# Semi-Automatic Ontology Development:

## Processes and Resources

Maria Teresa Pazienza
*University of Roma Tor Vergata, Italy*

Armando Stellato
*University of Roma Tor Vergata, Italy*

Information Science
REFERENCE

| | |
|---|---|
| Managing Director: | Lindsay Johnston |
| Senior Editorial Director: | Heather Probst |
| Book Production Manager: | Sean Woznicki |
| Development Manager: | Joel Gamon |
| Development Editor: | Michael Killian |
| Acquisitions Editor: | Erika Gallagher |
| Typesetter: | Milan Vracarich, Jr. |
| Cover Design: | Nick Newcomer, Lisandro Gonzalez |

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Semi-automatic ontology development : processes and resources / Maria Teresa Pazienza and Armando Stellato, editors.
p. cm.
Includes bibliographical references and index.
Summary: “This book includes state-of-the-art research results aimed at the automation of ontology development processes and the reuse of external resources becoming a reality, thus being of interest for a wide and diversified community of users”-- Provided by publisher.
ISBN 978-1-4666-0188-8 (hardcover) -- ISBN 978-1-4666-0189-5 (ebook) -- ISBN 978-1-4666-0190-1 (print & perpetual access) 1. Semantic Web. 2. Ontologies (Information retrieval) 3. Semantic networks (Information theory) 4. Automatic data collection systems. 5. Knowledge acquisition (Expert systems) I. Pazienza, Maria Teresa. II. Stellato, Armando, 1975-
TK5105.88815.S468 2012
006.3’3--dc23
2011051816

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

# Chapter 6
# OntoWiktionary:
## Constructing an Ontology from the Collaborative Online Dictionary Wiktionary

**Christian M. Meyer**
*Technische Universität Darmstadt, Germany*

**Iryna Gurevych**
*Technische Universität Darmstadt, Germany*

## ABSTRACT

*The semi-automatic development of ontologies is an important field of research, since existing ontologies often suffer from their small size, unaffordable construction cost, and limited quality of ontology learning systems. The main objective of this chapter is to introduce Wiktionary, which is a collaborative online dictionary encoding information about words, word senses, and relations between them, as a resource for ontology construction. The authors find that a Wiktionary-based ontology can exceed the size of, for example, OpenCyc and OntoWordNet. One particular advantage of Wiktionary is its multilingual nature, which allows the construction of ontologies for different languages. Additionally, its collaborative construction approach means that novel concepts and domain-specific knowledge are quick to appear in the dictionary.*

*For constructing their ontology OntoWiktionary, the authors present a two-step approach that involves (1) harvesting structured knowledge from Wiktionary and (2) ontologizing this knowledge (i.e., the formation of ontological concepts and relationships from the harvested knowledge). They evaluate their approach based on human judgments and find their new ontology to be of overall good quality. To encourage further research in this field, the authors make the final OntoWiktionary publicly available and suggest integrating this novel resource with the linked data cloud as well as other existing ontology projects.*

DOI: 10.4018/978-1-4666-0188-8.ch006

## INTRODUCTION

To date, many knowledge-based tasks utilize ontologies as a source of background knowledge. This includes, for example, the calculation of semantic relatedness, automatic word sense disambiguation, or machine translation systems. Ontologies also represent the backbone of the Semantic Web (Berners-Lee et al., 2006). It turns, however, out that existing ontologies are either small or show only limited quality, which prompts further research in this direction. In particular, the (semi-)automatic development of ontologies is still a significant challenge and a yet unsolved research question.

Recent developments in the World Wide Web actuate a large number of collaborative online projects, such as Wikipedia. These collaborative resources have the potential to form huge ontologies, since they can attract a large community of contributors. At the same time, they also ensure a reasonably good quality, as their content has been defined and verified by humans. It has been found that this type of resource can surmount the shortcomings of both expert-built resources, which are often fairly small and hard to keep up to date, and of data-driven ontology learning approaches, which are usually prone to noise and errors.

In this particular work, we focus on constructing an ontology from Wiktionary, which is a freely available, collaboratively built online dictionary. Although Wiktionary is still dramatically underresearched, it has proven to have enormous potential within a natural language processing system measuring the semantic relatedness between words (Zesch, et al., 2008a). Wiktionary encodes a huge number of words, word senses, and semantic relations, which is an ideal basis for constructing ontologies. Therefore, we will explore how large amounts of knowledge can be harvested from Wiktionary and how this knowledge can be "ontologized"—i.e., transformed into an ontological structure. As a result of our work, we present OntoWiktionary, which is a novel ontology consisting of concepts and relations harvested from Wiktionary. OntoWiktionary has several advantages over existing ontologies, as it contains a large number of concepts and lexicalizations, which include both commonly used ones as well as rare and domain-specific ones. Additionally, the collaborative construction process of Wiktionary allows it to quickly reflect usage trends and newly occurring concepts. The multilingual nature of Wiktionary moreover puts us in the position of constructing ontologies for a large number of languages. We make OntoWiktionary publicly available to foster integration with existing ontologies, as well as the development of knowledge-rich applications that can benefit from employing it as a source of background knowledge.

The remainder of this chapter is structured as follows: We will first discuss previous work in the area of ontology construction in general as well as using collaboratively created resources in particular. We then provide a comprehensive introduction to Wiktionary and the knowledge encoded therein. In order to harvest ontological knowledge from Wiktionary, we first need to discuss the structure of Wiktionary articles and explain how to deal with structural errors and inconsistencies pertinent to Wiktionary data. Then, we describe how we construct and evaluate OntoWiktionary by ontologizing the extracted Wiktionary knowledge. The ontologizing step consists of three tasks—namely, the anchoring of relations, the formation of ontological concepts, and the formation of relations between these concepts. We conclude our chapter with a discussion of our findings and outline some open issues and future research directions.

## BACKGROUND

Before taking a deeper look at Wiktionary and our ontology construction architecture, we introduce the notation used throughout the chapter and relate our approach to previous work in the area of both ontology construction in general and particularly in the context of collaborative resources.

## Notation

Over time, a broad variety of terms has emerged in the area of ontology construction. We therefore introduce the notation that we will use throughout the chapter. Our definitions are mainly based on the work by Guarino et al. (2009). By *ontology*, we refer to a computer artifact that is able to model everything that exists in a certain universe (not necessarily the real world). The process of building ontologies (i.e., the definition and population of the computer artifact with knowledge) is usually called *ontology construction* or *ontology development*. A more specialized term is *ontology learning*, which focuses on automatic ontology construction, usually using methods from machine learning or information extraction. Note that these terms are sometimes used synonymously or defined differently, so we will use the term *ontology construction* henceforth to denote the general building process of ontologies, including manual, semi-automatic, and fully automatic approaches.

According to Guarino et al. (2009), the building blocks of an ontology are *concepts* and *relations*. The former is a conceptualization of a phenomenon observed in the universe. An example is the idea of a dog — i.e., the animal of the genus *Canis*. Note that this concept 'Dog' comprises all dogs observed in the universe. Individual dogs (like 'Lassie') are, in contrast, called *instances*, which can also be modeled in an ontology. In the following, we will not consider instances any further, but focus on concepts. While a concept has a certain meaning, it can be referred to by multiple words of our language, which we call *lexicalizations*. The 'Dog' concept might, for instance, have the lexicalizations 'dog' and 'hound.' The backbone of an ontology are subsumption relations between concepts — i.e., a relationship that forms a hierarchy of concepts, which is also known as *generalization*, *specialization*, or *taxonomy*. The concept 'Dog' can, for example, be subsumed by a superconcept 'Animal' that represents any type of animal.

Note that we use single-quoted words starting with an upper case letter (e.g., 'Dog') to identify concepts, and single-quoted words starting with a lower case letter to refer to lexicalizations (e.g., 'dog'). Since word senses are used as lexicalizations in our approach, we use the same markup for them. Sets of concepts, lexicalizations, and word senses are denoted by curly brackets; for example, {dog, hound}. Relations between concepts, lexicalizations, and word senses are, in contrast, surrounded by round brackets — e.g., (dog, hound), which denotes a certain relation between 'dog' and 'hound.'

## General Ontology Construction Approaches

In the past, very different ways of constructing ontologies have been proposed and, accordingly, there has been a variety of classifications and surveys on this topic. Following Russel and Norvig (2010), we distinguish four general approaches based on:

1. the *manual modeling of experts*, such as lexicographers, ontology engineers, or domain specialists, which is the case with (inter alia) Cyc (Lenat, 1995) and OpenCyc.[1]
2. *information extraction* from large amounts of unstructured documents—e.g., using the TextRunner system (Banko, et al., 2007) on a large corpus of Web documents. An overview of such systems can be found in Maynard et al. (2008).
3. existing *(semi-)structured resources* that are either restructured to form a novel ontology, or used to populate an ontological model, or aligned with existing ontologies (Prévot, et al., 2005). Such resources can be, for example, linguistic resources (Gangemi,

et al., 2003), or domain-specific resources (Reed & Lenat, 2002).

4. a *collaborative annotation effort*, such as the OpenMind project (Singh, 2002), which provides a platform for non-experts to propose machine-readable common-sense knowledge on a voluntary basis.

Each of these approaches has its unique advantages and limitations that we discuss in the following and summarize in Table 1:

Expert-built ontologies, as described in (1), can be very consistent and of high quality; their size, however, is usually subject to time and budget considerations. This often yields rather small ontologies. OpenCyc 2.0 encodes, for example, only 56,000 concepts, although its creation required an enormous effort for years. Another problem with the manual construction process is the need for continuous revisions and updates. Human language is constantly changing and evolving, which introduces new concepts and lexicalizations that are not yet represented within an ontology. Expert-built ontologies are usually released at certain fixed dates and thus unable to integrate novel concepts until their next release.

The parsing of unstructured document collections that is proposed in (2) allows the construction of huge ontologies, though often of limited quality. The main reason is the lack of structure and ontological properties within the variety of documents used for the information extraction method, which causes noise in the resulting ontology. 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. They try to infer semantic knowledge from a large set of input data, while only a small fraction of it contains evidence (e.g., for a certain relation). A well-known example is the TextRunner system (Banko, et al., 2007). Although such systems have recently shown impressive progress in their precision, they still cannot reach the quality of human judgments. An additional problem is that unstructured document collections might be highly biased to certain topics, styles, registers, or genres. The same applies to the Web as a corpus, which is known to contain errors, sublanguages, and topics that are predominant within the World Wide Web (Kilgarriff & Grefenstette, 2003).

While ontology learning systems operating on a large amount of unstructured text data usually yield low precision, better structured resources as in (3) appear to be a viable option. Most of these structured or semi-structured resources that have been proposed for creating or populating an ontology are very focused on a certain purpose or domain and thus ill-suited for constructing a general ontology. Reed and Lenat (2002) report, for instance, on the integration of the Open Directory Project,[2] the CIA World Factbook,[3] and the Unified Medical Language System[4] into

*Table 1. Summary of advantages and limitations of different ontology construction approaches*

| | (1) Manual modeling of experts | (2) Information extraction | (3) Semi-structured resources | (4) Collaborative effort |
|---|---|---|---|---|
| Size | – | + + | o | + + |
| Quality of contents | + + | – | + | + |
| Development effort | – – | + + | – | + |
| Coverage of novel concepts | – – | + | – | + + |
| Coverage of domain and rare concepts | + | o | + + | + + |
| Available languages | – | + + | – | + + |

Cyc, which is a good starting point for enriching general ontologies. Their integration, however, still requires the judgment of experts in order to identify overlapping concepts. This human effort might be feasible for a small number of resources, but does not scale to a larger resource collection. Apart from this, changes in the resources, such as new categories within the Open Directory Project, require new judgments, which turns out to be a very time-consuming process in the long run.

