest point of something | apex | 0 | acme | Mature age; full bloom of life | synonymy | |

| source definition | source | is correct | target | target definition | relation type | comment |
|---|---|---|---|---|---|---|
| Aubergine, the edible fruit of the "Solanum melongena" | eggplant | 0 | aubergine | (British) an Asian plant, "Solanum melongena", cultivated for its edible purple, green, or white ovoid fruit | synonymy | fruit vs. plant |
| Aubergine, the edible fruit of the "Solanum melongena" | eggplant | 1 | aubergine | (British) the fruit of this plant, eaten as a vegetable | synonymy | both are fruits |
| Aubergine, the edible fruit of the "Solanum melongena" | eggplant | 0 | aubergine | a dark purple colour; eggplant | synonymy | fruit vs. color |

| source definition | source | is correct | target | target definition | language | comment |
|---|---|---|---|---|---|---|
| Obtainable without payment | free | 1 | umsonst | ohne Gegenleistung, ohne Geld bezahlen zu müssen | DE | without payment |
| Obtainable without payment | free | 0 | umsonst | ohne Erfolg, vergebens | DE | not successful |
| Obtainable without payment | free | 0 | umsonst | verneint: nicht grundlos, nicht ohne einen Zweck | DE | without purpose |

**Final remarks:**

- Do not imitate an algorithm, rely on your intuition.
- For each relation/translation, multiple "1"s are possible, i.e. more than one word sense of the target might be appropriate.
- Sometimes no "1" is suitable for a relation/translation, i.e. the correct word sense is not listed or the source and the target are not related.
- It is possible that the type of relation is not "suitable" – e.g., too course-grained synonyms (considering a *motorbike* and a *moped* synonymous) or too general hypernyms (an *apple tree* is a *plant* rather than a *tree*) – this kind of error should not be taken into account.

Just concentrate on finding the correct word senses for the relations/translation rather than judging their consistency.
- Sometimes the definitions are trimmed or normalized by our software library and thus do not clearly describe the word sense. It is allowed to consult Wiktionary for the original definition.
- As the task gets monotonous at some point, you should consider making breaks during the annotation process. Depending on your personal speed, it should take about 2–3 hours.
- Please do not change the order of the rows/cols or their IDs. We'll need this information for further processing. Of course, you are allowed to change the size of the columns to exactly fit on your screen.
- If you feel like explaining a certain decision, you can use the comment column. This is optional; you do not have to leave a comment for every pair!

After you finished your annotations, please fill out the last row with the annotation time needed in minutes. There is no correct solution for this task, so please, always judge the sense pairs based on your own understanding. We will measure the inter-annotator agreement of the submissions to find out, how often the participants agree or disagree.

## B.2 Concept formation gold standards

**Goal.** Concepts are the basic building blocks for ontologies. Each concept should denote a single entity of world that is modeled by the ontology. The term "Entity" includes real world objects, abstract ideas, processes, states, etc. Besides a textual description, a concept can be represented by lexicalizations, i.e. certain terms or expressions that directly refer to the concept. The concept ›DOG‹ could, e.g., be modeled for representing all instances that are denoted by the word *dog* in our world. The noun *dog* (in the animal sense) thus serves as a lexicalization of ›DOG‹, which might also be represented by a second lexicalization using the noun *hound*.

The goal of this annotation study is to validate the consistency of semi-automatically learned concepts. The concepts are represented by different lexicalizations that are explained by a short textual definition. The annotation study is intended to analyze the overall quality of the creation approach and if errors rather occur in the lexicalizations or their definitions (i.e. their meaning).

**Setup**:

- The dataset is given as an Excel sheet.
- The data is organized in sections.
  - Each section starts with lexicalizations in the first column and their textual definition in the second column – one pair of lexicalization and definition per row.

  - After the list of lexicalizations/definitions, you're asked to answer the question "Is the above concept consistent?" (see further explanation below). Please type your answer in the first column (before the question).
  - You can also leave a comment for each lexicalization/definition in the third column.
- Concepts whose lexicalizations AND definitions are consistent, i.e. they all belong to the same entity, should be marked with "1" (see Example #1).
- Concepts whose lexicalizations but NOT definitions are consistent, i.e. there is a meaning for all lexicalizations that belong to the same entity (although at least one of them is associated with the wrong definition), should be marked with "2" (see Example #3 – there is another meaning for the word *bass* that is commonly used in the English language).
- Inconsistent concepts should be marked with "0" (see Example #2).