Amongst others, Gangemi et al. (2003) suggest linguistic resources, like dictionaries, lexicons, or semantic networks, as a source for constructing ontologies. They usually cover general language and are thus not limited to certain domains. The Princeton WordNet (Fellbaum, 1998) is the de facto standard resource in the natural language processing community, and it is straightforward to use this resource for populating an ontology. WordNet has the advantage of being clearly structured, which avoids noise in the ontology construction process. Gangemi et al. (2003) present a semi-automatic method for constructing an ontologized version of WordNet called OntoWordNet. A different approach is introduced by Martin (2003), who proposes multiple transformation steps within WordNet to allow using it as an ontology directly. But since WordNet has been created by a small group of linguists, it shows — although encoding more concepts than OpenCyc—the same problems as the manually created ontologies described in (1), such as the time-consuming development and update process, which is restricted to a fixed release cycle. For languages other than English, the problems pertinent to expert-built resources are even more severe, as resources such as EuroWordNet (Vossen, 1998) are usually a lot smaller (if they exist at all).

A promising and emerging field of research makes use of collaboratively constructed knowledge resources as described in (4). While the phenomenon of collective intelligence—often denoted as the 'wisdom of the crowds'—has been found to be competitive to expert knowledge (Surowiecki, 2005), the advent of the socio-semantic Web gave rise to a large number of Web projects fostering collaborative text and knowledge editing, including blogs, forums, social tagging sites, and wikis. Such collaborative resources have the potential to be a source of extensive ontological knowledge due to the usually large user communities. At the same time, they ensure fairly good quality, as their content has been explicitly defined by humans rather than automatically extracted from heterogeneous text collections. The broad variety of authors in collaborative resources opens up new opportunities for harvesting knowledge from multiple languages, including both general and domain-specific concepts, as well as rare ones. Additionally, the construction costs of an ontology based on collaborative resources are rather small, since their contents can be freely accessed. This brings us to focus on this approach to constructing ontologies here. In the following section, we will review previous approaches from this strand of research in more detail and illustrate how Wiktionary can surmount limitations of alternative resources.

## Collaborative Resources as a Source for Ontologies

The most prominent types of collaborative resources are blogs, forums, social tagging websites, and wikis. Naturally, there are large differences amongst such projects, which we will discuss in the following and summarize in Table 2.

Regarding blogs and forums, which are mainly based on free text, automatic information extraction methods can be used for constructing ontologies. This raises again the problem of noise and errors discussed in the previous section. The inference of relations and the identification and disambiguation of concepts are the main source for errors here. The use of folksonomies (i.e., social tagging websites such as Del.icio.us[5] or Flickr[6] that encourage people to tag images, places, bookmarks, etc. with keyword tags)

*Table 2. Summary of advantages and limitations of different ontology construction approaches based on collaborative resources*

| | **Blog- and forum-based ontologies** | **Folksonomy-based ontologies** | **Collaborative ontology projects** | **Wikipedia-based ontologies** | **Wiktionary-based ontologies** |
|---|---|---|---|---|---|
| Community size | + + | + | – – | + + | + |
| Ontology size | + + | + | – | + + | + |
| Sense-disambiguated concepts | – | – | o | + | + |
| Instances | o | + | o | + | – |
| Abstract concepts | o | o | + | – | + |
| Lexicalizations | – | + | o | o | + |
| Clear-cut subsumption hierarchy | – | – | + | o | + + |

poses similar challenges: Gruber (2007) examines the differences between an ontology and a folksonomy and proposes a general ontological model for them, which is populated by (among others) Echarte et al. (2007). In folksonomies, the tags can be processed automatically without the necessity of information extraction methods. However, no explicit relations between tags are usually encoded, and the individual tags are not per se sense disambiguated (Mika, 2007). The tag 'tree' can, for instance, be used for tagging both botany-related objects as well as computer science-related ones.

Singh (2002) presents the collaborative ontology OpenMind,[7] whose website asks volunteers to add machine-readable common-sense knowledge that can directly be used for creating ontologies. The users first choose a predefined relationship and then insert the concepts for this relationship (e.g., that 'shoes' are made of 'leather'). This directly models ontological relations without the necessity of an extraction or learning step that would introduce noise. A problem is, though, that OpenMind does not really model concepts, but rather uses individual words only. Thus, there might be different relations for the synonymous words 'pullover' and 'sweater,' which denote a single concept. An additional problem is ambiguity. Relationships including, for instance, 'bass' do not distinguish the concept of a fish from that of the music instrument. Moreover, the community of the platform is rather small, which might be due to the specialized focus of the project. Ordinary Web users can hardly benefit from the knowledge encoded in OpenMind and might thus be less motivated to contribute.

Large wikis, such as Wikipedia and Wiktionary have, in contrast, become more and more popular and manage to attract a huge community of contributors. The ease of editing the content of a wiki page and the direct usefulness of the encoded contents for the users are crucial for their success. In the following sections, we will therefore focus on this type of collaborative resource and discuss Wikipedia-based as well as Wiktionary-based ontologies, which will be the main objective of this chapter.

## Wikipedia-Based Ontologies

Of particular research interest in both the natural language processing community and the Semantic Web community is Wikipedia,[8] which quickly became the largest encyclopedia in the world. Since Wikipedia is consulted by thousands of Web users every day, many people are motivated to contribute to the project by writing new articles or editing and correcting existing ones. The Wikipedia community is indeed three orders of magnitude larger than the OpenMind community and provides over three million articles in English that can be used to represent the concepts of an ontology. But not only is the size of such an ontology so huge; Wikipedia has also been found to be of competitive quality to expert-defined encyclopedias (Giles, 2005).

The most influential works in the area of Wikipedia-based ontologies are YAGO (Suchanek, et al., 2008) and DBpedia (Bizer, et al., 2009). The goals of these works are the transformation of Wikipedia into an ontology and the interlinking of its concepts with the Linked Data cloud (Bizer, et al., 2009a) — i.e., the transformation of Wikipedia data into standardized RDF models and relating it to other Linked Data by means of unique URIs. Both YAGO and DBpedia are nowadays well known and have been successfully used in various applications. However, there is some potential for improvement regarding these works: Typically, redirects, disambiguation pages, hyperlinks, categories, geographic coordinates, and infoboxes serve as a source for extracting the relationships between concepts. In this context, category labels are used to create a subsumption hierarchy. Although this yields a densely connected taxonomy of concepts, Ponzetto and Strube (2007) point out that the Wikipedia categories "do not form a taxonomy with a fully-fledged subsumption hierarchy." Both the YAGO and the DBpedia concept 'Iron (appliance)' is, for instance, not only a subsumption of 'Home appliance,' but also of 'Laundry.' This is not a generalization of 'Iron,' but represents the domain the concept is used in.

Another problem of a Wikipedia-based ontology lies in the lexicalizations of concepts. In order to reduce redundancy, each concept is encoded only once within Wikipedia and thus described within the article with the most common lexicalization of the concept. The concept 'Iron' in the sense of the 26th chemical element is, for example, described within the article 'Iron.' Additionally, Wikipedia allows one to define redirects from one article title to another; for example, the redirects from 'Fe,' 'Ferryl,' and 'Element 26' to the article 'Iron.' These redirects are also used as lexicalizations of the concept in DBpedia. Although such lexicalizations are generally correct, redirects are not always used for defining synonymous terms, but also for spelling errors (e.g., 'Iorn') and related concepts (e.g., 'Iron rope' or 'Iron compounds') that should not serve as lexicalizations for the concept 'Iron.' Of the fifteen redirects to the article 'Iron' in the current version of the English Wikipedia, only six represent valid lexicalizations of this concept.

## Wiktionary-Based Ontologies

Wiktionary[9] is a free online dictionary that is organized similarly to Wikipedia, but which focuses on linguistic rather than encyclopedic knowledge. It encodes knowledge about words, word meanings, and semantic relations between them (Zesch, et al., 2008). Wiktionary is much more structured than Wikipedia, which allows us to harvest the encoded knowledge in a more precise way than is the case with Wikipedia. In particular, synonymous terms and subsumption relations between word meanings are explicitly encoded in Wiktionary and can thus be acquired more accurately. Such relations are crucial for constructing ontologies with rich lexicalizations, which we will discuss in this chapter.

Additionally, Wiktionary is similar to WordNet as both resources encode linguistic knowledge in the form of word meanings and semantic relations, such as synonymy, hyponymy, and hypernymy. Wiktionary, however, comes with four major advantages over WordNet:

1. Wiktionary is far larger in size than WordNet. The English Wiktionary edition currently encodes knowledge for over 375,000 English words,[10] while WordNet's lexicon contains only about 155,000 words.
2. The data of Wiktionary is constantly updated by its community and thus, rather than re-

lying on certain fixed release dates as it is the case for WordNet, neologisms and new concepts are quick to appear in the resource.

3. Wiktionary has been found to encode a large number of domain-specific entries (Meyer & Gurevych, 2010a), which enables the creation of domain-specific ontologies or the enrichment of a general ontology with very specialized concepts from science, medicine, sports, etc.
4. Wiktionary is available in over 145 languages and thus can yield ontologies for languages where no expert-defined ontology or wordnet is yet available. This is particularly valuable for research in the context of machine translation and cross-lingual natural language processing.

As mentioned above, WordNet has been used as a basis to construct the ontology OntoWordNet (Gangemi, et al., 2003). We follow this principle and construct the novel ontology OntoWiktionary from linguistic knowledge harvested from Wiktionary. We expect our ontology to improve OntoWordNet with regard to (1)—(4) and especially focus on the formation of concepts, which are associated with a large number of lexicalizations, and on a more accurate subsumption hierarchy than present in DBpedia and YAGO. Besides the English Wiktionary, we will employ the German and Russian Wiktionary editions to demonstrate the possibility of constructing Wiktionary-based ontologies for a large number of languages. We make OntoWiktionary publicly available on our website to encourage other researchers to build upon this ontology, integrate it with other knowledge repositories, and utilize it in different natural language processing applications. In the next section, we introduce Wiktionary in more detail and outline the architecture of our Wiktionary-based ontology construction method.

## ONTOWIKTIONARY

In order to use Wiktionary as a source for constructing a new ontology, we first need to understand what kind of knowledge is encoded therein and how this resource is structured. In this section, we therefore provide an overview of Wiktionary's basic organization and some example entries illustrating it. Then, we introduce our architecture for constructing an ontology from Wiktionary.

### Wiktionary: A Collaborative Resource for Linguistic Knowledge

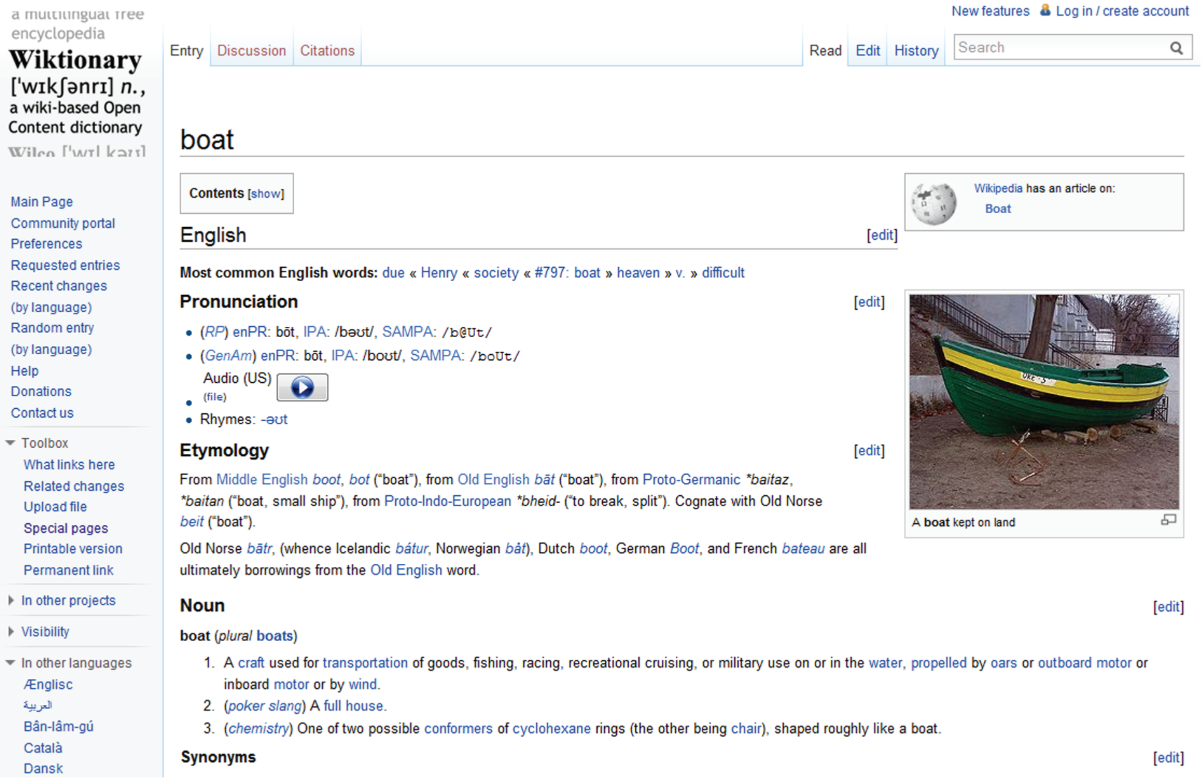
The goal of Wiktionary is to create a large, multilingual online dictionary that is both freely available and editable by volunteers. The project started in 2002 with the English Wiktionary. By 2004, the community began to set up Wiktionary editions for other languages. Since there are no special requirements for contributing to the project, the community of Wiktionary editors grew very quickly — by the beginning of 2011, about 460,000 users have created over 2,200,000 articles in the English edition.

Currently, there are 145 active language editions of Wiktionary.[11] The primary building blocks of each Wiktionary edition are article pages that contain lexical semantic information about a certain word or phrase—e.g., 'boat,' 'sleep,' or 'trace element.' Figure 1 shows the article 'boat' of the English Wiktionary as an example. A single Wiktionary language edition is not limited to encoding only those words of its own, native language. It is rather the vision of Wiktionary that every language edition contains information about words of any language. In each article, multiple language entries can thus be distinguished. The article 'sensible' within the English Wiktionary, for example, encodes linguistic knowledge about the corresponding English and French words. It should be noted that these two words only share the same written form rather than the same meaning (the French *sensible* means sensitive in English). For each language entry, there are multiple sections for encoding the word's part of speech, etymology, pronunciation, grammatically inflected word forms, and lots of other linguistic information.

Most important for our purpose is the section encoding a word's meaning, which is represented as a list of different word senses for the word described by the article. The enumeration of distinct word senses corresponds to common practice in printed and electronic dictionaries where words are divided into a number of distinct senses for pragmatic reasons (Atkins & Rundell, 2008). Each word sense is represented by a short definition text that might be accompanied by some example sentences or quotations illustrating the usage of the word sense. Meyer and Gurevych (2010a) note that the nature of word senses in Wiktionary is unique, since the collaborative construction approach leads to constant revision and discussion about the composition of word senses. This yields a consolidation of the different opinions of the speakers. The granularity of a word sense definition—i.e., where to split or lump two nuances of the meaning—is an open discussion that has previously almost solely been the province of a small number of expert lexicographers but is now transferred to a large community of ordinary speakers of a language. Constructing an ontology from these collaboratively defined word senses can help us to understand the different semantics of collaborative language resources such as Wiktionary and expert-built resources like WordNet.

In Wiktionary, there are also sections for encoding semantic relations, such as "Synonyms," "Hypernyms," "Hyponyms," "Meronyms," "De-

*Figure 1. Wiktionary article for the word 'boat'*



boat

English

**Most common English words:** due « Henry « society « #797: boat » heaven » v. » difficult

**Pronunciation**

- (*RP*) enPR: bōt, IPA: /bəʊt/, SAMPA: /b@Ut/
- (*GenAm*) enPR: bōt, IPA: /boʊt/, SAMPA: /boUt/
- Audio (US) (file)
- Rhymes: -əʊt

**Etymology**

From Middle English *boot*, *bot* ("boat"), from Old English *bāt* ("boat"), from Proto-Germanic **baitaz*, **baitan* ("boat, small ship"), from Proto-Indo-European **bheid-* ("to break, split"). Cognate with Old Norse *beit* ("boat").

Old Norse *bātr*, (whence Icelandic *bátur*, Norwegian *båt*), Dutch *boot*, German *Boot*, and French *bateau* are all ultimately borrowings from the Old English word.

A **boat** kept on land

**Noun**

**boat** (*plural* **boats**)

1. A craft used for transportation of goods, fishing, racing, recreational cruising, or military use on or in the water, propelled by oars or outboard motor or inboard motor or by wind.
2. (*poker slang*) A full house.
3. (*chemistry*) One of two possible conformers of cyclohexane rings (the other being chair), shaped roughly like a boat.

**Synonyms**

rived from," "See also," etc. Semantic relations are represented by a hyperlink to another Wiktionary article. The article 'boat' contains, for example, a link to the article 'vessel' within its "Synonyms" section, and a link to the article 'canoe' within its "Hyponyms" section. This notation puts us in the position of harvesting relations between concepts that we will explain in the main part of the chapter.

Besides words, word senses, and semantic relations, which we use for the construction of OntoWiktionary, there is lots of other linguistic information that is attached to the encoded words and word senses. This includes a word's pronunciation, hyphenation, etymology, alternative spellings, or rhyme schemes, as well as a word sense's semantic domain, translation, or image that illustrates the meaning. Such information can be used to enrich the ontology. Translations in particular offer interesting future research questions in the context of interlinking ontologies across multiple languages. Since we focus on the general ontology construction process here, we will, however, not discuss this kind of information in detail.

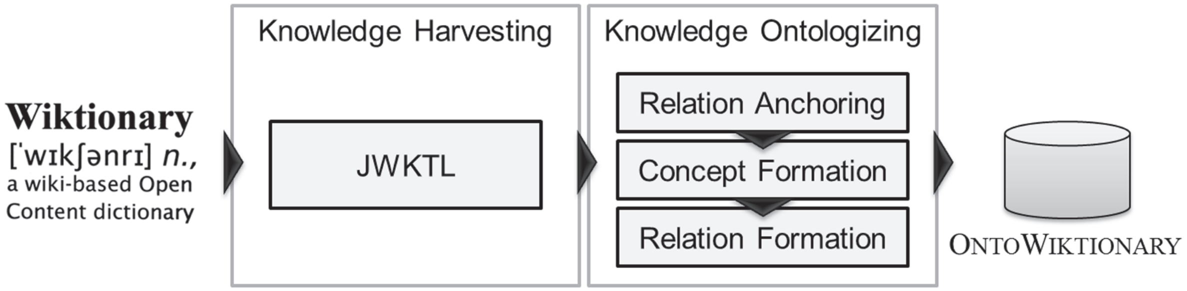
## Architecture of OntoWiktionary's Ontology Construction Process

In the previous section, we have seen that Wiktionary offers a large amount of linguistic knowledge that is relevant for ontology construction. Wiktionary is, however, essentially a dictionary for human readers rather than an ontology. We thus need an ontology construction process to transform the knowledge encoded in Wiktionary into the concepts and relations of an ontology. Figure 2 outlines our architecture, which will be explained within the subsequent sections.

Following Pantel and Pennacchiotti (2008), we divide our process into two parts: (1) *harvesting knowledge*, and (2) *ontologizing knowledge*. The former addresses obtaining the data from Wiktionary and extracting its knowledge in a structured and machine-readable manner. Since Wiktionary is a semi-structured resource, a carefully crafted system needs to be developed that is able to deal with noise induced by errors of the data extraction process on the one hand and with constant changes by the community pertinent to Wiktionary on the other hand. In particular, we address the extraction of words, word senses, and semantic relations from Wiktionary, which are required in our ontologizing step. Therefore, we use the JWKTL (Zesch, et al., 2008) software for processing the English and the German Wiktionary and extend the software for also harvesting knowledge from the Russian language edition. We will explain our approach of harvesting knowledge from Wiktionary in detail within the next section.

The latter part addresses the "ontologizing" of the extracted knowledge. This includes the *formation of concepts* and the *formation of relations* between them. We will particularly discuss how

*Figure 2. Ontology construction architecture for OntoWiktionary*



Wiktionary word senses can be used to induce ontological concepts and how hyperlinks between different Wiktionary articles may be treated as conceptual relations. A central point is thereby the *relation anchoring*—i.e., the association of a relation's endpoint with the correct concepts of the ontology—which is done in a separate preprocessing step. The explanation of the knowledge ontologizing part will be the subject of the section after the next one.

## HARVESTING KNOWLEDGE FROM WIKTIONARY

Wiktionary is intended to fulfill linguistic information needs of humans—i.e., to provide information about words, word senses, and semantic relations amongst them in a similar way as printed dictionaries do. Therefore, a focus has been put on providing a graphical user interface that is optimized for human perception rather than automatic data processing. Harvesting the knowledge encoded in Wiktionary thus raises the challenge of creating extraction software that processes Wiktionary's semi-structured content and transforms it into a machine-readable format for further processing. In the following section, we will first discuss the problem of noise pertinent to the data extraction process and how to deal with it, and then introduce existing software libraries and our extensions to them.

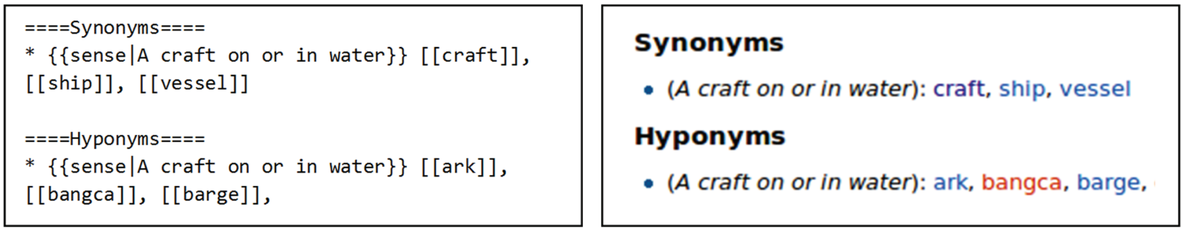
### Dealing with Noise and Errors

A main characteristic of Wiktionary is its openness—that is, the possibility for every Web user to add, modify, and delete content from the articles. While this openness is a key for the success of Wiktionary, it also presents a major challenge for the computational exploitation of this resource. The structural openness in particular turns out to be very challenging, as this includes missing sections, constant restructuring of the articles, malformation, and spam, as well as previously unseen types of knowledge, such as totally new sections.

An important feature of Wiktionary is the notion of templates. Templates are reusable patterns that can be defined in a central place and then invoked by a large number of articles. Each template is identified by a unique name. Invoking a template means that this name is added to the article text and enclosed by two curly brackets. Upon formatting the article to HTML, which is done when reading the article on the Wiktionary website, the invoked template is substituted with the template's text. The template may be further parameterized with different user inputs. For example, invoking the template {{rfe}} on an article page, causes the insertion of a box "This entry lacks etymological information […]" when the article is formatted, as well as a category tag "Requests for etymology" to allow searching for such entries easily. Another example is the template {{sense|<reference>}}, which is used to associate semantic relations to a certain word sense. This "sense" template is parameterized with a <reference> to the corresponding word sense, which is a shortened version of the sense's textual description. Figure 3 shows an example usage of (a) the "sense" template for the article 'boat,' and (b) the HTML-formatted result of this syntax. We will take a deeper look at the "sense" template later in the chapter when the semantic relations are anchored.