**Please note**:

- Do not imitate an algorithm!
- Sometimes lexicalizations from different parts of speech are put together. You should accept ("1" or "2") these concepts if the meaning of these lexicalizations belong to the same entity (see Example #4).
- Reject concepts ("0") that have at least one lexicalization that is very broad or very narrow. In Example #2, the *bass* is a certain type of a *singer*, so one would expect two different concepts here.
- You should ignore very subtle differences in the lexicalizations. The concept "statement that does not conform to the truth" can, e.g., be lexicalized as *lie* and *misrepresentation*. A lie, for instance, usually infers deceiving someone, while a misconception can be simply due to ignorance. This kind of subtle difference should be accepted ("1" or "2").
- Sometimes the textual definitions are trimmed or normalized by our extraction method. Additionally, there are some special characters or format commands within the definitions. This should be ignored for the judgment.
- You are allowed to use any additional resource for grounding your judgment, including dictionaries, lexicons, the Web, and particularly Wiktionary (`http://www.wiktionary.org`), which was used as a source for the textual definitions.
- Synsets within the Princeton WordNet are similar to the concepts that are to be judged in our study (besides the restriction to one part of speech). Experience in the work with WordNet can thus be used to judge the concepts. Caveat: Not every lexicalization in our study is part of Word-Net or is always in a consistent synset. Do not directly compare to WordNet, but stick to your own judgment.
- The annotation process should take between 1–3 hours. Please write the time you needed for the study at the end of the sheet.

| Lemma | Definition | Comment |
|---|---|---|
| bass | [N] A male singer who sings in the bass range. | |
| basso | [N] A bass singer, especially in opera. | |
| **Is the above synset consistent?** | **► "1"** | **(Example #1)** |
| bass | [N] A male singer who sings in the bass range. | |
| basso | [N] A bass singer, especially in opera. | |
| singer | [N] person who sings, is able to sing, or earns a living by singing. | *too broad* |
| **Is the above synset consistent?** | **► "0"** | **(Example #2)** |
| bass | [N] The perch; any of various marine and freshwater fish resembling the perch, all within the order of Perciformes. | *wrong sense* |
| basso | [N] A bass singer, especially in opera. | |
| **Is the above synset consistent?** | **► "2"** | **(Example #3)** |
| singer | [N] person who sings, is able to sing, or earns a living by singing. | |
| sing | [V] To produce harmonious sounds with one's voice. | |
| **Is the above synset consistent?** | **► "1"** | **(Example #4)** |

## B.3 Word sense alignment gold standard

**Task description**. With this annotation guidebook, you receive a spreadsheet that can be opened in Microsoft Excel or OpenOffice (please contact the authors in case of any problems). In the spreadsheet, you'll find a list of lemmas – one lemma per row. Each lemma is accompanied with two word senses that are represented by an id, a sense gloss (i.e., a definition text), and maybe some example sentences (written in quotes, where available). The word senses on the left hand side are taken from the English WordNet, the word senses on the right hand side from the English Wiktionary. For each lemma, all encoded word senses from WordNet are paired with all encoded word senses in Wiktionary in an ordered manner (i.e., all combinations are shown). In between you'll find a column "Same Sense", which is to be annotated. Your task is to make a binary decision for each pair: Write a

0 if both word senses are different, or
1 if both word senses represent the same or a highly similar meaning

into the "Same Sense" column. The result will be a word sense alignment of WordNet and Wiktionary on a sample of lemmas.

## Examples:

| Lemma | WordNet | Same Sense | Wiktionary | Comment |
|---|---|---|---|---|
| intersect | meet at a point | 1 | to cross; to cut | *clearly the same* |
| intersect | meet at a point | 0 | (mathematics) two sets intersect if and only if they have at least one common element | *Wiktionary sense is domain specific* |

| Lemma | WordNet | Same Sense | Wiktionary | Comment |
|---|---|---|---|---|
| people | the body of citizens of a state or country; "the Spanish people" | 1 | (plural: peoples) Persons forming or belonging to a particular group, such as a nation, class, ethnic group, country, family, etc; folk; community. | *although not every item is mentioned in WordNet (such as class), one could consider the pair as highly similar* |
| people | (plural) any group of human beings (men or women or children) collectively; "old people"; "there were at least 200 people in the audience" | 1 | (used as plural of person); a body of human beings considered generally or collectively; a group of two or more persons. | |
| people | (plural) any group of human beings (men or women or children) collectively; "old people"; "there were at least 200 people in the audience" | 0 | One's colleagues or employees. | *The Wiktionary sense is a variant of the WordNet sense. Additionally, the previous sense fits nicely* |