While the "RFE" template is primarily used to abbreviate an oft-used structure, the "sense" template is obviously used for different reasons, since the template syntax is actually longer than adding the formatted result directly. It is rather Wiktionary's way of adding structure and encouraging consistent encoding of the entries. Templates such as "sense" are not only useful for the community to quickly modify the formatting of all entries at once (e.g., using square brackets for the sense reference instead of round ones), but also allow for easy perception of the encoded knowl-

*Figure 3. Wiki syntax of (a) the "sense" template, and (b) its corresponding HTML format*



```
====Synonyms====
* {{sense|A craft on or in water}} [[craft]],
[[ship]], [[vessel]]

====Hyponyms====
* {{sense|A craft on or in water}} [[ark]],
[[bangca]], [[barge]],
```

(a)

**Synonyms**

- (*A craft on or in water*): craft, ship, vessel

**Hyponyms**

- (*A craft on or in water*): ark, bangca, barge,

(b)

edge, since all semantic relations follow a similar notation. Each article page in Wiktionary is usually a composition of many different templates and thus follows a common layout. At the same time, templates are a viable option to extract the data automatically, because it is easier to identify the "sense" template in front of a hyperlink than to rely on a certain combination of brackets and font styles that might be slightly varying for each article page.

In general, there is, however, no set of rigid rules for what an article page should look like. Rather, an author can extend or manipulate the proposed structure to better fit his or her needs. The Wiktionary guidelines[12] explain: "You may experiment with deviations, but other editors may find those deviations unacceptable, and revert those changes. They have just as much right to do that as you have to make them." An extraction system for Wiktionary is therefore required to deal with additional, modified, or missing structures. In the following section, we will introduce different software systems that can be used to harvest knowledge from Wiktionary.

## Obtaining and Extracting Wiktionary Data

Human readers use Wiktionary's Web front end to browse the encoded contents. An obvious way of obtaining the data automatically would thus be to crawl it from the front end. This would, however, imply extracting knowledge from the formatted HTML pages, which causes a loss of the encoded entry's structure in form of templates that we introduced in the previous section. As shown in Figure 3, templates such as "sense" are important to associate semantic relations to word senses. A formatted HTML page, on the other hand, contains only a remark in round brackets, which may be interpreted as a "sense" template or any other remark added to a semantic relation, such as a label denoting a register of language (such as "formal," or "colloquial"). Fortunately, the Wiktionary data is also available as an XML database dump, which contains the original wiki markup in the form of templates.[13] Although these dumps were originally intended for developing alternative user interfaces and hosting mirror sites, they are also an ideal starting point for extracting the encoded knowledge and using it to construct ontologies.

To date, we are aware of four software libraries that allow extracting Wiktionary's knowledge based on the XML dump files: The *Java-Based Wiktionary Library*[14] (JWKTL) introduced by Zesch et al. (2008), the *Wiki tool kit*[15] (Wikokit) by Krizhanovsky and Lin (2009), *WIktionarieS Improvement by Graphs-Oriented meTHods*[16] (WISIGOTH) introduced by Sajous et al. (2010), and *Zawilinski*[17] by Kurmas (2010). We discuss the differences between these software libraries very briefly here, but refer the reader to the original works for further details. Aside from Zawilinski, which concentrates on the extraction of inflected word forms, all software libraries are able to extract words, word senses, and semantic relations, which represent the required information to construct OntoWiktionary. An important property of Wiktionary is that each language edition has its own structure and format.[18] It is thus necessary to create a new extraction system for each language edition—or at least to adapt an existing one. As this is a time-consuming process, it is not surprising that each of the software libraries focuses on certain language editions: JWKTL allows one to process the English and the German Wiktionaries, Wikokit is able to process the Russian and English editions, WISIGOTH is suitable for the French and English editions, while Zawilinski has been built for the Polish Wiktionary.

For the construction of OntoWiktionary described in this chapter, we utilize JWKTL for processing the English and German Wiktionaries. Additionally, we create a novel JWKTL adapter to Wikokit and thus also extract the knowledge from the Russian Wiktionary using the same software system. This choice of languages allows for studying the ontology construction process for a very large Wiktionary (the English one), for a medium-sized Wiktionary (the German one), and for a Wiktionary having a script different from Latin (namely the Cyrillic alphabet in the Russian Wiktionary). As future work, we also plan to include other Wiktionaries, such as the French edition that is also one of the largest ones. At the time of writing, WISIGOTH is, however, subject to revision and thus prevents us from analyzing the French Wiktionary. The XML dump files processed with JWKTL are from February 2, 2011 for the English Wiktionary; February 1, 2011 for the German Wiktionary; and April 4, 2011 for the Russian Wiktionary. Any numbers reported in this chapter refer to these dates unless otherwise indicated.

Although each Wiktionary language edition encodes words from multiple languages, we focus on only those words that are "native" to a language edition (i.e., the English words in the English edition, the German words in the German edition, etc.). According to Meyer and Gurevych (2010a), these native entries represent the vast majority of a language edition. There is, for instance, an entry about the German word 'Boot' (English 'boat') within the English Wiktionary which is not considered by our approach.

## ONTOLOGIZING THE KNOWLEDGE IN WIKTIONARY

In the previous section, we focused on the knowledge harvesting step (i.e., obtaining and extracting the knowledge from Wiktionary). In order to build OntoWiktionary, we need to transform this knowledge into ontological structures—that is, to define concepts and relations between them. Pantel and Pennacchiotti (2008) call this step "ontologizing" the harvested knowledge.

The basic building blocks of OntoWiktionary are concepts and relations between them. We therefore address the formation of concepts and relations as two separate tasks of the ontologizing step. But before being able to form the concepts and relations, we need to apply a necessary preprocessing step that aims at associating the encoded hyperlinks to word senses. This process is called relation anchoring and will be the subject of the following subsection.

### Relation Anchoring

An important type of linguistic information encoded in Wiktionary is semantic relations between word senses. A Wiktionary entry may contain sections labeled "Synonyms," "Hypernyms," "Hyponyms," "Derived terms," etc. that allow for the inclusion of hyperlinks to other articles. The noun entry of 'boat,' for example, encodes a link to the article 'ship' within the "Synonyms" section, since 'boat' and 'ship' denote (roughly) the same meaning. Additionally, it contains a link to 'canoe' within the "Hyponyms" section, because a canoe is a special kind of boat. While these relations are linguistically motivated to denote words with the same (synonym), a broader (hypernym), or a narrower (hyponym) meaning, we note that they are an ideal basis for forming concepts based on synonymy links and for defining subsumption relations between them based on hypernymy and hyponymy links.

An inherent problem of Wiktionary's encoding format for these semantic relations is that the hyperlinks connect words rather than word senses. For example, from a hypernymy link pointing from 'flower' to 'plant,' it remains unspecified whether 'flower' is a narrower term of 'plant' in the biological sense or in the sense of an industrial facility. Accordingly, 'flower' not only refers to the botanical organism, but is also used to denote the finest part of something, as in the phrase "in the flower of her youth." This is especially a problem if chains of relations are considered: from the two hyponymy relations (smallmouth bass, bass) and (bass, music instrument) one could infer that 'smallmouth bass' and 'music instrument' are closely related, which is obviously wrong. In order to create a precise ontology, this issue needs to be tackled by our approach, which requires the anchoring of the encoded hyperlinks. This means that the correct word senses connected by the semantic relation need to be identified from the corresponding words in Wiktionary.

The necessity of anchoring relations has been observed before by Pantel and Pennacchiotti (2008). In particular, they mine a large amount of ontological relations from the Web using their Espresso system. Both the source and the target (i.e., the two endpoints of a relation) are thereby words that need to be "ontologized." In their approach, all possible word senses from WordNet serve as candidates for the relation's source and target word sense. These candidates are then disambiguated using measures based on distributional similarity. In this setting, the anchoring of relations is a fairly complex task, since both the source and the target word senses 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 anchoring method.

In the previous section, we introduced Wiktionary's template mechanism, which is commonly used by the Wiktionary community to associate a relation link with a so-called sense marker. A sense marker is—depending on the language of the Wiktionary edition—a numerical index or a shortened version of the textual description of a word sense, which identifies the corresponding word sense of a relation. The hyponym link 'canoe' in the article 'boat' is, for example, preceded by the sense marker "(a craft on or in water)," which associates this hyponymy relation with the first word sense of 'boat,' namely "a craft used for transportation of goods, [...]." The German Wiktionary uses numerical indices as sense markers instead. The hyponymy link 'Kanu' (English 'canoe') in the German Wiktionary is, for instance, preceded by the sense marker "[1]," which associates it with the first word sense of 'Boot' (English 'boat').

By considering the sense markers, we are able to extract the word sense of a relation's source directly from the encoded data. Given the word sense of the source, only the word sense of the relation's target remains to be found. Following this procedure, we are able to simplify the relation anchoring task by one degree of freedom, as only the two word senses of the target word 'canoe' need to be processed for anchoring our example relation (boat, canoe). This approach not only reduces the computational complexity, but also allows for a higher quality at the same time: Since the sense markers are defined by humans, no automatic disambiguation task is involved, which would introduce noise to the relation anchoring results. In the following subsection, we describe our approach to anchor Wiktionary's semantic relations. The evaluation of this approach is then discussed in the subsequent subsection.

## Word Sense Disambiguation-Based Relation Anchoring

Meyer and Gurevych (2010) introduce a word sense disambiguation method for Wiktionary relations, which we can directly apply for the anchoring of relations. In the following, we will briefly review this work, before we apply it to our setting. The hypothesis of the method is that the textual description of the target word sense is semantically related to the description of the source word sense. This is a direct consequence of the relatedness of the source and target word senses themselves. Consider the hyponymy relation between 'boat' and 'canoe'; the corresponding textual definitions are:

### Boat:

1. A craft used for transportation of goods, fishing, racing, recreational cruising, or military use on or in the water, propelled by oars or outboard motor or inboard motor or by wind.

### Canoe:

1. A small long and narrow boat, propelled by one or more people (depending on the size of canoe), using single-bladed paddles. The paddlers face in the direction of travel, in either a seated position, or kneeling on the bottom of the boat. Canoes are open on top, and pointed at both ends.
2. (slang) An oversize, usually older, luxury car.

From this example, we immediately observe that a disambiguation method based on word overlap (i.e., choosing the target word sense with the highest number of shared words) will not work very well, since only the word 'propelled' is present in more than one description (we ignore stop words such as 'the,' 'of,' etc. in the following). We though observe many pairs of related words (e.g., 'water' and 'boat,' 'oar' and 'paddle,' 'transportation' and 'people'). that are shared by the first word senses of 'boat' and 'canoe.' This motivates the application of methods based on the semantic relatedness of each pair of textual descriptions. It should be noted that there are also some related words shared by the first word sense of 'boat' and the second word sense of 'canoe,' like 'transportation' and 'car,' 'motor' and 'car,' etc. There is thus a need for carefully evaluating this method, which we will address in the next section.