| Lemma | WordNet | Same Sense | Wiktionary | Comment |
|---|---|---|---|---|
| week | any period of seven consecutive days; "it rained for a week" | 1 | A period of seven days. | |
| week | any period of seven consecutive days; "it rained for a week" | 0 | The seven days beginning with Sunday or Monday. | |
| week | any period of seven consecutive days; "it rained for a week" | 0 | A subdivision of the month into longer periods of work days punctuated by shorter weekend periods of days for markets, rest, or religious observation such as a sabbath. | |
| week | hours or days of work in a calendar week; "they worked a 40-hour week" | 0 | A period of seven days. | |
| week | hours or days of work in a calendar week; "they worked a 40-hour week" | 0 | The seven days beginning with Sunday or Monday. | |
| week | hours or days of work in a calendar week; "they worked a 40-hour week" | 0 | A subdivision of the month into longer periods of work days punctuated by shorter weekend periods of days for markets, rest, or religious observation such as a sabbath. | *Somehow similar, but WordNet clearly focuses on the (western) calendar week as a reference, while Wiktionary regards other calendar systems* |
| week | a period of seven consecutive days starting on Sunday | 0 | A period of seven days. | *There is a better match* |
| week | a period of seven consecutive days starting on Sunday | 1 | The seven days beginning with Sunday or Monday. | |

**Please note**:

- Do not imitate an algorithm, rely on your intuition.
- For each lemma, multiple "1"s are possible.
- Sometimes no "1" is suitable for a lemma, i.e. the word senses are complementary.
- If one of the word senses covers only a small part of the other word sense and there is a better correspondence available, this "smaller" word sense is considered a variant and should be annotated with "0".
- It is possible that one of the word senses seems to be "wrong", i.e. there is no such meaning for this lemma, the sense definition is erroneous, or not exhaustive enough to clearly separate the sense from others. This fact should not be taken into account, i.e. take the word sense for granted and focus on the decision whether there is a corresponding word sense in the other resource.
- Sometimes the glosses are trimmed or normalized by JWKTL and thus do not clearly describe the word sense. It is allowed to consult Wiktionary for the original gloss. You can easy query the Wiktionary by typing `http://en.wiktionary.org/wiki/<LEMMA>` in your browser.
- The same applies to WordNet, where the word sense definitions might not be enough to infer the meaning. You are allowed to consult the Web front-end of WordNet at: `http://wordnetweb.princeton.edu/perl/webwn`
- As the task gets monotonous at some point, you should consider making breaks during the annotation process. Depending on your personal speed, it should take about 2–3 hours.
- Please do not change the order of the rows/cols or their IDs. we'll need this information for further processing. Of course, you are allowed to change the size of the columns to exactly fit on your screen.
- If you feel like explaining a certain decision, you can use the comment column. This is optional; you do not have to leave a comment for every pair!

After you finished your annotations, please fill out the last row with the annotation time needed in minutes. There is no correct solution for this task, so please, always judge the sense pairs based on your own understanding. We will measure the inter-annotator agreement of the submissions to find out, how often the participants agree or disagree.

## B.4 Marketing blunder dataset

**Introduction**. Large companies increasingly advertise and sell their products in international markets. Developing a marketing campaign for a new country requires tremendous translation efforts in order to bridge language-related and cultural boundaries. A particular problem often occurs if an established product or company name is used in a new foreign market with-

out being adapted to local habits and language use. This regularly yields offensive, embarrassing, or (at best) funny results causing excessive remedial cost and maybe even the withdrawal of the product from the new market.

A common example is the word *mist*, which is used to describe fabulous, enigmatic, lightweight, or mystic things in English. The word has, for instance, been used by a British car manufacturer to advertise their *Silver Mist* model. In German, the homograph *Mist* means, however, dung or manure, and it is a frequently used slang expression to describe a futile, cheap, or broken product, nonsense, or an annoying, tedious situation. This pejorative meaning has caused the car manufacturer to rename its product. Hereafter, we will concentrate on such cross-lingual marketing blunders.

**Task description**. Along with this annotation guidebook, you'll receive a spreadsheet that can be opened in Microsoft Excel or OpenOffice (please contact the authors in case of any problems with opening the file). The spreadsheet contains the following columns:

- **Blunder text**: a product/company name or marketing slogan that has failed in a certain language community (e.g., Silver Mist). You can find the full list of marketing blunders and explanations on why each blunder failed in the appendix of this document.
- **Lexeme**: one part of the blunder text or an orthographic variant of it (e.g., mist).
- **Language**: the language, the lexeme is in (e.g., German).
- **Translation**: a translation of the lexeme into German or English (e.g., manure).
- **Annotation**: the column, we ask you to fill in your decision (as described below).
- **Definition**: A short definition of the lexeme and its translation in German or English.