For calculating the semantic relatedness between each possible pair of textual descriptions, we use Explicit Semantic Analysis (Gabrilovich & Markovitch, 2007), which is a state-of-the-art method for this task. In a preprocessing step, each textual description is tokenized and lemmatized using Helmut Schmid's (1994) TreeTagger. To avoid noise in the relatedness calculation, we also remove stop words from the descriptions. Then, we represent each token $t$ as a concept vector—i.e., a vector $c(t) = (w_i(t))$, where $w_i(t)$ denotes the degree of how well $t$ is represented by a concept $i$. Note that the concepts used here can be taken from any semantic space. Following Gabrilovich and Markovitch (2007), we use Wikipedia as a semantic space here, which has shown very good results on reference datasets for semantic relatedness, such as the WordSimilarity-353 collection (Finkelstein, et al., 2002). Thus, $w_i(t)$ denotes the degree of how well $t$ is represented by the $i^{th}$ article in Wikipedia. The values of $w_i(t)$ are calculated using the term frequency—inverse document frequency (tf–idf) schema. The token 'boat,' for instance, would receive a high $w_i(t)$ for the Wikipedia concepts 'Boat,' 'Ship,' 'Watercraft rowing,' 'Lighthouse,' etc., as it appears frequently on the corresponding article pages, and a low $w_i(t)$ for the articles 'Syntax,' 'Trumpet,' or 'Formula' that do not contain 'boat.' In order to obtain a semantic relatedness score $r(A, B)$ for two textual descriptions $A$ and $B$, we add up the concept vectors for all tokens $t_{A,i} \in A$ and $t_{B,j} \in B$, and calculate the cosine of the angle between them within our semantic space:

$$r(A,B) = \frac{c_A \cdot c_B}{\left\| c_A \right\| \cdot \left\| c_B \right\|}$$

with

$$c_A = \sum\nolimits_{t_{A,i} \in A} c(t_{A,i})$$

and

$$c_B = \sum\nolimits_{t_{B,j} \in B} c(t_{B,j}) \,.$$

Using Explicit Semantic Analysis, the descriptions of the maritime word senses of ‘boat’ and ‘canoe’ have nearly the same concept vectors and thus a high relatedness score *r*(*A, B*). The word sense of the relation’s target word with the highest semantic relatedness score is returned by the method and serves as the target word sense of our anchored relation. Note that this approach goes substantially beyond word-based cosine similarity in which the tokens are not represented in a semantic space.

## Evaluation

To evaluate our approach, we have randomly chosen 250 relations from the “Synonymy,” “Hyponymy,” and “Hypernymy” section of Wiktionary, and annotated each of the 920 possible pairs of word senses as positive (the two word senses are directly related by means of the semantic relation) or negative (the two word senses are not directly related—i.e., there should not be a semantic relation between them). The annotators were also allowed to annotate multiple target word senses for a given source word sense as positive, provided a relation holds between more than one pair of word senses. An example for such a case is the hypernymy relation from ‘drinking water’ to ‘water,’ for which the two word senses “mineral water” and “a serving of water” are suitable relation targets. The dataset has been annotated independently by two human raters.

In order to ensure the reliability of our annotations, we measured the inter-rater agreement, which turned out to be $A_O = 0.88$ in a non-chance-corrected setting and $\kappa = 0.72$ using the chance-corrected kappa measure. Since almost two thirds of our dataset (597 items) are marked with a negative annotation, the dataset is skewed, which, in general, causes lower kappa values (Artstein & Poesio, 2008). Therefore, we also measured the agreement in a set-based setting using Krippendorff’s $\alpha$ and the MASI distance function (Passonneau, 2006). This approach compares the annotations of both raters for each of the 250 relations rather than the 920 annotation pairs. For each relation, the set of positively annotated word senses is compared. Using this third measure of inter-rater agreement, we measured $\alpha = 0.86$, which indicates good agreement and allows us to draw conclusions from our results (Krippendorff, 1980).

We refrained from removing or re-annotating those cases where no agreement was found to preserve the hard cases that our relation anchoring method needs to tackle. Therefore, we are not providing precision and recall values but rather the inter-rater agreement between our approach (denoted by M in the following) and the individual human raters (denoted by A and B). This also allows the comparison of our method’s result with the agreement amongst the human raters that serves as an upper bound for our algorithm. As a baseline approach (denoted by 0), we always choose the first word sense of the target word, which is usually the most frequently used one. This kind of baseline is common practice in word sense disambiguation evaluations and is known to be difficult to surpass.

*Table 3. Evaluation results of our relation anchoring method of Wiktionary relations*

| | 0–A | 0–B | M–A | M–B | A–B |
|---|---|---|---|---|---|
| $A_O$ | 0.791 | 0.780 | 0.820 | 0.791 | 0.886 |
| $\kappa$ | 0.498 | 0.452 | 0.567 | 0.480 | 0.728 |
| $\alpha$ | 0.679 | 0.620 | 0.726 | 0.649 | 0.866 |

Table 3 shows the agreement of our method compared to the baseline and the upper bound. Our method exceeds the baseline in every case. There is, however, still room for improvement with respect to the upper bound A–B. In our error analysis, we observe large differences in the length of the textual descriptions. Although the semantic relatedness scores are normalized, this can significantly influence the performance. Very short descriptions in particular have been found to often yield errors. We also observed differences in the textual descriptions for each part of speech, which we plan to analyze in a separate study using a well-balanced dataset that covers each part of speech and relation type equally well. Another type of error is due to references to other word senses within the textual descriptions. The second word sense of 'tomato' (the fruit), for example, refers to its first sense (the plant): "[2] the fruit of [1]." Such references limit the number of words that can be used for calculating our semantic relatedness score. A future approach should take these cases into account by either augmenting them with words from the referenced description or by treating the distinctive feature (like "the fruit of sth.") in a special way. We also notice that the agreement of our method and rater A is systematically higher than the agreement with rater B. It turns out that rater A tended to rate a relation target as positive when in doubt, while rater B tended to rate the target as negative. Although the overall agreement between the two raters is fairly good, subsequent annotation studies of Wiktionary relations should further improve the annotation guidelines based on these results.

We now use the described method to anchor all harvested Wiktionary relations. This is a necessary preprocessing step for the formation of concepts and ontological relations in OntoWiktionary that we describe in the following.

## The Formation of Concepts

The data encoded in Wiktionary is based on the notion of word senses. The noun 'dog' has, for instance, the word senses "An animal, member of the genus *Canis* [...]" and "(*slang*) A coward." The basic building blocks of an ontology are, in contrast, concepts — i.e., a model of an entity observed in the universe. A concept also has a certain meaning, but might be represented by multiple words, which we call lexicalizations. The concept 'Dog' could, for example, be modeled for representing all instances that are denoted by the word 'dog' in our universe. The noun 'dog' (in the animal sense) then serves as a lexicalization of 'Dog.' Additionally, 'Dog' might also be represented by a second lexicalization using the noun 'hound' (in a general word sense).

From this example, we observe that both the Wiktionary word senses of 'dog' and 'hound' should be combined to form a concept 'Dog' with the two lexicalizations 'dog' and 'hound.' We thus need a method for identifying word senses representing the same meaning in order to form the concepts of our novel ontology OntoWiktionary. We will outline our approach in the following section.

## Concepts Based on Synonymy Links

In linguistics, word senses with the same meaning are considered to be synonyms. This also applies to ‘dog’ and ‘hound’ (in their sense of a member of the genus *Canis*). The definition of synonymy can directly be used to form concepts — namely, by combining those word senses that are connected by a synonymy relation. This approach has been followed for the construction of the Princeton WordNet, which organizes its contents in so-called synsets—i.e., sets of synonymous word senses. The synsets in WordNet may directly be used as the concepts of an ontology, as in OntoWordNet (Gangemi, et al., 2003), for instance.

Synonymy relations are also present in Wiktionary. They are defined within the “Synonyms” section by means of hyperlinks from one article to another. There is, for example, a hyperlink within the article ‘dog’ pointing to the article ‘hound.’ In the previous section, we have seen that these synonymy hyperlinks need to be anchored—i.e., associated with the correct word senses. We accomplish this task by extracting sense markers and disambiguating the link target using a method based on the semantic relatedness of short texts as explained above. Our idea is now to form ontological concepts using these anchored synonymy relations from Wiktionary.

In WordNet, the synonymy relation is assumed to be transitive—that is, if *a* and *b* are synonymous, and *b* and *c* are synonymous, then *a* and *c* are likewise synonymous. This is accounted for by the fact that *a*, *b*, and *c* are in the same synset. For instance, ‘CV’ is a synonym of ‘curriculum vitae,’ which in turn is a synonym of ‘resume.’ Consequently, ‘CV’ and ‘resume’ can also be considered synonymous. Additionally, it is obvious that WordNet’s definition of synonymy is also symmetric: If ‘CV’ is a synonym of ‘resume,’ then ‘resume’ is also a synonym of ‘CV.’

In Wiktionary, there is no such synset structure, which would make the synonymy-based formation of concepts a trivial task. Rather, synonyms are encoded for each word sense individually and thus are not necessarily required to have a symmetric or transitive counterpart. There is, for example, a synonymy link from ‘curriculum vitae’ to ‘CV,’ but not vice-versa. A viable option, therefore, is to first create a synset-like structure and then use these synsets as the concepts for OntoWiktionary. We obtain this synset structure by adding the missing symmetric and transitive counterparts of the synonymy relation. This makes Wiktionary’s synonymy relation an equivalence relation, whose transitive closure contains all inferred symmetric and transitive relations. There are, for instance, the synonymy relations (island, oasis), (oasis, island), and (oasis, refuge) that can be found in Wiktionary. By considering the transitive closure, the three additional relations (refuge, oasis), (island, refuge), and (refuge, island) are added. The corresponding concepts can now be formed from the equivalence classes of this transitive closure. In our example, the set {island, oasis, refuge} represents one equivalence class and thus forms a concept with three lexicalizations within OntoWiktionary.

Table 4 shows the number of concepts in OntoWiktionary generated from the synonymy relations encoded in the English, German, and Russian Wiktionaries. The largest ontology is obtained from the English Wiktionary. This is not surprising, since the English Wiktionary edition is currently the largest available one.[19] With its 456,638 concepts, the English OntoWiktionary is about three times larger than OpenCyc (153,920 concepts) and WordNet (117,659 synsets), as well as seven times larger than OntoWordNet (about 60,000 concepts).[20] The Wikipedia-based ontology DBpedia contains about 1.6 million entries, which are, however, mostly instances (i.e., proper names like places, organizations, people, etc.) Wiktionary focuses on common words rather than proper names and thus encodes a different type of concepts.

From the German and Russian Wiktionaries, a considerably lower number of concepts can be

*Table 4. Number of concepts, lexicalizations, and relationships within* OntoWiktionary

| | English Wiktionary | German Wiktionary | Russian Wiktionary |
|---|---:|---:|---:|
| Ontologized concepts | 456,638 | 64,335 | 72,390 |
| Lexicalizations | 469,025 | 72,157 | 80,618 |
| Ontologized relations | 8,026 | 153,685 | 66,192 |

formed. These language editions are much smaller than the English Wiktionary: there are 2.3 million articles in the English edition, but only about 158,000 in the German and 289,000 in the Russian edition, so there are many more word senses available for the formation of the concepts in the English OntoWiktionary. However, we observe a greater number of synonymy relations in both the German and Russian Wiktionaries. This yields a higher number of lexicalizations provided for each concept: while a concept has only 1.03 lexicalizations on average in our English ontology, there are 1.11 in the Russian version and 1.12 in the German OntoWiktionary. Since the English Wiktionary is rather sparse in the number of encoded synonymy relations, we plan to incorporate systems for synonymy identification into the concept formation step as part of our future work.

## Evaluation

The induction of a synset-like structure in Wiktionary might introduce errors into our final ontology that can be traced back to either errors in the relation anchoring step or inconsistencies in the encoded synonymy relations, which are not as rigidly structured as it is the case for WordNet. An evaluation of the relation anchoring step was presented in the previous section. Although the vast majority of relations could successfully be associated to the correct word senses, errors of this approach also affect the concept formation.

Since the synonymy relations in Wiktionary are added by humans rather than by an automatic system, we expect them to be generally correct. However, we can still expect to encounter errors due to extraction errors within the knowledge harvesting step or differences in the granularity of the word sense definition. For instance, Wiktionary encodes two word senses for the term ‘New York,’ namely the state within the U.S. and the city therein. Accordingly, there are synonyms listed for each sense: ‘Empire State’ and ‘New York State’ for the former and ‘Big Apple,’ ‘New York City,’ and ‘NYC’ for the latter. The abbreviation ‘N.Y.’ that is being used to refer to both of them—depending on the context—is additionally listed for both word senses. We would require (and expect) to find two word senses for ‘N.Y.’ denoting the abbreviation for the state on the one hand and the abbreviation for the city on the other hand. But Wiktionary encodes only a single word sense that covers both meanings. This distracts our word sense disambiguation algorithm, which chooses this more general word sense for anchoring both ‘N.Y.’ synonymy relations. The result is a lumped concept with the lexicalizations {Empire State, New York State, Big Apple, New York City, NYC, N.Y.}, which is clearly wrong.

Therefore, in order to analyze the quality of our concept formation step, we carried out another evaluation experiment that relies on human judgments. We have chosen 100 concepts from the English version of OntoWiktionary and 100 concepts from the corresponding German version. We considered only those concepts with at least three lexicalizations, because concepts with fewer lexicalizations are not influenced by the problem of lumped concepts described above; rather, they are directly formed from independent, explicitly encoded synonymy links and thus inherently correct. For both datasets, we asked two human raters to annotate the concepts as "consistent" (1), "lexically consistent" (2), or "inconsistent" (0). A concept is thereby represented by its lexicalizations, which consist of the corresponding word and the textual definition that is extracted for the corresponding word sense. Consider the following three examples:[21]

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.

In example (1), both lexicalizations refer to the same meaning, namely a singer in the bass range, although there are subtle differences, such as that 'basso' is used especially when talking about opera. As Hirst (1995) points out, many words that occur to be synonyms at first sight turn out at closer examination to be plesionyms (i.e., near-synonyms). A "statement that does not conform to the truth" can, for instance, be lexicalized as 'lie,' 'falsehood,' 'untruth,' 'fib,' or 'misrepresentation,' which have—although they share the same meaning—subtle differences. A 'lie,' for example, usually implies deceiving someone, while a 'misconception' can be simply due to ignorance (Hirst, 1995). For our concept formation step, we asked the annotators to ignore these subtle differences in order to obtain a rather coarse-grained ontology. The words 'lie,' 'falsehood,' 'untruth,' 'fib,' and 'misrepresentation' should thus form a single concept with multiple lexicalizations. We therefore asked the raters to judge (1) as "consistent" (1).

Example (2) contains the same lexicalizations as (1), but has an additional lexicalization 'singer.' A 'bass' is a certain kind of 'singer'; we would thus not expect to find both lexicalizations as a representation for the same concept. Humans would rather model two independent concepts {bass, basso} and {singer} that are connected by a subsumption relation. We hence asked the raters to judge such cases as "inconsistent" (0).

Regarding (3), the textual definitions would indicate an "inconsistent" concept, as the fish 'bass' and the singer 'basso' do clearly not represent the same meaning. We, however, asked the raters to judge such concepts as "lexically consistent" (2), since there is a different word sense for 'bass' that refers to the male singer. For judging a concept as "lexically consistent," the rater should hence ignore the textual description and rather judge if the words ('bass' and 'basso' in this case) refer to the same concept. While "inconsistent" concepts yield errors in our final ontology, "lexically consistent" concepts are still useful, as they represent valid lexicalizations of a concept. A concept that is lexicalized as {bass, basso} induces a clear, consistent meaning regardless of the textual definitions mined from Wiktionary. Such a concept can particularly be used in a subsumption relation to, for example, {singer} without introducing inconsistencies per se.

Each rater had previous experience in linguistic annotation studies. The annotation task was explained in an annotation guidebook that contains multiple examples illustrating the task. The annotators were also encouraged to consult other knowledge resources such as books or the Web, but were not supposed to discuss the items with each other. Wiktionary in particular could be used to better grasp the possible meanings of the lexicalizations. To allow for reproducibility, we make the dataset and the guidebook available on our website.

In order to ensure the reliability of our data, we measure the inter-rater agreement of each dataset; this is shown in Table 5. We observe a slight trend or bias of rater A annotating a concept as "consistent" or "lexically consistent," while rater B seems to use "lexically consistent" or "inconsistent" more often. This caused us to look more closely at the agreement between the annotations of both raters. For the English dataset, we observed an overall agreement of $A_O = 0.89$ and likewise $A_O = 0.87$ for the German dataset. While the observed agreement $A_O$ considers the absolute number of concepts that were annotated with the same class, some of these matches might be due to chance. We therefore also measured the chance-corrected inter-rater agreements $\kappa = 0.79$ for the English dataset and $\kappa = 0.71$ for the German dataset using Cohen's kappa (Artstein & Poesio, 2008). Both agreement scores are well above 0.67, which indicates substantial agreement and allows tentative conclusions to be drawn (Krippendorff, 1980).

As already mentioned for the anchoring of the relations, kappa is known to yield smaller values if the distribution of categories is skewed. From the distribution of the annotation categories shown in Table 5, we observe that most concepts have been rated as "consistent" (1), which indicates a skewed distribution of categories. We therefore analyzed each annotation category separately by measuring the observed agreement $A_{O,i}$ per category $i$ and the kappa per category $\kappa_i$ that has been introduced by Fleiss (1971). With the exception of the "lexically consistent" category of the German dataset, all $\kappa_i$ values are above 0.7; the "inconsistent" category of the English dataset is even above 0.9, which indicates perfect agreement. Hence, we consider our annotated dataset reliable.

Besides the inter-rater agreement, Table 5 also shows the actual annotations per class. As can be seen from the table, the vast majority of concepts (59–70% in the English and 65–77% in the German dataset) are judged as "consistent," which demonstrates the validity of our new ontology. Apart from that, the majority of the concepts not judged as "consistent" are considered "lexically consistent" by the raters. In the English dataset, 83% (rater A) and 80% (rater B) are annotated as either "consistent" or "lexically consistent." For the German dataset, even 94% (rater A) and 90% (rater B) of the concepts fall in these categories. As noted above, we only evaluated concepts with at least three lexicalizations. Concepts with only one lexicalization can be seen as consistent per se, as only one word sense is involved and concepts with two lexicalizations are at least "lexically consistent," since they only depend on the quality of the relation anchoring step. From these observations, we conclude that the concepts in OntoWiktionary are of good quality.

Our error analysis showed that most ill-formed concepts are due to errors in the relation anchoring step. Consider example (3) from the annotation task definition above. This concept is created from a synonymy relation between 'basso' and

*Table 5. Evaluation of our concept formation step*

| | Rater A | Rater B | $A_O$ | $A_{O,i}$ | $\kappa$ | $\kappa_i$ |
|---|---|---|---|---|---|---|
| **English data** | **100** | **100** | **0.890** | | **0.791** | |
| consistent (1) | 70 | 59 | | 0.915 | | 0.760 |
| lexically consistent (2) | 13 | 21 | | 0.765 | | 0.717 |
| inconsistent (0) | 17 | 20 | | 0.919 | | 0.901 |
| **German data** | **100** | **100** | **0.870** | | **0.712** | |
| consistent (1) | 77 | 65 | | 0.915 | | 0.709 |
| lexically consistent (2) | 17 | 25 | | 0.762 | | 0.699 |
| inconsistent (0) | 6 | 10 | | 0.750 | | 0.728 |

'bass,' whereby the relation target has been detected wrongly — i.e., the fish sense of 'bass' has been used rather than the 'singer' sense. Future improvements should thus concentrate on the relation anchoring step.

## The Formation of Relationships

We have already observed that there are different types of relations encoded in a Wiktionary article. Besides synonymy relations, this includes, amongst others, hyponymy and hypernymy relations, which are particularly useful for constructing an ontology. Hyponyms are "narrower" terms: for example, 'canoe' is a hyponym of 'boat,' since it is a special kind of boat. Conversely, hypernyms are "broader" terms, such as 'vessel,' which is a hypernym of 'boat.' These relations are capable of creating a subsumption hierarchy of the concepts in OntoWiktionary.

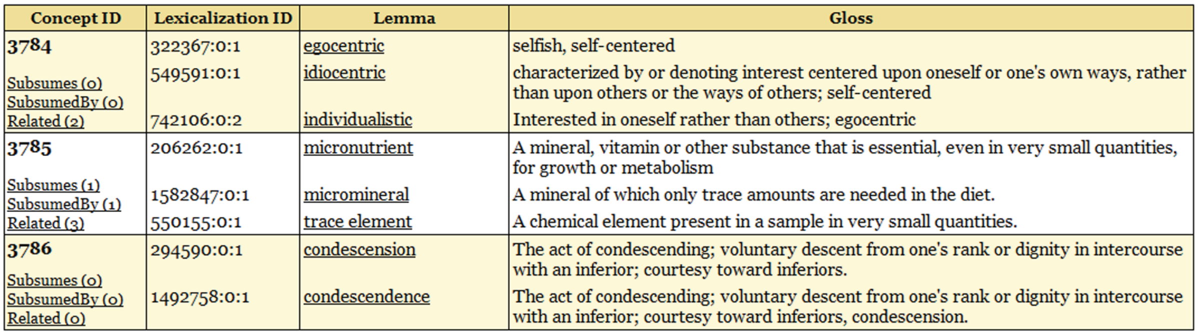
In order to incorporate them, we need to ontologize Wiktionary's semantic relations, since they are defined between word senses rather than concepts. This can be done directly based on the previous steps of our ontologizing process: As discussed in the previous section, the concepts in OntoWiktionary consist of individual word senses—i.e., they have been defined as the equivalence classes of the transitive closure of the synonymy relation. We can thus infer the concepts unambiguously from the word senses of a relation. Consider the hypernymy relation (submarine, boat). After applying our ontologizing approach, we are able to add a subsumption relation ({submarine, U-boat}, {boat, craft, ship}) to OntoWiktionary, since the 'boat' word sense is included in the concept {boat, craft, ship} and likewise for 'submarine.'

While the synonymy relation is usually considered to be symmetric, the hyponymy and hypernymy relations are invertible. If a hypernymy relation holds between 'submarine' and 'boat,' then an inverse hyponymy relation should hold between 'boat' and 'submarine' (and vice-versa). The inverse counterpart of a relation is not always explicitly defined in Wiktionary. This is why we generate them by flipping the relation's source and target word sense, as well as inverting the relation type.

In addition to hyponymy and hypernymy relations, Wiktionary encodes hyperlinks to derived words, words with opposite meaning (antonymy), or words that appear often with another word (collocation). We also extract these relations, and anchor them within our ontology as related concepts. Each concept thus contains relations to concepts that it subsumes, that it is subsumed by, and that it is related to. The concept {micronutrient, micromineral, trace element}, for example, subsumes the concept {vitamin}, is subsumed by the concept {nutrient}, and is related to the concepts {electrolyte} and {macronutrient}. Figure 4 shows an excerpt of OntoWiktionary, which illustrates its structure.