We call each row in the spreadsheet a *clue*. Your task is to judge if a clue is relevant or irrelevant for identifying a marketing blunder. That is to say, we ask you to decide if the given clue informs about a pejorative, vulgar, embarrassing, or otherwise distractive meaning which causes the product/company name (i.e., the blunder text) to fail. For the example of *Silver Mist* introduced above, a clue telling you that *Mist* means *manure* in German is relevant, since *manure* has a negative connotation and is considered vulgar. A clue explaining that *mist* means *fog* in Swedish is irrelevant, because *fog* does not yield negative associations when used as a car name. Thus, please mark each clue with a

1 if it is irrelevant and with a
2 if it is relevant.

Use the "Annotation" column for that.

Some of the clues refer to vulgar or embarrassing expressions that appear to be a relevant hint for detecting a marketing blunder. However, please always take the word form (i.e., the "lexeme" column) into account. You should only mark a clue as relevant if you consider the

lexeme "similar enough" to cause such a negative association when you read the product/company name (i.e., the blunder text column).

Consider the term *stubit*, for instance. This word form is very similar to the English *stupid* and thus provides a relevant clue for detecting a marketing blunder for *stubit*. As opposed to that, *stupid* would not be a relevant clue for the term *stepid*, although the two forms appear rather similar. The reasons are the differences in the pronunciation and that *stepid* reminds of the English *step it* in the first place, which has no negative connotation.

As a rule of thumb, consider you read and/or hear the blunder text on a newly occurred product in a shop in a country using the specified language. Decide (a) if the given lexeme would potentially come to your mind and (b) if the meaning of the lexeme might cause embarrassment or astonishment that might cause the customers to reject the product. Annotate these cases with a "1".

**Please note**:

- Do not imitate an algorithm, rely on your intuition.
- Please read the explanations of the marketing blunders in the appendix before starting with your annotations.
- For each marketing blunder there can be multiple relevant clues. Likewise, it is possible that none of the clues is relevant.
- Some clues are relevant to detect a marketing blunder in a language different to the one of the original blunder explained in the appendix. Please mark those clues also as relevant. If you want, you can use a "2" instead of a "1" to mark these clues.
- You may use external references such as dictionaries, encyclopedias, etc. to make your decision, but we ask you not to discuss your decisions with other annotators.
- As the task gets monotonous at some point, you should consider making breaks during the annotation process.
- Please do not change the order of the rows/cols or their IDs. We'll need this information for further processing. Of course, you are allowed to change the size of the columns to exactly fit on your screen.

There is no correct solution for this task, so please, always judge the clues based on your own understanding. We will measure the inter-annotator agreement of the submissions to find out, how often the participants agree or disagree.

# Index

# Wissenschaftlicher Werdegang des Verfassers[¶]

| | |
|---|---|
| 2003–2009 | Studium der Informatik<br>Technische Universität Darmstadt |
| 2006 | Abschluss als Bachelor of Science<br>Bachelor-Thesis: „Motivvariation in komplexen Netzwerken und ihre Auswirkung auf den Informationsgehalt“ aus dem Bereich Bioinformatik<br>Referent: Dr. habil. Matthias Müller-Hannemann |
| 2009 | Abschluss als Master of Science<br>Master-Thesis: „Combining Answers from heterogeneous Web Documents for Question Answering“ aus dem Bereich Sprachtechnologie<br>Referenten: Prof. Dr. Iryna Gurevych, Dr. Delphine Bernhard, Kateryna Ignatova |
| seit 2009 | Wissenschaftlicher Mitarbeiter am Ubiquitous Knowledge Processing Lab<br>Technische Universität Darmstadt |

# Ehrenwörtliche Erklärung[‡]

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades „Doktor-Ingenieur“ mit dem Titel *„Wiktionary: The Metalexicographic and the Natural Language Processing Perspective“* selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 29. August 2013

Christian M. Meyer

[¶] Gemäß § 20 Abs. 3 der Promotionsordnung der Technischen Universität Darmstadt.

[‡] Gemäß § 9 Abs. 1 der Promotionsordnung der Technischen Universität Darmstadt.

# Publikationsverzeichnis des Verfassers

Andrea Abel und **Christian M. Meyer**: 'The dynamics outside the paper: user contributions to online dictionaries', in: *Proceedings of the 3rd Biennial Conference on Electronic Lexicography* (eLex), Tallinn, Estonia, erscheint 2013.