Table 4 shows the number of relations in OntoWiktionary. The English Wiktionary encodes the fewest number of relations, which is surprising as it is the largest yet available Wiktionary. The reason is that the English Wiktionary's community has put a focus on the encoding of words and word senses for a long time. However, the recently started initiative Wikisaurus[22] addresses exactly the encoding of semantic relations. The Wikisaurus is a special part of the English Wiktionary which contains a list of hyperlinks to terms that are related to each other. The Wikisaurus entry for 'mountain' contains, for example, links to the synonymous terms 'mount' and 'hill,' as well as a hyponymy link to 'volcano' and many other links that are good semantic relations. However, since Wikisaurus is fairly new, it cannot be extracted by any of the Wiktionary extraction systems. We thus leave the inclusion of Wikisaurus to future work. The German OntoWiktionary encodes the most semantic relations, although it is the smallest Wiktionary edition regarding the

*Figure 4. An excerpt of OntoWiktionary showing three concepts with their different lexicalizations. Both the OntoWiktionary data and the user interface shown in the figure are publicly available from our website.*



| Concept ID | Lexicalization ID | Lemma | Gloss |
|---|---|---|---|
| **3784**<br>Subsumes (0)<br>SubsumedBy (0)<br>Related (2) | 322367:0:1 | egocentric | selfish, self-centered |
| | 549591:0:1 | idiocentric | characterized by or denoting interest centered upon oneself or one's own ways, rather than upon others or the ways of others; self-centered |
| | 742106:0:2 | individualistic | Interested in oneself rather than others; egocentric |
| **3785**<br>Subsumes (1)<br>SubsumedBy (1)<br>Related (3) | 206262:0:1 | micronutrient | A mineral, vitamin or other substance that is essential, even in very small quantities, for growth or metabolism |
| | 1582847:0:1 | micromineral | A mineral of which only trace amounts are needed in the diet. |
| | 550155:0:1 | trace element | A chemical element present in a sample in very small quantities. |
| **3786**<br>Subsumes (0)<br>SubsumedBy (0)<br>Related (0) | 294590:0:1 | condescension | The act of condescending; voluntary descent from one's rank or dignity in intercourse with an inferior; courtesy toward inferiors. |
| | 1492758:0:1 | condescendence | The act of condescending; voluntary descent from one's rank or dignity in intercourse with an inferior; courtesy toward inferiors, condescension. |

number of concepts. In our future work, we plan to add additional relations to OntoWiktionary by, for example, integrating our ontology with previously existing ones.

## FUTURE RESEARCH DIRECTIONS

There are multiple future research directions concerning the ontology construction process of OntoWiktionary. One crucial point is the development of the knowledge harvesting step, which relies on complex software systems that interpret Wiktionary's encoding format. Although much effort lies in the engineering of robust extraction software, emerging research in the field of Web-based information extraction and wrapper induction can help to improve the software components. This includes automatic detection of format changes and new sections, as well as the adaptation of the extraction software to a Wiktionary edition of another language. Particularly the latter is an important challenge, since Wiktionary currently utilizes different structures and guidelines for each language edition.

Furthermore, as regards the knowledge ontologizing step, our error analysis of the relation anchoring approach reveals some room for improvements in our algorithm. A future improvement we are working on involves the transfer of state-of-the-art methods in word sense disambiguation to this problem. Refining the relation anchoring approach also enables other usages of Wiktionary which rely on lexical semantic information, such as machine translation, and thus increases the need for high-quality results for this task. Apart from that, the concept and relation formation steps offer multiple future directions. We observe that the English OntoWiktionary contains rather few lexicalizations for each concept, which is caused by a fewer number of semantic relations in the English Wiktionary compared to other language editions. The information encoded in the Wikisaurus part of Wiktionary might prove helpful here, as well as the incorporation of methods for synonymy mining.

Besides improvements to our ontology construction process, we want to point out several other future directions in the context of OntoWiktionary and ontologies in general. We see an important research question in the integration of different ontologies. As discussed in the "background" section, the different ontology construction approaches have their individual limitations, which might be alleviated by an integrated ontology. The integration of OntoWiktionary with OpenCyc, OntoWordNet, and DBpedia appears to be a promising option, since Wiktionary encodes a huge amount of lexical semantic information that cannot be found within the other ontologies. Another important task is the integration of OntoWiktionary into the linked data cloud which has proven to be an excellent platform for combining heterogeneous ontologies. A main challenge therein will be the modeling of stable identifiers in Wiktionary. Since word senses in Wiktionary are subject to change at basically any time, their indices in the scope of a Wiktionary page might change and thus the index is not a good identifier for integrating OntoWiktionary with other ontologies.

Finally, we see lots of applications in the field of natural language processing that can benefit from OntoWiktionary. Such applications might be (1) foundational algorithms like calculating semantic relatedness or performing word sense disambiguation, as well as (2) applications to real world problems (e.g., question answering, automatic summarization systems, or semantic search engines) that usually require huge ontologies as a source of background knowledge. We refer to some interesting applications that started to discover Wiktionary within our additional reading section. We are, however, not aware of any work that transforms Wiktionary data into an ontology. Thus, we expect substantial impact from providing a huge Wiktionary-based ontology of considerable quality.

## CONCLUSION

The aim of this chapter is to explore the potential of the free online dictionary Wiktionary for constructing ontologies in a (semi-)automatic manner. Wiktionary is a collaborative wiki collecting knowledge about words, word senses, and semantic relations between words. The large community of voluntary editors has collected millions of individual facts in over 145 languages, which makes Wiktionary a hidden treasure for developing ontologies.

Employing Wiktionary as a source for ontology construction has several advantages: (1) it is larger than many other existing resources such as OpenCyc or OntoWordNet; (2) its data is constantly edited and extended by the community, which allows it to quickly reflect trends and emerging topics; (3) it has good coverage of domain-specific terminology that can be used to develop domain ontologies; and (4) its data is multilingual and thus can be used to develop ontologies for resource-poor languages. As opposed to automatically created ontologies using information extraction systems, Wiktionary's knowledge can be of fairly high quality, since it has been explicitly encoded by humans, and is constantly reviewed by its community. Our intuition is that Wiktionary has a similar potential as Wikipedia, which gained great attention within the Semantic Web community. Wiktionary shows similar properties as its encyclopedic companion, although focusing on linguistic knowledge. The harvesting of lexicalizations of ontological concepts and a clear-cut taxonomy of subsumption relations are two main strengths that we observe in the Wiktionary data and that we exploit for constructing ontologies.

In this chapter, we proposed a two-step approach to construct the novel ontology OntoWiktionary that contains concepts, their lexicalizations, and relations harvested from Wiktionary. The first step applies a software system to transform a Wiktionary dump into a structured database. After reviewing different existing software systems for this purpose and the challenges they need to tackle, we used JWKTL and developed a new adapter to Wikokit, which allows us to extract data from the English, German, and Russian Wiktionary editions.

The second step addresses the ontologizing of the harvested knowledge—i.e., the formation of concepts and relations within our ontology. A necessary preprocessing task for achieving this goal is the anchoring of Wiktionary's relations. For this task, each relation needs to be associated with the correct word sense (which is used to form the concepts later on). For our setting, we adapted a method by Meyer and Gurevych (2010), which works well for our purposes. The error analysis reveals some remaining issues that we leave to future work, such as references within the textual definitions of senses. From the anchored synonymy relations, we then formed the concepts of OntoWiktionary. In particular, we followed the approach of OntoWordNet that uses the synsets of WordNet as a basis for the concepts of their ontology. Wiktionary has no explicitly encoded synsets, so we induced a synset-like structure by considering the synonymy relation as an equivalence relation and using its transitive closure as automatically induced concepts for OntoWiktionary. We evaluated our method by asking human raters whether the concepts we formed are consistent and found that about three quarters are considered consistent, while between 80% and 94% are at least lexically consistent (i.e., they have consistent lexicalizations) although their textual definition might not be fully correct due to errors in the relation anchoring preprocessing step. Finally, we augmented the encoded taxonomic relations between senses by adding subsumption relations and related concepts to OntoWiktionary.

From our evaluation, we conclude that OntoWiktionary is of good quality. At the same time, it contains a large number of concepts and thus surmounts the size of OntoWordNet and OpenCyc in terms of concepts. The final OntoWiktionary for the English, German, and Russian language is available from our website.[23] The ontological data can be browsed using the Web-based user interface shown in Figure 4 or downloaded as a simple XML file for offline use. By making OntoWiktionary publicly available, we want to foster research in the field of Wiktionary and ontologies in general. One particular future task will be the integration of OntoWiktionary with existing ontologies as well as the linked data cloud.

## ACKNOWLEDGMENT

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant Nº I/82806. We would like to thank the anonymous reviewers for their valuable comments, our student Yevgen Chebotar for his contributions to this work, as well as our colleagues at the Ubiquitous Knowledge Processing Lab for many fruitful discussions about this chapter. Moreover, we thank Andrew Krizhanovsky for developing the Wikokit software library to access the Russian Wiktionary.

## REFERENCES

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596. doi:10.1162/coli.07-034-R2

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford, UK: Oxford University Press.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* (pp. 2670–2676). IEEE Press.

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., & Weitzner, D. J. (2006). Creating a science of the web. *Science*, *313*(5788), 769–771. doi:10.1126/science.1126902

Bizer, C., Heath, T., & Berners-Lee, T. (2009a). Linked data – The story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. doi:10.4018/jswis.2009081901

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia – A crystallization point for the web of data. *Journal of Web Semantics*, *7*(3), 154–165. doi:10.1016/j.websem.2009.07.002

Echarte, F., Astrain, J., Córdoba, A., & Villadangos, J. (2007). *Ontology of folksonomy: A new modeling method*. Paper presented at the Semantic Authoring, Annotation and Knowledge Markup Workshop. Whistler, Canada.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, *20*(1), 116–131. doi:10.1145/503104.503110

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–381. doi:10.1037/h0031619

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* (pp. 1606–1611). IEEE Press.

Gangemi, A., Navigli, R., & Velardi, P. (2003). The OntoWordNet project: Extension and axiomatization of conceptual relations in WordNet. In Meersman, R., Tari, Z., & Schmidt, D. C. (Eds.), *On the Move to Meaningful Internet Systems* (pp. 820–838). Berlin, Germany: Springer. doi:10.1007/978-3-540-39964-3_52

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*(7070), 900–901. doi:10.1038/438900a

Gruber, T. (2007). Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems*, *3*(1), 1–11. doi:10.4018/jswis.2007010101

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In Staab, S., & Studer, R. (Eds.), *Handbook on Ontologies* (pp. 1–7). Berlin, Germany: Springer. doi:10.1007/978-3-540-92673-3_0

Hirst, G. (1995). Near-synonymy and the structure of lexical knowledge. In *Proceedings of the AAAI Spring Symposium Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity,* (pp. 51–56). Menlo Park, CA: The AAAI Press.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, *29*(3), 333–347. doi:10.1162/089120103322711569

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage Publications.

Krizhanovsky, A., & Lin, F. (2009). Related terms search based on WordNet / Wiktionary and its application in ontology matching. In *Proceedings of the 11th Russian Conference on Digital Libraries,* (pp. 363–369). RCDL Press.

Kurmas, Z. (2010). *Zawilinski: A library for studying grammar in Wiktionary*. Paper presented at the 6th International Symposium on Wikis and Open Collaboration. Gdańsk, Poland.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, *38*(11), 33–38. doi:10.1145/219717.219745

Martin, P. (2003). Correction and extension of WordNet 1.7. In *Conceptual Structures for Knowledge Creation and Communication: 11th International Conference on Conceptual Structures,* (pp. 160–173). Berlin, Germany: Springer.

Maynard, D., Li, Y., & Peters, W. (2008). NLP techniques for term extraction and ontology population. In Buitelaar, P., & Cimiano, P. (Eds.), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (pp. 107–127). Amsterdam: IOS Press.

Meyer, C. M., & Gurevych, I. (2010a). Worth its weight in gold or yet another resource – A comparative study of Wiktionary, OpenThesaurus and GermaNet. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing: 11th International Conference,* (pp. 38–49). Berlin, Germany: Springer.

Meyer, C. M., & Gurevych, I. (2010b). *How Web communities analyze human language: Word senses in Wiktionary*. Paper presented at the Second Web Science Conference. Raleigh, NC.

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science*. *Services and Agents on the World Wide Web*, *5*(1), 5–15. doi:10.1016/j.websem.2006.11.002

Pantel, P., & Pennacchiotti, M. (2008). Automatically harvesting and ontologizing semantic relations. In Buitelaar, P., & Cimiano, P. (Eds.), *Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (pp. 171–198). Amsterdam: IOS Press.