**Christian M. Meyer** und Iryna Gurevych: 'Der lexikographische Prozess im deutschen Wiktionary', in Vera Hildenbrandt (Hrsg.): *Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie"*, OPAL – Online publizierte Arbeiten zur Linguistik, Mannheim: Institut für Deutsche Sprache, erscheint 2013.

Michael Matuschek, **Christian M. Meyer** und Iryna Gurevych: 'Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Translation Applications', *Translation: Computation, Corpora, Cognition: Special Issue "Language Technology for a Multilingual Europe"* 3 (1): 87–118, Juli 2013.

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek und **Christian M. Meyer**: 'UBY-LMF – Exploring the Boundaries of Language-Independent Lexicon Models', in Gil Francopoulo (Hrsg.): *LMF: Lexical Markup Framework*, Kapitel 10, S. 145–156, London: Wiley-ISTE, März 2013.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, **Christian M. Meyer** und Tri Duc Nghiem: 'UBY – A Large-Scale Unified Lexical-Semantic Resource', in: *Book of Abstracts of the 23rd Meeting of Computational Linguistics in the Netherlands* (CLIN), S. 81, Enschede, Niederlande, Januar 2013.

**Christian M. Meyer** und Iryna Gurevych: 'To Exhibit is not to Loiter: A Multilingual Sense-Disambiguated Wiktionary for Measuring Verb Similarity', in: *Proceedings of the 24th International Conference on Computational Linguistics* (COLING), Vol. 4, S. 1763–1780, Mumbai, Indien, Dezember 2012.

**Christian M. Meyer** und Iryna Gurevych: 'Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography', in Sylviane Granger und Magali Paquot (Hrsg.): *Electronic Lexicography*, Kapitel 13, S. 259–291, Oxford: Oxford University Press, November 2012.

Iryna Gurevych, Michael Matuschek, Tri Duc Nghiem, Judith Eckle-Kohler, Silvana Hartmann und **Christian M. Meyer**: 'Navigating Sense-Aligned Lexical-Semantic Resources: The Web Interface to UBY', in: *Proceedings of the 11th Conference on Natural Language Processing: Empirical Methods in Natural Language Processing* (KONVENS), S. 194–198, Wien, Österreich, September 2012.

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek und **Christian M. Meyer**: 'UBY-LMF – A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), S. 275–282, Istanbul, Türkei, Mai 2012.

Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek und **Christian M. Meyer**: 'The Open Linguistics Working Group', in: *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC), S. 3603–3610, Istanbul, Türkei, Mai 2012.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, **Christian M. Meyer** und Christian Wirth: 'UBY – A Large-Scale Unified Lexical-Semantic Resource Based on LMF', in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL), S. 580–590, Avignon, Frankreich, April 2012.

**Christian M. Meyer** und Iryna Gurevych: 'OntoWiktionary – Constructing an Ontology from the Collaborative Online Dictionary Wiktionary', in Maria Teresa Pazienza und Armando Stellato (Hrsg.): *Semi-Automatic Ontology Development: Processes and Resources*, Kapitel 6, S. 131–161, Hershey, PA: IGI Global, Februar 2012.

**Christian M. Meyer** und Iryna Gurevych: 'What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage', in: *Proceedings of the 5th International Joint Conference on Natural Language Processing* (IJCNLP), S. 883–892, Chiang Mai, Thailand, November 2011.

**Christian M. Meyer** und Iryna Gurevych: 'How Web Communities Analyze Human Language: Word Senses in Wiktionary', in: *Proceedings of the Second Web Science Conference* (WebSci), Raleigh, NC, USA, April 2010.

**Christian M. Meyer** und Iryna Gurevych: 'Worth its Weight in Gold or Yet Another Resource – A Comparative Study of Wiktionary, OpenThesaurus and GermaNet', in Alexander Gelbukh (Hrsg.): *Computational Linguistics and Intelligent Text Processing: Proceedings of the 11th International Conference* (CICLing), Lecture Notes in Computer Science Vol. 6008, S. 38–49, Berlin/Heidelberg: Springer, März 2010.

**Christian M. Meyer**: *Combining Answers from heterogeneous Web Documents for Question Answering*, Master Thesis, Technische Universität Darmstadt, April 2009.

**Christian M. Meyer**: *Motivvariation in komplexen Netzwerken und ihre Auswirkung auf den Informationsgehalt*, Bachelor Thesis, Technische Universität Darmstadt, August 2006.