Passonneau, R. J. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation,* (pp. 831–836). ACL Press.

Ponzetto, S. P., & Strube, M. (2007). Deriving a large-scale taxonomy from Wikipedia, In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence,* (pp. 1440–1445). Menlo Park, CA: AAAI Press.

Prévot, L., Borgo, S., & Oltramari, A. (2005). Interfacing ontologies and lexical resources. In *Proceedings of the IJCNLP 2005 Workshop Ontologies and Lexical Resources,* (pp. 91–102). IJCNLP Press.

Reed, S. L., & Lenat, D. B. (2002). Mapping ontologies into Cyc. In *Proceedings of the AAAI 2002 Workshop Ontologies and the Semantic Web,* (pp. 1–6). AAAI Press.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Upper Saddle River, NJ: Prentice Hall.

Sajous, F., Navarro, E., Gaume, B., Prévot, L., & Chudy, Y. (2010). Semi-automatic endogenous enrichment of collaboratively constructed lexical resources: Piggybacking onto Wiktionary. In H. Loftsson, E. Rögnvaldsson, & S. Helgadóttir (Eds.), *Advances in Natural Language Processing: Proceedings of the 7th International Conference on NLP,* (pp. 332–344). Berlin, Germany: Springer.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing,* (pp. 44–49). ICLP Press.

Singh, P. (2002). The public acquisition of commonsense knowledge. In *Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access,* (pp. 47–52). Menlo Park, CA: The AAAI Press.

Suchanek, F., Kasneci, G., & Weikum, G. (2008). YAGO – A large ontology from Wikipedia and WordNet. *Web Semantics: Science*. *Services and Agents on the World Wide Web*, *6*(3), 203–217. doi:10.1016/j.websem.2008.06.001

Surowiecki, J. (2005). *The wisdom of crowds*. New York, NY: Anchor Books.

Vossen, P. (1998). Introduction to EuroWordNet. *Computers and the Humanities*, *32*(2–3), 73–89. doi:10.1023/A:1001175424222

Zesch, T., Müller, C., & Gurevych, I. (2008a). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation,* (pp. 1646–1652). ACL Press.

Zesch, T., Müller, C., & Gurevych, I. (2008b). Using Wiktionary for computing semantic relatedness. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence,* (pp. 861–867). AAAI Press.

## ADDITIONAL READING

Bernhard, D., & Gurevych, I. (2009). 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 Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing,* (pp. 728–736). ACL Press.

Bouchard-Côté, A., Liang, P., Griffiths, T. L., & Klein, D. (2007). 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,* (pp. 887–896). ACL Press.

Buitelaar, P., & Cimiano, P. (Eds.). (2008). *Ontology learning and population: Bridging the gap between text and knowledge*. Amsterdam: IOS Press.

Burfoot, C., & Baldwin, T. (2009). Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers,* (pp. 161–164). ACL Press.

Chesley, P., Vincent, B., Xu, L., & Srihari, R. (2006). Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of the AAAI Spring Symposium Computational Approaches to Analysing Weblogs,* (pp. 27–29). Menlo Park, CA: The AAAI Press.

De Melo, G., & Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management,* (pp. 513–522). New York, NY: ACM.

De Melo, G., & Weikum, G. (2010). Providing multilingual, multimodal answers to lexical database queries. In N. Calzolari, et al. (Eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation,* (pp. 348–355). ACL Press.

Descy, D. E. (2006). The Wiki: True Web democracy. *TechTrends*, *50*(1), 4–5. doi:10.1007/s11528-006-7569-y

Etzioni, O., Reiter, K., Soderland, S., & Sammer, M. (2007). *Lexical translation with application to image search on the Web*. Paper presented at the Machine Translation Summit XI. Copenhagen, Denmark.

Fišer, D., & Sagot, B. (2008). Combining multiple resources to build reliable wordnets. In *Proceedings of the 11th International Conference on Text, Speech and Dialogue,* (pp. 61–68). Berlin, Germany: Springer.

Garoufi, K., Zesch, T., & Gurevych, I. (2008). *Graph-theoretic analysis of collaborative knowledge bases in natural language processing*. Paper presented at the Poster Session of the 7th International Semantic Web Conference. Karlsruhe, Germany.

Gurevych, I., & Wolf, E. (2010). Expert-built and collaboratively constructed lexical semantic resources. *Language and Linguistics Compass*, *4*(11), 1074–1090. doi:10.1111/j.1749-818X.2010.00251.x

Kann, V., & Rosell, M. (2006). Free construction of a free Swedish dictionary of synonyms. In S. Werner (Ed.), *Proceedings of the 15th Nordic Conference on Computational Linguistics,* (pp. 105–110). ACL Press.

Kulkarni, A., & Callan, J. (2008). Dictionary definitions based homograph identification using a generative hierarchical model. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* (pp. 85–88). ACL Press.

Kurmas, Z. (2010). *Encouraging language students to contribute inflection data to Wiktionary.* Paper presented at the 6th International Symposium on Wikis and Open Collaboration. Gdańsk, Poland.

Matuschek, M., & Gurevych, I. (2010). *Beyond the synset: Synonyms in collaboratively constructed semantic resources*. Paper presented at the Workshop on Computational Approaches to Synonymy at the Symposium on Re-Thinking Synonymy. Helsinki, Finland.

Maxwell, M., & Hughes, B. (2006). Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the COLING/ACL 2006 Workshop Frontiers in Linguistically Annotated Corpora,* (pp. 29–37). ACL Press.

Medero, J., & Ostendorf, M. (2009). *Analysis of vocabulary difficulty using Wiktionary*. Paper presented at the ISCA International Workshop on Speech and Language Technology in Education. Warwickshire, UK.

Müller, C., & Gurevych, I. (2009). Using Wikipedia and Wiktionary in domain-specific information retrieval. In C. Peters, et al. (Eds.), *Evaluating Systems for Multilingual and Multimodal Information Access: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum,* (pp. 219–226). Berlin, Germany: Springer.

Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., et al. (2009). Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the ACL 2009 Workshop The People's Web Meets NLP: Collaboratively Constructed Semantic Resources,* (pp. 19–27). ACL Press.

Perera, P., & Witte, R. (2005). A self-learning context-aware lemmatizer for German. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing,* (pp. 636–643). ACL Press.

Richman, A. E., & Schone, P. (2008). Mining Wiki resources for multilingual named entity recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* (pp. 1–9). ACL Press.

Sagot, B., & Fišer, D. (2008). Building a free French wordnet from multilingual resources. In *Proceedings of the LREC 2008 Workshop Ontologies and Lexical Resources,* (pp. 14–19).

Schlippe, T., Ochs, S., & Schultz, T. (2010). Wiktionary as a source for automatic pronunciation extraction. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association,* (pp. 2290–2293). ISCA Press.

Walther, G., Sagot, B., & Fort, K. (2010). *Fast development of basic NLP tools: Towards a lexicon and a POS tagger for Kurmanji Kurdish*. Paper presented at the 29th International Conference on Lexis and Grammar. Belgrade, Serbia.

Wandmacher, T., Ovchinnikova, E., Krumnack, U., & Dittmann, H. (2007). Extraction, evaluation and integration of lexical-semantic relations for the automated construction of a lexical ontology. In T. Meyer & A. C. Nayak (Eds.), *Proceedings of the AI 2007 Workshop Third Australasian Ontology Workshop,* (pp. 61–69). AAOW Press.

Weale, T., Brew, C., & Fosler-Lussier, E. (2009). Using the Wiktionary graph structure for synonym detection. In *Proceedings of the ACL 2009 Workshop The People's Web Meets NLP: Collaboratively Constructed Semantic Resources,* (pp. 28–31). ACL Press.

Weber, N., & Buitelaar, P. (2006). *Web-based ontology learning with ISOLDE*. Paper presented at the ISWC 2006 Workshop Web Content Mining with Human Language. Athens, GA.

Zesch, T. (2010). What's the difference? Comparing expert-built and collaboratively-built lexical semantic resources. In N. Calzolari, P. Baroni, M. Monachini, & C. Soria (Eds.), *Proceedings of the 2nd European Language Resources and Technologies Forum Language Resources of the Future / the Future of Language Resources,* (pp. 91–92). ACL Press.

## KEY TERMS AND DEFINITIONS

**Concept:** A model for objects/entities observed in a world (not necessarily the real world). Concepts are the building blocks of ontologies.

**Hypernymy:** A semantic relation between two word senses, whereby the target sense is broader (i.e., more general) than the source.

**Hyponymy:** A semantic relation between two word senses, whereby the target sense is narrower (i.e., more specific) than the source.

**Ontologizing:** The process of transforming knowledge into ontological structures – i.e., finding or creating concepts and relationships based on the given knowledge.

**Semantic Relation:** A binary relation between word senses that consists of a source, target, and relation type, which denotes a certain semantic relationship between the source and the target.

**Synonymy:** A semantic relation between two word senses that have an equivalent meaning.

**Synset:** A set of synonymous word senses — i.e., a set in which each pair of word senses are in a synonymy relation to each other.

**Wiki:** A software for collaborative text editing in the World Wide Web that is known for its simple and easy-to-use interface.

**Word Sense:** A certain aspect of meaning of a word that is usually found in dictionaries where it is defined by a brief textual description.

## ENDNOTES

[1] Cyc and OpenCyc – http://www.cyc.com

[2] Open Directory Project – http://www.dmoz.org

[3] CIA World Factbook – https://www.cia.gov/library/publications/the-world-factbook

[4] Unified Medical Language System – http://www.nlm.nih.gov/mesh/umlsforelis.html

[5] Del.icio.us – http://www.delicious.com

[6] Flickr – http://www.flickr.com

[7] OpenMind – http://www.openmind.org

[8] Wikipedia – http://www.wikipedia.org

[9] Wiktionary – http://www.wiktionary.org

[10] Note that we only count entries about English words here, which largely deviates from the number of articles in the entire English Wiktionary (2.3 million). For the distinction between word, entry, and article; see section "Wiktionary: A Collaborative Resource for Linguistic Knowledge."

[11] We only count active Wiktionary editions according to the list on http://meta.wikimedia.org/wiki/Wiktionary (April 1, 2011). There are about 35 additional editions, which have been newly created or are not maintained anymore, and are thus not considered an "active" Wiktionary edition

[12] Wiktionary: Entry layout explained – http://en.wiktionary.org/wiki/Wiktionary:ELE (February 10, 2011)

[13] Wikimedia database backup dumps – http://dumps.wikimedia.org

[14] Java-based Wiktionary Library (JWKTL) – http://www.ukp.tu-darmstadt.de/software/jwktl

[15] Wiki tool kit (Wikokit) – http://code.google.com/p/wikokit

[16] WIktionarieS Improvement by Graphs-Oriented meTHods (WISIGOTH) – http://redac.univ-tlse2.fr/wisigoth

[17] Zawilinski – http://www.cis.gvsu.edu/~kurmasz

[18] This is an important difference from Wikipedia, whose individual language editions are very similar.

[19] Note that the English and the French Wiktionary editions are head-to-head. While the French edition has been the largest one for several years, the English edition currently contains about 400,000 articles more than the French edition, cf. http://meta.wikimedia.org/wiki/Wiktionary (June 7, 2011)

[20] It should be noted that Meyer and Gurevych (2010a) mention that Wiktionary also encodes entries for inflected word forms, which are not part of comparable resources like OpenCyc or OntoWordNet.

[21] For brevity, we also use concepts with only two lexicalizations in our examples.

[22] Wikisaurus – http://en.wiktionary.org/wiki/Wiktionary:Wikisaurus

[23] OntoWiktionary – http://www.ukp.tu-darmstadt.de/data/lexical-resources/